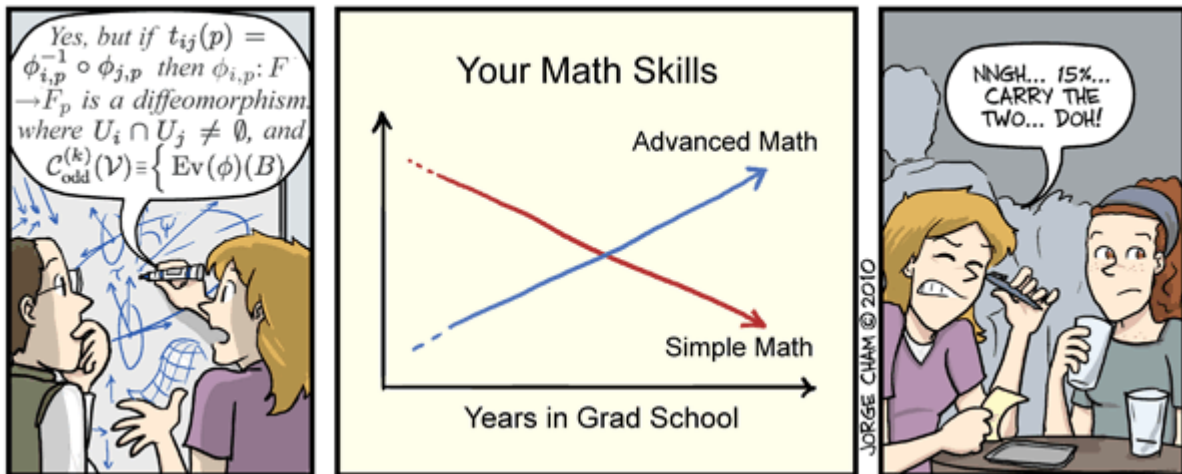


# ECNS 561: Econometrics I

## COURSE NOTES



ANTON BEKKERMAN

*Any errors in this document are the responsibility of the author. Corrections and comments regarding any material in this text are welcomed and appreciated. The material in this document is intended to be supplementary (not as a substitute) to attending lectures regularly.*

*Direct correspondence to: [anton.bekkerman@montana.edu](mailto:anton.bekkerman@montana.edu)*

Copyright © 2011 by Anton Bekkerman.

*This page intentionally left blank.*

# Contents

<b>Class Syllabus</b>	<b>8</b>
<b>1 Introduction to Econometrics</b>	<b>14</b>
1.1 Economic vs. Econometric Models . . . . .	15
1.1.1 Economic models . . . . .	15
1.1.2 Econometric models . . . . .	16
1.2 Data types and issues . . . . .	18
1.2.1 Data sources . . . . .	21
1.2.2 Functional form issues . . . . .	21
1.3 Conducting an econometric project . . . . .	23
1.4 Additional resources . . . . .	23
<b>2 Foundation of Probability Theory</b>	<b>24</b>
2.1 Random Variables . . . . .	24
2.1.1 Discrete random variables . . . . .	25
2.1.2 Continuous random variables . . . . .	25
2.2 Depicting Random Variables . . . . .	25
2.2.1 Discrete random variables . . . . .	26
2.2.2 Continuous random variables . . . . .	27
2.3 Cumulative Density Function . . . . .	28
2.3.1 Discrete random variables . . . . .	28
2.3.2 Continuous random variables . . . . .	29
2.3.3 Properties of CDFs . . . . .	29
2.4 Describing Random Variables . . . . .	30
2.4.1 Central tendency . . . . .	30
2.5 Measures of Dispersion . . . . .	33
2.5.1 Variance . . . . .	33
2.5.2 Standard deviation . . . . .	35
2.6 Joint Distributions . . . . .	35
2.6.1 Covariance . . . . .	37
2.6.2 Correlation . . . . .	38

2.7	Practice with Computing Summary Statistics . . . . .	39
2.8	Conditional Probabilities . . . . .	40
2.8.1	Conditional expectations . . . . .	41
2.8.2	Properties of conditional expectations . . . . .	43
2.9	Common Distributions . . . . .	44
2.9.1	Normal distribution . . . . .	44
2.9.2	Standard normal distribution . . . . .	46
<b>3</b>	<b>Mathematical Statistics</b>	<b>48</b>
3.1	Statistics of Samples . . . . .	49
3.2	Sampling distribution . . . . .	50
3.2.1	Unbiasedness . . . . .	50
3.2.2	Sample variance and efficiency . . . . .	51
3.3	Asymptotic Properties . . . . .	53
3.3.1	Asymptotic unbiasedness . . . . .	53
3.3.2	Consistency – probability limit . . . . .	53
3.4	Confidence intervals . . . . .	54
3.4.1	Confidence intervals with small samples . . . . .	56
3.5	Hypothesis Testing . . . . .	58
3.5.1	Devising a hypothesis test . . . . .	59
3.5.2	$p$ -values . . . . .	61
<b>4</b>	<b>Matrix algebra</b>	<b>63</b>
4.1	Basic Definitions . . . . .	63
4.2	Matrix Properties and Manipulations . . . . .	64
4.3	Linear independence . . . . .	68
4.3.1	Matrix rank . . . . .	69
<b>5</b>	<b>Simple linear regression</b>	<b>70</b>
5.1	Ordinary least squares estimation . . . . .	71
5.1.1	Estimating parameters of a linear model . . . . .	73
5.1.2	Sample variance of the estimator . . . . .	78
5.1.3	Consistency of the estimator . . . . .	79
5.2	OLS in Matrix Algebra . . . . .	79
5.2.1	Setting up the model . . . . .	80
5.2.2	Solving for the coefficients . . . . .	81
5.2.3	Projection and “residual maker” matrices . . . . .	82
5.3	Properties of the OLS Estimator . . . . .	84
5.3.1	Unbiasedness of the OLS estimator . . . . .	84
5.3.2	Variance of the OLS estimator . . . . .	85
5.3.3	Standard errors . . . . .	87

5.3.4	Distribution of estimators and error terms . . . . .	87
5.4	Gauss-Markov Assumptions for Linear Regression Models . . . . .	88
5.4.1	Proof: Gauss-Markov Theorem . . . . .	90
5.5	Asymptotic Properties of OLS Estimators . . . . .	91
5.5.1	Consistency . . . . .	91
5.5.2	Asymptotic normality . . . . .	93
5.5.3	Asymptotics: Why do we care? . . . . .	93
5.6	Inferences from OLS Estimators . . . . .	94
5.6.1	Goodness-of-fit statistics . . . . .	94
5.6.2	Hypothesis testing . . . . .	96
5.6.3	Confidence intervals . . . . .	97
5.7	Functional Forms . . . . .	98
5.7.1	Quadratic relationships . . . . .	99
5.7.2	Logarithmic relationships . . . . .	99
5.8	Maximum Likelihood Estimation Approach . . . . .	100
5.8.1	Likelihood function . . . . .	101
5.8.2	Properties of the ML estimator . . . . .	103
5.8.3	Linear regression MLE . . . . .	104
5.8.4	MLE: Example . . . . .	105
<b>6</b>	<b>Multiple regressor estimation</b>	<b>107</b>
6.1	Motivating the Use of Multiple Regressor Models . . . . .	107
6.2	Estimating Multiple Regressor Models . . . . .	109
6.2.1	Partiallying out . . . . .	110
6.3	Variable Selection . . . . .	111
6.3.1	Bias by Omission . . . . .	111
6.3.2	Inclusion of Irrelevant Variables . . . . .	113
6.3.3	Strategies for Selecting Regressors . . . . .	113
6.4	Indicator/Dummy Variables . . . . .	114
6.4.1	Interpreting Dummy Variables in a Double-log Model . . . . .	116
6.4.2	Dummy Variable Trap . . . . .	116
6.4.3	Dummy Variables Describing Multiple Categories . . . . .	117
6.5	Interaction Terms . . . . .	118
6.6	Inferences . . . . .	119
6.6.1	Single linear restriction test: a <i>t</i> -test . . . . .	119
6.6.2	General linear restrictions . . . . .	120
6.6.3	Test statistic and criterion: Wald Test . . . . .	122
6.7	Testing for Structural Breaks . . . . .	127
6.7.1	<i>F</i> test approach . . . . .	128
6.7.2	Lagrange multiplier test approach . . . . .	129

<b>7</b>	<b>Issues with OLS Estimations</b>	<b>134</b>
7.1	Multicollinearity . . . . .	135
7.1.1	Perfect Collinearity . . . . .	135
7.1.2	Imperfect Multicollinearity . . . . .	136
7.1.3	Identifying Multicollinearity . . . . .	137
7.1.4	Dealing with Multicollinearity . . . . .	138
7.2	Outliers . . . . .	138
7.2.1	Testing for Outliers . . . . .	139
7.2.2	Dealing with Outliers . . . . .	140
7.3	Heteroskedasticity . . . . .	140
7.3.1	Consequences of Heteroskedasticity . . . . .	142
7.3.2	Detecting Heteroskedasticity . . . . .	143
7.3.3	Dealing with Heteroskedasticity – Generalized Least Squares . . . . .	146
7.4	Autocorrelation / Serial Correlation . . . . .	151
7.4.1	Brief Introduction to Time-Series Data . . . . .	152
7.4.2	Violation of a Gauss-Markov Assumption . . . . .	153
7.4.3	The AR(1) Process . . . . .	154
7.4.4	Consequences of Autocorrelation . . . . .	157
7.4.5	Detecting Autocorrelation . . . . .	157
7.4.6	Dealing with Autocorrelation . . . . .	161
	<b>Econometric References and Resources</b>	<b>163</b>
	<b>Appendix 1: Statistical Tables</b>	<b>166</b>
	Cumulative Areas Under the Standard Normal Distribution . . . . .	166
	Critical Values of the $t$ Distribution . . . . .	167
	Critical Values of the $F$ Distribution . . . . .	167
	<b>Appendix 2: SAS Basics</b>	<b>170</b>
	Overview . . . . .	170
	Starting SAS . . . . .	170
	Initial View . . . . .	170
	Importing Data . . . . .	172
	Coding in SAS . . . . .	173
	Calculating Summary Statistics . . . . .	174
	Summary Statistics Using SAS . . . . .	175
	Implementing Graphical Analysis . . . . .	175
	Graphical Analysis Using SAS . . . . .	176
	SAS Code Samples . . . . .	178

## Class Syllabus

### Course Information

<b>Instructor:</b>	Dr. Anton Bekkerman	<b>Class days:</b>	Tue, Thur
<b>Office:</b>	205 Linfield Hall	<b>Class times:</b>	2:10 p.m. – 3:25 p.m.
<b>Phone:</b>	406-994-3032	<b>Classroom:</b>	Linfield 109

**Email:** anton.bekkerman@montana.edu

**Office hours:** Tue, Thur: 8 a.m. to 10:30 a.m. and by appointment

**Course website:** <http://www.montana.edu/bekkerman/ecns561f11.html>

**Optional readings:** *A Guide to Modern Econometrics (Marno Verbeek)*  
(Wiley; ISBN: 9780470517697)

*Introductory Econometrics (J.M. Wooldridge)*  
(South-Western; ISBN: 9780324660548)

*Statistics and Econometrics (Ashenfelter, Levine, Zimmerman)*  
(Wiley; ISBN: 9780470009451)

### Course Description

This course is intended to introduce important concepts in econometrics at the masters level. We will review probability theory and matrix algebra, and apply these to developing econometric analyses. The linear regression model will be discussed in detail, and students will learn how to use linear regression analysis to examine data and interpret results. Statistical properties of the linear regression model will be discussed in detail. Students who complete this course should be comfortable with using econometrics to answer important economic questions.

### Class Expectations

My commitment as a professor is to present relevant information, help you with challenging topics, and do as much as I can for you to be successful in this class. I have office hours – this is time that is devoted to my students. Please use them. There are numerous ways that you can schedule a meeting with me: (1) email or call me; (2) talk to me after class; (3) use the Google Calendar on the class website to view available times and/or make appointments (if you have a Google account).

Your commitment as a student is to put in the effort to understand the presented information, be inquisitive, and provide feedback. Feedback is extremely important

because it makes class more interactive, helps me understand whether you are understanding the material, and allows me to improve lectures and class materials. There are several ways that you can provide feedback:

1. Ask and answer questions in class.
2. Email me with questions and/or suggestions.
3. Leave anonymous comments – a link is provided on the class website. This is where you can tell me that I'm the greatest thing since Nutella or if today's class seemed like I brought lecture notes from another course and presented them in Russian. If I don't know that something is wrong, I can't change it.

### *Academic Integrity*

It is my expectation and that of the university that students follow guidelines described in the Montana State University Conduct Code.

### *Academic Misconduct*

Includes cheating, plagiarism, forgery, falsification, facilitation or aiding academic dishonesty; multiple submissions; theft of instructional materials or tests; unauthorized access to, manipulation of, or tampering with laboratory equipment, experiments, computer programs, or animals without proper authorization; alteration of grades or files; misuse of research data in reporting results; use of personal relationships to gain grades or favors; or otherwise attempting to obtain grades or credit through fraudulent means.

### *Disabled Student Services*

If you have a documented disability for which you are or may be requesting an accommodation(s), you are encouraged to contact me and Disabled Student Services as soon as possible.

<http://www.montana.edu/wwwres/disability/index.shtml>

## **Graded Opportunities**

You will be provided numerous opportunities to demonstrate your comprehension of the material. It is in your best interest to take advantage of all graded opportunities.

Homeworks provide you with an opportunity to practice concepts that we go over during lectures. Exams will give me an opportunity to evaluate how well you can apply your understanding of learned material.

Homeworks will be in the form of problem sets. Problem sets will include both theoretical (pencil and paper) and applied (analysis using statistical software) problems. For the latter, you may use any statistical software that you wish. However, given that you will have had some experience with SAS (and I can provide the most assistance with SAS programming questions), you may find that using SAS for this course may be the easiest. SAS is installed in the student computer lab located in 232 Linfield Hall.

Homework problems are an important part of this course because they allow you to practice the concepts taught during lectures. To provide an incentive for you to complete all of the problems, I will grade randomly selected problems from each homework. It is in your best interest, therefore, to work through all of the problems, as we will discuss them in class. I encourage that you work on problem sets individually and in groups.

There will be two midterm exams and a final exam. Every exam will be cumulative, because everything that we learn is a building block for the next topic. Material from the assigned readings and homeworks will be used as a basis for questions that will be on exams.

*Policy for turning in homeworks*

Due dates will be announced in class and may change depending on our progress. Homeworks must be turned in by 5 p.m. on the day they are due. If you wish, you can turn them in during class; otherwise there are several methods by which you can provide a copy of your work:

1. Drop it off at my office or in my mailbox.
2. Email a typed copy.
3. Scan and email a written copy.

On days when homeworks are due, I will be in my office until at least 5 p.m. You can always turn in your homework early, but if I don't get your copy by the due date and time it will not be accepted – *no exceptions*. At the end of the semester, your lowest homework grade will be dropped.

Research question and data collection

In order to make your thesis process more efficient and effective, this class provides an opportunity for you to think about a research question and compile a data set that can be used to start answering that question. Because relevant, interesting, and realizable research can be quite difficult, this step of your master's thesis is crucial. I will set up at least two meetings outside of class with each one of you in order to talk about your interests, research topics you may be interested in exploring, whether exploring these topics is realistic, and where you might find necessary data.

You will then be responsible to collect and manipulate the data such that they can be used to perform econometric analyses. Again, you may use any software to do the manipulation, but I recommend that you consider using SAS, as it is a powerful tool for combining and arranging data. At the end of the semester, you will be required to write a brief summary of your question/topic, sources from which you've retrieved data, and basic summary statistics of your compiled data set. This should be between two and three pages in length – the goal is for you not to write a paper, but rather to come up with an interesting question and compile a data set, both of which can be used as a foundation of your thesis.

## Grading Outline

---

---

Graded Opportunity	Weight
Homework	10%
Exam 1	20%
Exam 2	20%
Research question and data collection	15%
Final	35%

---

---

Incomplete Grades

Assigning of an *Incomplete* grade is in accordance with the guidelines of Montana State University, as outlined in the Course Catalog. This is as follows:

“The University takes the position that when students register, they commit themselves to completing their academic obligations as their primary responsibility. Therefore, the instructor may assign an *I* grade only in cases when students have suffered extreme personal hardship or in unusual academic situations.”

## Class Schedule

The outline of topics, associated chapters, and exam dates are provided below. If you know that you have an academically relevant scheduling conflict (e.g. job interview), please let me know *at least* one week in advance.

### Course Outline

Topic	Readings and Assignments
Class overview	Read syllabus
Introduction to econometrics	Wooldridge 1, 19
Foundations of probability theory	Wooldridge App. A, B Verbeek App. B Problem Set 1
Mathematical statistics	Wooldridge App. C Schedule first meeting
Matrix algebra	Wooldridge App. D Verbeek, App. A Problem Set 2
<b>Exam 1</b>	
Simple linear regression analysis	Verbeek 2.1 - 2.2 Wooldridge 2
Linear regression in matrix form	Wooldridge App. E.1 Problem Set 3
Linear regression properties	Verbeek 2.3-2.4, 3.1 Wooldridge 2, App. E.2 Schedule second meeting
Multiple regressor estimation	Wooldridge 3 Problem Set 4
Multiple regressor properties	Wooldridge 3

---

Course Outline – Continued

---

<b>Topic</b>	<b>Readings and Assignments</b>
<b>Exam 2</b>	
Multiple regressor inferences	Verbeek 2.5, 3.2 Wooldridge 4 Problem Set 5
Multiple regressor asymptotics	Verbeek 2.6 Wooldridge 5
Multiple regressor specification issues	Verbeek 3.3  Wooldridge 6 Problem Set 6
<b>Research question and data collection write-up</b>	
<b>Final exam* – December 12, 2011 (2 p.m. - 4 p.m.)</b>	

---

\* The final exam will be held in the same room as our regular class.

# Chapter 1

## Introduction to Econometrics

Econometrics is the empirical side of economics. An economist seeks to ask relevant and timely questions; econometrics provides tools that can be used to quantify answers to these questions. We can loosely generalize econometrics to be a tool for the following:

- Quantifying economic relationships
  - How much does library use increase as patrons become more affluent?
  - How do the prices of corn affect planting decisions for soybeans?
  - How does the size of a cattle's rib-eye area affect the cattle's sale price?
- Test competing hypothesis
  - Does one economic model explain the world more accurately than another economic model?
- Answer questions about policies
  - How does a tax cut affect unemployment?
  - Can a policy increase social welfare?
  - Did the No Child Left Behind policy positively or adversely affect children's education levels?
- Forecast
  - What will the price of corn be in two months?
  - How might unemployment rates behave in the next three quarters?
  - How do inflation rates behave in good and bad economic states?

## 1.1 Economic vs. Econometric Models

While economic models provide us with a theoretical representation of a phenomenon, econometric models offer an opportunity to empirically measure the phenomenon.

### 1.1.1 Economic models

Economic models are simplified representations of reality, which are often described through mathematical equations. Let's consider several economic models:

- Single-equation models: the behavior of a dependent variable  $Y$  is described by one or more independent variables  $X$ . That is:

$$Y = f(X_1, X_2, \dots, X_n)$$

Example :  $Beef = f(Cattle, Income, Price_{beef}, Price_{pork}, \dots)$

- Multi-equation models: the behavior of a dependent variable  $Y$  is described by several independent variables  $X$ , which can be themselves explained by other equations. That is:

$$Y = f(X_1, X_2, \dots, X_n)$$
$$X_1 = f(X_2, Z_1, Z_2, \dots, Z_n)$$

Example :  $Beef = f(Cattle, Income, Price_{beef}, Price_{pork}, \dots)$   
 $Income = f(Education, Age, \dots)$

- Time-series models: the behavior of a dependent variable  $Y$  is described by past values of the dependent variable  $Y_{t-1}$  as well as independent variables  $X$ . That is:

$$Y_t = f(Y_{t-1}, Y_{t-2}, X_1, X_2, \dots)$$

Example :  $Price_{beef,t} = f(Price_{beef,t-1}, Cattle, Income, Price_{pork}, \dots)$

When we construct economic models, we conceptualize a conductible experiment. For example, consider the following questions and the approach that we might take to quantify answers:

1. What is the effect of advertisement on attendance of farmers' markets?

Hypothesis: increasing advertisement will increase the number of people visiting farmers' markets.

Experiment: choose two identical locations; use a higher amount of advertisement in one location; compare the attendance at each farmer's market.

Although it would be almost impossible to find two identical locations (each may have differences in education levels, per capita income, ability to attend farmer's market, etc.) in order to conduct the experiment, it is useful to conceptualize how the experiment might be conducted.

2. Is the adoption of new technology positively related to education?

Hypothesis: more educated individuals will adopt new technology faster than less educated individuals.

Experiment: randomly survey individuals to determine their education level and how quickly they acquire a new technological tool after it is introduced.

In both cases, data from either the "experiment" in (1) or the survey results in (2) can be used to estimate an *econometric model*.

### 1.1.2 Econometric models

When we construct economic models, we assume that the relationship between the dependent variable  $Y$  and the independent (or lagged dependent) variables is *deterministic*. That is, given a value of  $X$ , we can exactly predict a value of  $Y$ . For example, consider a linear model:

$$Y = \beta_0 + \beta_1 X_1$$

Suppose that  $X_1 = 0$ . Then the exact value of the dependent variable is  $Y = \beta_0$ . If  $X_1 = 1$ , then  $Y = \beta_0 + \beta_1$ .

Life would be simple if the deterministic economic models depicted the truth. However, there is a lot of randomness that exists in life. For example, if you roll a die twice, you

may get two different outcomes. Or, by increasing  $X_1$  from 0 to 1 may not necessarily increase the value of  $Y$  by  $\beta_1$  units (e.g., increasing fertilizer will increase yield in year 1, but there is no effect in year 2 because of different weather conditions).

In an econometric model, we account for this randomness. That is, an econometric model indicates that the behavior of  $Y$  may not fully be determined by certain independent (or lagged dependent) variables. Rather, there is always a chance that increasing  $X_1$  from 0 to 1 will change  $Y$  by more or less than  $\beta_1$  units.

What causes random behavior in  $Y$ ?

- *Observable differences* – for example, differences in age, experience, and other characteristics.
- *Unobservable differences* – for example, ability levels, motivation.
- *Unpredictable events* – for example, weather events, space invasions.

Because we are unable to characterize the real world as being purely deterministic, econometrics is used to quantify the randomness. Consider how the economic model is re-written in order to account for randomness:

$$Y = f(X_1, X_2, X_3, \dots, X_n, \varepsilon)$$

where

$$f(X_1, X_2, X_3, \dots, X_n) \equiv \text{deterministic, observable components}$$

$$\varepsilon \equiv \text{random or unobservable component}$$

Because introducing the random component means that any particular realization of  $Y$  can also be random, we treat  $Y$  as a random variable. This entails making assumptions about the true distribution from which each observed value of  $Y$  was drawn.

## 1.2 Data types and issues

Data are the backbone to any empirical economic study because it allows us to measure relationships about variables, test hypotheses, and forecast values. There are three basic structures of data:

1. Cross-sectional data: information about entities (e.g. individuals, firms, households, farms) at a particular point in time. That is, these data are a “snapshot” of an economic situation. An important assumption is that each observation in a cross-sectional data set is *independent* of all other observations in that data set.

Example: Bull sale prices at a 2009 auction.

Bull ID	Sale price	Birth weight	365-day weight	Age	Daily gain
1	2500	76	1122	381	2.958
2	3000	72	1224	381	3.3361
3	3000	76	1171	370	3.0504
4	2500	92	1173	443	3.4622
5	2750	77	1262	446	3.0924
6	2250	86	1277	425	3.9664
7	2750	88	1360	457	3.6303
8	3500	75	1219	455	3.3025
9	2500	81	1264	441	3.9916
10	2500	78	1241	442	3.4874
11	1250	74	1122	433	3.1933

2. Time-series data: information on the same entity over a period of time. Often, each observation is dependent on preceding observations.

Example: Wheat planting and yield data for a county in Montana.

<b>Commodity</b>	<b>Year</b>	<b>Planted Acres</b>	<b>Yield</b>
WHEAT	1998	173100	30.6
WHEAT	1999	175400	29.9
WHEAT	2000	158900	29.9
WHEAT	2001	154800	30.9
WHEAT	2002	153500	24.6
WHEAT	2003	158500	27.8
WHEAT	2004	162800	29.2
WHEAT	2005	152700	32.2
WHEAT	2006	160300	23.9
WHEAT	2007	161000	26
WHEAT	2008	182500	16.5

3. Panel data: combination of cross-sectional and time-series types. Provides information about a variety of entities over a time period. This data type is becoming more readily available as technological advances have allowed for better data collection.

Example: Information on libraries over several years.

<b>Library ID</b>	<b>Year</b>	<b>Hrs. Open</b>	<b># Visits</b>	<b>Adult circulation</b>	<b>Child circulation</b>
AK0001	2004	1300	4500	6657	1039
AK0001	2005	1300	4600	6407	1218
AK0001	2006	1560	4950	7819	1354
AK0001	2007	1404	7075	9545	1653
AK0001	2008	1404	7357	13094	1336
AK0002	2004	11280	963000	1011765	487481
AK0002	2005	11352	966154	988804	450083
AK0002	2006	12292	948583	1014758	429566
AK0002	2007	13416	872493	1092235	450565
AK0002	2008	12765	840113	1087636	445026
AK0003	2004	486	3260	3992	1199
AK0003	2005	490	2978	2453	875
AK0003	2006	480	3329	4432	1628
AK0003	2007	480	2820	1643	2848
AK0003	2008	500	2028	1400	707

### 1.2.1 Data sources

Data are typically collected using original surveys and from existing data sets. Because there may be significant costs in collecting your own data, it is typically the case that much research uses data that has already been collected. Examples of data sources include:

- Bureau of Labor Statistics
- U.S. Census
- Current Population Survey (CPS)
- Agricultural Census
- USDA Risk Management Agency (RMA)
- USDA National Agricultural Statistics Service (NASS)

It is important to realize that *all* data must be handled with care. You must spend time inspecting and cleaning your data – being careless can lead to unexpected and often incorrect results. For example, suppose that you examine prices of shoes over the past 50 years. However, you forget to adjust for inflation. As a result, you forecast that the price of shoes in 2015 will be \$500/pair.

### 1.2.2 Functional form issues

When specifying an econometric model, an important realization is that there may be numerous relationships between a dependent variable  $Y$  and a particular independent variable  $X$ . For example, a simple assumption is to assume that  $Y$  is a linear function of  $X$ ; that is,  $Y = \beta_0 + \beta_1 X_1$ . This implies that for each additional unit of  $X_1$ ,  $Y$  increases at the same rate. Alternatively, the dependence of  $Y$  on  $X$  may be non-linear. For example, suppose that the relationship between  $Y$  and  $X$  is as follows:  $Y = \beta_0 + \beta_1 \ln X_1$ . These relationships are shown in figure 1.1 and figure 1.2.

The choice of a functional form is important because it represents your hypothesis about the relationship between  $Y$  and each  $X$ . Often, you can get an idea of these relationships by examining the data (e.g., plotting the relationship between the dependent and independent variables). However, imposing a wrong functional relationship can lead to poor inferences.

Figure 1.1:  $Y$  is a linear function of  $X$

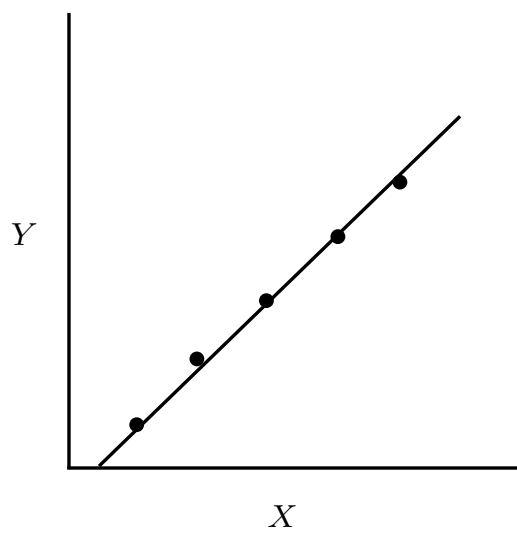
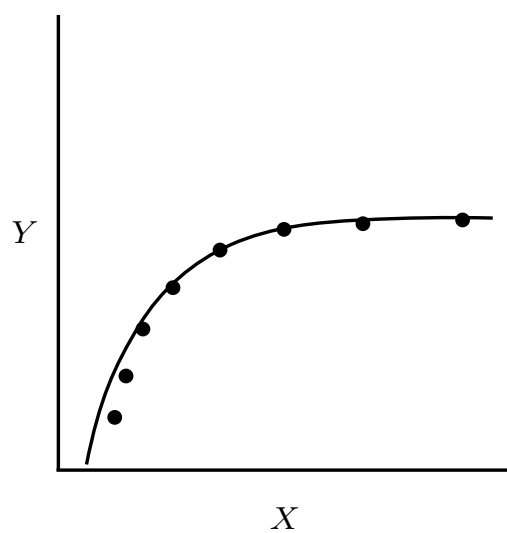


Figure 1.2:  $Y$  is a non-linear function of  $X$



## 1.3 Conducting an econometric project

1. *Pick a topic / pose a question* – when completing this step, you should remember to ask yourself why you are choosing a particular topic. That is, why does answering this question matter? Who cares that this question be answered? Is it possible to answer this question? Is there data available for answering the question?
2. *Construct a theoretical economic model* to answer the question.
3. *Specify an econometric model.*
4. *Gather data.*
5. *Estimate the econometric model.* The estimation should be such that the results are precise and there is a low margin of error.
6. *Interpret the results and perform statistical tests.*
7. *Evaluate whether econometric results are congruent with the theoretical model.* That is, do your outcomes make economic sense?

## 1.4 Additional resources

Several textbooks were used as references for the material presented in these lecture notes. If you are interested to find more in-depth information on topics found in these notes as well as additional issues in econometrics, I encourage you to refer to these texts.

The list found in the *Econometric References and Resources* section is by no means comprehensive, but it is a good start.

# Chapter 2

## Foundation of Probability Theory

Most empirical economics is non-experimental. That is, researchers do not control treatments. Rather, treatments, explanatory variables, and associated outcomes are simply observed. However, it is possible to come up with a framework that can relate the observed treatments and explanatory variables to outcomes.

What's the catch? There is almost never 100% *a priori* certainty that a particular treatment and/or explanatory variable will lead to a particular outcome. That is:

$$Outcome = f(\text{explanatory variables, treatments}) + \text{random error}$$

Because of the error term, repeating an experiment using the identical explanatory variables and treatments may not yield the same outcome as a preceding iteration of the experiment. In other words, the *Outcome* variable is a random variable.

### 2.1 Random Variables

Random variable: a variable whose values are unknown prior to carrying out an experiment. Another way to think about is that a random variable can have different outcomes depending on a “state of nature.”

There are two types of random variables: *discrete* and *continuous*.

### 2.1.1 Discrete random variables

Discrete random variables are ones that take on a limited number of distinct values. That is, the outcomes are countable. Examples include: yes/no replies; number of cattle on a ranch; number of children.

Recall the question we asked earlier: What is the effect of advertisement attendance of farmers' markets? We can set up an advertisement campaign and *count* the number of additional people attending the farmer's market with the advertisement campaign relative to the farmer's market without the promotion.

What is the outcome variable? Why is the outcome variable a *random variable*?

### 2.1.2 Continuous random variables

Continuous random variables are ones that can take on an infinite number of outcomes. For example, a continuous random variable can be any value on the interval  $[0,1]$ . Examples include: farm yields, prices, birth weights.

Suppose that we ask the following question: How does attendance of a farmer's market affect the price of tomatoes sold at the market? Again, we can measure the number of people attending the farmer's market and the prices of tomatoes.

What is the outcome variable? Why is the outcome variable a *random variable*?

## 2.2 Depicting Random Variables

Because random variables cannot be predicted with certainty, it is useful to have a way to "describe" these variables. We do this by assigning *probabilities* to certain outcomes. That is, although no outcome is certain to occur every single time, some outcomes may be more probable than others.

Example 1: If you have six-sided fair die, what is the probability of rolling a "one?" What about rolling a "two?"

### 2.2.1 Discrete random variables

Let's consider another example. Suppose you are measuring attendance of women at a farmer's market. A general way to indicate whether the next person at the farmer's market is a woman is:

$$P[\text{Woman} = \text{Yes}] = \phi$$

$$P[\text{Woman} = \text{No}] = 1 - \phi$$

where  $\phi$  is some probability that a person at a farmer's market is a woman.

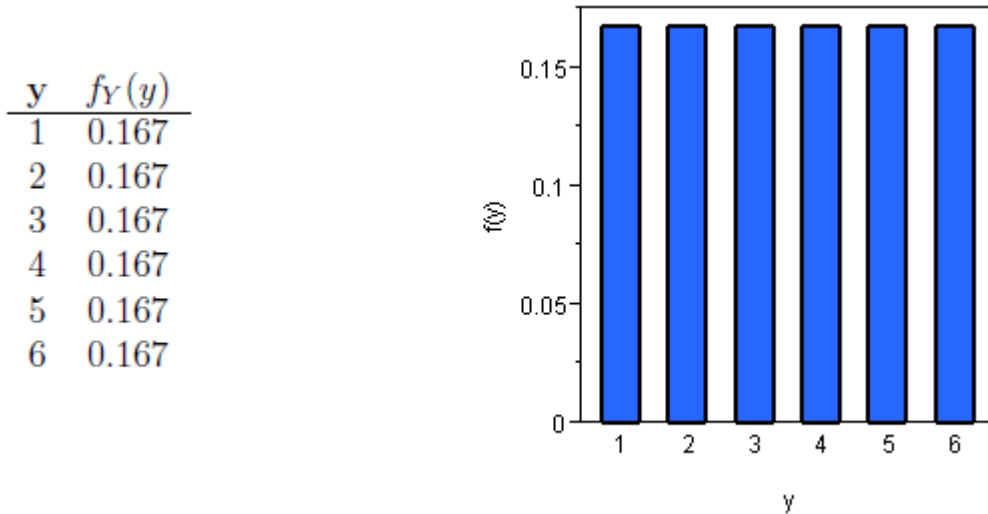
We can summarize this information using a *probability density function* (pdf). In the case of a discrete random variable, the pdf describes the probability that a random variable will take on a particular value. In notation:

$$\text{pdf} = f_Y(y) = P[Y = y]$$

where  $Y$  is the random variable and  $y$  is the outcome that the random variable takes on.

Two ways that a pdf can be described are with a table or a plot are presented in figure 2.1.

Figure 2.1: Discrete Probability Density Function



Example: Calculate  $P[Y \leq 3]$ .

$$P[Y \leq 3] = P[Y = 1] + P[Y = 2] + P[Y = 3] = \frac{1}{2}$$

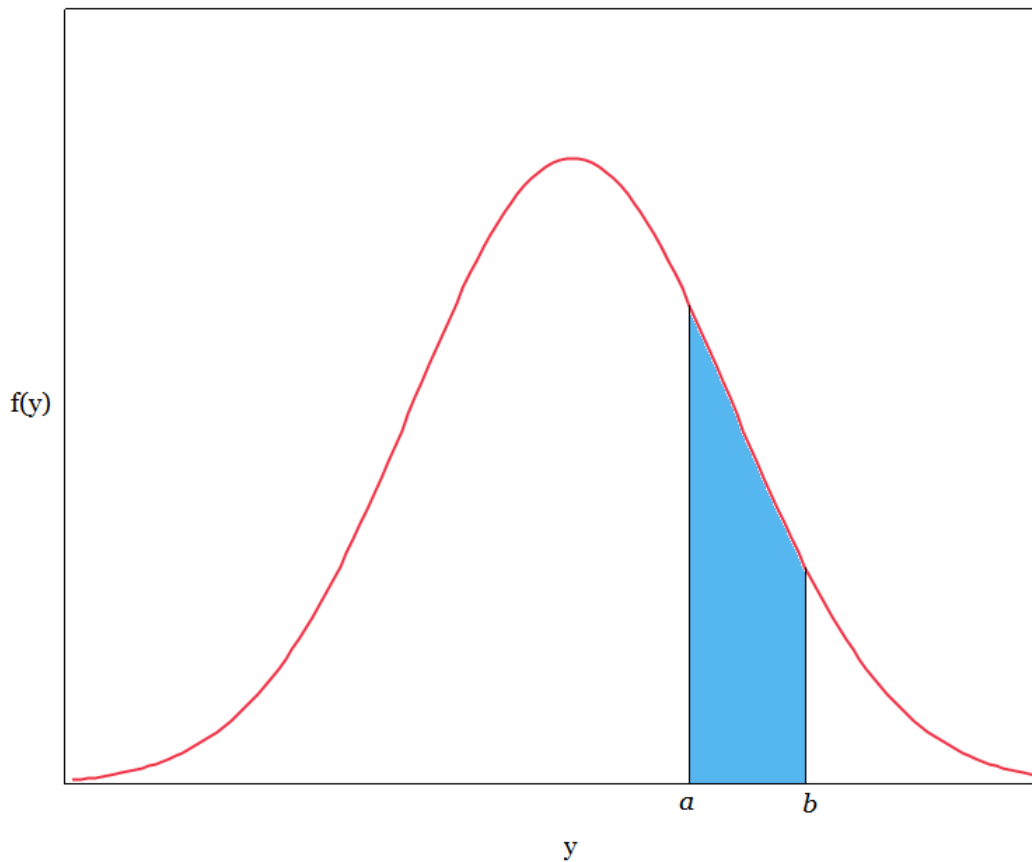
### 2.2.2 Continuous random variables

Summarizing continuous random variables is a bit less intuitive. Because a random variable  $Y$  can take on an unlimited set of values, the probability that  $Y$  takes on any unique value is zero. So, instead of attempting to identify the probability of  $Y$  attaining a particular value, we determine the probability that  $Y$  takes on a value within a range of values. In notation:

$$\text{pdf} = P[a \leq Y \leq b] = \int_a^b f_Y(y)dy$$

The integration sign indicates that you are trying to find the area under the function  $f_Y(y)$  between the points  $y = a$  and  $y = b$ . This is shown in figure 2.2.

Figure 2.2: Continuous Probability Density Function



### Properties of continuous density functions

1. The probability that an outcome  $Y$  is between some range of the distribution must be greater than zero. That is:

$$P[a \leq Y \leq b] = \int_a^b f_Y(y)dy \geq 0$$

2. The total area under a pdf curve is equal to one. In other words, the probability that you will observe a value over the entire range of outcomes is 100%. That is:

$$P[-\infty \leq Y \leq \infty] = \int_a^b f_Y(y)dy = 1$$

## 2.3 Cumulative Density Function

When working with continuous random variables, we can use the cumulative density function (CDF) to describe the random variable. A CDF describes the probability that the random variable  $Y$  assumes a value that is less than or equal to  $y$ .

$$F_Y(y) \equiv P[Y \leq y]$$

### 2.3.1 Discrete random variables

In the case of discrete random variables, the CDF for some outcome  $y$  is the sum of all pdf values such that  $\tilde{y} \leq y$ . For example, consider the pdf table from above. The CDF for each value is as follows:

$y$	$f_Y(y)$	$F_Y(y) \equiv P[Y \leq y]$
1	0.167	0.167
2	0.167	0.334
3	0.167	0.501
4	0.167	0.668
5	0.167	0.835
6	0.167	1

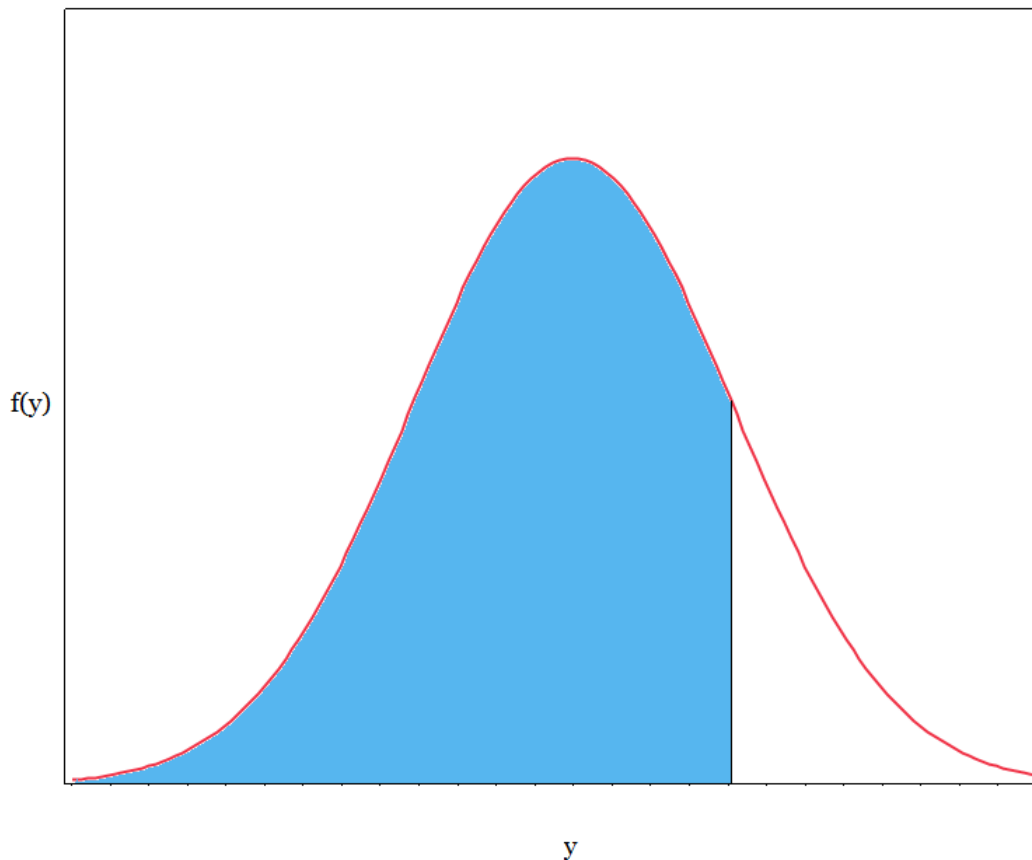
### 2.3.2 Continuous random variables

In the case of continuous random variables, the CDF is the area under the pdf to the left of the an outcome  $y$ . Consider the continuous pdf from above. The CDF can be calculated as follows:

$$\text{CDF} = F_Y(y) = \int_{-\infty}^y f_Y(u) du$$

where  $f_Y(u)$  is the pdf of  $Y$ . Graphically, this is shown in figure 2.3.

Figure 2.3: Continuous PDF



### 2.3.3 Properties of CDFs

1.  $F_Y(y)$  is simply a probability. Thus,  $0 \leq F_Y(y) \leq 1$ .

2. The CDF is an increasing function of  $y$ . That is, if  $y_1 \leq y_2$ , then  $F_Y(y_1) \leq F_Y(y_2) \equiv P[Y \leq y_1] \leq P[Y \leq y_2]$ .
3. For any number  $c$ , it is equivalent to state that:  $P[Y > c] = 1 - P[Y \leq c] \equiv 1 - F_Y(c)$ .
4. For any numbers  $a$  and  $b$  such that  $a \leq b$ , the following holds:

$$P[a \leq Y \leq b] = F_Y(b) - F_Y(a)$$

## 2.4 Describing Random Variables

Probability and cumulative density functions provide a large amount of information about a random variable. However, at times we are simply interested in simpler descriptive statistics that can give us a feeling about what the random variable's distribution is without giving us the entire distribution. This is especially helpful when there are a large number of observations, which is typical of many data sets. There are three common descriptive statistics: central tendency, dispersion, and correlation/covariance.

### 2.4.1 Central tendency

The central tendency describes typical outcomes of a random variable. For example, if you look at the ages of 1,000 college students, you might find that the central tendency of age is 19.5 years of age. That is, if you asked a group of 1,000 college students their age, you will find that 19.5 years old is the answer that you will receive with highest probability.

The most common descriptor of central tendency is *expected value*. Expected value is a weighted average of all possible outcomes of a random variable  $Y$ . In other words, you weigh each possible outcome by the probability of that outcome occurring.

#### Discrete random variables

To calculate the central tendency for a discrete random variable  $Y$ , compute:

$$E(Y) = \sum_{n=1}^N (y_n \cdot P[Y = y_n]) \equiv \sum_{n=1}^N (y_n \cdot f_Y(y_n))$$

Example: Suppose that you have the grades from an undergraduate econometrics exam and the probability of each grade occurring. Compute the central tendency.

<b>y = Score</b>	<b><math>f_Y(y)</math></b>
100	0.2
95	0.2
90	0.2
88	0.2
76	0.2

What do we typically use as the expected value of a random variable? That is, if we know nothing about the probability of the occurrence of a certain outcome, how do we determine its central tendency?

### Continuous random variables

The analogous calculation for determining expected values of a continuous variable involves an integration under the pdf:

$$E(Y) = \int_{-\infty}^{\infty} y \cdot f_Y(y) dy$$

Example: Suppose the  $Y \sim \text{uniform}(0, 5)$ . The pdf for a uniform distribution is:  $f_Y(y) = \frac{1}{b-a}$  if  $a < y < b$ , and  $f_Y(y) = 0$  otherwise. Calculate the expected value of  $Y$ .

$$E(Y) = \int_0^5 \left( y \cdot \frac{1}{5-0} \right) dy = \frac{1}{5} \int_0^5 y \cdot dy$$

$$\frac{1}{5} \cdot \left[ \frac{y^2}{2} \right]_0^5 = \frac{1}{5} \cdot \frac{25}{2} = 2.5$$

### Properties of expected values

1. The expected value of a constant  $c$  is the constant itself (e.g., if you get the same grade on every exam, what is your average across exams?)

$$E(c) = c$$

2. The expectation operator can be applied to each term in a linear function.

$$E(cY + d) = cE(Y) + d$$

More generally, if  $c_i$  are constants and  $Y_i$  are random variables, then:

$$E(c_1Y_1 + c_2Y_2 + \dots + c_nY_n) = c_1E(Y_1) + c_2E(Y_2) + \dots + c_nE(Y_n)$$

$$E\left(\sum_{i=1}^n c_i Y_i\right) = \sum_{i=1}^n (c_i E(Y_i))$$

3. Jensen's inequality: for a non-linear function of a random variable  $h(Y)$ , you cannot assume that  $E(h(Y)) = h(E(Y))$ . That is, you cannot assume the expected value of the function of  $Y$  can be determined by the expected value of  $Y$ . Specifically, for a concave function of  $Y$  ( $h''(Y) = 0$ ), Jensen's inequality holds:

$$E(h(Y)) \leq h(E(Y))$$

Example: consider the function of  $Y$ ,  $h = Y^2$ . Suppose that you can have three outcomes (1, 2, 3), each with probability 0.33. Determine the expected value of  $h$ .

$$E(h(Y)) = E(Y^2) \neq (E(Y))^2$$

That is, you cannot approximate the expected value of  $Y^2$  by squaring the expectation of  $Y$ .

$$E(Y^2) = \frac{1}{3} \cdot (1^2 + 2^2 + 3^2) = \frac{14}{3}$$

$$E(Y)^2 = \left(\frac{1}{3} \cdot (1 + 2 + 3)\right)^2 = 4$$

Clearly,  $E(Y^2) \neq E(Y)^2$

## 2.5 Measures of Dispersion

The measure of dispersion for a random variable describes the *variance* (or the *spread*). Variance is a complement to the central tendency because knowing the central tendency and the dispersion, we can develop a good idea of the random variable's properties. Along with variance, we can use *standard deviation* as a descriptor of dispersion.

### 2.5.1 Variance

The variance of a distribution describes the expected measure of how far an outcome of  $Y$  is from the expected value of  $Y$ . In notation, variance is defined as:

$$\text{Var}(Y) = \sigma_Y^2 = E([Y - E(Y)]^2)$$

In other words, variance is the weighted average of squared difference between outcomes of a random variable and the random variable's expected value. Squaring insures that positive differences do not eliminate negative differences. It should be noted that  $\text{Var}(Y)$  is itself a random variable, because it depends on the random outcomes of  $Y$ .

Example: Show that  $\sigma_Y^2 = E([Y - E(Y)]^2) = E(Y^2) - E(Y)^2$

$$\begin{aligned} & E([Y - E(Y)]^2) \\ & E(Y^2 - 2YE(Y) + E(Y)^2) \\ & E(Y^2) - 2E(YE(Y)) + E(Y)^2 \end{aligned}$$

Recall that  $E(Y)$  is the expected value, which is a constant (typically the mean). Let's notate  $E(Y) = \mu$ .

$$\begin{aligned} & E(Y^2) - 2E(Y\mu) + \mu^2 \\ & E(Y^2) - 2\mu E(Y) + \mu^2 \\ & E(Y^2) - 2\mu \cdot \mu + \mu^2 \\ & E(Y^2) - 2\mu^2 + \mu^2 \\ & E(Y^2) - \mu^2 \\ & E(Y^2) - E(Y)^2 \end{aligned}$$

**Properties of variance**

1. Variance is always nonnegative.
2. The variance of a constant is zero:  $\text{Var}(c) = 0$ . Remember, constants don't vary.
3. Adding a constant to a random variable does not change the variance. However, multiplying the random variable by a constant changes the dispersion, which also changes the variance of the random variable.

$$\text{Var}(cY + d) = c^2\text{Var}(Y)$$

Example: Show that  $\text{Var}(cY + d) = c^2\text{Var}(Y)$ .

- (a) First, let's show that adding the constant does not change the variance.  
 Let's specify a new random variable  $Z = (Y + d)$ .

$$\begin{aligned} \text{Var}(Y + d) &= \text{Var}(Z) \\ \text{Var}(Z) &= E(Z^2) - E(Z)^2 \\ &= E((Y + d)^2) - [E(Y + d)]^2 \\ &= E(Y^2 + 2dY + d^2) - [E(Y) + d]^2 \\ &= E(Y^2) + 2dE(Y) + d^2 - E(Y)^2 - 2E(Y)d - d^2 \\ &= E(Y^2) - E(Y)^2 = \text{Var}(Y) \end{aligned}$$

- (b) Next, let's show that multiplying by a constant changes the variance.  
 Again, let's specify  $Z = cY$ .

$$\begin{aligned} \text{Var}(cY) &= \text{Var}(Z) \\ \text{Var}(Z) &= E(Z^2) - E(Z)^2 \\ &= E(c^2Y^2) - [E(cY)]^2 \\ &= c^2E(Y^2) - [cE(Y)]^2 \\ &= c^2E(Y^2) - c^2E(Y)^2 \\ &= c^2[E(Y^2) - E(Y)^2] \\ &= c^2\text{Var}(Y) \end{aligned}$$

## 2.5.2 Standard deviation

The second measure of dispersion is *standard deviation*. Standard deviation is simply the square root of the variance. We take the square root because it allows us to have a more intuitive measure of an outcome's average distance from the expected value.

$$\text{SD}(Y) = \sigma_Y = \sqrt{\sigma_Y^2}$$

### Properties of standard deviation

1. The standard deviation of a constant is zero:  $\text{SD}(c) = 0$ .
2. Adding a constant to a random variable does not change the standard deviation. However, multiplying the random variable by a constant changes the dispersion, which also changes the standard deviation of the random variable.

$$\text{SD}(cY + d) = |c| \cdot \text{SD}(Y)$$

When  $c$  is nonnegative,  $\text{SD}(cY) = c \cdot \text{SD}(Y)$ .

## 2.6 Joint Distributions

So far, we have looked only at the properties of a single random variable. Assuming that we can explain economic relationships with a single variable is similar to saying that incomes of economics professors are not related to any other variables. A more realistic and more interesting scenario would be to examine the relationship between two or more random variables.

For example, suppose that you wanted to examine exam scores for an undergraduate econometrics class. You have a pdf of the grades, and you also have the pdf of student ages. Suppose that you would like to determine the *joint probability* that a certain grade occurs and that the grade is earned by a student of a certain age.

For two discrete random variables  $Y$  and  $Z$ , the joint probability density function of  $(Y, Z)$  is as follows:

$$f_{Y,Z}(y, z) = P[Y = y, Z = z]$$

To determine the joint probability of an outcome, a simplifying assumption that we will make is that the random variable  $Y$  is *independent* from the random variable  $Z$ . That is, the outcome of  $Y$  does not change the probability of an outcome of  $Z$ , and vice versa. This assumption can often be made in even relatively complex situations because there are many factors that can and can't change the probability of an outcome. So, it is often reasonable to assume that an outcome of a certain event will not change the occurrence probability of another event.

For example, consider rolling a fair die. There is always a  $1/6$  chance that you roll a six. Now consider that you want to determine the joint probability that out of 100 undergraduate and graduate college students, a graduate student is chosen and she rolls a six. The chances of choosing a graduate student does not change the probability that a six is rolled. That probability is still  $1/6$ .

Using this assumption, we can define the joint probability density of  $(Y, Z)$  to be:

$$f_{Y,Z}(y, z) = f_Y(y)f_Z(z) = P[Y = y] \cdot P[Z = z]$$

Note that each individual pdf ( $f_Y(y)$ ) is known as the *marginal distribution*.

Example: Suppose that you have the following information about die outcomes and students:

y = Die outcome	$f_Y(y)$	z = Student	$f_Z(z)$
1	0.167	Undergraduate	0.6
2	0.167	Masters	0.3
3	0.167	Ph.D.	0.1
4	0.167		
5	0.167		
6	0.167		

1. Determine the joint probability that a masters student rolls a 3.

$$f_{Y,Z}(3, M) = f_Y(3) \cdot f_Z(M) = P[Y = 3] \cdot P[Z = M]$$

$$f_{Y,Z}(3, M) = (0.167) \cdot (0.3) = 0.0501$$

2. Construct a table of the joint pdf,  $f_{Y,Z}(y, z)$ .

To construct the table, simply find the other joint probabilities:

		<b>Z</b>			
		<b>z</b>	<b>Undgrd</b>	<b>Masters</b>	<b>PhD</b>
		$f_Z(z)$	0.6	0.3	0.1
<b>y</b>	$f_Y(y)$				
1	0.167		0.1002	0.0501	0.0167
2	0.167		0.1002	0.0501	0.0167
<b>Y</b>	3	0.167	0.1002	0.0501	0.0167
	4	0.167	0.1002	0.0501	0.0167
	5	0.167	0.1002	0.0501	0.0167
	6	0.167	0.1002	0.0501	0.0167

Note that adding across a particular row or column of joint probabilities yields the marginal probability for that row or column.

Additionally, the assumption of independence among random variables allows us to come up with a general notation for many random variables. That is, for a series of random variables  $\{Y_1, Y_2, \dots, Y_n\}$ , the joint pdf under the independence assumption can be written as follows:

$$f_{Y_1, Y_2, \dots, Y_n}(y_1, y_2, \dots, y_n) = f_{Y_1}(y_1)f_{Y_2}(y_2) \cdots f_{Y_n}(y_n)$$

### 2.6.1 Covariance

Covariance is an indicator of how, on average, a random variable varies with variation in another random variable. That is, covariance provides an indicator of the direction that a random variable's outcomes move as the outcomes of another random variable move. It is important to note that covariance *does not* reveal the magnitude of co-movements. This is because covariances depend on the units of each random variables. So scaling the units would also change the magnitude of the covariance, but would not change the direction.

Covariance between two random variables  $Y$  and  $Z$  is defined as follows:

$$\text{Cov}(Y, Z) = \sigma_{Y,Z} = E[(Y - E(Y))(Z - E(Z))]$$

$\sigma_{Y,Z} > 0$  implies that when  $Y > E(Y)$ , then  $Z$  tends be greater than  $E(Z)$ .

$\sigma_{Y,Z} < 0$  implies that when  $Y < E(Y)$ , then  $Z$  tends be greater than  $E(Z)$ .

Example: Show that  $\sigma_{Y,Z} = E[(Y - E(Y))(Z - E(Z))] = E(YZ) - E(Y)E(Z)$ .

$$\begin{aligned}\sigma_{Y,Z} &= E[(Y - E(Y))(Z - E(Z))] \\ &= E[YZ - YE(Z) - ZE(Y) + E(Y)E(Z)] \\ &= E(YZ) - E(Y)E(E(Z)) - E(Z)E(E(Y)) + E(E(Y)E(Z))\end{aligned}$$

Recall that we can specify  $E(Y) = \mu_Y$  and  $E(Z) = \mu_Z$  because these are constant means.

$$\begin{aligned}E(YZ) - E(Y)E(E(Z)) - E(Z)E(E(Y)) + E(E(Y)E(Z)) \\ E(YZ) - \mu_Y\mu_Z - \mu_Z\mu_Y + E(\mu_Y\mu_Z) \\ E(YZ) - \mu_Y\mu_Z - \mu_Z\mu_Y + \mu_Y\mu_Z \\ E(YZ) - \mu_Y\mu_Z \\ E(YZ) - E(Y)E(Z)\end{aligned}$$

### Properties of covariances

1. Multiplying a random variable by a constant will change the covariance.

$$\text{Cov}(aY + c, bZ + d) = ab\text{Cov}(Y, Z)$$

2. Variance of a sum of random variables is as follows:

$$\text{Var}(aY + bZ) = a^2\sigma_Y^2 + b^2\sigma_Z^2 + 2ab\sigma_{Y,Z}$$

It is important to note that if  $Y$  and  $Z$  are independent, then  $\text{Cov}(Y, Z) = 0$ .

### 2.6.2 Correlation

An important downside to measuring directional relationships between random variables using covariance is that the covariance is dependent on the units of each random variable. Measuring the relationship between planted area and yields depends on whether acres are measured in acres or hectares.

A unit-less measure used to examine the relationship between variables is *correlation*. Correlation can reveal the strength of the linear relationship between random variables, because it is unit-less. The correlation coefficient is defined as follows:

$$\text{Corr}(Y, Z) = \rho_{Y,Z} = \frac{\text{Cov}(Y, Z)}{\text{SD}(Y)\text{SD}(Z)} = \frac{\sigma_{Y,Z}}{\sigma_Y\sigma_Z}$$

### Properties of correlation

1.  $\rho_{Y,Z} \in [-1, 1]$
2. If  $\rho_{Y,Z} = 0$ , then the two random variables are not correlated. (Note that a correlation of zero *does not* imply independence).
3. For any constants  $a, b, c$  and  $d$ :

$$\begin{aligned} \text{Corr}(aY + c, bZ + d) &= \text{Corr}(Y, Z), \text{ if } a \cdot b > 0 \\ \text{Corr}(aY + c, bZ + d) &= -\text{Corr}(Y, Z), \text{ if } a \cdot b < 0 \end{aligned}$$

## 2.7 Practice with Computing Summary Statistics

Recall the example from above in we were interested in the effect of advertisements on attendance at a farmers' markets. Suppose that you observe the number of posters/flyers in area advertising the farmer's market as well as the attendance. Additionally, you are provided with the information about associated probabilities.

<b>y = Attendance</b>	$f_Y(y)$	<b>z = Posters</b>	$f_Z(z)$
50	0.025	20	0.45
60	0.175	35	0.2
45	0.2	10	0.1
75	0.35	25	0.25
80	0.2		
120	0.05		

Calculate the following:

1. Expected values of  $Y$  and  $Z$ .
2. Variances of  $Y$  and  $Z$ .
3. Correlation between  $Y$  and  $Z$ .

## 2.8 Conditional Probabilities

Correlation and covariances explain the co-movement relationships between two random variables. However, we are often interested in observing the behavior of one random variable *conditional* on an outcome of one or more other random variables. That is, we would like to know the probability distribution of  $Y$  conditional on the knowledge that  $Z$  has attained a particular outcome. For a discrete random variance, this is denoted as follows:

$$f_{Y|Z}(y|z) = P[Y = y|Z = z] = \frac{f_{Y,Z}(y, z)}{g_Z(z)}$$

The second equality tells us that a conditional distribution can be interpreted as the ratio of the joint probability between  $Y$  and  $Z$  and the marginal distribution of  $Y$ .

Specifying continuous conditional distributions is similar:

$$f_{Y|Z}(y|z) = \frac{f_{Y,Z}(y, z)}{g_Z(z)} = \frac{\int \int f_{Y,Z}(y, z) dy dz}{\int f_Z(z)}$$

Example: suppose that you are interested in determining the probability of higher attendance at farmers' markets given that there is advertisement of the market. That is, we would like to know the distribution of attendance outcomes conditional on the number of advertisement fliers posted around town.

$$f_{At|Ad}(at|ad) = P[At = at|Ad = ad]$$

Consider the following pdf that describes the number of extra people attending farmers' markets and number of posted advertisements:

<b>At = Attendance</b>	$f_{At}(at)$	<b>Ad = Ads</b>	$f_{Ad}(ad)$
50	0.025	20	0.45
60	0.175	35	0.2
45	0.2	10	0.1
75	0.35	25	0.25
80	0.2		
120	0.05		

We can use our knowledge of joint probabilities to calculate the joint pdf for these two random variables. Let's suppose that we are interested in the probability that there will be 60 extra people conditional on 20 additional fliers. Assuming independence:

$$P[At = 60, Ad = 20] = P[At = 60] \cdot P[Ad = 20] = (0.175 \times 0.45) = 0.0785$$

Now we can calculate the conditional probability by dividing the joint probability by the probability that the number of posted fliers will be 20. That is:

$$f_{At|Ad}(at|ad) = \frac{P[At = 60, Ad = 20]}{P[Ad = 20]} = \frac{0.0785}{0.45} = 0.175$$

(Note that the conditional probability of 60 additional people is the same as the unconditional probability of this higher attendance. This only holds if two random variables are independent.)

### 2.8.1 Conditional expectations

As with unconditional random variables, we would like to find a way to describe conditional random variables. That is, we may be interested in finding the expected values of a random variable. For example, suppose that we want to know the central tendency of yields given a certain level of fertilizer. Or, we would like to know the most likely number of wins for a baseball team conditional on the total player payroll of that team.

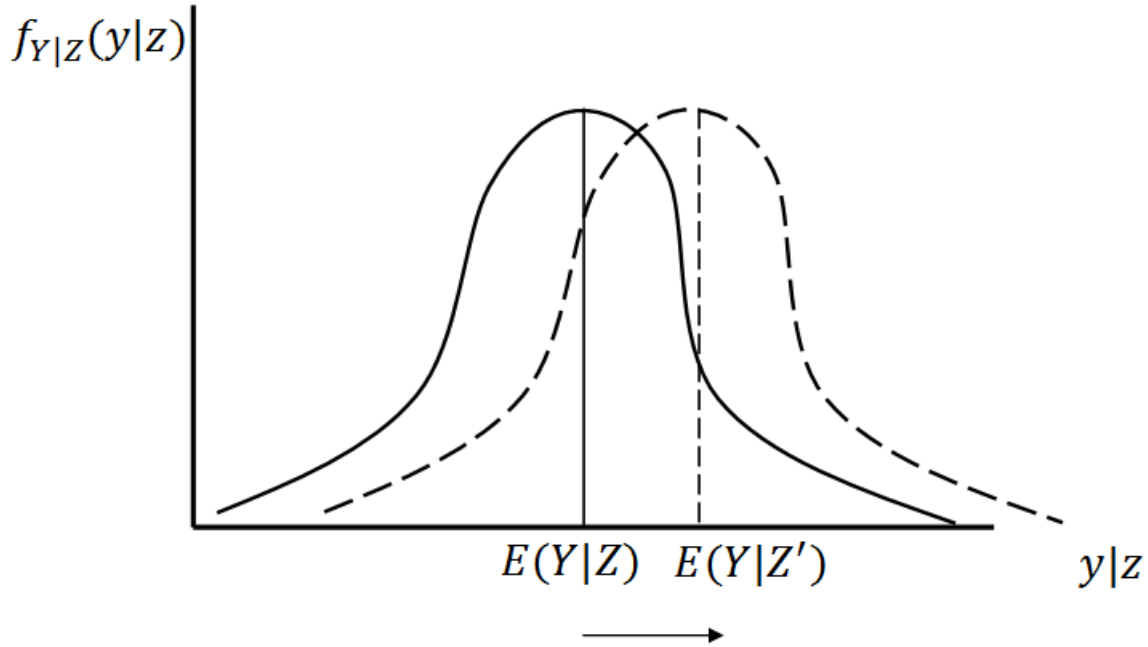
This *conditional expectation* is also often known as the *conditional mean*, and an integral concept for discussing least squares regressions. Another intuitive way of thinking about the conditional mean asking the question: how does the mean level of  $Y$  change with changes in  $Z$ ? Figure 2.4 illustrates a *location* shift the conditional mean of  $Y$  given a change in  $Z$ .

For discrete random variables, we can formalize the conditional mean to be as follows:

$$E(Y|Z = z) = \sum_{i=1}^n y_i f_{Y|Z}(y_i|z) = \sum_{i=1}^n y_i \frac{f_{Y,Z}(y_i, z)}{f_Z(z)}$$

This shows that the expected value of  $Y$  is still a weighted average (as the case for unconditional expectations), but now the weights depend on the outcome of the random variable  $Z$ .

Figure 2.4: Location Shift of the Conditional Distribution



Similarly, the continuous case is as follows:

$$E(Y|Z = z) = \int_{-\infty}^{\infty} y f_{Y|Z}(y|z) dy = \int_{-\infty}^{\infty} y \frac{f_{Y,Z}(y_i, z)}{f_Z(z)} dy$$

Example: suppose that you find that attendance at farmers' markets and advertisements are not independent (for example, organizers of the farmers market may put up more fliers if less new people come). After carefully searching information sources, you find the following joint probability function of attendance and advertisement:

		<b>ad</b>	10	20	30
		$f_{Ad}(ad)$	0.45	0.35	0.2
<b>at</b>	$f_{At}(at)$				
50	0.65		0.2925	0.2275	0.13
60	0.25		0.1125	0.0875	0.05
75	0.1		0.045	0.035	0.02

You're interested in determining the expected value of additional attendance given that there are 20 posted advertisements. That is,  $E(At|Ad = 20)$ . This is as follows:

$$\begin{aligned}
 E(At|Ad = 20) &= \sum_{i=1}^n at_i \frac{f_{At,Ad}(at_i, ad = 1)}{f_{Ad}(ad = 1)} \\
 &= \frac{(50 \times 0.2275) + (60 \times 0.0875) + (75 \times 0.02)}{0.35} \\
 &= \frac{17.25}{0.35} \\
 &\approx 52
 \end{aligned}$$

That is, if there are 20 fliers posted for advertisements, you are expected to attract approximately 52 additional farmers' market patrons.

### 2.8.2 Properties of conditional expectations

1. The conditional expectation of a function of  $Y$  is just the function of  $Y$ :

$$E(h(Y)|Y) = h(Y)$$

2. For a linear function with several random variables, applying the conditional expectation operator of  $Y$  only affects random variables other than  $Y$ :

$$E([h(Y)Z + k(Y)]|Y) = h(Y)E(Z|Y) + k(Y)$$

3. If two random variables are independent, then the conditional expectation of one random variable given another is the same as the unconditional expectation:

$$E(Y|Z) = E(Y)$$

4. Law of Iterated Expectations – the unconditional expected value of  $Y$  is the expectation of its conditional expectations:

$$E[E(Y|Z)|Z] = E(Y)$$

5. Law of Iterated Expectations – expanded form:

$$E[E(Y|Z, W)|Z] = E(Y|Z)$$

Example: suppose we want to determine corn yields based on observable characteristics  $Z$  (e.g. amount of fertilizer, farmer's experience and education, etc) and unobserved characteristics  $W$  such as ability of the farmer. We can't observe the farmer's ability level, but we can retrieve its expected value provided observable variables. That is:

$$\begin{aligned} E(Y|Z) &= E[E(Y|Z, W)|Z] \\ &= E[(\beta_1 z_1 + \beta_2 z_2 \dots + \omega W)|Z] \\ &= \beta_1 z_1 + \beta_2 z_2 \dots + \omega E(W|Z) \\ &= \beta_1 z_1 + \beta_2 z_2 \dots + \omega(\phi_1 z_1 + \phi_2 z_2 \dots) \\ &= \beta_1 z_1 + \beta_2 z_2 \dots + \omega\phi_1 z_1 + \omega\phi_2 z_2 \dots \\ &= (\beta_1 + \omega\phi_1)z_1 + (\beta_2 + \omega\phi_2)z_2 + \dots \\ &= \eta_1 z_1 + \eta_2 z_2 \dots \eta_n z_n \end{aligned}$$

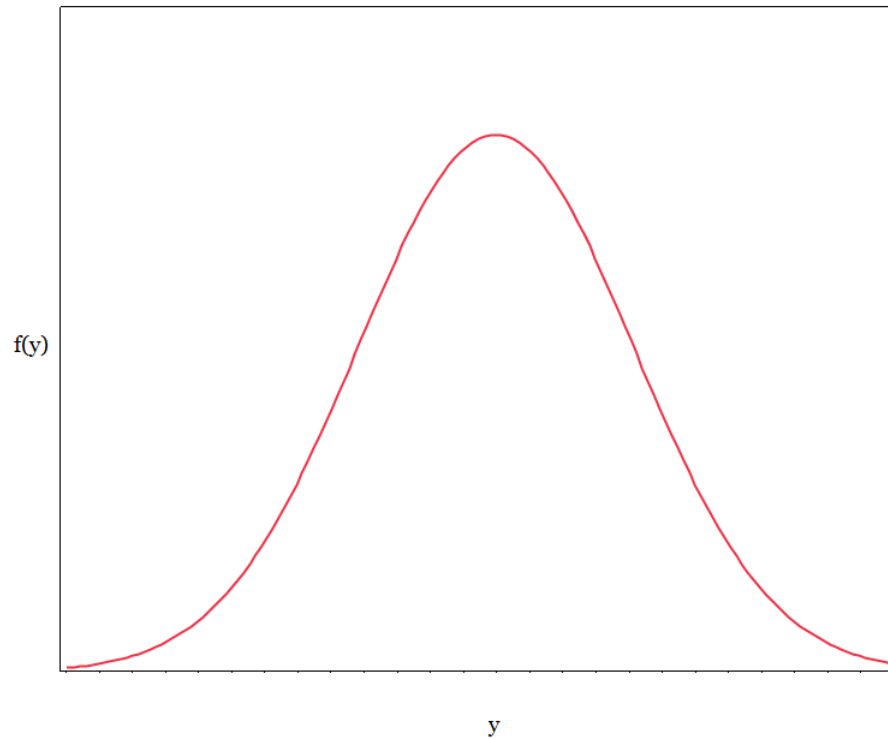
## 2.9 Common Distributions

Many random variables in empirical economics are continuous. There are numerous probability density functions that exist for describing these random variables. The one that is most often used is the *normal distribution*. Additionally, the *central limit theorem* provides a powerful result that makes the normal distribution the crux of many statistical analyses.

### 2.9.1 Normal distribution

A normally distributed random variable is one that is continuous and can take on any value. The normal distribution can also be called a *Gaussian distribution*, after the statistician C.F. Gauss. The distribution has a bell shape, as shown in figure 2.5.

Figure 2.5: Normal Distribution



The pdf of a random variable  $Y$  is defined as follows:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp \left\{ \frac{-(y - \mu_Y)^2}{2\sigma_Y^2} \right\}$$

To specify that a random variable  $Y$  is distributed normally, we use the notation:  $Y \sim N(\mu_Y, \sigma_Y^2)$ , where  $\mu_Y$  is the mean and  $\sigma_Y^2$  is the variance.

When a random variable is not normally distributed, it may be possible to transform the variable such that its distribution becomes normal. For example, there are many cases in which taking the natural log of outcomes for a particular random variable will cause the logged distribution to be normal. This is called a *log-normal distribution*.

### Properties of normal distributions

1. If  $Y \sim N(\mu_Y, \sigma_Y^2)$ , then  $aY + c \sim N(a\mu_Y + c, a^2\sigma_Y^2)$

2. Any linear combination of independent, identically distributed (i.i.d.) normal random variables has a normal distribution.
3. The normal distribution is symmetric about its mean. That is, mean = median = mode.

## 2.9.2 Standard normal distribution

The standard normal distribution is a special case of the normal. If a variable is distributed with a standard normal, it is notated as:  $Y \sim N(0, 1)$ . This implies that the mean of the random variable is zero and the variance (and standard deviation) is one. The pdf of a standard normal distribution typically denoted by  $\phi$  and is as follows:

$$f_Y(y) = \phi_Y(y) = \frac{1}{\sqrt{2\pi}} \exp \left\{ \frac{-y^2}{2} \right\}$$

### Central Limit Theorem

There are two important properties that make the normal distribution very powerful in statistics and econometrics.

1. If  $Y \sim N(\mu_Y, \sigma_Y^2)$ , then  $\frac{(Y-\mu_Y)}{\sigma_Y} \sim N(0, 1)$ . That is, we can turn any normal distribution into a standard normal distribution by standardizing the random variable (i.e. subtracting its mean and dividing through by the standard deviation).
2. The *Central Limit Theorem* (CLT) states that if a random variable is standardized, then *any* random variable will have a standard normal distribution as the number of observations (sample size,  $n$ ) goes to infinity.

Suppose that a random variable  $Y$  has outcomes  $(y_1, y_2, \dots, y_n)$  which are distributed *i.i.d.*  $\sim (\mu_Y, \sigma_Y^2)$ . Additionally, you only have a sample of outcomes (e.g. you are unable to survey all students from a university, but you've surveyed 100 as a representative sample). Define  $\bar{Y}$  be the mean of the sample. Lastly, suppose that we don't know the true distribution of  $Y$ . Then we can standardize the sample observations as follows.

$$Z = \frac{\bar{Y} - \mu_Y}{\sqrt{\text{Var}(\bar{Y})}} \equiv \frac{\bar{Y} - \mu_Y}{\text{SD}(\bar{Y})/\sqrt{n}}$$

As  $n \rightarrow \infty$ ,  $Z$  converges to  $Z \sim N(0, 1)$ . This is a powerful property, because it allows us to simplify many statistical tests.

Note:  $\sqrt{\text{Var}(\bar{Y})} \equiv \text{SD}(\bar{Y})/\sqrt{n}$  because:

$$\begin{aligned}\sqrt{\text{Var}(\bar{Y})} &= \text{Var}\left(\frac{\sum y_i}{n}\right) \\ &= \frac{1}{n^2} \text{Var}(\sum y_i)\end{aligned}$$

Because we've assumed that  $(y_1, y_2, \dots, y_n)$  are distributed *i.i.d.* (independently), the variance of the sum is simply the sum of the variance. That is:

$$\begin{aligned}\frac{1}{n^2} \text{Var}(\sum y_i) &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(Y_i) \\ &= \frac{1}{n^2} n \sigma_Y^2 \\ \frac{1}{n} \sigma_Y^2 &= \sqrt{\frac{\sigma_y^2}{n}} \\ \text{SD}(\bar{Y})/\sqrt{n}\end{aligned}$$

# Chapter 3

## Mathematical Statistics

Probability density functions and summary statistics are important foundations for statistical inferences of a random variable. However, when we talked about each particular topic, we assumed that we knew the entire *population*; that is, all possible entities with a set of characteristics. At times, gathering information about the population is not difficult. For example, suppose you wanted to study test taking behaviors of students in this class. Collecting information from 8 or 9 people would not be very time consuming. However, imagine studying behaviors of U.S. mining companies. In order to study the population, you would need to find information about *all* mining companies located in the U.S. This may turn out to be an extremely time consuming and difficult task.

Because population data are difficult to gather, applied economists focus on studying a representative *sample* of a population. The term “representative” is important, because a good data sample should be a “miniature” population. This would allow you to make appropriate inferences about the population. One of the easiest methods to sample is using a *random sample* technique. This technique has the following properties:

1. Each element of the population has an equal probability of being selected into the sample.
2. The selection of one element does not affect the probability of another element being selected – *independent*.
3. All selected sample observations are from the same population pdf – *identically distributed*.

From these properties, there are two inferences:

1. Sampled observations are *independent, identically distributed (i.i.d.)*.
2. Two different random samples from the same population may yield different inferences about the population.

### 3.1 Statistics of Samples

Suppose that there is a simple random sample consisting of  $n$  observations (random variables)  $\{Y_1, Y_2, \dots, Y_n\}$ . The sample is drawn from a population that is characterized by some parameter  $\theta$ . That is,  $\theta$  can be the mean of the population pdf. Because we don't know the actual value of  $\theta$ , we must come up with an *estimator* for the value. If we know that a simple random sample is a "miniature representation" of the population, then our best guess as to the mean of the population would be the mean of the sample.

To calculate the sample mean, we simply average the outcomes of the sample random variables  $\{Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n\}$ :

$$\hat{\theta} = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

where  $\hat{\theta}$  is the estimator of  $\theta$ . It is necessary to note that  $\hat{\theta} = \bar{Y}$  should *not* be interpreted in the same way as a population mean, which we estimated in section 2.4.1. While the population mean is the actual central tendency of the population, the sample mean is a *point estimator* of the population mean.

Of course, we could also choose some other estimation of the parameter  $\theta$ . For example, we might say that:

$$\hat{\theta} = \frac{X_{max} + X_{min}}{2}$$

So, which  $\hat{\theta}$  is the best approximation of  $\theta$ ?

## 3.2 Sampling distribution

The sample mean is the expected value of drawn samples, which are random variables. Thus, the sample mean is itself a random variable, because the actual outcomes of each random variable will vary with each sample. This implies that the sample statistics also have probability distributions. *Sampling distributions* describe the estimator  $\hat{\theta}$  across different random samples. Using sampling distributions, we can create rules that can reveal which estimator is best in describing the true descriptor of the population distribution.

### 3.2.1 Unbiasedness

An estimator is *unbiased* if the expected value of the estimator is equal to the population parameter being estimated. That is:  $E(\hat{\theta}) = \theta$ . Be careful: unbiasedness does *not* imply that any one estimator is equal to the population parameter. Rather, unbiasedness means that if you indefinitely draw random samples and calculate  $\hat{\theta}$ , the average of these  $\hat{\theta}$  would be equal to  $\theta$ .

Using the definition of unbiasedness, we can define *bias* as follows:

$$Bias = E(\hat{\theta}) - \theta$$

Example: Show that  $\hat{\theta} = \frac{Y_1 + Y_2}{2}$  is an unbiased estimator of the population mean, where  $Y \sim N(\mu_Y, \sigma_Y^2)$ .

To show unbiasedness, we need to take the expected value of  $\hat{\theta}$ .

$$E(\hat{\theta}) = E\left(\frac{Y_1 + Y_2}{2}\right)$$

$$\frac{1}{2}(E(Y_1) + E(Y_2))$$

$$\frac{1}{2}(\mu_Y + \mu_Y)$$

$$E(\hat{\theta}) = \mu_Y$$

Thus, this estimator is an unbiased estimator of the population mean.

Example: Using the above result, generalize to show that the sample average is an unbiased estimator of the population mean for *any* underlying population distribution.

$$\begin{aligned} E(\hat{\theta}) &= E(\bar{Y}) = E\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) \\ &= \frac{1}{n} E\left(\sum_{i=1}^n Y_i\right) \end{aligned}$$

Because the sample observations are *i.i.d.*, the expected value of the sum is the sum of the expected value.

$$\begin{aligned} &\frac{1}{n} E\left(\sum_{i=1}^n Y_i\right) \\ &= \frac{1}{n} \sum_{i=1}^n E(Y_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mu_Y \\ &= \frac{1}{n} n \mu_Y \\ E(\hat{\theta}) &= E(\bar{Y}) = \mu_Y \end{aligned}$$

### 3.2.2 Sample variance and efficiency

Unbiasedness provides one method to determine the appropriateness of an estimator. However, we may find that there are more than one unbiased estimator. This brings us back to our original dilemma, with a twist: which unbiased estimator is the best? To answer this question, we need another measure: *sample variance*.

Consider the properties of two sample estimators:

- Estimator  $\hat{\theta}_1$ 
  - Expected value of the estimator is unbiased.
  - After repeating several trials of resampling and calculating the estimator, you find that some of the estimators were very far from  $E(\hat{\theta})$  and some were very close to  $E(\hat{\theta})$ .

- Estimator  $\hat{\theta}_2$ 
  - Expected value of the estimator is unbiased,  $E(\hat{\theta}) = \theta$
  - After repeating several trials of resampling and calculating the estimator, you find that although some deviations from  $E(\hat{\theta})$  occurred, these deviations were not very large.

Which estimator is better,  $\hat{\theta}_1$  or  $\hat{\theta}_2$ ? It's intuitive that the second estimator is likely better because there is a better chance that if you were to draw another random sample, the estimator value will be closer to the true population mean. That is, using  $\hat{\theta}_2$ , there is less chance that you will get a very "wrong" representation of the population.

To determine the *sampling variance* around the sample mean, we can use the following:

$$\text{Var}(\bar{Y}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{\sigma^2}{n}$$

That is, the sampling variance is equal to the population variance divided by the sample size,  $n$ . An unbiased estimator for  $\sigma^2$  is:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Note that the variance of the sampling mean goes to zero as the sample size goes to infinity. That is, the more observations you have, the more accurate the estimator of the population mean.

### Efficiency

Comparing sampling variance of estimators leads to the concept of *estimator efficiency*. For any two estimators  $\hat{\theta}_1$  and  $\hat{\theta}_2$ ,  $\hat{\theta}_1$  is more efficient than  $\hat{\theta}_2$  if:

$$\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2)$$

Note that comparing variances is only useful for unbiased estimators. Otherwise, we could simply choose an estimator whose sampling variance is zero.

### 3.3 Asymptotic Properties

If you find an estimator to be unbiased and efficient, it remains unbiased and efficient regardless of the sample size. It is often the case that estimator sampling properties depend on the sample size, and you may have sample sizes at which the calculated estimator is neither unbiased nor inefficient. However, as the sample size increases, the particular estimator possesses the appropriate properties. That is, the sampling distribution of the estimator may change as the sample size increases. In this case, the sampling distribution is referred to as a *asymptotic distribution*.

With finite sample estimators, we were interested in unbiasedness and efficiency. For asymptotic estimators, we consider *asymptotic unbiasedness*, *consistency*, and *asymptotic efficiency*.

#### 3.3.1 Asymptotic unbiasedness

An estimator is asymptotically unbiased if the following holds:

$$\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta$$

That is, as the number of observations grows infinitely larger, the estimator converges to the true, population value of  $\theta$ .

#### 3.3.2 Consistency – probability limit

Consistency is related to unbiasedness. The concept states that as the number of observations increase, the probability that the bias is equal to zero. We state this as follows:

$$\lim_{n \rightarrow \infty} P[|\hat{\theta}_n - \theta| > \varepsilon] = 0$$

That is, as  $n$  grows infinitely in size, the probability that there is a bias ( $|\hat{\theta}_n - \theta| > \varepsilon$ ) is zero. We can also express the above property as:  $plim \hat{\theta}_n = \theta$ .

We can think about the probability limit as the phenomenon in which as the number of observations in a sample increase, the sampling distribution becomes more and more concentrated around the true population parameter,  $\theta$ . If an estimator is not consistent, then we are unable to learn about the population parameter no matter how many observations we obtain.

### Probability limit properties

1. Unbiasedness *does not* imply consistency. Only estimators whose variances tend to zero as  $n$  grows are consistent. That is:

$$\text{If } \lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta \text{ and } \lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) \rightarrow 0$$

then

$$\text{plim } \hat{\theta}_n = \theta$$

2. *Law of Large Numbers*: to estimate the population mean,  $\mu$ , we can get infinitely close by using the sample average and increasing the number of observations in the sample. That is:

$$\text{plim}(\bar{Y}_n) = \mu$$

3. For a continuous function  $g(\cdot)$ :

$$\text{plim } g(\theta_n) = g(\text{plim } \theta)$$

We can use this property to specify a consistent estimator of the population standard deviation.

$$S_n = \sqrt{S_n^2} = \sqrt{\left( \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \right)}$$

4. Central limit theory (see section 2.9.2)

## 3.4 Confidence intervals

In the preceding sections, we have examined *point estimations*. That is, we examined properties of an estimate for the population parameter. However, we said nothing about the probability that the estimator is likely to be the true parameter.

Typically, we assess the degree of uncertainty about the estimator with the sampling standard deviation,  $S$ . That is, we describe the sampling distribution around the estimate,

which provides for a notion of how well we can describe the population parameter with using another sample. However, we still do not know where the true population parameter is located relative to the sample estimate. We can describe this relationship using a *confidence interval*.

The confidence interval is a numerical range within which we have some degree of certainty  $(1 - \alpha)$  that the true population parameter resides. That is, the interval around the population mean is a lower and upper bound between which we *expect* the population mean to exist with a degree of  $(1 - \alpha)$  certainty. We define  $\alpha$  to be the *significance level*. That is, the confidence interval can be described as follows:

$$LB \leq \mu \leq UB$$

Example: If  $\alpha = 0.05$ , then for 95%  $((1 - \alpha) = 0.95)$  of all random samples, the confidence interval  $LB \leq \mu \leq UB$  contains the true parameter value  $\mu$ . That is:

$$P[LB \leq \mu \leq UB] = 0.95$$

From our discussion of the central limit theorem (CLT), we know that the sampling distribution tends to be normal if (a) the population is normal; or (b) the sample size is large. That is:

$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$$

Of course, we can standardize this to be distributed with a standard normal distribution:

$$Z = \frac{\bar{Y} - \mu_Y}{\sigma/\sqrt{n}} \sim N(0, 1)$$

(a) the population is normal; or (b) the sample size is large. Now, we can write the confidence interval for the standardized  $\bar{Y}$  to be:

$$P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{Y} - \mu_Y}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

Rearranging:

$$P(\bar{Y} - z_{1-\frac{\alpha}{2}} \cdot \sigma/\sqrt{n} \leq \mu_Y \leq \bar{Y} + z_{1-\frac{\alpha}{2}} \cdot \sigma/\sqrt{n}) = 1 - \alpha$$

The term  $z_{1-\frac{\alpha}{2}}$  is the value of  $Z$  for which  $F(Z \leq z_{1-\frac{\alpha}{2}}) = 1 - \alpha/2$ , where  $F(\cdot)$  is the CDF of a standard normal distribution. The division of  $\alpha$  is due to the fact that we are performing a two-tailed test and the normal distribution is symmetric around the mean.

The confidence interval implies that there is a  $(1 - \alpha)$  probability that the standardized sample mean is between the lower and upper bounds.

Example: Suppose that we would like to construct a 95% confidence interval of  $\mu_Y$ . Using a statistical table that describes the cumulative area under the standard normal distribution (table 1 of the appendix), we find that  $-z_{1-\frac{\alpha}{2}} = F(Z \geq 0.9750) = -1.96$  and  $z_{1-\frac{\alpha}{2}} = F(Z \leq 0.9750) = 1.96$ . Thus, the confidence interval is:

$$P(\bar{Y} - 1.96 \cdot \sigma/\sqrt{n} \leq \mu_Y \leq \bar{Y} + 1.96 \cdot \sigma/\sqrt{n}) = 0.95$$

Equivalently, we can write the interval as:

$$[\bar{Y} - 1.96 \cdot \sigma/\sqrt{n}, \bar{Y} + 1.96 \cdot \sigma/\sqrt{n}]$$

This implies that there is a 95% probability that  $\mu_Y$  is in the confidence interval. It is important to note that the interval itself is a *random* and can change with each random sample. That is, *before* the random sample is drawn, there is a 95% probability that  $\mu$  is contained in the interval above. It is *not* the case that the probability that  $\mu$  is in the interval is 95%. Rather, if you draw 100 random samples from a population, 95 times out of 100, the confidence interval around each random sample's  $\bar{Y}$  will contain the true population mean,  $\mu$ .

### 3.4.1 Confidence intervals with small samples

The above calculation of the confidence relied on two assumptions: (a) the population is normal; or (b) the sample size is large. Thus, we could approximate the standard deviation to be  $\sigma = s = \sqrt{(\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2)}$  and use the standard normal CDF to determine the critical values.

If  $n$  is not very large, we can still approximate  $\sigma = s = \sqrt{(\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2)}$  (assuming that the true population is normal, which is a pretty good assumption if you don't know

anything about the population), but we must use critical values from a  $t$ -distribution. That is:

$$\frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

where the subscript of  $t$  denotes the degree of freedom for determining the critical values.

Example: Construct a 95% ( $\alpha = 0.05$ ) confidence interval around  $\mu$  when  $n = 30$ .

We are still interested in a two-tailed test, and we know that there are  $n - 1 = 29$  degrees of freedom. The confidence interval can be constructed as:

$$P(\bar{Y} - t_{1-\frac{\alpha}{2}} \cdot s/\sqrt{n} \leq \mu_Y \leq \bar{Y} + t_{1-\frac{\alpha}{2}} \cdot s/\sqrt{n}) = 1 - \alpha$$

Using table 2, the critical values for  $F(t_{1-\frac{0.05}{2}})$  are  $\pm 2.045$ . So, the interval is:

$$[\bar{Y} - 2.045 \cdot s/\sqrt{n}, \bar{Y} + 2.045 \cdot s/\sqrt{n}]$$

As the sample size increases, the  $t$ -distribution approaches the standard normal. So, if you have a substantially large sample size, you can simply use the critical values from the standard normal CDF.

Note that  $s/\sqrt{n}$  is the *standard error* of a point estimate. The standard error will become an important concept during estimation. However, we use it to show that a confidence interval can be written as follows:

$$[\bar{Y} - t_{1-\frac{\alpha}{2}} \cdot \text{SE}(\bar{Y}), \bar{Y} + t_{1-\frac{\alpha}{2}} \cdot \text{SE}(\bar{Y})]$$

## 3.5 Hypothesis Testing

Point estimators and confidence intervals allow us to determine information about the true population and find out the general accuracy of our information. Additionally, we may be interested in using this information to answer important yes or no questions. Methods used in answering these questions are known as *hypothesis testing*.

Hypothesis tests seeks to measure how strong is the evidence from a particular sample against a particular hypothesized value. To set up a hypothesis test, we need two hypotheses:

1. *Null hypothesis*:  $H_0 : \hat{\theta} = \theta$
2. *Alternative hypothesis*:
  - $H_a : \hat{\theta} < \theta$
  - $H_a : \hat{\theta} > \theta$
  - $H_a : \hat{\theta} \neq \theta$

We presume that we cannot reject the null hypothesis until there is strong statistical evidence that we should reject it. However, in setting up and testing hypothesis, there are two errors that we can make:

1. *Type I error* (false negative): we reject the null hypothesis even though it is actually true in the population.

Example:  $H_0$  : Increased sugar intake leads to poorer dental health.

A Type I error would result in the conclusion that increased sugar intake does not lead to poor dental health.

2. *Type II error* (false positive): we fail to reject the null hypothesis even though it is actually not true in the population.

Example:  $H_0$ : Your entry form for an all expense paid getaway to a tropical island was selected as the winner.

A Type II error would result in the conclusion that you were selected for the trip even though it was actually someone else.

In both cases, we want to avoid making the two error types. However, it is often difficult (if not impossible) to completely eliminate the chance that an error occurs. So we want to minimize the probability that a particular error occurs rather than trying to eliminate the probability all together.

To minimize the probability of committing Type I and II errors, we maximize the *power* of the hypothesis test while minimizing the *significance level* (probability) of a Type I error. That is, first we choose a probability level at which we feel comfortable committing a Type I error. Because we want this probability to be relatively small (i.e., minimize our chances of getting a false negative), we want to choose a low significance level. However, if we choose a very low significance level, then we may never be able to reject the null hypothesis. The significance level is as follows:

$$\alpha = P[\text{Reject } H_0 | H_0 \text{ is true}]$$

That is,  $\alpha$  is the probability that the null hypothesis is rejected, given that the null hypothesis is actually true. A typical value of  $\alpha$  is 0.10, 0.05, and 0.01.

After choosing the significance level, we want to minimize the probability of a Type II error. That is, maximize the power of the hypothesis test in order to maximize the probability of not rejecting a correct hypothesis. The power of the hypothesis is how well it can find deviations from the null hypothesis. It is defined as follows:

$$\pi(\theta) = P[\text{Reject } H_0 | \theta] = 1 - P[\text{Type II error} | \theta]$$

That is, the probability that we reject the null hypothesis when it is actually false. With large samples, Type II errors are less likely because there is more precision in finding deviations from the null,  $H_0$ .

### 3.5.1 Devising a hypothesis test

Once we know the null and alternative hypotheses, we must calculate a test statistic and a critical value used to evaluate the test statistic. In other words, we need to calculate a statistic using a random sample, and then determine whether to reject or fail to reject the null hypothesis, given the critical value that sufficiently minimizes both Type I and Type II errors.

A *test statistic* is a function of a random sample (and is therefore also a random variable). We can think of one example to be the estimator of the population mean. Next, we

determine the *critical value* against which to test the statistic. The rejection or failure of rejection of the null hypothesis will depend on whether the test statistic is above or below the value of the critical value. The critical value is determined by the distribution of the test statistic.

Example: Suppose that the true mean of the distribution is from a normal distribution,  $N(\mu_0, \sigma^2)$ . How would test whether an estimator is equal to the true mean.

1. Set up a null hypothesis (it should always be an equality):

$$H_0 : \bar{Y} = \mu_0$$

2. Define the rejection rule. That is, specify the alternative hypothesis:

- $H_a : \bar{Y} > \mu_0$  (one-sided test)
- $H_a : \bar{Y} < \mu_0$  (one-sided test)
- $H_a : \bar{Y} \neq \mu_0$  (two-sided test)

3. Set up the test statistic:

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{\bar{Y} - \mu_0}{\text{SE}(\bar{Y})}$$

Note that the test statistic is exactly the same as what we calculated for the confidence interval. Similarly, it is distributed as a  $t_{n-1}$ .

4. Choose a significance level,  $\alpha$ . The critical value is  $c = t_{n-1}(1 - \alpha)$ . For a two-tailed test, the critical value is  $c = t_{n-1}(1 - \frac{\alpha}{2})$ .
5. Specify the rejection rule. That is, if the rejection rule is satisfied, then you must reject the null hypothesis in favor of the alternative. Given the three different alternative hypotheses, we have three rejection criteria:
  - $t > c$
  - $t < -c$
  - $|t| > c$
6. Determine if the statistic matches any of the rejection criteria.

Example: Suppose that you are examining the affect of planting a genetically modified crop on the yield of that crop. After collecting data from 20 farms that adopted the GMO crop, you find that the average difference in traditional versus GMO crop yields is 10 bushels per acre. Additionally, you find that the standard deviation is 7.5 bushels per acre. Determine if GMO crops change yields.

1. Null hypothesis –  $H_0$  : no change in yields,  $\Delta y = 0$ .
2. Alternative hypothesis –  $H_a$  : yields have changed,  $\Delta y \neq 0$ .
3. Set up the test statistic:

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{10}{7.5/\sqrt{20}} = 5.96$$

4. Choose a significance level and calculate critical value.  
95% significance  $\rightarrow c = t_{19}(0.025) = 2.093$
5. Specify rejection rule:  $|t| > c$
6. Determine if the absolute value of the test statistic is greater than the critical value.  
The test statistic, 5.96, is greater than the critical value, 2.093. So, we must reject the null hypothesis of no yield change.

Example: Use the preceding information to test whether GMO crops increase yields with a 99% significance level.

This is a one-sided test. With a 99% significance level, the critical value is  $c = t_{19}(0.01) = 2.539$ . Again, because 5.96 is greater than 2.539, we must reject the null hypothesis of no change in favor of the alternative hypothesis that yields have increased.

### 3.5.2 $p$ -values

An alternative method to carry out hypothesis tests is to ask the question: if the null hypothesis is true, then what is the probability that we calculate a sample statistic as large as was calculated by a particular sample? The answer to this is the  $p$ -value.

$$p\text{-value} = P[|T| > c|H_0] = 1 - F(c)$$

The  $p$ -value is the probability that we observe a value that is equal to larger than  $T$  when the null hypothesis is true. When the  $p$ -value is less than the desired significance level, we reject the null hypothesis. Generally, small  $p$ -values provide evidence *against* the null hypothesis. Conversely, high  $p$ -values indicate that there is a high probability that the null cannot be rejected. That is, you can find another statistic that will not reject  $H_0$ . Low  $p$ -values indicate that there is little chance of finding another such statistic.

Example: Suppose that you are testing whether there is a change in precipitation between 1990 and 2010 in the Northern and Southern Great Plains. You use a statistical analysis software to test the difference of inches of rain, and the software spits out the following summary table:

	Northern GP	Southern GP
Difference, inches	0.45	1.98
$p$ -value	0.16	0.01

How do you interpret the results?

In the Northern GP region, the difference in precipitation over the 20 years is 0.45 inches, while in the Southern GP the difference is 1.98 inches. The  $p$ -values is 0.01 for the difference in the Southern GP, which implies that we would there is only a 1% chance that the difference is only due to random error. This indicates that at a 99% significance level, the difference of precipitation in the Southern GP region is different than zero. Alternatively, the  $p$ -value for the Northern GP is 0.16, indicating that there is a 16% chance that this difference is due to random error. That is, there is an 84% significance level that the difference in rain is not zero. This is a pretty low significance level, implying that we cannot reject the null hypothesis that there was no change in precipitation over the last 20 years in the Northern GP.

# Chapter 4

## Matrix algebra

Matrix algebra is a powerful tool for analyzing statistics. Most of the statistical analyses performed by statistical analysis software are performed using matrices. When we begin to learn about linear regressions, matrices will be used extensively. This chapter is intended as an overview and reference guide for using matrices to perform linear regression analysis.

### 4.1 Basic Definitions

- *Vector*: is a row or column of entities (e.g. values, variables). A column vector of dimension  $(n \times 1)$  and a row vector of dimension  $(1 \times m)$  are as follows:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{(n \times 1)} \qquad \mathbf{y} = [y_1 \ y_2 \ \dots \ y_m]_{(1 \times m)}$$

- *Matrix*: a rectangular array. A matrix of dimension  $(n \times m)$  is as follows:

$$\mathbf{y} = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1m} \\ y_{21} & y_{22} & \dots & y_{2m} \\ \vdots & \ddots & & \vdots \\ y_{n1} & y_{n2} & \dots & y_{nm} \end{bmatrix}_{(n \times m)}$$

Notation: we will specify the number of rows in a matrix/vector by  $n$ , and the number of columns in a matrix/vector by  $m$ . **You should always write out the dimensions of matrices/vectors.**

Example: what are the dimensions of the following matrix?

$$\mathbf{y} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}$$

There are two rows and three columns. Thus, the dimensions are  $(2 \times 3)$ .

- *Scalar*: a matrix of dimension  $(1 \times 1)$ . A scalar is a constant.

$$\mathbf{y} = [y_{11}]_{(1 \times 1)}$$

- *Special matrices*: there are a number of matrices that can be identified directly by their name.

- *Square matrix*: a matrix in which the number of rows is the same as the number of columns. That is, it is a matrix of dimension  $(n \times n)$ .
- *Diagonal matrix*: a square matrix in which only the diagonal elements are non-zero. All off-diagonal terms are zero. An  $(n \times n)$  diagonal matrix is as follows:

$$\mathbf{y} = \begin{bmatrix} y_{11} & 0 & \dots & 0 \\ 0 & y_{22} & \dots & 0 \\ \vdots & \ddots & \vdots & \\ 0 & 0 & \dots & y_{nn} \end{bmatrix}_{(n \times n)}$$

- *Identity matrix*: a diagonal matrix with ones on the diagonal and zeros on the off-diagonal. An  $(n \times n)$  identity matrix is as follows:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \\ 0 & 0 & \dots & 1 \end{bmatrix}_{(n \times n)}$$

## 4.2 Matrix Properties and Manipulations

- *Matrix dimensional equality*: two matrices are dimensionally equal if each matrix has the same number of rows and columns as the other matrix. That is, a matrix of

dimensions ( $n \times m$ ) is dimensionally equivalent to any other matrix of dimensions ( $n \times m$ ). Two dimensionally equal matrices are as follows:

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}_{(2 \times 3)} \quad == \quad \begin{bmatrix} 4 & 14 & 32 \\ 22 & 44 & 64 \end{bmatrix}_{(2 \times 3)}$$

- *Matrix addition/subtraction*: elementwise addition/subtraction of dimensionally equal matrices. The resulting matrix is of the same dimensions as the two original matrices. An example of matrix addition is as follows:

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}_{(2 \times 3)} \quad + \quad \begin{bmatrix} 4 & 14 & 32 \\ 22 & 44 & 64 \end{bmatrix}_{(2 \times 3)} \quad = \quad \begin{bmatrix} 5 & 16 & 35 \\ 25 & 48 & 69 \end{bmatrix}_{(2 \times 3)}$$

- *Scalar multiplication/division*: elementwise multiplication of a matrix by a constant. An example of scalar multiplication is as follows:

$$2 \cdot \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}_{(2 \times 3)} \quad = \quad \begin{bmatrix} 2 & 4 & 6 \\ 6 & 8 & 10 \end{bmatrix}_{(2 \times 3)}$$

- *Matrix multiplication*: requires that two matrices are of *appropriate* dimensions. That is, for two matrices  $\mathbf{y}$  and  $\mathbf{z}$ , the product  $\mathbf{yz}$  requires that the number of columns in the matrix  $\mathbf{y}$  is equal to the number of rows in matrix  $\mathbf{z}$ . For example, the following matrices can be multiplied:

$$\mathbf{y} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}_{(2 \times 3)} \quad \mathbf{z} = \begin{bmatrix} 1 & 3 \\ 2 & 1 \\ 3 & 0 \end{bmatrix}_{(3 \times 2)}$$

The first matrix is of dimension ( $2 \times 3$ ) and the second is of dimension ( $3 \times 2$ ). It is always easy to know whether you can multiply two matrices by placing the dimensions next to each, and seeing if the “inside” numbers match. That is, the dimensions ( $2 \times 3$ ) ( $3 \times 2$ ) can be multiplied because the “inside” numbers 3 match. The resulting matrix will have the dimensions of the “outside” numbers. That is, the resulting matrix product  $\mathbf{yz}$  will be of dimensions ( $2 \times 2$ ).

To multiply matrices, the method to remember is “row 1 times column 1 plus row 2 times column 2 plus . . .” That is, a new element in the product matrix  $\mathbf{yz}$  is obtained by multiplying the element in the  $i^{\text{th}}$  row in  $\mathbf{y}$  by the element in the  $j^{\text{th}}$  column of  $\mathbf{z}$ , and adding these products together. Consider how to multiply the following matrices:

$$\begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \end{bmatrix}_{(2 \times 3)} \cdot \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \\ z_{31} & z_{32} \end{bmatrix}_{(3 \times 2)} = \begin{bmatrix} (y_{11} \cdot z_{11} + y_{12} \cdot z_{21} + y_{13} \cdot z_{31}) & (y_{11} \cdot z_{12} + y_{12} \cdot z_{22} + y_{13} \cdot z_{32}) \\ (y_{21} \cdot z_{11} + y_{22} \cdot z_{21} + y_{23} \cdot z_{31}) & (y_{21} \cdot z_{12} + y_{22} \cdot z_{22} + y_{23} \cdot z_{32}) \end{bmatrix}_{(2 \times 2)}$$

Example: Suppose you have matrices  $\mathbf{y}$  and  $\mathbf{z}$ . Find the matrix product  $\mathbf{yz}$ .

$$\mathbf{y} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}_{(2 \times 3)} \quad \mathbf{z} = \begin{bmatrix} 1 & 3 \\ 2 & 1 \\ 3 & 0 \end{bmatrix}_{(3 \times 2)}$$

$$\mathbf{yz} = \begin{bmatrix} (1 \cdot 1 + 2 \cdot 2 + 3 \cdot 3) & (1 \cdot 3 + 2 \cdot 1 + 3 \cdot 0) \\ (3 \cdot 1 + 4 \cdot 2 + 5 \cdot 3) & (3 \cdot 3 + 4 \cdot 1 + 5 \cdot 0) \end{bmatrix}_{(2 \times 2)} = \begin{bmatrix} 14 & 5 \\ 26 & 13 \end{bmatrix}_{(2 \times 2)}$$

Properties of Matrix Addition and Multiplication

Assume that  $\alpha$  and  $\beta$  represent any scalar, and  $\mathbf{x}$ ,  $\mathbf{y}$  and  $\mathbf{z}$  are matrices.

1.  $(\alpha + \beta)\mathbf{y} = \alpha\mathbf{y} + \beta\mathbf{y}$
2.  $\alpha(\mathbf{y} + \mathbf{z}) = \alpha\mathbf{y} + \alpha\mathbf{z}$
3.  $(\alpha\beta)\mathbf{y} = \alpha(\beta\mathbf{y})$
4.  $\alpha(\mathbf{yz}) = (\alpha\mathbf{y})\mathbf{z}$
5.  $\mathbf{y} + \mathbf{z} = \mathbf{z} + \mathbf{y}$
6.  $(\mathbf{yz})\mathbf{x} = \mathbf{y}(\mathbf{zx})$
7.  $\mathbf{y}(\mathbf{z} + \mathbf{x}) = \mathbf{yz} + \mathbf{yx}$
8.  $\mathbf{Iy} = \mathbf{yI} = \mathbf{y}$
9. Generally,  $\mathbf{yz} \neq \mathbf{zy}$ , even if  $\mathbf{y}$  and  $\mathbf{z}$  are both square matrices.

Carefully note the property (9). This is an important property that you need to remember.

Example: Recall from our previous example that  $\mathbf{yz} = \begin{bmatrix} 14 & 5 \\ 26 & 13 \end{bmatrix}_{(2 \times 2)}$ . Now, calculate  $\mathbf{zy}$  and show that  $\mathbf{yz} \neq \mathbf{zy}$ .

$$zy = \begin{bmatrix} 10 & 14 & 18 \\ 5 & 8 & 11 \\ 3 & 6 & 9 \end{bmatrix}_{(3 \times 3)} \neq \begin{bmatrix} 14 & 5 \\ 26 & 13 \end{bmatrix}_{(2 \times 2)}$$

- *Transpose*: the interchanging of rows and columns in a matrix. You can also think of a transpose as flipping the matrix along the diagonal. The transpose is denoted by either a tilde ' or by a superscript  $T$ .

$$\mathbf{y} = \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \end{bmatrix}_{(2 \times 3)} \quad \mathbf{y}' = \mathbf{y}^T = \begin{bmatrix} y_{11} & y_{21} \\ y_{12} & y_{22} \\ y_{13} & y_{23} \end{bmatrix}_{(3 \times 2)}$$

As you can see, the first row of  $\mathbf{y}$  became the first column of  $\mathbf{y}'$ . The second row of  $\mathbf{y}$  became the second column of  $\mathbf{y}'$ .

Example: Find the transpose of the matrix  $\mathbf{y} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}_{(2 \times 3)}$

$$\mathbf{y} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \end{bmatrix}_{(2 \times 3)} \quad \mathbf{y}' = \begin{bmatrix} 1 & 3 \\ 2 & 4 \\ 3 & 5 \end{bmatrix}_{(3 \times 2)}$$

#### Properties of Matrix Transposes

1.  $(\mathbf{y}')' = \mathbf{y}$
  2.  $(\mathbf{yz})' = \mathbf{z}'\mathbf{y}'$
  3.  $(\mathbf{y} + \mathbf{z})' = \mathbf{y}' + \mathbf{z}'$
  4. If  $\mathbf{x}$  is an  $(n \times 1)$  vector, then  $\mathbf{x}'\mathbf{x} = \sum_{i=1}^n x_i^2$
  5. If a matrix is square and  $\mathbf{y}' = \mathbf{y}$ , then  $\mathbf{y}$  is a symmetric matrix.
- *Inverse*: only square matrices can be inverted. The inverse of an  $(n \times n)$  matrix  $\mathbf{y}$  is identified as  $\mathbf{y}^{-1}$ . The inverse of a matrix has the same dimensions as the original matrix.

Example: suppose you have a  $2 \times 2$  matrix  $\mathbf{y} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ . Find the inverse of  $\mathbf{y}$ .

For a  $2 \times 2$  matrix, the inverse is as follows:

$$\mathbf{z}^{-1} = \frac{1}{\det(\mathbf{z})} \cdot \begin{bmatrix} z_{22} & -z_{12} \\ -z_{21} & z_{11} \end{bmatrix}$$

where  $\det(\mathbf{z}) = (z_{11}z_{22} - z_{12}z_{21})$  is the determinate of the matrix  $\mathbf{y}$ . Therefore, for the provided matrix, the inverse can be solved for as follows:

$$\mathbf{y}^{-1} = \frac{1}{(1 \cdot 4 - 2 \cdot 3)} \cdot \begin{bmatrix} 4 & -2 \\ -3 & 1 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 1.5 & -0.5 \end{bmatrix}$$

### Properties of Matrix Inverses

1. For a matrix  $\mathbf{y}$  of any dimension ( $n \times m$ ), you can take the inverse of:

(a)  $\mathbf{y}'\mathbf{y}$

(b)  $\mathbf{y}\mathbf{y}'$

2.  $(\mathbf{y}\mathbf{z})^{-1} = \mathbf{z}^{-1}\mathbf{y}^{-1}$

3.  $\mathbf{y}^{-1}\mathbf{y} = \mathbf{y}\mathbf{y}^{-1} = I$

4.  $(\mathbf{y}^{-1})' = (\mathbf{y}')^{-1}$

- *Idempotent matrix:* any matrix for which  $\mathbf{y}\mathbf{y} = \mathbf{y}$  is true. This concept will be important when we discuss projection matrices.

## 4.3 Linear independence

Linear independence is a crucial issue in econometrics. We will discuss the reasons later, but for now, let's explore how linear independence (and lack thereof) may affect matrix manipulations.

Consider a set of ( $n \times 1$ ) vectors,  $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k\}$ , and all scalars in the set  $\{c_1, c_2, \dots, c_k\}$  are zero. The vectors are *linearly independent* if and only if:

$$c_1\mathbf{y}_1 + c_2\mathbf{y}_2 + c_3\mathbf{y}_3 + \dots + c_k\mathbf{y}_k = \mathbf{0}$$

If this equation holds when any of the scalars are non-zero, then the set of vectors are *linearly dependent*. Another way to think about linear dependence is asking whether any  $\mathbf{y}_k$  is a linear combination of any other two vectors. That is,  $\mathbf{y}_k$  is a linear combination of  $\mathbf{y}_1$  and  $\mathbf{y}_2$  if:

$$\mathbf{y}_k = c_1\mathbf{y}_1 + c_2\mathbf{y}_2$$

### 4.3.1 Matrix rank

The *rank* of a matrix reveals the maximum number of linearly independent columns are in a matrix. If the matrix  $\mathbf{y}$  has the dimensions  $(n \times m)$  and the  $rank(\mathbf{y}) = m$ , then the matrix  $\mathbf{y}$  is *full rank*.

Example: What is the rank of the following two matrices?

$$\mathbf{y} = \begin{bmatrix} 1 & 2 & 13 \\ 3 & 4 & 2 \end{bmatrix}_{(2 \times 3)} \quad \mathbf{z} = \begin{bmatrix} 1 & 3 & 2.5 \\ 2 & 4 & 4 \\ 3 & 5 & 5.5 \end{bmatrix}_{(3 \times 2)}$$

The rank of  $\mathbf{y}$  is  $rank(\mathbf{y}) = 3$  because there is no column that is a linear combination of other columns. However, the rank of  $\mathbf{z}$  is  $rank(\mathbf{z}) = 2$  because the third column is a linear combination of columns 1 and 2 ( $col_1 + 0.5 \cdot col_2 = col_3$ ). Thus,  $\mathbf{y}$  is full rank, but  $\mathbf{z}$  is not full rank.

Note that if a square matrix is *not full rank*, then it is not invertible (because essentially, it is not a square matrix).

# Chapter 5

## Simple linear regression

Linear regression models are the most widely used estimation approach in economics. They are easy to implement and the results are relatively easy to interpret. The focus of linear regression models is to explain the relationship between two or more variables. Results of regression models summarize the probability distribution of a variable of interest using the estimation results.

Let's consider an economic model of bull weights:

$$W = \beta_0 + \beta_1 F$$

where  $W$  represents weight and  $F$  is amount of feed. Recall that this representation of bull weights is deterministic. That is, a change in feed  $F$  by one unit will exactly change the bull's weight by  $\beta_1$ . However, we know that there might be other, unobservable factors that can affect a bull's weight. For example, the size of a bull's predecessors may determine whether one bull that gains more weight for each pound of feed than another bull. We can introduce this uncertainty as with the term  $\varepsilon$ . That is, the econometric model is specified as follows:

$$W = \beta_0 + \beta_1 F + \varepsilon$$

Now, assume that on average, deviations from the deterministic relationship  $W = \beta_0 + \beta_1 F$  is zero. That is, if we take the entire population of bulls and weigh them, then we will find that all deviations cancel out. Practically, however, we must understand that depending on each chosen sample, these deviations vary. Knowing these two factors, we state that:

$$\varepsilon_W \sim (0, \sigma^2)$$

So, knowing the distribution of the error term, we can determine the expected value of the random variable  $W$ . However, we must take the expectation of  $W$  conditional on a particular value of  $F$ . That is,  $E(W|F)$ . Taking this expectation yields:

$$E(W|F) = E(\beta_0 + \beta_1 F + \varepsilon)$$

$$E(W|F) = \beta_0 + \beta_1 E(F|F) + E(\varepsilon|F)$$

The expectation of any variable conditional on itself is the variable. Additionally, we will assume that the expected value of the error term does not depend on the independent variable  $F$ . That is, if you take different slices of the population of  $F$ , the average of the random deviations from  $W$  is zero. Thus:

$$E(W|F) = \beta_0 + \beta_1 \bar{F}$$

This is the population regression function. The function tells us that on average, a one-unit increase in feed will change the expected weight by a value of  $\beta_1$ .

Note that this does not imply that for *all* observations in a population a one-unit increase in feed changes weight by  $\beta_1$ . It only reveals the *average* change in weight, conditional on a change in feed.

## 5.1 Ordinary least squares estimation

From the above discussion, we see that a linear regression equation has two components:

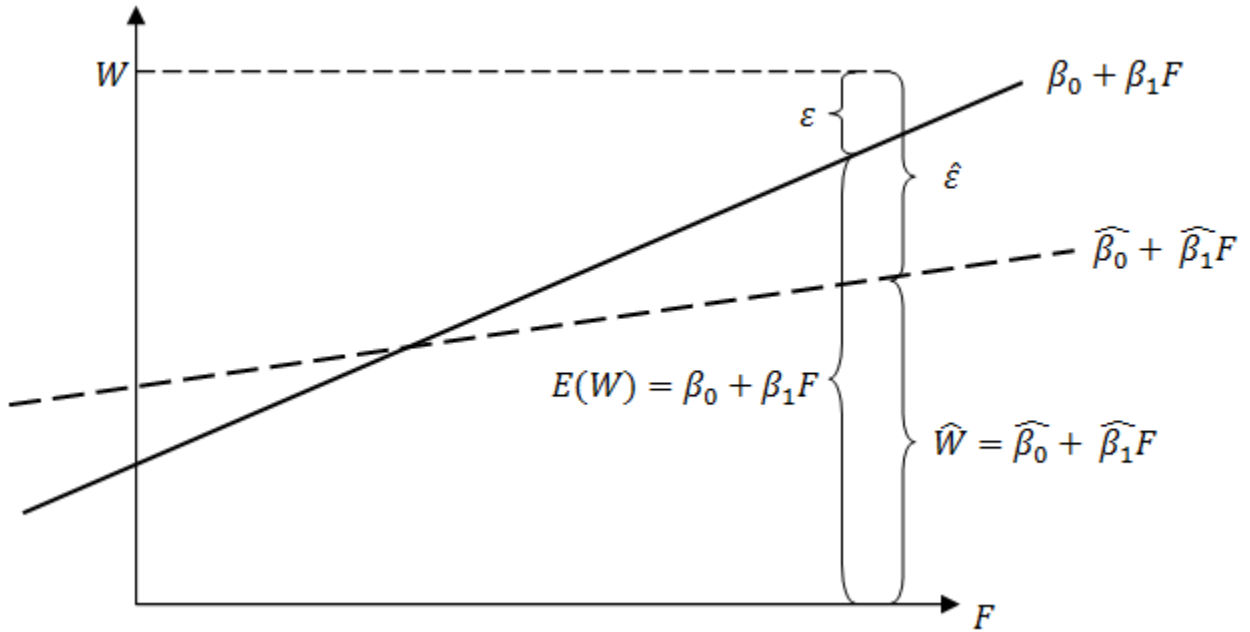
1. *Systematic*: portion of the dependent variable that is explained by the regressors. That is,  $E(W|F)$ . Let's refer to this as  $\mu$ , or the population mean.

2. *Unsystematic*: the unexplained portion of the dependent variable,  $\varepsilon$ . Because we have the unsystematic part of the regression model, the expected value of  $W$  will have a variance around the mean. Let's refer to this as  $\sigma_W^2$ , or the population variance.

If we had the true population parameters,  $[\beta_0, \beta_1]$ , we would be able to define the true population  $\mu$  and  $\sigma_W^2$ . However, we don't know these parameters, implying that we need to estimate  $[\beta_0, \beta_1]$  using a sample, and then get an estimate of the mean and variance. That is,  $W \sim (\mu, \sigma_W^2)$ , and we need to estimate the component  $\mu$  and  $\sigma_W^2$ .

The difference between knowing the population parameters  $[\beta_0, \beta_1]$  and the estimated sample values  $[\hat{\beta}_0, \hat{\beta}_1]$  is illustrated in figure 5.1.

Figure 5.1: Population vs. Sample Regression



That is, if we knew the entire population, we could find the population parameters  $[\beta_0, \beta_1]$  and predict  $W$  with an error  $\varepsilon$ . However, we only have a representative sample of the population, we retrieve the parameters  $[\hat{\beta}_0, \hat{\beta}_1]$ , which we can use to predict  $\widehat{W}$ . Then, we can use these estimates to determine  $\hat{\mu}$  and  $\hat{\sigma}^2$ . The downside is that now our error is  $\hat{\varepsilon}$ .

The goal is to find an estimation method which minimizes the difference between  $\varepsilon$  and  $\hat{\varepsilon}$ . One method that has been shown to have good statistical properties and minimizes this difference is *ordinary least squares*.

### 5.1.1 Estimating parameters of a linear model

In order to estimate  $\mu$ , we must retrieve the unknown components that make up  $\mu$ . That is, we need to get an estimate for  $[\beta_0 \ \beta_1]$ .

Suppose that you have a sample of bull weights and associated feed information for  $n$  bulls. That is, your data set is as follows:

$$\{(w_1, f_1), (w_2, f_2), \dots, (w_n, f_n)\}$$

This implies that each weight can be modeled as  $w_i = \beta_0 + \beta_1 f_i + \varepsilon_i$ , for  $i = 1, \dots, n$ . So you can imagine  $n$  such equations:

$$\begin{aligned}w_1 &= \beta_0 + \beta_1 f_1 + \varepsilon_1 \\w_2 &= \beta_0 + \beta_1 f_2 + \varepsilon_2 \\&\vdots \\w_n &= \beta_0 + \beta_1 f_n + \varepsilon_n\end{aligned}$$

**Conceptual idea:** we must choose the set of coefficients in the matrix  $\boldsymbol{\beta} = [\beta_0 \ \beta_1]$  that minimize the unsystematic component in each equation. That is, we want to choose  $\boldsymbol{\beta}$  that allows us to explain as much behavior in weight variability across animals using variations in feed.

To express the fact that we wish to minimize the random part of the linear equation, we can state for each observation:

$$\min(\varepsilon_i) = \min(w_i - E(w_i|f_i)) = \min(w_i - \beta_0 - \beta_1 f_i)$$

This equation indicates that we wish to choose  $\boldsymbol{\beta}$  that minimizes the difference between the actual observation and a predicted value, which uses feed to predict weight.

Now, what we are actually interested is to choose  $\boldsymbol{\beta}$  that minimizes the error component over the entire sample. That is, we want to minimize the sum of all errors:

$$\min S(\boldsymbol{\beta}) = \sum_{i=1}^n (w_i - E(w_i|f_i))$$

There is one problem with this approach: if we have both positive and negative errors, they may cancel out leading to a better-than-actual conclusion about the estimated values in  $\boldsymbol{\beta}$ . One way to avoid this problem is to square the errors prior to summing. That is:

$$\min S(\boldsymbol{\beta}) = \sum_{i=1}^n (w_i - E(w_i|f_i))^2$$

In this case,  $S$  is an objective function that represents the minimum of the sum of squared deviations in the expected weight  $E(W|F)$  from the actual weight  $W$ .

### Solving for the coefficients

Estimating the components that will help us retrieve an estimate  $\mu$  can be done using calculus. Assuming that  $S$  is a well-behaving, convex function of  $\boldsymbol{\beta}$  (i.e., we can find a minimum), we can find the minimum of this function by taking the first-order conditions with respect to each of the parameters:<sup>1</sup>

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_0} = -2 \sum_{i=1}^n (w_i - \hat{\beta}_0 - \hat{\beta}_1 f_i) = 0$$

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_1} = -2 \sum_{i=1}^n f_i (w_i - \hat{\beta}_0 - \hat{\beta}_1 f_i) = 0$$

Note that sample estimates of  $\beta_0$  and  $\beta_1$  are denoted by  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .

Now, the two first-order conditions above define a two-equation, two-unknowns system of equations. This implies that we can retrieve unique solutions for  $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0 \hat{\beta}_1]$ . Let's take a look at how we can solve for each coefficient by solving each of the first-order conditions separately.

---

<sup>1</sup>Convexity can be verified using the second-order conditions.

From first-condition 1:

$$\frac{1}{n} \sum_{i=1}^n w_i - \hat{\beta}_0 - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n f_i = 0$$

$$\hat{\beta}_0 | f = \bar{w} - \hat{\beta}_1 \bar{f}$$

From first-condition 2:

$$\frac{1}{n} \sum_{i=1}^n f_i (w_i - \hat{\beta}_0 - \hat{\beta}_1 f_i) = 0$$

$$\frac{1}{n} \sum_{i=1}^n f_i w_i - \hat{\beta}_0 \frac{1}{n} \sum_{i=1}^n f_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n f_i f_i = 0$$

$$\frac{1}{n} \sum_{i=1}^n f_i w_i - (\bar{w} - \hat{\beta}_1 \bar{f}) \bar{f} - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n f_i^2 = 0$$

$$\frac{1}{n} \sum_{i=1}^n w_i f_i - \bar{w} \bar{f} - \hat{\beta}_1 \left( \frac{1}{n} \sum_{i=1}^n f_i^2 - \bar{f}^2 \right) = 0$$

$$\hat{\beta}_1 | f = \frac{\frac{1}{n} \sum_{i=1}^n w_i f_i - \bar{w} \bar{f}}{\left( \frac{1}{n} \sum_{i=1}^n f_i^2 - \bar{f}^2 \right)}$$

Now, we can show that  $\left( \frac{1}{n} \sum_{i=1}^n f_i^2 - \bar{f}^2 \right) = \frac{1}{n} \sum_{i=1}^n (f_i - \bar{f})^2$ :

$$\begin{aligned}
 \frac{1}{n} \sum_{i=1}^n (f_i - \bar{f})^2 &= \frac{1}{n} \sum_{i=1}^n (f_i^2 - 2f_i\bar{f} + \bar{f}^2) \\
 &= \frac{1}{n} \sum_{i=1}^n f_i^2 - 2\bar{f} \frac{1}{n} \sum_{i=1}^n f_i + \bar{f}^2 \\
 &= \frac{1}{n} \sum_{i=1}^n f_i^2 - 2\bar{f}^2 + \bar{f}^2 \\
 &= \frac{1}{n} \sum_{i=1}^n f_i^2 - \bar{f}^2
 \end{aligned}$$

Note that this result can also be used to show that:

$$\left( \frac{1}{n} \sum_{i=1}^n w_i f_i - \bar{w} \bar{f} \right) = \frac{1}{n} \sum_{i=1}^n (w_i - \bar{w})(f_i - \bar{f})$$

Using these properties we can summarize the formulas for the estimators of  $\beta_0$  and  $\beta_1$  as following:

$$\begin{aligned}
 \hat{\beta}_0 | f &= \bar{w} - \hat{\beta}_1 \bar{f} \\
 \hat{\beta}_1 | f &= \frac{\frac{1}{n} \sum_{i=1}^n (w_i - \bar{w})(f_i - \bar{f})}{\frac{1}{n} \sum_{i=1}^n (f_i - \bar{f})^2}
 \end{aligned}$$

Notice that  $\hat{\beta}_1$  is nothing more than the sample covariance between  $w$  and  $f$ , divided by the sample variance of  $f$ :

$$\hat{\beta}_1 = \frac{\text{Cov}(w, f)}{\text{Var}(f)}$$

This indicates that if weight and feed are positively correlated in the sample, then  $\hat{\beta}_1$  is positive; conversely, a negative correlation between weight and feed leads to a negative  $\hat{\beta}_1$ .

**Point:** now that we have estimates of  $\beta_0$  and  $\beta_1$ , we can use them to determine an estimate of the random variable weight's  $\mu$  parameter. That is,  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are used to determine  $\hat{\mu}$ .

### Unbiasedness of the estimator

Recall that an important desired property of an estimator is unbiasedness. That is, on average, the value of the estimator is equal to the true value of the parameter:  $E(\hat{\beta}_1) = \beta_1$ .

Let's start with showing that  $E(\hat{\beta}_1) = \beta_1$ :

$$E(\hat{\beta}_1|f) = E\left(\frac{\sum_{i=1}^n (w_i - \bar{w})(f_i - \bar{f})}{\sum_{i=1}^n (f_i - \bar{f})^2}\right)$$

Note that  $(w_i - \bar{w}) = (\beta_0 + \beta_1 f_i + \varepsilon_i) - (\beta_0 + \beta_1 \bar{f}) = \beta_1(f_i - \bar{f}) + \varepsilon_i$ . Inserting this into the equation above yields:

$$\begin{aligned} E(\hat{\beta}_1|f) &= E\left(\frac{\sum_{i=1}^n \beta_1(f_i - \bar{f})^2 + (f_i - \bar{f})\varepsilon_i}{\sum_{i=1}^n (f_i - \bar{f})^2}\right) \\ &= \beta_1 + E\left(\frac{\sum_{i=1}^n (f_i - \bar{f})\varepsilon_i}{\sum_{i=1}^n (f_i - \bar{f})^2}\right) \end{aligned}$$

Now, if we assume (as we did before) that  $E(\varepsilon|F) = 0$  (that is, the covariance between the explanatory variables and the error term is zero), then the second term goes to zero. This yields the result that  $E(\hat{\beta}_1) = \beta_1$ .

Next, we show that  $E(\hat{\beta}_0) = \beta_0$  as follows:

$$\begin{aligned} E(\hat{\beta}_0|f) &= E(\bar{w} - \hat{\beta}_1 \bar{f}) \\ &= E([\beta_0 + \beta_1 \bar{f} + \bar{\varepsilon}] - \hat{\beta}_1 \bar{f}) \\ &= \beta_0 + E([\beta_1 - \hat{\beta}_1] \bar{f}) + E(\bar{\varepsilon}) \\ &= \beta_0 \end{aligned}$$

The last step follows because we already showed that  $E(\hat{\beta}_1) = \beta_1$ . Thus,  $\beta_1 - E(\hat{\beta}_1) = 0$ .

### 5.1.2 Sample variance of the estimator

Now that we have determined the sample estimator,  $\hat{\beta}$ , we would like to know whether this estimator is the most efficient. That is, out of a set of all possible estimators, is the variance of  $\hat{\beta}$  smallest?

Our first step is, of course, finding the sample of the estimator. Recall the definition of variance:

$$\text{Var}(Y) = E([Y - E(Y)]^2) = E(Y^2) - E(Y)^2$$

Using this definition, we would like to determine the variance of  $\hat{\beta}$ . That is,  $\text{Var}(\hat{\beta})$ . This is as follows:

$$\text{Var}(\hat{\beta}|f) = E([\hat{\beta} - E(\hat{\beta})]^2) = E([\hat{\beta} - \beta]^2),$$

where  $E(\hat{\beta}|f) = \beta$  from previously showing unbiasedness of the OLS estimator.

Generally, we only care about the variance of  $\hat{\beta}_1$ , because we are mostly interested in understanding the relationship between the dependent and independent variables, and how well our estimator allows us to explain that relationship.

For  $\hat{\beta}_1$ , the sample variance is as follows:

$$\begin{aligned} \text{Var}(\hat{\beta}_1|f) &= \frac{\sum_{i=1}^n (f_i - \bar{f})^2 \cdot \text{Var}(\varepsilon_i)}{\sum_{i=1}^n (f_i - \bar{f})^4} \\ &= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (f_i - \bar{f})^2} \end{aligned}$$

Note that the variance of the estimator depends on the population variance of the error term and a “pseudo-variance” of the explanatory variables (dividing the denominator term by  $n - 1$  yields the variance of  $F$ ). However, we don't know the population variance,  $\sigma_\varepsilon^2$ . Thus, we'll need to approximate it using the sample.

One intuitive way to do this is to replace the population variance with the variance of the *residuals* ( $\hat{\varepsilon}$ ) from the regression. That is, an appropriate approximation is:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-k} \sum_{i=1}^n \hat{\varepsilon}_i^2,$$

where  $k$  is the number of parameters estimated in the OLS regression. Specifically, it is the degrees of freedom. For example, the two-regressor model above implies that  $k = 2$ .

The term  $\frac{1}{n-k}$  is used to appropriately weigh the estimator. Without this weight, the estimator of  $\hat{\sigma}_\varepsilon^2$  would be biased. Following from this estimator, the variance estimator for  $\hat{\beta}$  is:

$$\widehat{\text{Var}}(\hat{\beta}_1|f) = \frac{\frac{1}{n-k} \sum_{i=1}^n \hat{\varepsilon}_i^2}{\sum_{i=1}^n (f_i - \bar{f})^2}$$

An important feature of  $\widehat{\text{Var}}(\hat{\beta})$  is that taking the square root yields the estimator's *standard error*. That is,  $(\widehat{\text{Var}}(\hat{\beta}))^{1/2} \equiv \text{se}(\hat{\beta})$  is an estimate of the standard deviation of the dependent variable  $W$  after the effect of  $F$  has been taken out. Lower standard errors indicate the accuracy of the estimator  $\hat{\beta}$ .

### 5.1.3 Consistency of the estimator

Recall that consistency is expressed as the *plim*  $\hat{\beta}_n = \beta$ . That is:

$$\text{If } \lim_{n \rightarrow \infty} E(\hat{\beta}_n) = \beta \text{ and } \lim_{n \rightarrow \infty} \widehat{\text{Var}}(\hat{\beta}_n) \rightarrow 0$$

We have already shown that  $E(\hat{\beta}_n) = \beta$ . Recalling that the  $\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\frac{1}{n-k} \sum_{i=1}^n \hat{\varepsilon}_i^2}{\sum_{i=1}^n (f_i - \bar{f})^2}$ , it is clearly the case that as  $n \rightarrow \infty$ , the variance estimator goes to zero. Thus, the OLS estimator is consistent.

## 5.2 OLS in Matrix Algebra

Using matrix algebra to find OLS estimators is much neater, faster, and may be more intuitive. Using matrix algebra, we “mimic” the summation approach to minimizing the square of squared residuals. Additionally, when we explore multiple regressor estimation, you will see the true power of matrix algebra.

### 5.2.1 Setting up the model

Recall a simple regression model:  $Y = \beta_0 + X\beta_1 + \varepsilon$ . Previously, we determined the value of  $\varepsilon$ , and then sought to determine  $\hat{\beta}$  by taking the square of  $\varepsilon$  for each observation, and then summing over the squares. Now, we'll do the same but with matrices.

For some  $n$  observations, we can set up a vector representing the dependent variable, a matrix representing the explanatory variables, a matrix representing the coefficients, and a vector representing the error term. That is:

$$Y_{(n \times 1)} = \mathbf{X}_{(n \times 2)} \cdot \boldsymbol{\beta}_{(2 \times 1)} + \boldsymbol{\varepsilon}_{(n \times 1)}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{(n \times 1)} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_n \end{bmatrix}_{(n \times 2)} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}_{(2 \times 1)} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{(n \times 1)}$$

To convince you that this is equivalent to the form that we've seen before, consider multiplying out the matrices.

$$Y_1 = \{(1 \cdot \beta_0) + (X_1 \cdot \beta_1)\} + \varepsilon_1$$

$$Y_1 = \beta_0 + X_1\beta_1 + \varepsilon_1$$

$$Y_2 = \{(1 \cdot \beta_0) + (X_2 \cdot \beta_1)\} + \varepsilon_2$$

$$Y_2 = \beta_0 + X_2\beta_1 + \varepsilon_2$$

$\vdots$

$$Y_n = \{(1 \cdot \beta_0) + (X_n \cdot \beta_1)\} + \varepsilon_n$$

$$Y_n = \beta_0 + X_n\beta_1 + \varepsilon_n$$

Clearly, the matrix form is just a representation of the  $n$  typical equations.

### 5.2.2 Solving for the coefficients

As before, we will need to set up a minimizing problem and solve for first-order conditions in order. Recall that we are trying to find a set of  $\hat{\beta}$  that minimizes the sum of squared  $\varepsilon$ . From the linear regression model above, we know that:

$$\varepsilon_{(n \times 1)} = (Y_{(n \times 1)} - \mathbf{X}_{(n \times 2)} \cdot \beta_{(2 \times 1)})$$

So the square of  $\varepsilon_{(n \times 1)}$  is:

$$\varepsilon^2 = (Y - \mathbf{X}\beta)_{(n \times 1)}^2$$

$$(\varepsilon'\varepsilon)_{(1 \times 1)} = (Y - \mathbf{X}\beta)'_{(n \times 1)}(Y - \mathbf{X}\beta)_{(n \times 1)}$$

Thus, we can specify the minimization problem as follows:

$$\min_{\{\hat{\beta}\}} S(\hat{\beta}) = \min_{\{\hat{\beta}\}} (Y - \mathbf{X}\hat{\beta})'(Y - \mathbf{X}\hat{\beta})$$

Now, let's take the partial derivative of the function  $S(\hat{\beta})$  with respect to  $\hat{\beta}$ , and then solve for  $\hat{\beta}$ :

$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}} = -2\mathbf{X}'(Y - \mathbf{X}\hat{\beta}) = 0$$

$$\mathbf{X}'Y - \mathbf{X}'\mathbf{X}\hat{\beta} = 0$$

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'Y$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$$

In these lectures notes, only one formula is in a box – this is to stress the fact that this is an important formula. Take note of it.

It should be noted that we can retrieve  $\hat{\beta}$  if and only if the matrix  $(\mathbf{X}'\mathbf{X})$  is invertible. Recall that invertibility requires two properties:

1. The matrix is square.
2. The matrix is full rank (i.e. all columns are linearly independent).

The first property is satisfied because the  $\mathbf{X}$  is pre-multiplied by its transpose, which implies that the matrix product is a square matrix. The second property is one that we will assume to be true. We will explore violations of linear independence later.

### 5.2.3 Projection and “residual maker” matrices

When we estimate a regression model, what are we actually trying to accomplish? We are attempting to explain the behavior of a random dependent variable using some weighted function of independent variables. In the linear regression case, it is a weighted sum.

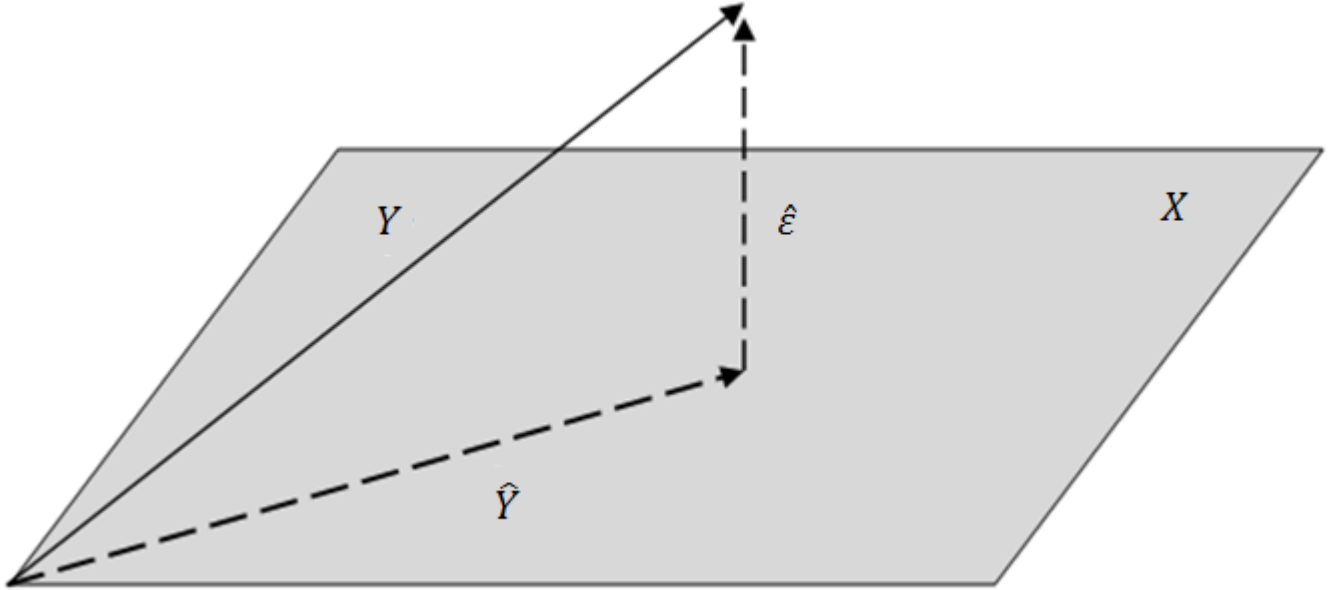
In a simple linear regression, we are attempting to take an independent variable  $X$  and use it to predict the dependent variable  $Y$  by assigning a weight  $\beta$  to  $X$ . In other words, we are “projecting”  $Y$  onto the column space of  $X$  in order to get a predicted  $Y$  by only using weighted sums of  $X$ .

Unfortunately, because  $Y$  is a random variable, we are unable to perfectly predict  $Y$  using a weighted  $X$ . This implies that we have some type of error. That is, our estimated regression is as follows:

$$Y = \hat{Y} + \hat{\varepsilon} = \mathbf{X}\hat{\beta} + \hat{\varepsilon}$$

We can visualize the projection of  $Y$  on the column space of  $X$  using figure 5.2. The figure illustrates an important assumption/property of a linear regression model: each column of  $X$  is orthogonal (perpendicular) to  $\hat{\varepsilon}$ . That is,  $\mathbf{X}'\hat{\varepsilon} = 0$ .

Figure 5.2: Projection of  $Y$  onto column space  $X$



Recall that  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$ . Using this knowledge, let's plug this back into the equation for  $Y$ .

$$\begin{aligned}
 Y &= \mathbf{X}\hat{\beta} + \hat{\varepsilon} \\
 &= \mathbf{X}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y) + \hat{\varepsilon} \\
 &= (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')Y + \hat{\varepsilon} \\
 &= P_X Y + \hat{\varepsilon}
 \end{aligned}$$

The matrix  $P_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  is known as the *projection* matrix. This matrix is symmetric and idempotent ( $P_X = P_X P_X$ ). This matrix “projects” the vector  $Y$  onto the column space  $X$ , which provides you with the fitted values using the least squares regression method. That is, by pre-multiplying the dependent variable vector  $Y$  by  $P_X$  provides you the estimated systematic portion of the linear regression model. A very important concept is that the column space  $X$  goes through the mean of  $Y$  (i.e., the estimated responses from changes in  $X$  are average marginal effects).

You can follow a similar process to determine the “*residual maker*” matrix.

$$\begin{aligned}
 \hat{\varepsilon} &= Y - \mathbf{X}\hat{\boldsymbol{\beta}} \\
 &= Y - \mathbf{X}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y) \\
 &= (I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')Y \\
 &= (I_n - P_X)Y \\
 &= M_X Y
 \end{aligned}$$

The “residual maker” matrix,  $M_X = (I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$  is also symmetric and idempotent. Using  $P_X$  and  $M_X$ , we can define the linear least squares regression model to be:

$$Y = P_X Y + M_X Y = \text{Projection} + \text{Residual}$$

## 5.3 Properties of the OLS Estimator

### 5.3.1 Unbiasedness of the OLS estimator

Recall that an important property of an estimator is that it is unbiased. That is,  $E(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$ . To show that the OLS estimator is an unbiased estimator of the true population parameter, consider the following:

$$E(\hat{\boldsymbol{\beta}}|\mathbf{X}) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y)$$

We can use the fact that  $Y = \mathbf{X}\boldsymbol{\beta} + \varepsilon$  to substitute into  $Y$ :

$$\begin{aligned}
 E(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})) \\
 &= E(((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})|\mathbf{X}) \\
 &= \boldsymbol{\beta}E((\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})|\mathbf{X}) + E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}|\mathbf{X}) \\
 &= \boldsymbol{\beta}E(I_n) + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}|\mathbf{X}) \\
 &= \boldsymbol{\beta} + 0
 \end{aligned}$$

In the second to last step, the term  $E(\boldsymbol{\varepsilon}|\mathbf{X}) = 0$  because of we've made the assumption that  $E \perp \boldsymbol{\varepsilon}$ , which implies that  $E(\boldsymbol{\varepsilon}|\mathbf{X}) = E(\boldsymbol{\varepsilon}) = 0$ . Thus, we can see that the OLS estimator is indeed an unbiased estimator of  $\boldsymbol{\beta}$ .

### 5.3.2 Variance of the OLS estimator

We would like to know the precision of the OLS estimator, which can be determined by examining the estimator's variance. The variance is defined as follows:

$$\begin{aligned}
 \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\
 &= E[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})'] \\
 &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]
 \end{aligned}$$

The middle term  $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X})$  is the variance of the error terms, which can be written as  $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$ . This indicates that the variance is just a diagonal matrix with the same value  $\sigma^2$  for each observation  $n$ . This assumption is known as *homoskedasticity*. It also implies that the covariances of error terms are zero, which implies that error terms are uncorrelated. Thus, because  $\sigma^2$  is a constant, the conditional expectation of a constant is the constant.

Continuing:

$$\begin{aligned}
 &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\
 &= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\
 \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\
 \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \sigma^2 \left( \sum_{i=1}^n x_i x_i' \right)^{-1}
 \end{aligned}$$

### Estimator of the OLS variance

The formula for the variance of the OLS estimator indicates that we must know the population variance  $\sigma^2$ . We don't know the population variance, so we must provide an unbiased estimator for  $\sigma^2$ .

A natural choice is the squared estimation residual,  $\hat{\varepsilon}^2$ . The unbiased estimator of  $\sigma^2$  is as follows:

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \left( \sum_{i=1}^n \hat{\varepsilon}_i^2 \right) = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k-1},$$

where  $n$  is the number of observations, and  $k$  is the number of independent variables (regressors). In our simple one regressor scenario,  $k = 1$ . Therefore, the unbiased variance of  $\hat{\boldsymbol{\beta}}$  is as follows:

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

Remember that  $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}|\mathbf{X})$  is a matrix. Each element on the diagonal indicates the variance of the associated estimated OLS parameter. That is, suppose that we were estimating the model:

$$Y_{(n \times 1)} = \mathbf{X}_{(n \times 2)} \cdot \boldsymbol{\beta}_{(2 \times 1)} + \boldsymbol{\varepsilon}_{(n \times 1)}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{(n \times 1)} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_n \end{bmatrix}_{(n \times 2)} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}_{(2 \times 1)} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{(n \times 1)}$$

The estimated parameters  $\hat{\beta} = [\hat{\beta}_0 \ \hat{\beta}_1]'$  will each have an associated variance. So the variance  $\widehat{\text{Var}}(\hat{\beta}|\mathbf{X})$  is a  $(2 \times 2)$  matrix as follows:

$$\widehat{\text{Var}}(\hat{\beta}|\mathbf{X}) = \begin{bmatrix} \widehat{\text{Var}}(\hat{\beta}_0) & \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1) \\ \widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_0) & \widehat{\text{Var}}(\hat{\beta}_1) \end{bmatrix}$$

So for the  $k^{\text{th}}$  OLS parameter, the associated variance is the  $kk^{\text{th}}$  value in the  $\widehat{\text{Var}}(\hat{\beta}|\mathbf{X})$  variance matrix.

### 5.3.3 Standard errors

Recall that the standard deviation of an estimator is the square root of its variance. So, if we had the variance matrix with the population  $\sigma^2$ , we would be able to retrieve the standard errors by solving  $\sqrt{\text{Var}(\hat{\beta}|\mathbf{X})}$ . However, because we only have  $\widehat{\text{Var}}(\hat{\beta}|\mathbf{X})$ , the square root of the estimated variance matrix provides the *standard errors* of each estimated parameter. That is:

$$\text{SE}(\hat{\beta}) = \sqrt{\widehat{\text{Var}}(\hat{\beta}|\mathbf{X})} = \begin{bmatrix} \sqrt{\widehat{\text{Var}}(\hat{\beta}_0|\mathbf{X})} & \cdot \\ \cdot & \sqrt{\widehat{\text{Var}}(\hat{\beta}_1|\mathbf{X})} \end{bmatrix}$$

### 5.3.4 Distribution of estimators and error terms

Having derived the statistical properties associated with the error term and estimators, we can now characterize the distributions of each. The error term is distributed normal (this is an assumption discussed in the next section) with mean zero and variance matrix  $\sigma^2 \mathbf{I}_n$ . That is:

$$\varepsilon \sim N(0, \sigma^2 \mathbf{I}_n).$$

A particular observation of the error term can be defined as follows:

$$\varepsilon_i \sim NID(0, \sigma^2),$$

where “NID” implies that each observation of the error term is an independent drawing from a normal distribution, with mean zero and variance  $\sigma^2$ .

Similarly, we have shown that the estimator of  $\hat{\beta}$  is unbiased with a variance matrix  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ . That is:

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

For a particular  $k^{\text{th}}$  estimated coefficient,  $\hat{\beta}_k$  is distributed as follows:

$$\hat{\beta}_k \sim N(\beta_k, \sigma^2(X'X)_{kk}^{-1}),$$

where  $(X'X)_{kk}^{-1}$  is the  $kk^{\text{th}}$  element of the  $(\mathbf{X}'\mathbf{X})^{-1}$  matrix.

## 5.4 Gauss-Markov Assumptions for Linear Regression Models

Although we have seen how to estimate linear regression models, we haven't discussed what assumptions we are making when OLS methods are used. Let us now specify what assumptions allow us to estimate models using ordinary least squares.

1. Linearity in parameters – the regression model must be linear in parameters:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

2. Full rank – there is no exact linear relationship between any of the independent variables. That is, all columns of matrix  $\mathbf{X}$  are linearly independent. This assumption allows us to invert  $\mathbf{X}'\mathbf{X}$  in order to determine the OLS parameters.
3. The expected value of the error terms is zero –  $E(\varepsilon) = 0$ .

4. Exogeneity of independent variables – the expected value of the error at any observation  $i$  is not a function of the independent variables at observation  $i$  and any other observation in the sample. This implies that the independent variables have no useful information for explaining  $\varepsilon$ .

$$E(\varepsilon_i|\mathbf{X}) = 0$$

5. Homoskedasticity – the variance of the error term at each observation  $i$  is the same and constant:

$$\text{Var}(\varepsilon_i) = \sigma^2, \quad \text{for all } i$$

$$\text{Var}(\varepsilon) = E(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2 I_n$$

6. Non-autocorrelation – the error term at an observation  $i$  is uncorrelated with the error term at any other observation:

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad \text{for all } i \neq j$$

Note that assumptions 5 and 6 imply that the error terms are known to be *spherical*.

7. Error terms are normally distributed – the distribution of  $\varepsilon$  is:

$$\varepsilon \sim N(0, \sigma^2)$$

The important implication of making the Gauss-Markov assumptions is that we can then state that the OLS estimator  $\hat{\beta}$  is the most efficient linear unbiased estimator out of all other linear estimators.

### 5.4.1 Proof: Gauss-Markov Theorem

Recall that  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ . Now suppose that we assume some other linear estimator:

$$\tilde{\beta} = \mathbf{A}'\mathbf{Y} ,$$

where  $\mathbf{A}$  is a matrix consisting of nonrandom numbers and functions of  $\mathbf{X}$ . We can re-write  $\tilde{\beta}$  to be as follows:

$$\tilde{\beta} = \mathbf{A}'(\mathbf{X}\beta + \varepsilon) = (\mathbf{A}'\mathbf{X})\beta + \mathbf{A}'\varepsilon$$

If we then take the expected value of the estimator  $\tilde{\beta}$ , we retrieve the following:

$$\begin{aligned} E(\tilde{\beta}|\mathbf{X}) &= \mathbf{A}'\mathbf{X}\beta + E(\mathbf{A}'\varepsilon|\mathbf{X}) \\ &= \mathbf{A}'\mathbf{X}\beta + \mathbf{A}'E(\varepsilon|\mathbf{X}) \\ &= \mathbf{A}'\mathbf{X}\beta \end{aligned}$$

The second step is due to the fact that  $\mathbf{A}$  is a function of  $\mathbf{X}$ , and the last step follows from the assumption that the error term is independent of all observations in  $\mathbf{X}$ . Now, because the expected value of  $\tilde{\beta}$  must equal  $\beta$  for the estimator to be unbiased, that would imply that the following must hold:  $\mathbf{A}'\mathbf{X} = \mathbf{I}_n$ .

Suppose that this holds. Now, we're interested in determining whether the alternative estimator is the most efficient. That is, we'd like to know whether  $\text{Var}(\tilde{\beta}|\mathbf{X}) \leq \text{Var}(\hat{\beta}|\mathbf{X})$ . The conditional variance of  $\tilde{\beta}$  is as follows:

$$\begin{aligned} \text{Var}(\tilde{\beta}|\mathbf{X}) &= \text{Var}((\mathbf{A}'\mathbf{X})\beta + \mathbf{A}'\varepsilon|\mathbf{X}) \\ &= \text{Var}(\mathbf{A}'\varepsilon|\mathbf{X}) \\ &= \mathbf{A}'\text{Var}(\varepsilon|\mathbf{X})\mathbf{A} \\ &= \sigma^2\mathbf{A}'\mathbf{A} \end{aligned}$$

Now, let's determine whether the difference between the new estimator ( $\tilde{\beta}$ ) and the OLS estimator ( $\hat{\beta}$ ) is zero (or negative).

$$\begin{aligned}
 \text{Var}(\tilde{\beta}|\mathbf{X}) - \text{Var}(\hat{\beta}|\mathbf{X}) &= \sigma^2 \mathbf{A}'\mathbf{A} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\
 &= \sigma^2[\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}] \\
 &= \sigma^2[\mathbf{A}'\mathbf{A} - \mathbf{A}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}] \\
 &= \sigma^2 \mathbf{A}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{A} \\
 &= \sigma^2 \mathbf{A}'\mathbf{M}_\mathbf{X}\mathbf{A}
 \end{aligned}$$

Note that the second step follows from the fact that  $\mathbf{I}_n = \mathbf{A}'\mathbf{X}$ . The result indicates that the matrix ( $\sigma^2 \mathbf{A}'\mathbf{M}_\mathbf{X}\mathbf{A}$ ) is positive semi-definite (i.e. analogous to a positive real number).<sup>2</sup> Therefore, the variance of the alternative estimator is greater than the variance of the OLS estimator, implying that it is a less efficient estimator.

## 5.5 Asymptotic Properties of OLS Estimators

The properties that we derived above (section 5.3) are based only on the fact that the Gauss-Markov assumptions hold. If we relax some of the assumptions, these finite-sample properties become unknown. Therefore, we must use large-sample properties to establish the quality of OLS estimators. That is, as the number of observations in our sample grows to infinity, is the OLS estimation method produce the most efficient and unbiased estimators of a linear regression function.

### 5.5.1 Consistency

Recall that consistency refers to the probability limit theorem:  $\text{plim } \hat{\beta} \rightarrow \beta$ . The two components of proving consistency is to show asymptotic unbiasedness and asymptotic efficiency.

---

<sup>2</sup>A positive semi-definite matrix is one for which the eigenvalues are all positive.

### Asymptotic unbiasedness

Recall that we can write the OLS estimator as follows:  $\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$ . We can multiply the last term by  $\frac{n}{n}$  and rewrite the estimator as follows:

$$\hat{\beta} = \beta + \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \frac{\mathbf{X}'\varepsilon}{n}$$

Taking the probability limit, we retrieve:

$$plim \hat{\beta} \rightarrow \beta + plim \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \cdot plim \frac{\mathbf{X}'\varepsilon}{n}$$

By the law of large numbers, the term  $plim \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \rightarrow \mathbf{Q}^{-1}$ , which is some positive definite matrix. Furthermore, assuming exogeneity, the term  $plim \frac{\mathbf{X}'\varepsilon}{n} \rightarrow 0$ . Therefore:

$$plim \hat{\beta} \rightarrow \beta$$

### Asymptotic efficiency

From the definition of  $plim \hat{\beta}$  above, we see that the variance will only involve the term  $\frac{\mathbf{X}'\varepsilon}{n}$ . This variance can be written as  $\text{Var}\left(\frac{\mathbf{X}'\varepsilon}{n}|\mathbf{X}\right) = E[\text{Var}\left(\frac{\mathbf{X}'\varepsilon}{n}|\mathbf{X}\right)] + \text{Var}[E\left(\frac{\mathbf{X}'\varepsilon}{n}|\mathbf{X}\right)]$ . (Note: This is the variance of a conditional random variable). The second term is zero because of the exogeneity assumption. The first term is as follows:

$$\begin{aligned} \text{Var}\left(\frac{\mathbf{X}'\varepsilon}{n}|\mathbf{X}\right) &= E\left[\left(\frac{\mathbf{X}'\varepsilon}{n}\right)\left(\frac{\mathbf{X}'\varepsilon}{n}\right)'|\mathbf{X}\right] \\ &= \frac{1}{n}\mathbf{X}'E[\varepsilon\varepsilon'|\mathbf{X}]\mathbf{X}\frac{1}{n} \\ &= \left(\frac{\sigma^2}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right) \\ \text{Var}\left(\frac{\mathbf{X}'\varepsilon}{n}|\mathbf{X}\right) &= \left(\frac{\sigma^2}{n}\right)E\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right) \end{aligned}$$

Because we have showed that the second term is just a positive definite matrix  $\mathbf{Q}$ , the product goes to zero as  $n \rightarrow \infty$ .

### 5.5.2 Asymptotic normality

Under small-sample properties, we assumed that the estimator is normally distributed. If we relax this assumption, we can still show normality in large-samples using the central limit theorem. With the proof left as an exercise to the reader (for an overview of the proof, see Greene (2003)), asymptotic normality indicates that:

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow N \left[ 0, \sigma^2 \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right) \right]$$

Therefore, if the error terms  $\varepsilon_i \sim i.i.d.(0, \sigma^2)$ , then the following is true:

$$\hat{\beta} \sim^a N \left[ \beta, \sigma^2 \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right) \right]$$

Because we are unable to retrieve the population error variance  $\sigma^2$ , we use the approximation defined in section 5.3.2. That is:

$$\widehat{\text{AVar}}(\hat{\beta}) = \hat{\sigma}^2 \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right)$$

### 5.5.3 Asymptotics: Why do we care?

The asymptotic properties reveal that even if the errors are not normally distributed, as long as observations are *i.i.d.*, the normality of the least squares estimator *does not* depend on the normality of the disturbance terms. The estimator,  $\hat{\beta}$ , will be asymptotically distributed normally, and its properties will converge to those implied by the Gauss-Markov assumptions. This is important because we are able to use standard inference testing to test the statistical validity of the estimator. Therefore, we do not need to impose strict distributional assumptions about the errors to test for the accuracy of OLS estimators. This is an extremely powerful outcome.

## 5.6 Inferences from OLS Estimators

We have now established that the OLS is an efficient and unbiased estimator. This is true even if some of the Gauss-Markov assumptions are relaxed, because OLS estimators are asymptotically consistent and normal.

Now, let's explore what statistical inferences we can learn about the quality of the independent variables to explain variation in the dependent variable, whether the estimated coefficients are statistically different from zero, and how we can construct confidence intervals around the population expected values of the estimators.

### 5.6.1 Goodness-of-fit statistics

Goodness-of-fit statistics provide a measure of how well the estimated regression line fits the observations. There are various measures, but we will discuss three: adjusted  $R^2$ , Akaike's information criteria, and the Bayesian information criteria. It is crucial to understand that when comparing models, the dependent variable and the number of observations (sample size) need to be the same. Otherwise, comparing goodness-of-fit measures is unintuitive.

#### Adjusted $R^2$

The most common measure of goodness-of-fit is the adjusted  $R^2$ . This measure reveals the proportion of variation in  $Y$  that is explained by the independent variables. An intuitive way to think about it is as follows:

1. Regress  $Y$  on all independent variables.
2. Regress  $Y$  on only the intercept term.
3. Compare how better the fit is in (1) relative to (2).

It is important to note that the  $R^2$  measure is simply the squared correlation between the variation in  $\hat{Y}$  and  $Y$ . That is, how well does the variation in  $\hat{Y}$  explain the variation in  $Y$ . Because this measure is a correlation, the  $R^2 \in [0, 1]$ .

The unadjusted  $R^2$  is defined as follows:

$$R^2 = (\text{Corr}[y_i, \hat{y}_i])^2 = \frac{(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y}))^2}{(\sum_{i=1}^n (y_i - \bar{y})^2) (\sum_{i=1}^n (\hat{y}_i - \bar{y})^2)}$$

If we are evaluating an OLS regression, then the special case of the above equation is as follows:

$$R^2 = 1 - \frac{\frac{1}{n-1} \sum_{i=1}^n \hat{\epsilon}_i^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$$

One problem with unadjusted  $R^2$  is that you get a higher value of this measure every time a new explanatory variable is added to the regression model. That is,  $\hat{\epsilon}^2$  always decreases as you add more covariates, so the unadjusted  $R^2$  always increases. This can be problematic when you attempt to add variables that have no true explanatory power. Therefore, a measure an adjusted  $R^2$  is used because it adds a penalty for including additional explanatory variables. Thus, if a new explanatory variable is added but it has no explanatory power, the adjusted  $R^2$  falls.

The adjusted  $R^2$  is as follows:

$$adj R^2 = 1 - \frac{\frac{1}{n-k} \sum_{i=1}^n \hat{\epsilon}_i^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2},$$

where  $k$  is the number of explanatory variables in the model. Adjusting for the explanatory variables reduces the degrees of freedom, thus reducing the goodness-of-fit measure if additional variables have no explanatory power.

It should be noted that the  $R^2$  is not necessarily a measure of the quality of the *statistical* model. It is simply a measure of the quality of the *linear approximation*. You cannot compare  $R^2$  measures of linear and non-linear models, because it is always the case that linear model  $R^2$  will be greater.

### Akaike's information criteria

As the case of the adjusted  $R^2$ , the AIC provides another goodness-of-fit measure that accounts for the trade-off between a better fit and having a simpler model (less explanatory variables). Intuitively, the AIC judges a model by how close its fitted values are to the true values. That is, the AIC penalizes models with more explanatory variables.

The AIC is as follows:

$$AIC = \log \left( \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2 \right) + \frac{2k}{n}$$

A lower AIC measure implies a better fit. It is important to note that AIC can only be used to compare models across a number of possible models; it cannot be used (as is the case with  $R^2$ ) to indicate the model fit directly.

### Bayesian information criteria

The Bayesian information criteria is similar to the AIC, but it places an even greater penalty on models with a high number of explanatory variables. The BIC is as follows:

$$BIC = \log \left( \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2 \right) + \frac{k}{n} \log(n)$$

As the case with AIC, a lower BIC implies a better fit.

## 5.6.2 Hypothesis testing

As we have seen before, we can design a hypothesis test to evaluate a sample statistic. This is the foundation for hypothesis testing in OLS models. The primary use of hypothesis tests is to evaluate whether estimated parameters are statistically significantly different from zero. That is, to test whether an explanatory variable has a statistically significant marginal effect in explaining variation within the dependent variable.

As before, we can set up a hypothesis test as follows:

1. Set up a null hypothesis:

$$H_0 : \beta_k = 0$$

2. Define the rejection rule. That is, specify the alternative hypothesis. Typically, we will simply use a two-sided test:

$$H_a : \beta_k \neq 0 \text{ (two-sided test)}$$

3. Set up the test statistic:

$$t_{stat} = \frac{\hat{\beta}_k - \beta_k}{se(\hat{\beta}_k)} = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)}$$

4. Choose a significance level,  $\alpha$ . The critical value is  $c = t_{n-k}(1 - \frac{\alpha}{2})$ . Note that we now have  $n - k$  degrees of freedom.
5. Specify the rejection rule. That is, if the rejection rule is satisfied, then you must reject the null hypothesis in favor of the alternative:  $|t| > c$
6. Determine if the statistic matches any of the rejection criteria. If so, reject the null hypothesis.

### 5.6.3 Confidence intervals

As before, we can also form a confidence interval around the population value of a parameter. The confidence interval tells us that for a significance level  $\alpha$ , in repeated sampling,  $(1-\alpha)\%$  of the intervals will contain the true value of  $\beta_k$ .

A confidence interval is as follows:

$$P(\hat{\beta}_k - t_{n-k,1-\alpha/2} \cdot se(\hat{\beta}_k) \leq \beta_k \leq \hat{\beta}_k + t_{n-k,1-\alpha/2} \cdot se(\hat{\beta}_k)) = 1 - \alpha$$

$$[\hat{\beta}_k - t_{n-k,1-\alpha/2} \cdot se(\hat{\beta}_k), \hat{\beta}_k + t_{n-k,1-\alpha/2} \cdot se(\hat{\beta}_k)]$$

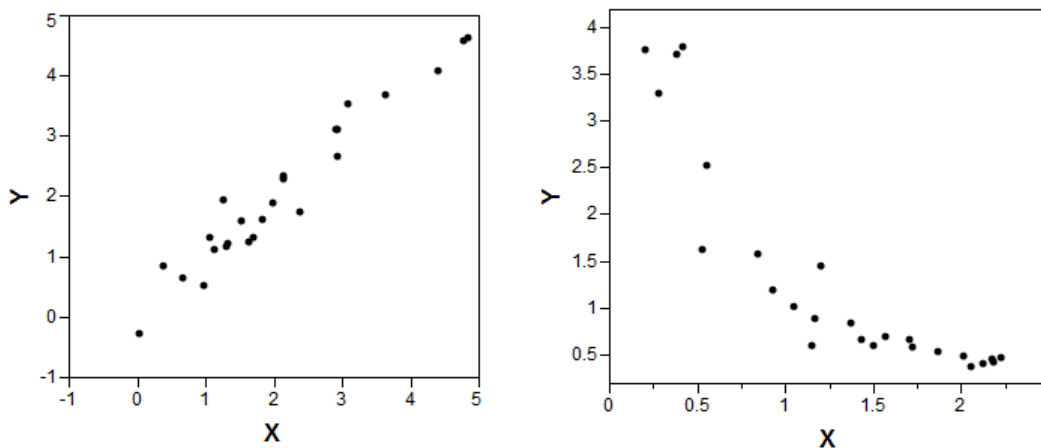
## 5.7 Functional Forms

A *functional form* refers to the depiction of the relationship between the dependent and independent variables. For example, in a basic linear model that we have explored, the marginal effect of independent variables is straightforward: a one-unit change in  $X$  will cause a  $\hat{\beta}$ -unit change in the expected value of  $Y$ .

$$\frac{\partial E(Y|X)}{\partial X} = \hat{\beta}$$

However, there are multiple factors that can alter how the relationship of the dependent and independent variables is specified. For example, you may plot a particular data set, and notice that a scatter plot of  $Y$  and  $X$  indicates a non-linear relationship. Figure 5.3 illustrates a linear and non-linear relationship of the dependent and independent variables.

Figure 5.3: Linear and Non-linear Relationships



Other factors that may cause a consideration of non-linear functional forms are theory, *a priori* information (experience), previous research, or the desire to test a hypothesis about the functional form.

We can specify non-linear relationships between  $Y$  and  $X$  in several ways. Two of the most common non-linear specifications are quadratic and logarithmic.

It should be noted that when we talk about non-linear relationships, we always talk about non-linearities in the dependent and/or independent variables. We cannot introduce non-linearities into the parameters, because we would no longer be able to estimate these models using OLS.

### 5.7.1 Quadratic relationships

An example of quadratic relationship is as follows:<sup>3</sup>

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

In this case, the estimated parameter  $\hat{\beta}_1$  and  $\hat{\beta}_2$  indicate that the variable  $X$  has a non-uniform marginal effect on  $Y$ . For example, if  $\hat{\beta}_1 > 0$  and  $\hat{\beta}_2 < 0$ , this implies that an additional unit of  $X$  increases  $Y$  but at a decreasing rate. So, the rate of the effect of  $X$  is dependent on the estimate of  $\hat{\beta}_2$ .

$$\frac{\partial E(Y|X)}{\partial X} = \hat{\beta}_1 + 2\hat{\beta}_2 X$$

This reveals that the marginal effect of  $X$  on  $Y$  is dependent on a particular observation of  $X$ .

### 5.7.2 Logarithmic relationships

In some cases, we may want to examine the marginal effect in elasticity terms. There are two approaches to estimating these relationships.

#### Log-linear models

The first type of model is a log-linear model. It also known as a *semi-elasticity* form.

$$\ln Y = \beta_0 + \beta_1 X + \varepsilon$$

The partial derivative is now of  $\ln Y$  with respect to  $X$ . That is:

$$\frac{\partial E(\ln Y|X)}{\partial X} = \hat{\beta}_1$$

If we multiply  $\hat{\beta}_1$  by 100, the marginal effect reveals the percentage change in  $Y$  with a one-unit change in  $X$ .

---

<sup>3</sup>Note that we typically model quadratic forms in a multiple regressor model. We will discuss these models later, but we will add one additional term to allow for more intuition of the squared regressor.

## Log-log models

The log-log model is a full elasticity model, which requires that you log both the dependent and independent variables.

$$\ln Y = \beta_0 + \beta_1 \ln X + \varepsilon$$

$$\frac{\partial E(\ln Y | \ln X)}{\partial \ln X} = \frac{\% \Delta Y}{\% \Delta X} = \hat{\beta}_1$$

Therefore, the estimated coefficient gives you the elasticity between  $Y$  and  $X$ . That is, it reveals the percentage change in  $Y$  due to a one-percent change in  $X$ .

### Example

Consider a Cobb-Douglas model:  $Q = \beta_0 K^{\beta_1} L^{\beta_2} e^{\varepsilon}$ . Obviously, this model cannot be directly estimated using OLS because it is non-linear in the parameters. However, take the logarithms of both sides yields an estimable linear model:

$$\ln Q = \ln \beta_0 + \beta_1 \ln K + \beta_2 \ln L + \varepsilon$$

Now, the estimated parameters  $\hat{\beta}_1$  and  $\hat{\beta}_2$  provide the elasticity measures of a one-percent change in capital and labor, respectively, on output.

## 5.8 Maximum Likelihood Estimation Approach

We will briefly depart from the ordinary least squares method to explore another technique for estimating linear regression models. More generally, the maximum likelihood estimation (MLE) method is a much more flexible and robust technique than least squares, because it can handle estimation of data that do not conform to the Gauss-Markov assumptions.

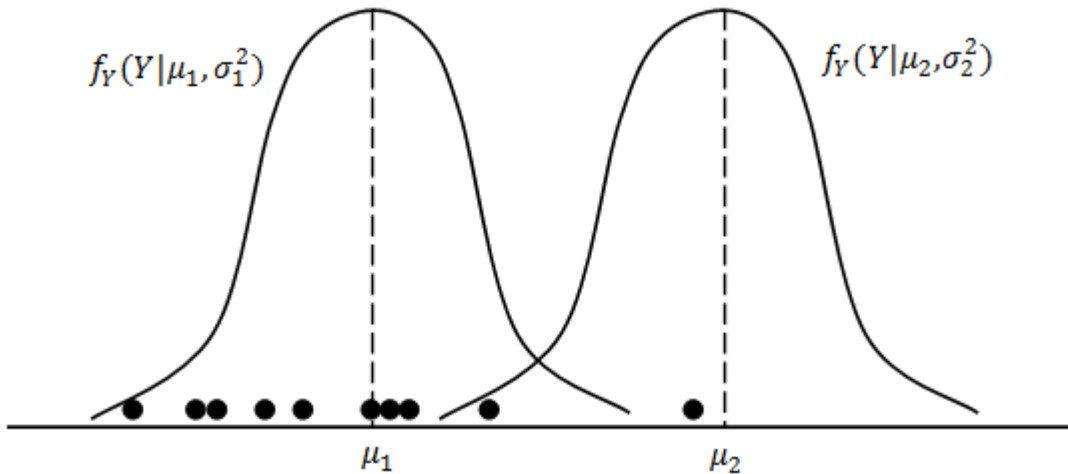
The principle of MLE is that the data we observe is associated with some probability distribution function, whose parameters are unknown. We then estimate these parameters based on the sample. That is, because we don't know the true population distribution, we must assume that the chosen random sample is our best approximation of the true

distribution. Therefore, we will attempt to find parameters of the pdf that maximize the likelihood that we will be able to draw the same sample from the estimated distribution.

For example, suppose that we have a sample of 30 recent college graduates, and we are interested in their starting salary. We assume that this is a random sample and, therefore, describes the population relatively well. Assuming that salary is distributed normally, we would like to determine the unknown mean and variance parameters that would specify a normal pdf which would be most likely to produce the same salary observations.

This principle can be illustrated in figure 5.4. The horizontal axis shows the observations in the sample and two possible normal pdfs with unknown parameters  $(\mu_1, \sigma_1^2)$  and  $(\mu_2, \sigma_2^2)$ , respectively. The pdf on the left (with mean  $\mu_1$  and variance  $\sigma_1^2$ ) appears to characterize the sample data much more accurately than the pdf on the right (mean  $\mu_2$  and variance  $\sigma_2^2$ ). Therefore, assuming that the sample data is the best representation of the population, we would choose the pdf with mean  $\mu_1$  and variance  $\sigma_1^2$  rather than the alternative. This pdf will maximize the likelihood that we will be able to draw the same sample again.

Figure 5.4: Illustration of pdfs Superimposed on Sample Points



However, the MLE method doesn't only look at two distributions. Rather, it considers all the possible pdfs that could be deduced from the sample data.

### 5.8.1 Likelihood function

Suppose that you observe a random sample  $\{Y_1, Y_2, Y_3, \dots, Y_n\}$  drawn from a distribution  $f_Y(Y|\mathbf{X}; \boldsymbol{\theta})$ , where  $\mathbf{X}$  denotes the set of regressors and  $\boldsymbol{\theta}$  represents the set of unknown

parameters that define the pdf (e.g. the mean and variance for a normal distribution). For each random variable, the marginal density is  $f_Y(y_n|x_n; \boldsymbol{\theta})$ ; the joint density for all of the observations in the sample is  $f(y_1|x_1, y_2|x_2, y_3|x_3, \dots, y_n|x_n; \boldsymbol{\theta})$ .

However, if we make a standard assumptions that observations in the sample independent, the joint density function can be written as follows:

$$f(y_1|x_1; \boldsymbol{\theta})f(y_2|x_2; \boldsymbol{\theta})f(y_3|x_3; \boldsymbol{\theta}) \cdots f(y_n|x_n; \boldsymbol{\theta}) = \prod_{i=1}^n f(y_i|x_i; \boldsymbol{\theta}).$$

This joint density is known as the *likelihood function*. The likelihood function can be written in a manner that allows us to maximize it over the set of unknown distribution parameters,  $\boldsymbol{\theta}$ .

$$\prod_{i=1}^n f(y_i|x_i; \boldsymbol{\theta}) = \max_{\{\boldsymbol{\theta}\}} \prod_{i=1}^n L_i(\boldsymbol{\theta}|y_i, x_i) = \max_{\{\boldsymbol{\theta}\}} L(\boldsymbol{\theta}|Y, \mathbf{X})$$

It is often much more convenient to evaluate the log of the likelihood function. The *log-likelihood function* is as follows:

$$\max_{\{\boldsymbol{\theta}\}} \sum_{i=1}^n \log L_i(\boldsymbol{\theta}|y_i, x_i)$$

### The score function

To maximize the log-likelihood function, we need to take the first-order condition with respect to  $\boldsymbol{\theta}$ :

$$s(\boldsymbol{\theta}) = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^n \frac{\partial \log L_i(\boldsymbol{\theta}|y_i, x_i)}{\partial \boldsymbol{\theta}} = 0$$

Solving for the parameters yields the vector  $\hat{\boldsymbol{\theta}}$ .

### The information matrix

The information is used to attain the covariance matrix of the ML estimator. Generally, the information matrix is as follows:

$$I_i(\boldsymbol{\theta}) = -E \left[ \frac{\partial^2 \log L_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]$$

The average information matrix for the entire sample is simply:  $I(\boldsymbol{\theta}) = \frac{1}{n} \sum_i I_i(\boldsymbol{\theta})$ . The asymptotic covariance matrix of the ML estimator is simply the inverse of the average information matrix:

$$\text{Var}(\boldsymbol{\theta}) = I(\boldsymbol{\theta})^{-1}$$

In practice, the approximate covariance matrix can be obtained by evaluating the information matrix at the estimated parameter values,  $\hat{\boldsymbol{\theta}}$ . That is:

$$\widehat{\text{Var}}(\boldsymbol{\theta}) = - \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log L_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \bigg|_{\hat{\boldsymbol{\theta}}} \right)^{-1}$$

Interestingly, we can also evaluate the covariance matrix using the first-order conditions; that is, using the product of the score functions.

$$\widehat{\text{Var}}(\boldsymbol{\theta}) = \left( \frac{1}{n} \sum_{i=1}^n s_i(\hat{\boldsymbol{\theta}}) s_i(\hat{\boldsymbol{\theta}})' \right)^{-1}$$

### 5.8.2 Properties of the ML estimator

The ML estimator has several attractive statistical properties:

1. The estimator is consistent:  $plim \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ .
2. The estimator is asymptotically efficient. That is, the ML estimator has the smallest variance among all consistent, asymptotically normal estimators.
3. The ML estimator is asymptotically normally distributed:  $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow N(0, \text{Var}(\boldsymbol{\theta}))$ .

Lastly, it should be noted that the MLE approach is a *parametric* estimation method. That is, we need to assume a marginal density prior to specifying a likelihood function.

### 5.8.3 Linear regression MLE

Suppose that we attempt to estimate the regression  $Y = \beta_0 + \beta_1 X + \varepsilon$ . In the maximum likelihood approach, we would seek to find  $\hat{\beta}_0$  and  $\hat{\beta}_1$  that maximize the likelihood of observation the observations in a sample.

Relying on the assumption that  $\varepsilon_i \sim N(0, \sigma^2)$ , we know that  $Y_i \sim N(\hat{\beta}_0 + \hat{\beta}_1 X, \sigma^2)$ . Therefore, each individual observation has a normal distribution, which can be described as follows:

$$f_Y(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right\}$$

We can define a likelihood function of the joint densities as follows:

$$L(\beta_0, \beta_1, \sigma^2 | Y, X) = f(y_1 | x_1; \beta, \sigma^2) \cdot f(y_2 | x_2; \beta, \sigma^2) \cdots (y_n | x_n; \beta, \sigma^2)$$

Plugging in the marginal densities, the likelihood function becomes:

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2 | Y, X) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(Y_1 - \beta_0 - \beta_1 X_1)^2}{2\sigma^2} \right\} \cdots \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(Y_n - \beta_0 - \beta_1 X_n)^2}{2\sigma^2} \right\} \\ &= \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \cdot \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2 \right\} \end{aligned}$$

The log-likelihood function follows:

$$\max_{\{\beta, \sigma^2\}} LL(\beta_0, \beta_1, \sigma^2 | Y, X) = n \cdot \ln \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2$$

Now, we can determine the first-order conditions by differentiating with respect to  $\beta_0$ ,  $\beta_1$ , and  $\sigma^2$ :

$$\left. \frac{\partial LL}{\partial \beta_0} \right|_{\hat{\beta}_0} = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial LL}{\partial \beta_1} \Big|_{\hat{\beta}_1} = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

$$\frac{\partial LL}{\partial \sigma^2} \Big|_{\hat{\sigma}^2} = \hat{\sigma}^2 - \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0$$

These are exactly the same first-order conditions as those derived using the ordinary least squares estimation approach. Therefore, for linear regression models, the OLS and MLE methods are almost equivalent. In the OLS approach,  $\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ , while the ML estimator does not account for the degrees of freedom. Asymptotically, this difference disappears because  $n$  becomes substantially larger than  $k$ . However, for small samples, the ML estimator of  $\sigma^2$  is biased. It should be noted that this is *not* a typical outcome when using ML estimations.

### 5.8.4 MLE: Example

Suppose that you are flipping a coin, such that there are two potential outcomes: heads ( $Y = 1$ ) or tails ( $Y = 0$ ). The pdf that would characterize this 0-1 event is Bernoulli, which is characterized by the function:

$$f_Y(y) = p^{y_i} (1 - p)^{(1-y_i)},$$

where  $p$  is the probability of attaining the outcome heads ( $Y = 1$ ). Therefore, the likelihood function is defined as follows:

$$L(p; Y) = f_Y(y_1) \cdot f_Y(y_2) \cdots f_Y(y_n) = p^{\sum_i y_i} (1 - p)^{\sum_i (1-y_i)}$$

Taking the log yields the log-likelihood function:

$$LL(p; Y) = \sum_i y_i \log(p) + (n - \sum_i y_i) \log(1 - p)$$

Suppose that we had the following sample:  $Y = \{1, 1, 0, 0, 1, 1, 1, 0, 0, 1\}$ . According to the principle of MLE, our best guess is that the sample is representative of the population,

and the parameter  $p$  which would maximize the probability of drawing the same sample is  $p = 0.6$ . However, let's prove this using two approaches.

First, take the first-order condition and evaluate it at  $\hat{p}$ . That is:

$$\frac{\partial LL(p; Y)}{\partial p} = \frac{\sum_i y_i}{p} - \frac{(n - \sum_i y_i)}{1 - p} \Big|_{\hat{p}} = 0$$

$$\hat{p} = \frac{1}{n} \sum_i y_i$$

Plugging in the values from the sample yields  $\hat{p} = 0.6$ , which is exactly what we had expected.

Suppose now that we wanted to use a grid-search method, where we chose starting values of  $p$  and determined which of these values maximized the log-likelihood. The following table describes several of these choices and outcomes:

$p = 0.3$	$LL = 6 \cdot \log(0.3) + 4 \cdot \log(0.7)$	$LL = -8.65$
$p = 0.5$	$LL = 6 \cdot \log(0.5) + 4 \cdot \log(0.5)$	$LL = -6.93$
$p = 0.6$	$LL = 6 \cdot \log(0.6) + 4 \cdot \log(0.4)$	$LL = -6.72$
$p = 0.7$	$LL = 6 \cdot \log(0.7) + 4 \cdot \log(0.3)$	$LL = -6.96$

Again, it the probability  $p = 0.6$  maximizes the log-likelihood function.

# Chapter 6

## Multiple regressor estimation

When we consider estimation of a simple linear regression, we implicitly assume that all other factors affecting  $Y$  are uncorrelated with the regressor  $X$ . This is often unrealistic. Therefore, we need to develop a mechanism that allows us to investigate how a number of factors affect variations in  $Y$ .

A *multiple regressor model* allows us to investigate the marginal effects of characteristics in addition to explicitly controlling for effects of other characteristics. This is important because we can better accommodate correlated explanatory variables.

Typically, adding more factors will result in a better explanation of variation in the dependent variable. Therefore, multiple regressor models are used to generate better predictions. Furthermore, we can use our knowledge of various functional forms to devise a better representation of the dependent variable.

### 6.1 Motivating the Use of Multiple Regressor Models

A natural question might be how does a multiple regressor estimation help us better understand marginal effects of explanatory variables? To start answering this question, let's recall the model that measured bull weight as a function of feed. However, knowing that bull weight is also determined by other factors, we believe that including these factors into the model will improve our predictions of bull weight from the modeled variables.

Suppose now we model bull weight as a function of feed and age. It is relatively easy to convince yourself that older bulls are likely to be heavier. However, because bull weight is

a random variable (there's that term again), neither feed nor wage can be used to perfectly predict weight. The two-variable model is as follows:

$$Weight = \beta_0 + \beta_1 \cdot Feed + \beta_2 \cdot Age + \varepsilon$$

This model tells us that weight is determined by two exogenous factors *Feed* and *Age*, as well as other unobservable components. You're still interested in understanding how *Feed* affects weight, but by including *Age* explicitly in the equation, you take that particular unobservable component *out* of the error term. That is, you are controlling for the effects of age on the weight.

Furthermore, because *Feed* and *Age* may be correlated (e.g., it's reasonable to assume that older calves are likely to consume more feed), adding the *Age* variable into the equation will affect the estimated marginal effect of *Feed*. Additionally, if *Age* is not included (and implicitly remains in the error term), then the error term becomes correlated with the regressor *Feed*. This causes bias of the estimator (more on this later).

The interpretation of estimated coefficients is also an important motivation for using multiple regressor models. The marginal effects in the above model are as follows:

$$\frac{\partial E[Weight|Feed, Age]}{\partial Feed} = \beta_1 \qquad \frac{\partial E[Weight|Feed, Age]}{\partial Age} = \beta_2$$

How do we interpret these coefficients? Carefully. That is, it is necessary to understand how to interpret each individual parameter, relative to all other factors. This is the notion of *ceteris paribus*, which loosely translates to "holding all other factors constant."

For example, the interpretation of  $\beta_1$  can be summarized as follows: the marginal effect of feed on the conditional expected value of weight, holding all other factors (in our case, age) constant. That is, if the age of all bulls is the same, then on average a one-unit increase in feed will change a bull's weight by  $\beta_1$  units. A similar statement can be made for  $\beta_2$ .

Therefore, when interpreting coefficients in a multiple regressor model, we examine marginal effects of one variable at a time. All others are *control variables*. A relatively intuitive example of how control variables can be interpreted is as follows:

Suppose that you are interested in explaining whether adding children's education programs would change the number of times patrons attend a library. However, you are worried that library attendance is also related to the annual income of patrons. That is, patrons with higher income may be, in general, less likely to attend libraries than patrons with lower income. Consider the following model:

$$Visits = \beta_0 + \beta_1 Program + \varepsilon$$

The information about income is implicitly included in the error term, so you are unable to explicitly understand how income affects library visitation. For example, increasing children's programs may affect visitation by a lot, but you might be capturing the effects in low income communities. Therefore, you cannot say anything about how programs affect visitation across all income groups. The following model does exactly this:

$$Visits = \beta_0 + \beta_1 Program + \beta_2 Income + \varepsilon$$

By including the *Income* variable you are able to *control* for the effects of income differences have on visitation. That is,  $\beta_1$  now tells you the marginal of adding children's programs after accounting for income differences. Therefore, this is a much better measure of children program effects on visitation.

We will return to a proof/illustration of this concept later in the chapter.

## 6.2 Estimating Multiple Regressor Models

We can generalize a simple linear regression to have as many explanatory variables as we choose:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_m X_m$$

where each  $\beta_i$  is the parameter associated with each explanatory variable,  $X_i$ . In matrix notation, however, this equation can be written exactly as we have seen before:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

The only change is that  $\mathbf{X}$  has dimensions  $(n \times m)$  instead of  $(n \times 2)$ , and the  $\boldsymbol{\beta}$  vector has dimensions  $(m \times 1)$ , rather than  $(2 \times 1)$ . This is the beauty of using matrix notation to find OLS estimators. To calculate  $\hat{\boldsymbol{\beta}}$ , calculate:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

This is exactly the same formula as that used for simple linear regression models, except that  $\hat{\boldsymbol{\beta}}$  has estimated coefficients for each column in the  $\mathbf{X}$  matrix.

The following properties of the OLS estimator can be calculated exactly as they were in a simple regression model:

Property	Equation
Regression residuals	$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$
Error variance	$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k}$
Variance-covariance	$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}} \mathbf{X}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$
Standard error	$\text{SE}(\hat{\beta}_m) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_m \mathbf{X})}$
t-statistic	$t_{\text{statistic}}(\hat{\beta}_m) = \frac{\hat{\beta}_m - \beta_m}{\text{SE}(\hat{\beta}_m)}$
Adjusted $R^2$	$\text{adj } R^2 = \left[ 1 - \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-k)}{(\mathbf{Y}-\hat{\mathbf{Y}})'(\mathbf{Y}-\hat{\mathbf{Y}})} \right]$

### 6.2.1 Partialling out

Let's return to the interpretation of the estimated coefficients, and attempt to understand why these are “partial” effects. We discussed the fact that partial effects help us understand the impact of a unit change in an explanatory variable after controlling for (holding constant) all other explanatory variables.

How does multiple regressor OLS account for this?

Consider the model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ . Now, suppose we perform the following steps:

1. Perform a simple regression:  $X_1 = \alpha_0 + \alpha_1 X_2 + u_1$ .
2. Retrieve the estimation residuals,  $\hat{u}_1$ .
3. Perform the regression  $Y = \gamma_0 + \beta_1 \hat{u}_1 + e$ .

By performing steps 1 and 2, we obtain the part of  $X_1$  that is uncorrelated with  $X_2$ . That is,  $\hat{u}_1$  is  $X_1$  after the effects of  $X_2$  have been partialled out/controlled for. Therefore, regressing  $Y$  on  $\hat{u}_1$  provides the marginal effect of  $X_1$  on  $Y$  *after* removing any effects that  $X_2$  may have had on  $Y$  through  $X_1$ .

For example, suppose that we wanted to measure the incomes of individuals of some ages and education levels. If we wanted to obtain the marginal effect of education levels after controlling for age, we would regress education on age, retrieve the residuals, and regress income on those residuals. The estimated coefficient is the marginal effect of education on income, after controlling for age (i.e., keeping age constant).

What happens if  $X_1$  and  $X_2$  are orthogonal (independent)? In this case, there is no correlation between the two independent variables and the following theorem of orthogonal partitioned regression is true:

*In a multiple linear least squares regression of  $\mathbf{Y}$  on  $\mathbf{X}_1$  and  $\mathbf{X}_2$ , if the two independent variables are orthogonal, then the marginal effects of each variable obtained from estimating the model:  $\mathbf{Y} = f(\mathbf{X}_1, \mathbf{X}_2) + \varepsilon$  are the same as performing two simple linear regressions:  $\mathbf{Y} = f(\mathbf{X}_1) + \varepsilon_1$  and  $\mathbf{Y} = f(\mathbf{X}_2) + \varepsilon_2$ .*

## 6.3 Variable Selection

An important step in specifying models is the choice of included regression variables. We have discussed the fact that not explicitly accounting for a variable correlated with the dependent variable (that is, leaving the variable in the error term) will bias the estimator. Alternatively, you may include too many variables in the model, therefore overspecifying the regression. Although the consequence of overspecification is not as dire as omitting variables, it reduces the efficiency of estimators, and should, therefore, be avoided.

### 6.3.1 Bias by Omission

Why does variable omission cause bias?

Suppose that we have a correctly specified model:  $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \varepsilon$ . However, you actually regress  $\mathbf{Y}$  on  $\mathbf{X}_1$  only. From this regression, you retrieve:

$$\hat{\boldsymbol{\beta}}_1 = \mathbf{X}'_1\mathbf{X}_1^{-1}\mathbf{X}'_1\mathbf{Y}$$

As we did before, we can use the definition of  $\mathbf{Y}$  to plug into the above equation, yielding:

$$\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1 + \mathbf{X}'_1\mathbf{X}_1^{-1}\mathbf{X}'_1\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{X}'_1\mathbf{X}_1^{-1}\mathbf{X}'_1\varepsilon$$

For the estimator of  $\beta_1$  to be unbiased, either of the following two events must occur:

$$(1) \quad \mathbf{X}'_1 \mathbf{X}_2 = 0$$

$$(2) \quad \beta_2 = 0$$

Otherwise,  $\beta_1$  is subject to *omitted variable bias*.

### Direction of omitted variable bias

The direction in which the estimator is biased depends on the correlation between  $\mathbf{X}_1$  and  $\mathbf{X}_2$ .

1. If  $\text{Corr}[\mathbf{X}_1, \mathbf{X}_2] > 0$ , then  $E[\tilde{\beta}] > \beta$ . This is known as *upward bias*, and causes the estimated coefficient to appear to have a larger marginal effect on  $\mathbf{Y}$  than the variable actually has.
2. If  $\text{Corr}[\mathbf{X}_1, \mathbf{X}_2] < 0$ , then  $E[\tilde{\beta}] < \beta$ . This is known as *downward bias*, and causes the estimated coefficient to appear to have a smaller marginal effect on  $\mathbf{Y}$  than the variable actually has.

For example, suppose that the true function of bull weight is pounds of feed and the bull's age. However, we estimate bull weight only as a function of feed. If  $\text{Corr}[\text{Feed}, \text{Age}] > 0$  (which seems reasonable, since you might feed an older bull more feed), then we would falsely believe that feed affects weight by more than its true effect. This might cause us to underfeed a bull.

### 6.3.2 Inclusion of Irrelevant Variables

Include variables into a model that are not relevant to explaining the dependent variable is known as *overidentification*. Typically, overidentification is less problematic than omitted variable bias, because the estimated parameters remain unbiased. That is, suppose a model  $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \varepsilon$ , in which  $\mathbf{X}_2$  is actually irrelevant to explaining variation in  $\mathbf{Y}$ . Because the matrix  $\mathbf{X}_1$  is used to control for all relevant factors,  $\hat{\boldsymbol{\beta}}_2 = 0$ . Therefore,  $E[\hat{\boldsymbol{\beta}}_1] = 0$ .

However, blindly overspecifying a regression model is still not a good strategy, because overspecification increases the variance of the estimator. That is, suppose that  $\hat{\boldsymbol{\beta}}_1$  represents the OLS estimator from a correctly specified model and  $\tilde{\boldsymbol{\beta}}_1$  is the OLS estimator from an overspecified model. It is always the case that  $[\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_1)] < [\widehat{\text{Var}}(\tilde{\boldsymbol{\beta}}_1)]$ , because  $\hat{\sigma}^2$  is always larger when  $k$  (the number of variables) increases.

For example, suppose that  $\hat{\varepsilon}'\hat{\varepsilon} = 4$  and  $n = 6$ . Then:

$$\text{Correctly specified model } (k = 2) : \hat{\sigma}^2 = \frac{1}{4} \times 4 = 1$$

$$\text{Overspecified model } (k = 3) : \hat{\sigma}^2 = \frac{1}{3} \times 4 = \frac{4}{3}$$

Therefore, the estimator from the overspecified model is less efficient.

### 6.3.3 Strategies for Selecting Regressors

The general approach for selecting modeling variables is economic theory. As a researcher, you need to investigate what is suggested by good economic intuition, what has been shown in previous literature, and what seems to be a reasonable hypothesis for a foundation of your empirical model.

In the past 20-30 years, there has been a fundamental shift regarding modeling methodologies. Before, the *simple-to-general* approach was most popular, because researchers would start with a very simple model, and continue to add variables. However, this approach is almost certainly doomed to fail, because any criteria for choosing whether to add a variable is tainted by the omitted variable bias.

Another technique may be data mining. That is, trying to estimate the model with every possible combination of variables until you find a model that either confirms a particular hypothesis or best fits the data (e.g., highest  $R^2$ ). However, this may not yield the most relevant model, because these types of methods are always subject to type I errors: rejecting the null hypothesis that the model is poor even though the model is actually poor. For example, statistical anomalies may cause data mining techniques to select a model that has very little economic intuition.

A technique that has emerged most recently is based on the principle of *general-to-specific*. The emergence has been substantially aided by the advances in both computer software and hardware, which help quickly and effectively estimate relatively complex models. The general-to-specific approach first seeks to specify the most complex model possible, and then attempt to simplify the model. That is, if a complex model does a good job of explaining reality, but a more parsimonious model does no worse, then the simple model should be selected.

For most researchers, the best approach may be to start “somewhere in the middle.” That is, using economic theory and intuition and develop a model that is between very complex and very simple. Then, consider two strategies for either expanding or simplifying the specification:

1. Test whether current assumptions and model restrictions are valid. For example, in an OLS regression, it would be useful to test whether the assumptions of homoskedasticity and no autocorrelation are valid.
2. Analyze whether additional restrictions could be made. For example, test whether certain parameters are not significantly different from zero and could be restricted to be zero.

## 6.4 Indicator/Dummy Variables

Dummy / categorical / indicator variables are used to model more qualitative relationships between the dependent and independent variables. For example, indicator variables can be used to specify that a particular person is a male or female, and whether he/she has a college degree or does not. By definition, dummy variables are mutually exclusive – a person cannot simultaneously have and not have a college degree. Therefore, dummy variables answer “either / or” questions.

In a regression, dummy variables are intercept shifters. That is, an estimated parameter associated with a dummy variable can be interpreted as follows:

By being in one group rather than another (e.g., being male rather than female), the expected conditional effect is given by the estimated coefficient value.

For example, suppose that we are measuring monthly turkey prices. We know that during the holiday season (say, October through December) there is a much higher demand for turkeys, which would raise turkey prices above their expected values throughout the remainder of the year. We can model this as follows:

$$Price_{turkey} = \beta_0 + \beta_1 HS + \varepsilon ,$$

where  $HS$  is a dummy variable which equals 1 when the month is October, November, or December, and 0 otherwise. To see the marginal effect of holiday seasons, consider the conditional expected values:

$$\begin{aligned} E[Price_{turkey}|HS = 1] &= \hat{\beta}_0 + \hat{\beta}_1 \\ E[Price_{turkey}|HS = 0] &= \hat{\beta}_0 \\ E[Price_{turkey}|HS = 1] - E[Price_{turkey}|HS = 0] &= \hat{\beta}_1 \end{aligned}$$

Therefore,  $\hat{\beta}_0$  explains the average price throughout the entire year and  $\hat{\beta}_1$  explains the difference between turkey prices during the holiday season and turkey prices during the rest of the year. In the preceding case, the benchmark/control is non-holiday months, and the treatment is holiday season.

Certainly, models can be generalized to include both continuous and indicator variables. For example, we know that turkey prices are likely affected by chicken prices, which is a continuous variable. This model can be written as follows:

$$Price_{turkey} = \beta_0 + \beta_1 HS + \beta_2 Price_{chicken} + \varepsilon$$

Now,  $\hat{\beta}_1$  tells us the difference between holiday and non-holiday turkey prices *after* controlling for variation in turkey prices due to variation in chicken prices.

### 6.4.1 Interpreting Dummy Variables in a Double-log Model

Suppose that you are estimating a constant elasticity (double-log) model of turkey prices:

$$\ln Price_{turkey} = \beta_0 + \beta_1 HS + \beta_2 \ln Price_{chicken} + \varepsilon$$

The estimated marginal effect  $\hat{\beta}_1$  can be interpreted as follows: during the holiday season, turkey prices are  $(\hat{\beta}_1 \times 100)\%$  higher (or lower) than turkey prices during other months of the year. For example, if  $\hat{\beta}_1 = 0.04$ , then during holiday seasons, turkey prices are 4% higher than during non-holiday months.

### 6.4.2 Dummy Variable Trap

It is necessary to caution against falling into the *dummy variable trap*.

Suppose that you wanted to explicitly control for turkey prices in holiday and non-holiday months. Then, you would set up the model as follows:

$$Price_{turkey} = \beta_0 + \beta_1 HS + \beta_2 NonHS + \varepsilon$$

However, this regression cannot be performed because the  $\mathbf{X}$  matrix is not full rank:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

In the above  $\mathbf{X}$  matrix, the intercept term is a linear combination of columns 2 and 3 (what does this imply about the rank of  $\mathbf{X}$ ?). You will not be able to invert the matrix  $(\mathbf{X}'\mathbf{X})$ . Therefore, you would need to remove the first (intercept) column from the  $\mathbf{X}$  matrix. However, doing so makes it difficult to interpret the estimated coefficients.

### 6.4.3 Dummy Variables Describing Multiple Categories

Dummy variables can also be used when you are describing more than two categories. For example, suppose that you are interested in estimating a linear probability model that predicts whether an individual is a registered voter.

$$Voter = \beta_0 + \beta_1 Income + \beta_2 Education + \varepsilon ,$$

where *Income* is a continuous variable describing the person's annual income and education is a categorical variable describing the level of education:

$$Education = \begin{cases} \text{Less than high school} \\ \text{High school} \\ \text{Bachelors} \\ \text{Masters} \end{cases}$$

This can be set up to be four yes/no indicator variables:

$$Edu\_NoHS = \begin{cases} 1, & \text{if Less than high school} \\ 0, & \text{otherwise} \end{cases}$$

$$Edu\_HS = \begin{cases} 1, & \text{if High school} \\ 0, & \text{otherwise} \end{cases}$$

$$Edu\_Ba = \begin{cases} 1, & \text{if Bachelors} \\ 0, & \text{otherwise} \end{cases}$$

$$Edu\_Ma = \begin{cases} 1, & \text{if Masters} \\ 0, & \text{otherwise} \end{cases}$$

Then, set up the estimation regression as follows:

$$Voter = \beta_0 + \beta_1 Income + \beta_2 Edu\_NoHS + \beta_3 Edu\_HS + \beta_4 Edu\_Ba + \beta_5 Edu\_Ma + \varepsilon$$

Setting up the equation as above would lead you to fall into the dummy variable trap. Therefore, you would need to exclude one of the indicator variables. Which one to exclude? Commonly, it is best to exclude the variable that has the most observations in the sample (population). For example, if the most common outcome was a person with a high school degree, then excluding *Edu\_HS* and comparing all other marginal effects relative to the high school degree control group would be most intuitive. That is,  $\hat{\beta}_4$  would reveal the marginal effect on the probability that an individual is a voter if that individual has a bachelors degree rather than a high school degree.

### An alternative specification

In the preceding example, using categorical dummy variables was useful, but it didn't necessarily give you an indication of how much the probability of being a voter changed with earning an additional degree. The set up in the preceding problem only revealed the difference in voter probability between different education groups.

An alternative method to set up a categorical variable regression is to specify all of the groups to which a person belongs. For example, if a person has a bachelors degree, then they also have a high school degree, but not a masters. A person with a masters has all possible degrees, and a person with no high school diploma has no degrees. The following table is an example of how to set up these types of variables:

Obs.	HS	Ba	Ma
1	1	1	0
2	1	0	0
3	1	1	1
4	0	0	0

In this type of model, each estimated coefficient indicates the marginal effect of having the additional degree. Therefore,  $\hat{\beta}_{Ba}$  would indicate the additional probability of being a voter from having a college education.

## 6.5 Interaction Terms

Interaction terms are useful in revealing marginal effects of joint relationships. For example, suppose we return to our example of measuring bull weights as a function of age and feed. Suppose we hypothesize that as both a bull's age and feed increase, the effect on weight will be greater. We can test this hypothesis by modeling the following regression equation:

$$Weight = \beta_0 + \beta_1 Age + \beta_2 Feed + \beta_3 (Age \times Feed) + \varepsilon$$

The estimated parameter  $\hat{\beta}_3$  tells us the joint impact on a bull's weight of increasing feed as the bull gets older.

You can define interaction terms as combinations of continuous variables, dummy variables, or both.

## 6.6 Inferences

In this section, we will take a look at additional insights we can learn from estimating a multiple regressor OLS model. Specifically, we will take close look at two examples of linear tests, which are both special cases of the Wald test.

It is important to note that the following tests only deal with nested models. Non-nested tests are out of the scope of this course.

### 6.6.1 Single linear restriction test: a $t$ -test

First, we will return to the simple  $t$ -test. Intuitively, the  $t$ -test allows us to ask the following question: after we have estimated a set of parameters, do these estimates come close to satisfying a particular linear restriction (e.g.,  $\hat{\beta} - \beta = 0$ ). That is, we test whether the failure to satisfy this linear condition is due to sampling error or if the failure is systematic. Recall that the  $t$ -statistic is calculated as follows:

$$t_{stat} = \frac{\hat{\beta} - 0}{SE(\hat{\beta})}$$

In a simple linear regression this was easy to interpret, because we only dealt with a single variable. Now, we must deal with multiple variables. However, this “hardship” is also what will allow us to perform more complex and interesting tests (discussed later).

Suppose we estimate a model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ . We are interested in learning whether  $\hat{\beta}_1 = 0$ . Well, we can impose a *linear restriction* on just the  $\beta_1$  coefficient as follows:

$$0 \cdot \beta_0 + 1 \cdot \beta_1 + 0 \cdot \beta_2 + 0 \cdot \beta_3 = 0 \equiv \beta_1 = 0$$

Alternatively, if we defined a vector  $\mathbf{R}_{J \times K} = [0 \ 1 \ 0 \ 0]$  and a vector  $\mathbf{q} = 0$ , then we can impose the same linear restriction in matrix form as:

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$$

Therefore, we can set up a  $t$ -test as follows:

$$\begin{aligned} \mathbf{R}\hat{\boldsymbol{\beta}} &= \mathbf{q} \\ \mathbf{R}\hat{\boldsymbol{\beta}} &\neq \mathbf{q} \end{aligned}$$

and the associated  $t$ -statistic is as follows:

$$t_{stat} = \frac{\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}}{\text{SE}(\mathbf{R}\hat{\boldsymbol{\beta}})} = \frac{\hat{\boldsymbol{\beta}} - 0}{\text{SE}(\mathbf{R}\hat{\boldsymbol{\beta}})} \sim t_{n-k, 1-\alpha/2}$$

## 6.6.2 General linear restrictions

Now, suppose that we were not only interested a single a variable against a scalar, but rather we wanted to explore more interesting dynamics. Let's consider three examples that illustrate some of the dynamics that you can test.

### Example 1: Equality of coefficients

Suppose that you were estimating wheat yields as a function of several brands of fertilizers. You were interested in knowing if the marginal effect of one fertilizer brand is the same as the marginal effect of another fertilize brand. That is, in the model above, you might be interested in determining whether  $\beta_1 = \beta_2$ . Alternatively, whether  $\beta_1 - \beta_2 = 0$ . To test this, set up a linear restriction as follows:

$$\mathbf{R}_{1 \times 4} = [0 \ 1 \ -1 \ 0] \quad \mathbf{q} = 0$$

where the null and alternative hypotheses are the same as in the  $t$ -test example. In this case, there are  $J = 1$  restrictions.

**Example 2: Multiple linear restrictions**

Suppose that we would like to test the null hypothesis of some joint relationships between coefficients. That is, we want to test whether fertilizer brands 1 and 2 affect yields identically, while simultaneously, the third fertilizer has no effect on yields. In this case, there are two restrictions:

$$\mathbf{R}_{2 \times 4} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{q} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

In this case, there are  $J = 2$  linear restrictions that can be tested.

**Example 3: Testing that all coefficients are zero (F-test)**

A common test statistic you will encounter in the output of most statistical packages is the  $F$ -test. This tests for the joint insignificance of all estimated coefficients. That is, if none of the estimated parameters are statistically significantly different from zero, then the regression model is not very helpful.

Therefore, we are interested in testing whether:  $H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$ . The alternative hypothesis is that at least one of the coefficients is statistically significant from zero. You will often see this test set up as follows:

$$F_{\text{statistic}} = \frac{R^2/(k-1)}{1-R^2/(n-k)} \sim F_{J, n-k}$$

where  $J$  is the number of coefficients being tested (typically, all but the intercept term) and  $k$  is all of the parameters in the model (including the intercept). If the null hypothesis cannot be rejected ( $|F_{\text{stat}}| < \text{critical value}$ ), then this suggests that the model is poorly specified.

Example

Suppose you estimate a linear regression model with 3 independent variables, one intercept, and 200 observations. The adjusted  $R^2$  is 0.75. To determine whether at least one of the independent variables explains variation in the dependent variable, you must perform the  $F$ -test.

First, set up the  $F$ -statistic:

$$F_{stat} = \frac{R^2/(k-1)}{1-R^2/(n-k)} = \frac{0.75/3}{0.25/196} = 196$$

Then, determine the  $F$  critical value using the  $F$  probability table.

$$F_{crit} = F_{J,n-k} = F_{4,196} = 2.37$$

Because  $|F_{stat}| > F_{crit}$ , you must reject the null hypothesis that none of the regressors explain variation in the dependent variable. Therefore, at least one regressor is relevant in explaining variation in the dependent variable. Note, however, that the  $F$  test *does not* tell us which regressors explain variation and which do not.

A more general method of specifying the  $F$ -test is by using the general linear restrictions approach. That is, the restrictions matrix is specified as follows:

$$\mathbf{R}_{J \times K} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \quad \mathbf{q} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

### 6.6.3 Test statistic and criterion: Wald Test

Now that we know how to set up linear restrictions, we are interested in using them to actually test for the relationships we are interested in. That is, we seek to test whether  $\mathbf{R}\boldsymbol{\beta} - \mathbf{q} = \mathbf{m}$ . The expected value of  $\mathbf{m}$  is zero, and we seek to find whether deviations of  $\mathbf{m}$  from zero are just sampling errors or whether they are significant to reject the null hypothesis.

To derive this, we start with the variance of  $\mathbf{m}$ , because we would like to know how “certain” we can be in rejecting the null (larger variance implies less precision and lack of evidence to reject the null).

$$\begin{aligned} \text{Var}(\mathbf{m}|\mathbf{X}) &= \text{Var}(\mathbf{R}\boldsymbol{\beta} - \mathbf{q}|\mathbf{X}) \\ &= \mathbf{R}(\text{Var}(\boldsymbol{\beta}|\mathbf{X}))\mathbf{R}' \\ &= \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' \end{aligned}$$

Using this variance, a test of the null hypothesis can be performed by testing a *Wald* criterion against a critical value from the Chi-squared distribution with  $J$  degrees of freedom ( $\chi_J^2$ ).

$$\begin{aligned} W &= \mathbf{m}'\{\text{Var}(\mathbf{m}|\mathbf{X})\}^{-1}\mathbf{m} \\ &= (\mathbf{R}\boldsymbol{\beta} - \mathbf{q})'\{\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\}^{-1}(\mathbf{R}\boldsymbol{\beta} - \mathbf{q}) \\ &= \frac{(\mathbf{R}\boldsymbol{\beta} - \mathbf{q})'\{\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\}^{-1}(\mathbf{R}\boldsymbol{\beta} - \mathbf{q})}{\sigma^2} \end{aligned}$$

Formally:

$$\frac{(\mathbf{R}\boldsymbol{\beta} - \mathbf{q})'\{\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\}^{-1}(\mathbf{R}\boldsymbol{\beta} - \mathbf{q})}{\sigma^2} \sim \chi_J^2$$

Of course, because we don't know the population error variance, we will again need to use the sampling error variance,  $\hat{\sigma}^2$ .

### The F-test, again

The general Wald statistic can be modified slightly to characterize the *F*-test:

$$F_{stat} = \frac{(\mathbf{R}\boldsymbol{\beta} - \mathbf{q})'\{\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\}^{-1}(\mathbf{R}\boldsymbol{\beta} - \mathbf{q})}{J\hat{\sigma}^2} \sim F_{J,n-k}$$

### Example 1

Suppose that you estimate the model:  $\mathbf{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ . You then retrieve the following results:

$$\begin{array}{ccc} \beta_0 & \beta_1 & \beta_2 \\ -141.2 & 1.27 & 3.60 \\ (33.38) & (3.11) & (0.90) \end{array}$$

$$\begin{array}{ccc} R^2 & \hat{\varepsilon}'\hat{\varepsilon} & n \\ 0.744 & 2120 & 19 \end{array}$$

The values in parentheses are standard errors. Also, the matrix  $(\mathbf{X}'\mathbf{X})^{-1}$  is as follows:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 8.41 & 0.103 & -0.156 \\ 0.103 & 0.073 & -0.017 \\ -0.156 & -0.017 & 0.006 \end{bmatrix}$$

Test:  $\beta_1 = \beta_2$

First, set up the  $\mathbf{R}$  and  $\mathbf{q}$  matrices:

$$\mathbf{R} = [0 \ 1 \ -1] \quad \mathbf{q} = [0]$$

So, because  $\mathbf{R}$  has only one row, there are  $J = 1$  restrictions. Now, set up the components of the  $F$  statistic:

$$F_{stat} = \frac{(\mathbf{R}\boldsymbol{\beta} - \mathbf{q})' \{ \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \}^{-1} (\mathbf{R}\boldsymbol{\beta} - \mathbf{q})}{J\hat{\sigma}^2}$$

1. Determine the  $\hat{\sigma}^2$  term:

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n - k} = \frac{2120}{19 - 3} = 132.5$$

2. Determine the difference  $(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})$ :

$$(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}) = [0 \ 1 \ -1] \cdot \begin{bmatrix} -141.2 \\ 1.27 \\ 3.60 \end{bmatrix} - [0] = -2.33$$

3. Determine the product  $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ :

$$\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' = [0 \ 1 \ -1] \cdot \begin{bmatrix} 8.41 & 0.103 & -0.156 \\ 0.103 & 0.073 & -0.017 \\ -0.156 & -0.017 & 0.006 \end{bmatrix} \cdot [0 \ 1 \ -1]' = 0.114$$

4. Calculate the  $F$  statistic:

$$F_{stat} = \frac{(-2.33)' \{0.114\}^{-1} (-2.33)}{1 \cdot 132.5} = 0.359$$

Now that you have the  $F$  statistic, you need to retrieve a critical value from the  $F$  distribution to determine whether you can or cannot reject the null hypothesis. At a 95% confidence level, the critical value is as follows:

$$F_{crit} = F_{1,16} = 4.49$$

Because  $|F_{stat}| < F_{crit}$ , you cannot reject the null hypothesis that  $\beta_1 = \beta_2$ .

### Example 2

Suppose that you estimate the model:  $\mathbf{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ . You then retrieve the following results:

$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$
0.7696	0.7846	-0.1308	0.0875
(0.7082)	(0.0912)	(0.0397)	(0.0367)

$R^2$	$\hat{\varepsilon}'\hat{\varepsilon}$	$n$
0.332	880.61	428

The values in parentheses are standard errors. Also, the matrix  $(\mathbf{X}'\mathbf{X})^{-1}$  is as follows:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.24 & -0.029 & 0.0005 & -0.006 \\ -0.02 & 0.004 & 0.0003 & 0.0003 \\ 0.0005 & 0.0003 & 0.0007 & -0.006 \\ -0.006 & 0.0003 & -0.0006 & 0.0006 \end{bmatrix}$$

Joint Test:  $\beta_0 = 0$  and  $\beta_1 + \beta_2 = 0$

First, set up the  $\mathbf{R}$  and  $\mathbf{q}$  matrices:

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \mathbf{q} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

So, because  $\mathbf{R}$  has two rows, there are  $J = 2$  restrictions. Now, set up the components of the  $F$  statistic:

$$F_{stat} = \frac{(\mathbf{R}\boldsymbol{\beta} - \mathbf{q})' \{ \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \}^{-1} (\mathbf{R}\boldsymbol{\beta} - \mathbf{q})}{J\hat{\sigma}^2}$$

1. Determine the  $\hat{\sigma}^2$  term:

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n - k} = \frac{880.61}{428 - 4} = 2.077$$

2. Determine the difference  $(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})$ :

$$(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0.7696 \\ 0.7846 \\ -0.1308 \\ 0.0875 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.7696 \\ -0.0433 \end{bmatrix}$$

3. Determine the product  $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ :

$$\{ \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \}^{-1} = \begin{bmatrix} 16.03 & 515.4 \\ 515.4 & 22344 \end{bmatrix}$$

4. Calculate the  $F$  statistic:

$$F_{stat} = 4.11$$

Now that you have the  $F$  statistic, you need to retrieve a critical value from the  $F$  distribution to determine whether you can or cannot reject the null hypothesis. At a 95% confidence level, the critical value is as follows:

$$F_{crit} = F_{2,424} = 3.00$$

Because  $|F_{stat}| > F_{crit}$ , you reject the null hypothesis that jointly  $\beta_0 = 0$  and  $\beta_1 + \beta_2 = 0$ .

It is important to note, however, that you cannot tell which one of the restrictions is causing the null hypothesis to be rejected. You can only make inferences about the joint test, *not the individual components*.

## 6.7 Testing for Structural Breaks

A useful and interesting question may be one that asks whether all parameters estimated using one subsample of data are different from parameters estimated using another subsample. That is, determining whether there is a *structural difference* between two or more subsamples.

For example, one might be interested in modeling airline ticket prices as a function of airline costs and other characteristics. However, after the events of September 11, 2001, we may expect that there was a structural shift in the magnitude or direction of the regressors' marginal effects. Alternatively, in our bull weight modeling example, we may be interested in determining whether the marginal effects of *Age* and *Feed* are different for the Angus breed and the Limousin breed.

Consider the general specification of the bull weight as follows:

$$\begin{bmatrix} \mathbf{Y}_A \\ \mathbf{Y}_L \end{bmatrix} = \begin{bmatrix} \mathbf{X}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_L \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{\beta}_A \\ \boldsymbol{\beta}_L \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_A \\ \boldsymbol{\varepsilon}_L \end{bmatrix}$$

Generalizing the matrices to be  $\mathbf{Y}$ ,  $\mathbf{X}$ , and  $\boldsymbol{\beta}$ , we can perform the typical OLS regression and retrieve the set of estimated parameters as follows:

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} = \begin{bmatrix} \mathbf{X}'_A \mathbf{X}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{X}'_L \mathbf{X}_L \end{bmatrix}^{-1} \cdot \begin{bmatrix} \mathbf{X}'_A \mathbf{Y}_A \\ \mathbf{X}'_L \mathbf{Y}_L \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_A \\ \hat{\boldsymbol{\beta}}_L \end{bmatrix}$$

In this case, you can calculate the residuals for each subsample to be  $\hat{\varepsilon}_A$  and  $\hat{\varepsilon}_L$ , and then determine the sum of squared residuals (SSR). The total SSR for the model is the sum of the individual SSRs:

$$\hat{\varepsilon}'\hat{\varepsilon} = \hat{\varepsilon}'_A\hat{\varepsilon}_A + \hat{\varepsilon}'_L\hat{\varepsilon}_L$$

Formally understanding whether all coefficients estimated with one subsample are statistically different from coefficients estimated the other subsample requires that you specify and test the null hypothesis:  $H_0 : \beta_A = \beta_L$ . Alternatively, this can be specified as:  $H_0 : \beta_A - \beta_L = 0$ .

### 6.7.1 $F$ test approach

One method to test for a structural break is using linear restrictions and testing an  $F$  statistic against the appropriate critical value. The appropriate  $\mathbf{R}$  matrix is as follows:

$$\mathbf{R} = [\mathbf{I} : -\mathbf{I}] = \begin{bmatrix} 1 & 0 & \cdots & 0 & \vdots & -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \vdots & 0 & -1 & \cdots & 0 \\ \vdots & & \ddots & \vdots & \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & \vdots & 0 & 0 & \cdots & -1 \end{bmatrix}$$

The matrix  $\mathbf{R}$  is  $[(k_A + k_L) \times (k_A + k_L)]$  dimensions matrix, where  $k_A$  and  $k_L$  correspond to the number of parameters in the  $\beta_A$  and  $\beta_L$  vectors, respectively. The associated  $\mathbf{q}$  is a  $[(k_A + k_L) \times 1]$  vector of zeros. That is,  $\mathbf{q} = \mathbf{0}_{(k_A+k_L) \times 1}$ . Lastly, because there are  $(k_A + k_L)/2$  rows in the  $\mathbf{R}$  matrix, the number of restrictions is  $J = (k_A + k_L)/2$ . The associated degrees of freedom are  $n = n_A + n_L - (k_A + k_L)$ .

From here, you can derive the appropriate  $F$  statistic and the associated  $F$  critical value.

## 6.7.2 Lagrange multiplier test approach

You may also develop an alternative test for determining structural breaks. Let's start with the intuition.

This type of hypothesis test focuses on the fit of regression models. The test questions whether choosing one model over another significantly reduces the fit of the regression line. Alternatively, whether imposing a particular restriction improves the model fit. In the case of the structural break analysis, we are interested in determining whether the fit of the regression line is reduced when using a general/combined set of parameters rather than different parameters for each subsample. This is the intuition foundation for a Lagrange multiplier test.

There is an important difference between a Wald and Lagrange multiplier tests. To use the Wald test, you first estimate the parameters, and then test whether these parameters statistically significantly satisfy the restriction described by the null hypothesis. The Lagrange multiplier test imposes an *a priori* restriction, and then examines whether the restriction has caused a statistically significant reduction in the model fit.

Recall that in an unrestricted (typical) OLS model, the estimated parameters minimize the function:

$$\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Now, suppose that you wish to impose a linear restriction and recalculate the parameters subject to this restriction. In the typical manner, we can specify the restriction to be  $\mathbf{R}\tilde{\boldsymbol{\beta}} = \mathbf{q}$ , where  $\tilde{\boldsymbol{\beta}}$  denotes the set of restricted parameters. Now, the minimization becomes:

$$\min_{\tilde{\boldsymbol{\beta}}} S(\tilde{\boldsymbol{\beta}}) = (\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \quad \text{s.t. } \mathbf{R}\tilde{\boldsymbol{\beta}} = \mathbf{q}$$

The associated Lagrangean function for this problem is as follows:<sup>1</sup>

$$\mathcal{L}(\tilde{\boldsymbol{\beta}}, \lambda) = (\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) + 2\lambda'(\mathbf{R}\tilde{\boldsymbol{\beta}} - \mathbf{q})$$

The term  $\lambda$  can be interpreted as the measure of how much the OLS fit will improve if the constraint is relaxed. If  $\lambda = 0$ , then the fit cannot be improved by relaxing (not imposing)

<sup>1</sup>The second term is scaled by 2 for convenience; because  $\lambda$  is not restricted, the scaling does not affect the general result.

the restriction, and so we cannot reject the null hypothesis  $H_0 : \mathbf{R}\tilde{\boldsymbol{\beta}} = \mathbf{q}$ . Therefore, the restriction must be imposed.

The intuition behind  $\lambda$  implies that we need to determine  $\lambda_*$  that satisfies the above problem and test if it is zero. To do this, we take the first-order conditions with respect to  $\tilde{\boldsymbol{\beta}}$  and  $\lambda$  are as follows:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \tilde{\boldsymbol{\beta}}} &= -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_*) + 2\mathbf{R}'\lambda_* = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= 2(\mathbf{R}\boldsymbol{\beta}_* - \mathbf{q}) = 0\end{aligned}$$

The terms  $\boldsymbol{\beta}_*$  and  $\lambda_*$  are those that will satisfy the first-order conditions. That is, the task is to determine the value  $\lambda_*$  at which the first-order conditions (FOCs) are zero. After dividing through by 2, we can re-write the FOCs in matrix form as follows:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{R}' \\ \mathbf{R} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_* \\ \lambda_* \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{q} \end{bmatrix}$$

Solving for the vector  $[\boldsymbol{\beta}_* \ \lambda_*]'$  requires that we invert the first matrix. If  $\mathbf{X}'\mathbf{X}$  is non-singular (full rank), then unique solutions exist for  $\boldsymbol{\beta}_*$  and  $\lambda_*$ . It can be shown that (see Greene (2003) for a full derivation) after inverting and solving,  $\lambda_*$  is as follows:

$$\lambda_* = \{\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\}^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})$$

Therefore, the term that we are seeking to test is depending on the estimated coefficients. The statistic for testing  $\lambda_* = 0$  can be compared with a critical value of an  $F$  distribution as follows:

It should be noted that the Lagrange multiplier test can be, as the Wald/F-test, used to test many different hypothesis. That is, you can define the  $\mathbf{R}$  and  $\mathbf{q}$  matrices to be as desired, and then test whether the restricted model produces a better fit.

### The Chow test

The Chow test is an application of the LM test procedure for testing structural breaks. The intuition rests in the idea that you compare the fit of an unrestricted model (two sets of parameters exist for two data subsamples) to a restricted model (parameters are not different across subsamples).

For example, suppose that we define the regression equation of bull weights as follows:

$$\mathbf{W} = \mathbf{X}\boldsymbol{\beta}_A + I_{\{L\}}\mathbf{X}\boldsymbol{\beta}_L + \varepsilon$$

The term  $I_{\{L\}}$  is an indicator variable that is interacted with the sample to specify which bulls are of the Limousin breed. The null hypothesis is that there are no differences between the Angus and Limousin bulls, then  $\boldsymbol{\beta}_L = 0$ , which would indicate support for the restricted model.

To test the hypothesis, we can derive an  $F$  statistic as follows:

$$F_{stat} = \frac{(SSR_R - SSR_U)/(k)}{SSR_U/(n - 2k)} \sim F_{J, n-2k}$$

The term  $SSR_R = \hat{\varepsilon}'\hat{\varepsilon}$  denotes the sum of squared residuals from the restricted model, which regresses bull weights on the entire data (without differentiating among Angus and Limousin breeds). The term  $SSR_U = SSR_A + SSR_L = \hat{\varepsilon}'_A\hat{\varepsilon}_A + \hat{\varepsilon}'_L\hat{\varepsilon}_L$  represents the sum of the SSR from regressing bull weights on only Angus bulls and SSR from regressing bull weights on only Limousin bulls.

If  $|F_{stat}| > F_{crit}$ , then we must reject the null hypothesis that  $\boldsymbol{\beta}_L = 0$ , indicating that there does exist a structural break and the two models should be estimated separately. Otherwise, we cannot reject a structural break, and the same set of parameter values applies to both Angus and Limousin bulls.

#### Example

Suppose that you model a 1-year-old bull's weight as a function of his genetic predisposition to gain weight and the bull's average daily weight gain:

$$Weight_{365} = \beta_0 + \beta_1 BW_{EPD} + \beta_2 ADG + \varepsilon$$

However, you want to test whether the parameter set  $\boldsymbol{\beta}$  is different for Angus and Limousin bulls. To test this, perform the following steps:

1. Perform the restricted regression in which you model the bull weights for all (Angus and Limousin) bulls. Suppose you retrieve the following statistical output:

$\beta_0$	$\beta_1$	$\beta_2$
703.59	7.04	161.18
(21.74)	(1.60)	(6.32)

$R^2$	$\hat{\varepsilon}'\hat{\varepsilon}$	$n$
0.481	3101851	770

2. Then you repeat the regression using only Angus bulls, and retrieve the following results:

$\beta_0$	$\beta_1$	$\beta_2$
752.69	4.61	149.79
(23.28)	(1.67)	(6.65)

$R^2$	$\hat{\varepsilon}'\hat{\varepsilon}$	$n$
0.424	2758422	715

3. Lastly, you repeat the regression using only Limousin bulls, and retrieve the following results:

$\beta_0$	$\beta_1$	$\beta_2$
704.74	4.85	143.58
(60.81)	(4.71)	(19.69)

$R^2$	$\hat{\varepsilon}'\hat{\varepsilon}$	$n$
0.504	140593	55

4. You can now compute the  $F$  statistic as follows:

$$F_{stat} = \frac{(SSR_R - SSR_U)/(k)}{SSR_U/(n - 2k)} = \frac{(3101851 - (2758422 + 140593))/(3)}{(2758422 + 140593)/(770 - 6)} = \frac{67612}{3794} = 17.82$$

5. Lastly, you'll need to determine the critical value from the  $F$  distribution with  $J = 3$  and 764 degrees of freedom:

$$F_{crit} = 2.10$$

6. Comparing the  $F$  statistic to the critical value, we find that  $|F_{stat}| > F_{crit}$ , which leads us to conclude that the null hypothesis must be rejected. That is, there is enough statistical evidence to indicate that two separate models must be estimated for the Angus and Limousin breeds.

# Chapter 7

## Issues with OLS Estimations

Using asymptotic properties, we have shown that we can relax some of the Gauss-Markov assumptions. However, we have, for the most part, assumed the Gauss-Markov theorem to be true. In this chapter, we turn to investigating scenarios in which we allow some assumptions to be relaxed. Specifically, we will explore several issues:

1. Multicollinearity
2. Outliers
3. Heteroskedasticity
4. Autocorrelation

These violations can cause estimators to be inefficient, and may lead you to inappropriately infer information about the regression. In each case, we will seek to understand what causes these problems to arise, what tools exist for detecting these problems, and what statistical methods can be used to appropriately correct for these violations.

## 7.1 Multicollinearity

Recall that when the multiple regressor model was introduced, the importance of correlation between independent variables was discussed. That is, the reason that multiple regressor models are estimated is to observe the marginal effects of a particular variable after accounting for any correlation between the variable and other regressors.

However, in some cases, the correlation between variables may be very high (and sometimes perfect). For example, suppose that you are modeling new housing starts as a function of gross national product (GNP) and population size. We would typically assume that these two variables are highly correlated, because a higher population implies higher total output, and vice versa. Let's consider two cases: one in which GNP and population are perfectly correlated (increasing the population by one person increases GNP by one unit), and a second case in which the variables are highly correlated (but not perfectly).

### 7.1.1 Perfect Collinearity

We have already discussed the consequences of perfect collinearity when we examined the issue of matrix rank. Perfect collinearity causes  $\mathbf{X}$  to not be full rank, implying that we cannot retrieve a unique solution of the OLS estimator. Recall that violation of the full rank implies that a column in the  $\mathbf{X}$  matrix is a linear combination of two or more other columns in that matrix.

When can perfect multicollinearity occur? There are several situations:

1. *Dummy variable trap*: when you specify indicator/dummy variables such that they add up to one, the intercept will be a linear combination of those two columns.

For example, if you specify both a Male and Female dummy in a regression model, then (Male + Female) = Intercept.

2. *Regressors that are functions of other regressors*: when you specify a regressor term that is a function of other regressors in the model.

For example, suppose that you are measuring wages as a function of age, years of schooling, and experience. However, you define experience to be (Age - Years of Schooling). Therefore, you specify the model as:

$$Wage = \beta_0 + \beta_1 Age + \beta_2 Schooling + \beta_3 (Age - Schooling) + \varepsilon$$

You can see that the  $\mathbf{X}$  matrix will not be full rank. Intuitively, you will not be able to distinguish the marginal effect of experience from marginal effects of age and schooling.

3. *No variability in regressor*: when a particular regressor does not vary across the entire sample. Intuitively, you would not be able to determine how this regressor affects the dependent variable.

For example, suppose that the Age variable in the above model is the same for all individuals in the sample. Then, you would not be able to determine how changes in age will affect wages.

Perfect multicollinearity may be either a specification and/or data problem. In case (1) and (2), you may be able to correct for perfect multicollinearity by re-specifying the regressors that are in your model (e.g., eliminate of the dummy variables or intercept term; remove a regressor). In the case of (3), this is a data problem and may need to be dealt with by gathering additional data or finding an alternative explanatory variable.

### 7.1.2 Imperfect Multicollinearity

This type of multicollinearity is a data problem, which results from an approximate linear relationship between explanatory variables. There are several undesired consequences that can arise from multicollinearity:

1. *Estimates are not robust*: small changes in the data, such as inclusion or exclusion of a few observations, can lead to wide swings of parameter estimates.
2. *Implausible estimate inferences*: estimated coefficients may have wrong (theoretically inappropriate) signs or magnitudes.
3. *Imprecise estimates*: estimated coefficients may have very high standard errors/low  $t$  statistics, even though they are jointly significant and there is a relatively high regression  $R^2$ .

Why do these consequences arise? Let's consider the regression:  $Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ . Suppose that  $X_1$  and  $X_2$  are highly correlated. Then the variance-covariance matrix can be written as follows:

$$\text{Var}(\boldsymbol{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{bmatrix} \sum X_{1i}X'_{1i} & \sum X_{1i}X'_{2i} \\ \sum X_{2i}X'_{1i} & \sum X_{2i}X'_{2i} \end{bmatrix}^{-1} =$$

$$\frac{\sigma^2}{\sum X_{1i}X'_{1i} \sum X_{2i}X'_{2i} - \sum X_{1i}X'_{2i} \sum X_{2i}X'_{1i}} \begin{bmatrix} \sum X_{2i}X'_{2i} & -\sum X_{2i}X'_{1i} \\ -\sum X_{1i}X'_{2i} & \sum X_{1i}X'_{1i} \end{bmatrix}$$

If there is strong correlation between  $X_1$  and  $X_2$ , then the two variables are either very large or very small together. (Consider, for example, the high correlation between elevation and inches of snow. The higher the elevation, the more snow, implying that when one number is large, the other is as well.) Therefore, the denominator gets smaller and the variance increases.

In general, multicollinearity will have little impact on the overall fit and prediction qualities of the model as a whole (that is, using all explanatory variables to predict the dependent variable). However, marginal effects might be quite inaccurate.

### 7.1.3 Identifying Multicollinearity

There are several ways that you can identify potential multicollinearity:

1. *Intuition and theory*: ask whether there is reason to believe that certain variables might be very highly correlated.
2. *Observe anomalies in regression inferences*: if your regression has a particularly high  $R^2$  (good prediction powers) but high standard errors (and low  $t$  statistics), you may have multicollinearity problems.
3. *Use statistical methods*:
  - Check the basic correlation between variables.
  - Examine how sensitive your regression model is to inclusion/exclusion of suspect variables. If excluding a variable causes  $t$  statistics to increase but not change the  $R^2$ , then the excluded variable may have contributed to multicollinearity.
  - Use the *variance inflation factor*:

- (a) Regress an independent variable that you believe to be collinear (linear combination of other regressors) on all other independent variables.
- (b) Determine the  $\tilde{R}^2$  and calculate:  $VIF = \frac{1}{1-\tilde{R}^2}$ .
- (c) If the VIF is high, then the variance of the parameter representing the regressed independent variable is inflated, relative to a hypothetical situation when no correlation between the regressed variable and all other independent variables exist.

### 7.1.4 Dealing with Multicollinearity

There are several possible ways to try alleviating the multicollinearity:

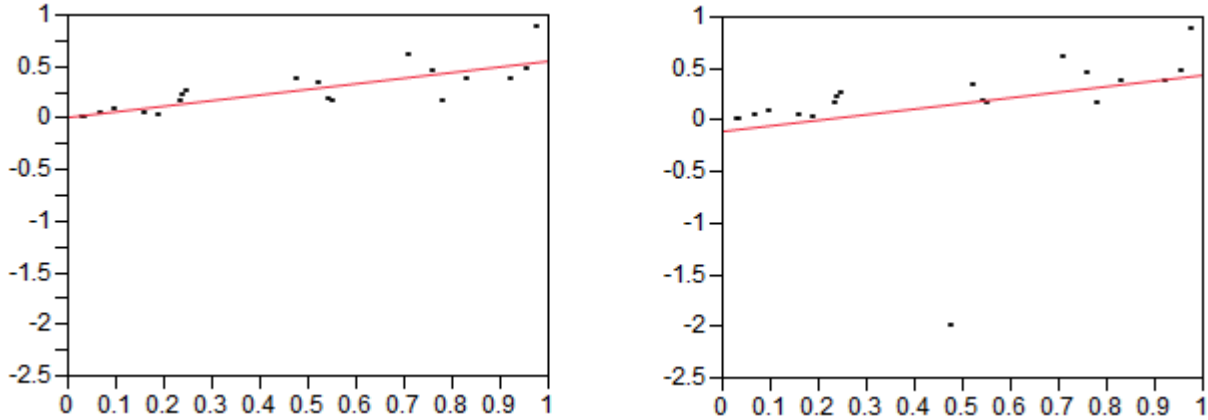
1. *Restrict or remove collinear variables*: this must be done with caution to avoid omitted variable bias. This approach may be most beneficial if you have overfit/overspecified the initial model.
2. *Transform collinear variables*: use another representation of variables to measure a similar effect. One example may be to first-difference the data.
3. *Increase sample size*: increasing  $n$  will reduce variance and standard errors, which will dampen the effects of multicollinearity.
4. *Amend the functional form*
5. *Do nothing*: if multicollinearity is small enough to not substantially affect the parameter estimates (magnitudes and theoretically expected signs) and standard errors, then the problem may not need to be explicitly addressed.

## 7.2 Outliers

Another potentially problematic data problem is the existence of outliers. Especially in regression models with a small sample size, outliers may substantially affect the regression line, causing it to be pulled (rotated) toward the influential observation point. Figure 7.1 illustrates the influence of an outlier on an OLS regression line.

The problem with outlier, with respect to the OLS estimation approach, is that minimizing the square of residuals causes large residuals (from outliers) to receive a large weight. Typically, if you have a large sample size, the potential for outliers causing inaccuracies diminishes.

Figure 7.1: Graph of Outliers



### 7.2.1 Testing for Outliers

One way to test for outliers (Belsley, Kuh, and Welsh 1980) is to standardize regression residuals and observe which standardized residuals lie outside the “normal” thresholds. To do this, we will use our knowledge of the “residual maker” matrix (see section 5.2.3). Specifically, complete the following:

1. Use OLS to estimate a model  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ .
2. Retrieve the residuals,  $\hat{\varepsilon}$ .
3. Calculate  $\hat{\sigma}^2$ .
4. Calculate the standardized residuals vector as follows:

$$\tilde{\varepsilon} = \frac{\hat{\varepsilon}}{\{\hat{\sigma}^2 \cdot \text{vecdiag}(\mathbf{M})\}^{1/2}},$$

where  $\text{vecdiag}(\mathbf{M})$  denotes the vector comprised of the elements on the main diagonal of the residual maker matrix,  $\mathbf{M}$ .

5. Check for any values in the standardized residual vector that are greater than 2 or less than -2. These will correspond to observations in the sample that substantially differ from expectations.

### 7.2.2 Dealing with Outliers

Because outliers may cause OLS parameter estimates to be inaccurate, it is at times beneficial to estimate models after eliminating the outliers. For cross-sectional data, this may be possible; however, when dealing with time series data, eliminating observations from the middle of a time series can be intuitively problematic.

It is necessary to note that outliers may not necessarily be anomalies. Each outlier may still have important information about the data, and so should be eliminated only after careful consideration.

There are also ways to deal with outliers using statistical approaches. One such approach is to use a *least absolute deviation* model (more generally, a LAD model is a special case of a *quantile regression model*). These models minimize the sum of deviations from the median (or quantile), which substantially diminishes the potential influence of outliers on estimation results. Why does this happen?

## 7.3 Heteroskedasticity

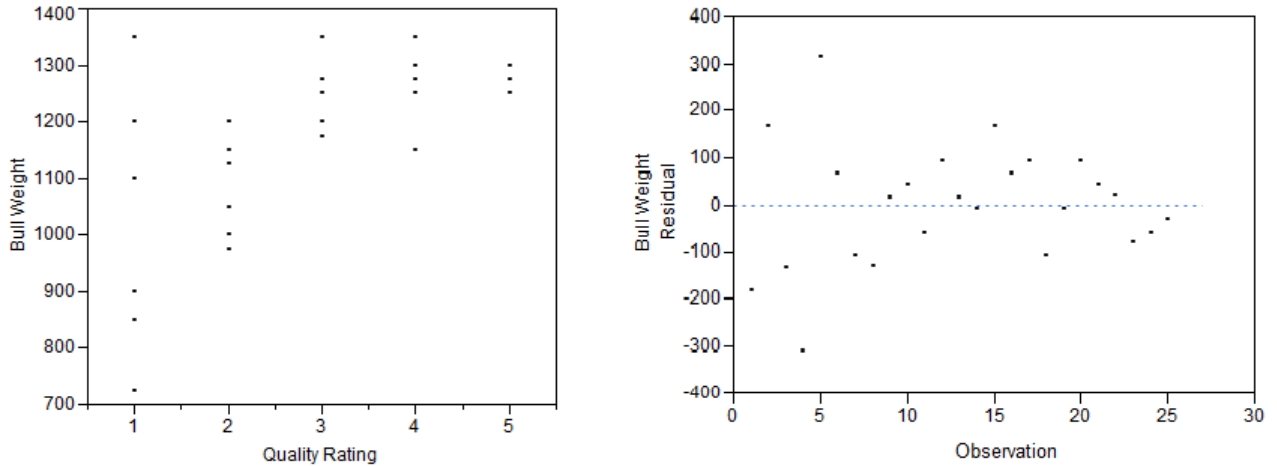
Recall Gauss-Markov assumption 5, described in section 5.4. Under the assumption, the variance of the error term at each observation is constant. That is, all  $\sigma_i^2$  are the same for all observations in the data:

$$\text{Var}(\varepsilon) = E(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2 I_n$$

However, suppose this isn't the case. It is possible to come up with numerous examples where this is true. High quality bulls may gain weight with more certainty (less variance) than bulls of low quality. Families with high incomes, the amount of money that they spend on food may be quite variable across families (some families dine at high-end restaurants, while others choose to cook at home). However, for low-income families, the variance in food expenditures might be much smaller (these families have less choice as to how they spend on food). This difference in variance across the sample is known as heteroskedasticity. Cross-sectional data are often more subject to heteroskedasticity issues.

Visually, heteroskedasticity can be shown in figure 7.2, which shows two alternative perspectives.

Figure 7.2: Illustration of Heteroskedastic Data



The first, left-most plot illustrates bull weights plotted against quality ratings. Bulls of the lowest quality (1) have the most variance among weights. However, bulls of the highest quality (5) have the least variance. The second, right-most plot shows the bull weight residuals at each observation after fitting a model:  $Weight = \beta_0 + \beta_1 Quality + \varepsilon$ . The plotted residuals show a reverse “fanning” effect, such that the residuals for lower quality bulls (first set of observations) are most spread out about the zero line. As the quality (observation number) increases, the residuals become much more concentrated. If you observe these types of phenomena in your data, it may signal existence of potential heteroskedasticity.

Mathematically, heteroskedasticity can be denoted as follows:

$$\text{Var}(\varepsilon_i | \mathbf{X}) = \sigma_i^2$$

$$\text{Var}(\varepsilon | \mathbf{X}) = E(\varepsilon \varepsilon' | \mathbf{X}) = \sigma^2 \Omega = \Sigma$$

Note that if  $\Omega = I_n$ , then we return to homoskedasticity. Otherwise, we assume that  $\Omega$  is a positive definite matrix.

In the case of heteroskedasticity, the disturbances are still assumed to be uncorrelated across observations (that is, no autocorrelation), which implies that we can write the variance matrix as follows:

$$\Sigma = \sigma^2 \Omega = \sigma^2 \cdot \begin{bmatrix} \omega_{11} & 0 & \dots & 0 \\ 0 & \omega_{22} & \dots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \dots & \omega_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

### 7.3.1 Consequences of Heteroskedasticity

Why do we care about heteroskedasticity? Because we would like to know whether non-constant error variance will affect the unbiasedness and/or efficiency of the OLS estimator. If these two properties are affected by heteroskedasticity, then the OLS estimator is no longer the most efficient, unbiased linear estimator, and we may need to consider alternative estimators.

#### Unbiasedness of OLS estimators

Let's first consider whether heteroskedasticity causes the OLS estimator to be biased. Recall that we can write the estimator as follows:

$$\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\varepsilon$$

Because we still assume that the regressors are exogenous to the disturbances,  $E(\varepsilon|\mathbf{X}) = 0$ , then we still find that  $E(\hat{\beta}|\mathbf{X}) = \beta$ . Therefore, heteroskedasticity does not bias the OLS estimator.

#### Efficiency of OLS estimators

The second goal is to determine whether heteroskedasticity leads to inefficient OLS estimators. Recall that in section 5.4.1, we proved that the OLS estimator is most efficient, because there was no other linear estimator whose variance was smaller. Now that the variance of the disturbances is not constant, let's reconsider the variance of the OLS estimator:

$$\begin{aligned}
 \text{Var}(\hat{\beta}|\mathbf{X}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'(\mathbf{X}'\mathbf{X})^{-1}] \\
 &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
 &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}
 \end{aligned}$$

Again, if  $\Omega = I_n$ , then the variance would collapse to the variance associated with the variance of homoskedastic data. However, it is clear that using a homoskedastic OLS estimator variance,  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ , for data that are heteroskedastic would be inappropriate, producing misleading statistical inferences (e.g., wrong standard errors,  $t$ - and  $F$ -test results).

The direction of the error in the variance depends on the pattern of the heteroskedasticity. As an example, consider a simple regression model:  $\mathbf{Y} = \beta_0 + \beta_1\mathbf{X} + \varepsilon$ . If the variance of disturbances increases with the level of  $\mathbf{X}$ , then the estimate  $\hat{\sigma}^2$  would be too small at higher levels of  $\mathbf{X}$  and too large at lower levels of  $\mathbf{X}$ . This is because the  $\hat{\sigma}^2$  associated with an OLS estimator that assumes homoskedasticity averages out variances across all levels of  $\mathbf{X}$ . Therefore, for example, because the estimated variance is too small at higher levels of  $\mathbf{X}$ , the standard errors associated with  $\hat{\beta}_1$  would be too small, biasing away from the null hypothesis. This will result in a greater likelihood of rejecting the null hypothesis  $H_0 : \beta = 0$ , when it shouldn't be rejected. The converse can be characterized when the variance of disturbances decreases with the level of  $\mathbf{X}$ .

What's the point? The OLS estimator is not most efficient when heteroskedastic data are analyzed.

### 7.3.2 Detecting Heteroskedasticity

When testing for heteroskedasticity, the null hypothesis is that the errors are homoskedastic. That is:

$$\begin{aligned}
 H_0 : & \text{Homoskedasticity} \\
 & \text{Var}(\varepsilon|\mathbf{X}) = \sigma^2 \\
 & E(\varepsilon^2|\mathbf{X}) = E(\varepsilon^2) = \sigma^2 \\
 H_a : & \text{Heteroskedasticity}
 \end{aligned}$$

Therefore, we would like to examine whether the expected squared error is related to one or more of the regressors. If heteroskedasticity is present, then  $\varepsilon^2$  is some function of  $\mathbf{X}$ . A simple approach is to assume that there is a linear relationship between  $\varepsilon^2$  and  $\mathbf{X}$ :

$$\varepsilon^2 = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \dots + \gamma_k X_k + \nu$$

Existence of homoskedasticity would imply that  $\gamma_1 = \gamma_2 = \dots = \gamma_k = 0$ . Therefore, we can use either an  $F$  or  $LM$  tests to determine if there is enough statistical evidence to reject the null; if so, then we must be concerned about heteroskedasticity.

Most tests for heteroskedasticity follow a strategy:

1. OLS estimators are unbiased, and residuals from an OLS model will resemble (although imperfectly) the potential heteroskedasticity of the true disturbances (e.g., see the figure 7.2 in section 7.3).
2. Test the OLS residuals for failing to provide statistical support for homoskedasticity.

### White test

The Lagrange multiplier approach of the above strategy is known as the *White test* for heteroskedasticity (White 1980). To implement the White test, perform the following steps:

1. Estimate an OLS model:  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ .
2. Calculate the vector of regression residuals,  $\hat{\boldsymbol{\varepsilon}}$ . Then, square each element in the regression residuals vector to retrieve  $\hat{\boldsymbol{\varepsilon}}^2$ .
3. Estimate an OLS model by regression  $\hat{\boldsymbol{\varepsilon}}^2$  on an intercept, all unique  $\mathbf{X}$ ,  $\mathbf{X}^2$ , and all cross-products of  $\mathbf{X}$ . Obtain the  $R$ -squared from this regression,  $\tilde{R}^2$ .

4. Compute the LM statistic:

$$LM_{stat} = n \cdot \tilde{R}^2 \sim \chi_k^2$$

$k$  is the number of regressors in the OLS model of step (3). Typically, this is all of the regressors.

5. Compare the test statistic to the appropriate critical value. If  $|LM_{stat}| > \chi_{crit}^2$ , then you reject the null hypothesis of homoskedasticity.

The intuition is that a higher  $\tilde{R}^2$  implies that the regressors help explain more variation in the squared residuals. This suggests that there is a stronger relationship between  $\mathbf{X}$  and  $\hat{\varepsilon}^2$ .

It should also be noted that you can certainly add more complex relationships between  $\mathbf{X}$  and  $\hat{\varepsilon}^2$ . That is, because we are simply interested in whether  $\hat{\varepsilon}^2$  is a function of  $\mathbf{X}$ , the OLS regression in step (3) of the White test procedure can have squared terms, logged terms, etc.

### Breusch-Pagan test

Although the White test is very general (because it doesn't assume any specific form of heteroskedasticity), this can also be a downfall of this test. That is, the White test may not actually reveal absence of homoskedasticity, but rather, some other model misspecification issue. Which result the test reveals is impossible to separate.

A more powerful test was developed by Breusch and Pagan (1979). The general idea is the same as the White test, but it is a more precise measure of heteroskedasticity. To carry out the BP test, perform the following steps:

1. Estimate an OLS model:  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ .
2. Calculate the vector of regression residuals,  $\hat{\boldsymbol{\varepsilon}}$ . Then, square each element in the regression residuals vector to retrieve  $\hat{\varepsilon}^2$ .
3. Koenker (1981) and Koenker and Bassett (1982) suggest the computation of the appropriate LM statistic as follows:

$$LM_{stat} = \left( \frac{1}{V} \right) (\hat{\varepsilon}^2 - \bar{u}\boldsymbol{\iota})' \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' (\hat{\varepsilon}^2 - \bar{u}\boldsymbol{\iota}) \sim \chi_k^2$$

where:

$$V = \frac{1}{n} \sum_{i=1}^n \left[ \hat{\varepsilon}_i^2 - \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n} \right]^2$$

$$\bar{u} = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}$$

$$\boldsymbol{\iota} = [1 \ 1 \ 1 \ \dots \ 1]'_{(n \times 1)}$$

This particular test statistic is slightly better than the original statistic proposed by Breusch and Pagan, because the BP statistic assumes that the errors are normally distributed. This modified LM statistic is as good when normality is assumed, and more powerful when normality is not assumed.

### Goldfeld-Quandt test

This particular test assumes that observations in a sample can be divided into two groups: one in which disturbances are homoskedastic and another in which disturbances are potentially heteroskedastic. To perform the test, follow these steps:

1. Order the data into two groups: one with similar disturbances and one without this property (e.g. suppose that prior to a particular technological breakthrough, corn yields were very similar on all farms; however, after the breakthrough, yields were substantially different based on who adopted the technology).
2. Perform OLS regressions on each subsample and compute the  $\hat{\sigma}^2$  for each subsample:  $\hat{\sigma}_a^2$  and  $\hat{\sigma}_b^2$ .
3. Compute the  $F_{stat}$  as follows:

$$F_{stat} = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_b^2} \sim F_{(n_1-k, n_2-k)}$$

4. If the  $F$  statistic is large enough to reject the null hypothesis, then we cannot assume homoskedasticity. That is, the variance of disturbances in the control group (homoskedastic errors) is different than the variances of disturbances in the treatment group.

The GQ test is helpful when you know which variables to use for separating data into subsamples. However, often it can be difficult to properly create the subsamples.

### 7.3.3 Dealing with Heteroskedasticity – Generalized Least Squares

Now that we know how to find and test for heteroskedasticity, let us turn to methods for appropriately performing estimation when heteroskedasticity exists. These estimations are necessary to retrieve correct standard errors, because we have already shown that when homoskedasticity cannot be assumed, then OLS is not the most efficient estimator among the class of unbiased linear estimators.

Recall that variance of the disturbances can be written generally as follows:

$$\text{Var}(\varepsilon|\mathbf{X}) = E(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2\Omega$$

where  $\Omega = h(\mathbf{X})$  is a function of the explanatory variables. At this point, we need to determine  $\text{Var}(\varepsilon|\mathbf{X})$ , which depends on whether we know the function  $h(\mathbf{X})$ . Once knowledge of this function is established, we can then produce an estimator that is more efficient than the OLS estimator. This new estimator is known as the *generalized least squares* (GLS) estimator, and can be expressed as follows:

$$\hat{\beta}^* = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{Y}$$

Generally, the GLS estimator is obtained by regressing:

$$\begin{bmatrix} y_1/\sqrt{\omega_1} \\ y_2/\sqrt{\omega_2} \\ \vdots \\ y_n/\sqrt{\omega_n} \end{bmatrix} \quad \text{on} \quad \begin{bmatrix} \mathbf{X}_1/\sqrt{\omega_1} \\ \mathbf{X}_2/\sqrt{\omega_2} \\ \vdots \\ \mathbf{X}_n/\sqrt{\omega_n} \end{bmatrix}$$

To estimate the transformed model using ordinary least squares, we specify appropriate weights for elements specified by  $\omega$ . That is, the *weighted least squares* (WLS) estimator is as follows:

$$\hat{\beta}^* = \left[ \sum_{i=1}^n \sqrt{h(\mathbf{X})} \mathbf{X}_i \mathbf{X}_i' \right]^{-1} \left[ \sum_{i=1}^n \sqrt{h(\mathbf{X})} \mathbf{X}_i y_i \right]$$

The intuition behind the weighted least squares estimates is that  $\hat{\beta}^*$  minimizes the weighted sum of squared errors, where each squared residual is weighted by  $1/h_i(\mathbf{X})$ . Therefore, less weight is applied to observations with higher disturbance variance. We will see why applying these weights lead to an efficient estimator in the next subsection.

It is necessary to note that the squared errors are weighted by  $1/h_i(\mathbf{X})$ , while the estimated parameters are weighted by  $1/\sqrt{h_i(\mathbf{X})}$ .

### WLS: Estimation with known weights

In some cases, you may be able to determine the explanatory variable(s) that can be used to specify weights in the WLS estimator. That is, you may be able to define the function  $h(\mathbf{X})$ .

For example, consider the model discussed earlier in the section that measures food expenditures as a function of income, number of children, and daily household caloric intake:

$$FoodExp = \beta_0 + \beta_1 Income + \beta_2 Children + \beta_3 Calories + \varepsilon$$

The potential for heteroskedasticity exists in the fact that low-income households may exhibit much smaller variability in food expenditure than high-income households, because lower income implies less choice about how much is spent on food. Therefore, the disturbance variance is *proportional* to the level of income – as income increases, we may observe a higher variance in food expenditure.

We can express the influence of income on variance as follows:

$$\text{Var}(\varepsilon_i | Income_i) = \sigma^2 \cdot Income_i$$

It is quite evident that the function  $h(\mathbf{X})$  can be described as:  $h(\mathbf{X}) = Income$ . How do we use this information to derive an efficient estimator? Let's again return to the definition of the conditional variance of disturbances,  $\text{Var}(\varepsilon | \mathbf{X}) = E(\varepsilon' \varepsilon | \mathbf{X}) = \sigma^2 h(\mathbf{X})$ . If  $h(\mathbf{X})$  is not an identity matrix, then we have heteroskedasticity.

If heteroskedasticity is present, then we need to devise a method by which we can return the variance of the disturbances to be  $\sigma^2$ . We can do so by dividing  $\varepsilon$  by  $\sqrt{h(\mathbf{X})}$  as follows:

$$\text{Var}(\varepsilon / (\sqrt{h(\mathbf{X})})^2 | \mathbf{X}) = E(\varepsilon' \varepsilon | \mathbf{X}) / h(\mathbf{X}) = \sigma^2 h(\mathbf{X}) / h(\mathbf{X}) = \sigma^2$$

Therefore, dividing by  $\sqrt{h(\mathbf{X})}$  leads the desired result, and implies that we have appropriately transformed the disturbances to be homoskedastic. Thus, the resulting OLS results using the transformed errors lead to the most efficient estimator.

With respect to the original linear equation, the transformation can be performed by dividing through by  $\sqrt{Income}$  as follows:

$$\frac{FoodExp_i}{\sqrt{Income_i}} = \frac{\beta_0}{\sqrt{Income_i}} + \beta_1 \frac{Income_i}{\sqrt{Income_i}} + \beta_2 \frac{Children_i}{\sqrt{Income_i}} + \beta_3 \frac{Calories_i}{\sqrt{Income_i}} + \frac{\varepsilon_i}{\sqrt{Income_i}}$$

$$FoodExp_i^* = \frac{\beta_0^*}{\sqrt{Income_i}} + \beta_1^* Income_i^* + \beta_2^* Children_i^* + \beta_3^* Calories_i^* + \varepsilon_i^*$$

The estimated parameters are different from the original OLS estimates, and represent marginal effects that account for heteroskedasticity in the disturbances. Furthermore,  $E(\varepsilon^*|\mathbf{X}^*) = 0$  and  $\text{Var}(\varepsilon^*|\mathbf{X}^*) = \sigma^2$ .

The interpretation of each estimated coefficient  $\beta_k^*$  is exactly the same as that in the original OLS regression. The adjustments only change the value of the estimate and the standard error, but do not change the interpretation.

### FGLS: Estimation with unknown weights

Often, the exact form of heteroskedasticity is unknown, and determining the function  $h(\mathbf{X})$  can be difficult. However, we can attempt to model a general form of  $h(\mathbf{X})$  and then use data to estimate the weight used to transform a linear regression model.

Assume that the variance of disturbances can be modeled as follows:

$$\text{Var}(\varepsilon|\mathbf{X}) = \sigma^2 \cdot \exp(\gamma_0 + \gamma_1 X_1 + \dots + \gamma_k X_k)$$

The exponential form is used to insure that the variance is positive. If the parameter vector  $\boldsymbol{\gamma}$  was known, then we would just use the WLS estimation approach. However, this is often not the case. Therefore, we will have to estimate these parameters and estimate a transformed, heteroskedasticity adjusted model using the following steps:

1. Using OLS, regress  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ . Obtain the vector of regression residuals,  $\hat{\boldsymbol{\varepsilon}}$ .
2. Square the elements in the vector  $\hat{\boldsymbol{\varepsilon}}$  and then take the natural log to create a vector:  $\log(\hat{\boldsymbol{\varepsilon}}^2)$ .
3. Regress  $\log(\hat{\boldsymbol{\varepsilon}}^2) = \mathbf{X}\boldsymbol{\gamma} + \mathbf{e}$ . Then, obtain the vector of fitted values,  $\hat{\mathbf{g}} = \mathbf{X}\hat{\boldsymbol{\gamma}}$ .
4. Exponentiate  $\hat{\mathbf{g}}$  to obtain  $\hat{\mathbf{h}} = \exp(\hat{\mathbf{g}})$ .
5. Use  $1/\sqrt{\hat{\mathbf{h}}}$  as weights to transform the original OLS regression and estimate it using WLS (as described in the previous section).

These steps describe estimation of feasible generalized least squares (FGLS). Asymptotically, FGLS is more efficient than OLS.

### White's heteroskedasticity-consistent standard errors

When you have no knowledge about the form of the variance term, it may be useful to estimate a heteroskedasticity-robust variance-covariance matrix for the OLS estimators. That is, adjust only the variance-covariance matrix, rather both the matrix and the estimated coefficients.

Recall again the structure of a heteroskedastic conditional variance matrix of the estimator:

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Therefore, if we knew the values of the  $\boldsymbol{\Sigma}$  matrix, then we would be able to retrieve the correct variance-covariance matrix for the estimator. White (1980) suggests that a consistent estimate of the  $\boldsymbol{\Sigma}$  matrix can be characterized as follows:

$$\hat{\boldsymbol{\Sigma}}_{(k \times k)} = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2 \mathbf{x}_i \mathbf{x}_i' = \frac{1}{n} \mathbf{X}' \cdot \text{vecdiag}(\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}') \cdot \mathbf{X}$$

After calculating the heteroskedasticity-robust variance matrix, you can infer statistical properties about the OLS estimators by using the standard errors (and associated  $t$ -stats) from the robust variance matrix.

Asymptotically, the estimator of the heteroskedasticity-robust variance is consistent:  $\text{plim} [\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}] = \mathbf{0}$ . However, in small samples, White's form of the estimator is too optimistic. That is, the variance matrix causes the  $t$ -stats to be too large.

Davidson and MacKinnon (1993) suggest two ways to adjust White's variance matrix calculation.

$$\begin{aligned}(1) \hat{\boldsymbol{\Sigma}}_{(k \times k)} &= \frac{n}{n-k} \sum_{i=1}^n \hat{\varepsilon}_i^2 \mathbf{x}_i \mathbf{x}_i' \\ (2) \hat{\boldsymbol{\Sigma}}_{(k \times k)} &= \frac{1}{n} \sum_{i=1}^n \left( \frac{\hat{\varepsilon}_i^2}{1 - \mathbf{x}_i' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i} \right) \cdot \mathbf{x}_i \mathbf{x}_i'\end{aligned}$$

In matrix form, these can be written as follows:

$$(1) \hat{\Sigma}_{(k \times k)} = \frac{n}{n-k} [\mathbf{X}' \cdot \text{vecdiag}(\hat{\varepsilon}\hat{\varepsilon}') \cdot \mathbf{X}]$$

$$(2) \hat{\Sigma}_{(k \times k)} = \frac{1}{n} \left[ \mathbf{X}' \text{vecdiag} \left( \frac{\hat{\varepsilon}_i^2}{1 - \mathbf{x}_i' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i} \right) \mathbf{X} \right]$$

It is important to note that if you know the form of the heteroskedasticity (that is, you know  $h(\mathbf{X})$ ), then FGLS may provide more efficient estimators.

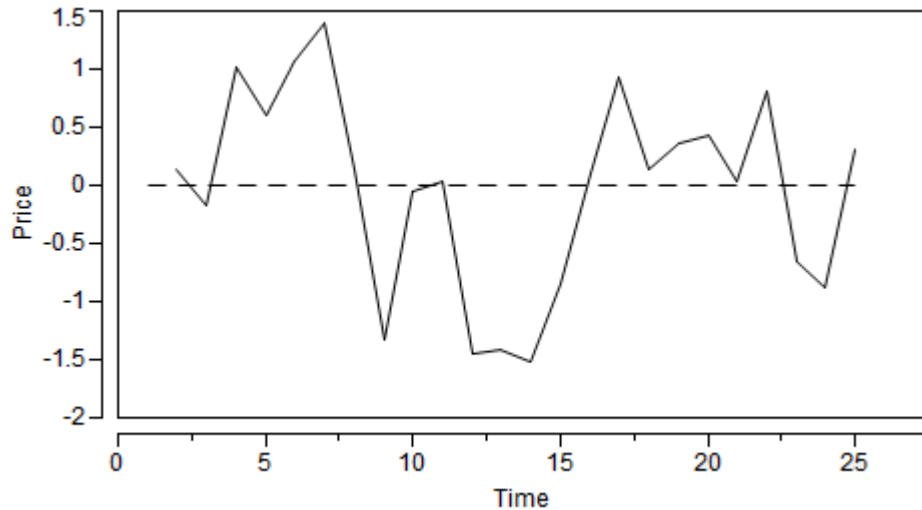
## 7.4 Autocorrelation / Serial Correlation

Now that we have discussed heteroskedasticity, it is relatively straightforward to examine another violation of the Gauss-Markov constant variance assumption. While heteroskedasticity is typically a problem in cross-sectional data (of course, exceptions exist; a primary example is autoregressive conditional heteroskedasticity), autocorrelation is most common in time-series data.

To provide some initial intuition into autocorrelation, let's consider when autocorrelation might occur in time-series and panel data. In a time-series data structure, autocorrelation is relatively easy to comprehend. The error terms (which contain all unobservable characteristics explaining variation in the dependent variable) may be related to each other over time. That is, the unobservable effects in period  $t$  may be affected by the unobservable effects in period  $t - 1$ , period  $t - 2$ , etc. For example, when studying daily stock prices, you may observe periods when prices move up and periods when prices move down. However, even after controlling for various factors, you still observe these movements in the residuals. This is due to unobservable factors from a previous period affecting the current period. An illustration of this is shown in figure 7.3.

Typically, autocorrelation is not present in pure cross-sectional data, because these data are constructed from a random sample. However, in a panel data structure, you may have cross-sectional observations that occur for the same entities over time. That is, because a "time-series" aspect is introduced into a cross-sectional data structure, the potential for autocorrelation now exists.

Figure 7.3: Illustration of Autocorrelated Data



### 7.4.1 Brief Introduction to Time-Series Data

Before we get into the depths of autocorrelation issues, it is useful to have a brief overview of time-series data. In a time-series model we seek to explain factors that influence the path of a variable  $Y_t$ . Typically, we would seek to explain this path using lagged (past) values of the variable, contemporaneous and/or lagged values of exogenous factors, and contemporaneous and lagged value of the disturbances. A general representation of a time-series model is as follows:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_t + \beta_3 X_{t-1} + \varepsilon_t + \varepsilon_{t-1}$$

The dependent variable,  $Y_t$ , represents a random event that occurs in time. For example, the end-of-day futures price for a wheat futures contract at the Kansas City Board of Trade, or the number of individuals who voted in Gallatin County in the 2000 presidential elections. The entire sequence of observations is the time-series process  $\{Y_t\}_{t=-\infty}^{t=\infty}$ , and is characterized by:

- Time ordering, such that if  $a < b$ , then the occurrence of the event in  $t = a$  is prior to the event in  $t = b$ .
- Systematic correlation between observations in the sequence.

An important sampling characteristic of time-series data is that data are generated not by random sampling from a population, but from sets of observations taken from a *time window*,  $t = 1, \dots, T$ . We can think about the asymptotics of time-series data being constructed based on an increasing time window rather than an increasing sample size.

### 7.4.2 Violation of a Gauss-Markov Assumption

The Gauss-Markov assumes the following about the disturbance term:

1. The expected value of the disturbance term is:  $E[\varepsilon_t] = 0$ .
2. The variance of the disturbance is:  $\text{Var}(\varepsilon_t) = \sigma^2\mathbf{\Omega}$ , where  $\mathbf{\Omega} = \mathbf{I}_n$ .
3. The covariance between disturbances is:  $\text{Cov}(\varepsilon_s, \varepsilon_t) = 0$ , for  $s \neq t$ .

Making these assumptions implies that the mean and variance of the disturbance are not changing over time. You will often find that disturbances that are characterized by the assumptions above are known as *white noise*.

However, in most time-series data, we will assume that  $\text{Cov}(\varepsilon_s, \varepsilon_t) \neq 0$ , for  $s \neq t$ . Because time-series are typically homoskedastic, we can write the conditional variance of the disturbances as follows:

$$\text{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) = \sigma_\varepsilon^2\mathbf{\Omega} = \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & & \ddots & \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix}$$

Therefore, if  $|\rho| < 1$ , then the strength of the relationship between disturbance terms in time dissipates across time. That is, the relationship between  $\varepsilon_t$  and  $\varepsilon_{t-1}$  is strong than then relationship between  $\varepsilon_t$  and  $\varepsilon_{t-10}$ .

All of this implies that, as with heteroskedasticity, an OLS estimator will not be the most efficient linear unbiased estimator. Furthermore, we can no longer use the typical  $t$  and  $F$  statistics for inferences.

### A slight diversion about stationarity

We will also assume that  $\Omega_{t,s}$  is a function of  $|t - s|$ , and not of  $t$  or  $s$  alone. This is an assumption of *weak stationarity*, and it implies that the relationship between disturbances is only a function of their distance apart, not a function of the time origins.

The concept of stationarity is very important. There are two types of stationarity: *strong stationarity* and *weak stationarity*. Strong stationarity states that for a time series process  $\{z_t\}_{t=-\infty}^{t=\infty}$ , the distribution of any  $z_t$  is the same as the distribution of  $z_s$ , and that the joint distribution probability distribution of any set of  $h$  observations in the sequence is the same regardless of the origin  $t$ . For example, the joint distribution between  $(z_1, z_2)$  is the same as the joint distribution  $(z_t, z_{t+1})$  for any  $t \geq 1$ . That is, the correlation structure among terms is the same across all periods, which allows us to predict future values based on past values.

Strong stationarity, however, is often not required in most empirical work, and we can focus on having a time series that is weakly stationary. Weak stationarity (also known as covariance stationary) states that the expected value and variance of  $z_t$  are finite and constant for all  $t$ , and the covariance of any two observations,  $\text{Cov}(z_t, z_{t-h})$  depends only on  $h$ , not  $t$ . The fact that  $\text{Corr}(z_t, z_{t-h})$  depends only on  $h$  follows directly. Again, stationarity is an assumption that allows us to model a stable relationship of a random variable across time.

In the context of the  $\Omega_{t,s}$  matrix, stationarity involves a stable relationship among disturbances as well as the fact that  $|\rho| < 1$ .

### 7.4.3 The AR(1) Process

Here we will use a concrete example. The most simple case of autocorrelation is a one-period autoregressive process of the disturbance term, AR(1). We can write this type of model as follows:

$$\mathbf{Y}_t = \mathbf{X}_t\boldsymbol{\beta} + \varepsilon_t$$

where

$$\varepsilon_t = \rho\varepsilon_{t-1} + e_t$$

The term  $e$  represents a white noise term. You can generalize the AR(1) process to an AR(k) process by adding  $k$  number of  $\varepsilon$  lags. However, empirical literature has shown overwhelming evidence of AR(1) processes for many time series.

As a quick side-note, let's consider repeated substitution:

$$\begin{aligned}
 \varepsilon_t &= \rho\varepsilon_{t-1} + e_t \\
 &= e_t + \rho(\rho\varepsilon_{t-2} + e_{t-1}) \\
 &= e_t + \rho e_{t-1} + \rho^2\varepsilon_{t-2} \\
 &\vdots \\
 &= e_t + \rho e_{t-1} + \rho^2 e_{t-2} + \rho^3 e_{t-3} + \dots \\
 &= \rho^s \varepsilon_{t-s} + \sum_{i=0}^{s-1} \rho^i u_{t-i}
 \end{aligned}$$

This implies that we can specify  $\varepsilon_t$  in a *moving-average* form, and directly observe that the disturbance is a function of all the information embodied by the history of the  $e$ 's. The most recent observations receive the most weight.

The moving-average form reiterates the fact that  $|\rho| < 1$ . Intuitively, if  $|\rho| \geq 1$ , then  $\varepsilon_t$  would explode (never converge), because you would assign a monotonically higher weight to later realizations of  $e_t$ .

The moments of  $\varepsilon_t$  for an AR(1) process can be derived as follows:

- Expected value:  $E[\varepsilon_t] = 0$ .
- Variance:

$$\begin{aligned}
 \text{Var}(\varepsilon_t) &= \text{Var}(e_t + \rho e_{t-1} + \rho^2 e_{t-2} + \rho^3 e_{t-3} + \dots) \\
 &= \rho^2 \text{Var}(\varepsilon_{t-1}) + \sigma_e^2 \\
 &= \frac{\sigma_e^2}{1 - \rho^2}
 \end{aligned}$$

The relationship  $\text{Var}(\varepsilon_{t-1}) = \text{Var}(\varepsilon_t)$  holds because of the stationarity assumption.

- Covariance:

$$\begin{aligned}
 \text{Cov}(\varepsilon_t, \varepsilon_{t-1}) &= E[\varepsilon_t \varepsilon_{t-1}] \\
 &= E[\varepsilon_{t-1}(\rho \varepsilon_{t-1} + e_t)] \\
 &= \rho \text{Var}(\varepsilon_{t-1}) \\
 &= \frac{\rho \sigma_e^2}{1 - \rho^2}
 \end{aligned}$$

Generally, covariance can be written as:  $\text{Cov}(\varepsilon_t, \varepsilon_{t-s}) = \frac{\rho^s \sigma_e^2}{1 - \rho^2}$ .

- Correlation:

$$\begin{aligned}
 \text{Corr}(\varepsilon_t, \varepsilon_{t-1}) &= \frac{\text{Cov}(\varepsilon_t, \varepsilon_{t-1})}{\sqrt{\text{Var}(\varepsilon_t)} \sqrt{\text{Var}(\varepsilon_{t-1})}} \\
 &= \frac{\text{Cov}(\varepsilon_t, \varepsilon_{t-1})}{\text{Var}(\varepsilon_t)} \\
 &= \frac{\rho \sigma_e^2 / 1 - \rho^2}{\sigma_e^2 / 1 - \rho^2} \\
 &= \rho
 \end{aligned}$$

Generally, the correlation can be written as:  $\text{Corr}(\varepsilon_t, \varepsilon_{t-s}) = \rho^s$ .

What does this all imply? It indicates that we are able to rewrite the variance structure of the disturbance term as follows:

$$\text{Var}(\boldsymbol{\varepsilon} | \mathbf{X}) = \sigma_\varepsilon^2 \boldsymbol{\Omega} = \frac{\sigma_e^2}{1 - \rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \dots & \rho^{T-3} \\ \vdots & & \vdots & & \ddots & \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \dots & \rho & 1 \end{bmatrix}$$

### 7.4.4 Consequences of Autocorrelation

The consequences are similar to those of heteroskedasticity, described in section 7.3.1. That is, the OLS estimator is still unbiased, but it no longer the most efficient among the class of unbiased, linear estimators. Specifically, the estimated variance of residuals,  $\hat{\sigma}^2$  is likely to be incorrect, because we do not account for the autocorrelation among residuals. This will cause us to calculate the incorrect estimator variance, standard errors,  $t$  and  $F$  statistics, and  $R^2$  statistic.

### 7.4.5 Detecting Autocorrelation

There are various heuristic and statistical methods that can be used to test for autocorrelation. We will discuss the heuristic graphical approach as well as the Durbin-Watson, Breusch-Godfrey, and Box-Pierce-Ljung statistical tests.

#### Graphical detection

There are several ways that you can use graphical analysis to get a general understanding of whether autocorrelation exists in the data. Consider three methods:

- Plot the OLS regression residuals,  $\hat{\varepsilon}_t$ , against time,  $t$ . If you visually detect a pattern, then your data may be subject to autocorrelation, which is shown in figure 7.4.
- Plot standardized OLS regression residuals,  $\tilde{\varepsilon}_t = \hat{\varepsilon}_t / \sqrt{\hat{\sigma}^2}$ , against time. Observe any patterns.
- To observe whether an AR(1) process exists, plot the OLS regression residuals,  $\hat{\varepsilon}_t$ , against lagged regression residuals,  $\hat{\varepsilon}_{t-1}$ . This is shown in figure 7.5.

Figure 7.4: Positive Autocorrelation

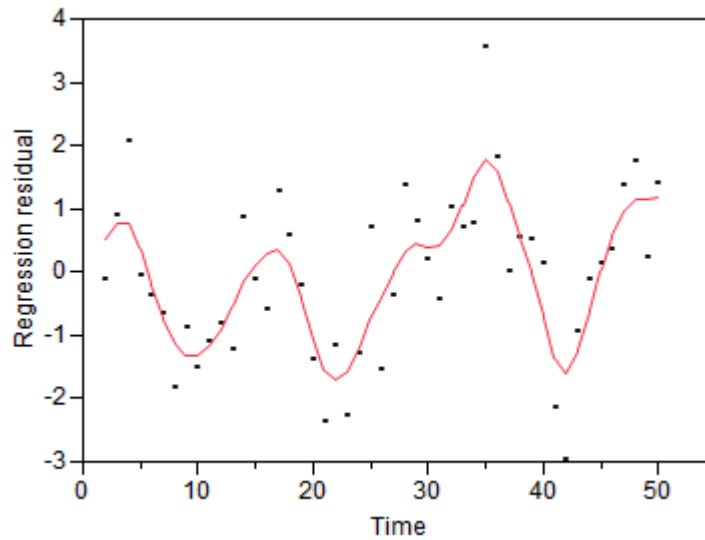
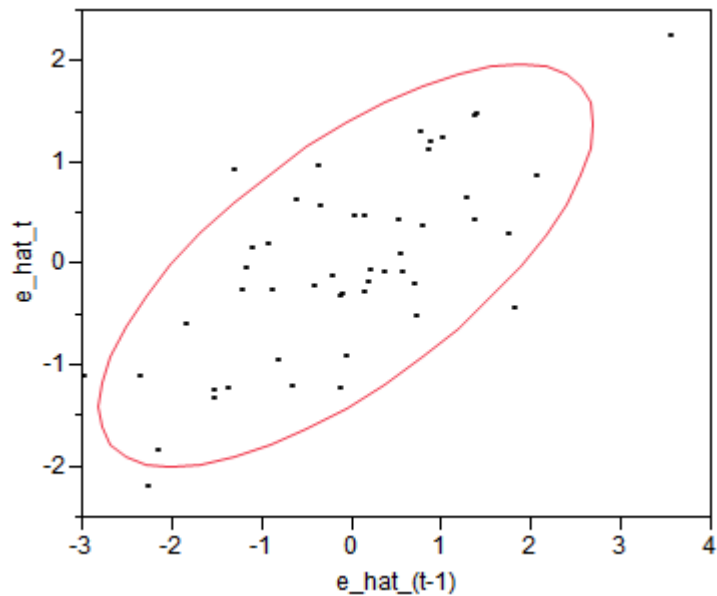


Figure 7.5: Plot of Residuals Against Lagged Residuals



### Durbin-Watson test

The test developed by Durbin and Watson (1950) was the first formal procedure for testing autocorrelation. The test uses OLS residuals, and the test statistic is as follows:

$$\begin{aligned}
 d &= \frac{\sum_{t=2}^T (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^T \hat{\varepsilon}_t^2} && \text{(sample correlation of regression residuals)} \\
 &= \frac{\sum_{t=2}^T \hat{\varepsilon}_t^2 + \sum_{t=2}^T \hat{\varepsilon}_{t-1}^2 - 2 \sum_{t=2}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\sum_{t=1}^T \hat{\varepsilon}_t^2} \\
 &= 2 - 2 \frac{\sum_{t=2}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\sum_{t=1}^T \hat{\varepsilon}_t^2} \\
 &\approx 2(1 - \hat{\rho})
 \end{aligned}$$

The last equality follows when the sample is reasonably large. The Durbin-Watson test statistic cannot be compared to a standard distribution. Alternatively, the authors have derived a table of values that can be used for comparison. One advantage of using these tables is that it can give you an idea of whether you may have negative, positive, or no autocorrelation.

That is, for a particular sample size and number of explanatory variables (excluding the intercept), the Durbin-Watson table provides a lower and upper significance points,  $d_L$  and  $d_U$ . If the calculated  $d$  test statistic falls below  $d_L$ , then there is evidence of positive first-order correlation. Conversely, if  $d > d_U$ , then there is evidence of negative autocorrelation. If neither case holds, then you cannot reject the null of no autocorrelation.

Most statistical packages will provide  $p$ -values for positive and negative autocorrelation tests. Finding small  $p$ -values for the test that  $d < d_L$  suggests a rejection of no autocorrelation in favor of positive autocorrelation. Small  $p$ -values of the test  $d > d_U$  suggests negative correlation. Large  $p$ -values for both tests is evidence against positive and negative autocorrelation.

### Breusch-Godfrey test

Two important downsides to the Durbin-Watson test for autocorrelation are as follows:

- The test is intended for AR(1) processes only.
- The test is not intended for models with lagged dependent variables. The results of the DW test are likely to be invalid when a lagged dependent variable is specified.

The BG test is able to overcome these shortfalls, and is able to test for higher-order AR or higher-order moving-average processes. Furthermore, lagged variables will not invalidate the test.

For an AR(p) process, the error structure is:  $\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \dots + \rho_p\varepsilon_{t-p} + e_t$ . The BG tests the null hypothesis  $H_0 : \rho_1 = \rho_2 = \dots = \rho_p = 0$  using the following procedure:

1. Estimate the model  $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$  with OLS. Then, retrieve the regression residuals,  $\hat{\boldsymbol{\varepsilon}}$ .
2. Regress  $\hat{\varepsilon}_t = \mathbf{X}\gamma + \hat{\varepsilon}_{t-1}\rho_1 + \hat{\varepsilon}_{t-2}\rho_2 + \dots + \hat{\varepsilon}_{t-p}\rho_p + e_t$ . Retrieve the  $\tilde{R}^2$ .
3. Construct the BG test statistic:

$$(n - p)\tilde{R}^2 \sim \chi_p^2$$

If  $|BG_{stat}| > BG_{crit}$ , then we reject the null hypothesis and conclude that at least one  $\rho$  is statistically significant. Therefore, we cannot conclude absence of autocorrelation.

The intuition behind the test is that if  $\tilde{R}^2$  is substantially large, then there is evidence that lagged errors are able to explain variation in the current period's error term.

An alternative (matrix algebra) technique to calculate the test statistic after retrieving the vector of estimated residuals from the OLS regression is as follows:

$$BG_{stat} = T \cdot \left[ \frac{\hat{\boldsymbol{\varepsilon}}' \mathbf{X}_* (\mathbf{X}_*' \mathbf{X}_*)^{-1} \mathbf{X}_*' \hat{\boldsymbol{\varepsilon}}}{\hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}} \right]$$

where  $\mathbf{X}_* = [\mathbf{X} \ \hat{\boldsymbol{\varepsilon}}_{t-1} \ \hat{\boldsymbol{\varepsilon}}_{t-2} \ \dots \ \hat{\boldsymbol{\varepsilon}}_{t-p}]$ .

### Box-Pierce-Ljung test

An alternative is the BPL test, which is asymptotically equivalent to the BG test when  $\rho = 0$  is true and when  $\mathbf{X}$  does not include lagged dependent variables. It is also known as the  $Q$  test. The  $Q$  statistic, augmented by Ljung and Box (1979), is as follows:

$$Q_{stat} = T(T + 2) \sum_{i=1}^p \frac{r_i^2}{T - i}$$

where  $r_i = (\sum_{t=i+1}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t-i}) / (\sum_{t=1}^T \hat{\varepsilon}_t^2)$  is the sample correlation among residuals.

However, because the BPL test does not explicitly condition on  $\mathbf{X}_t$ , the test may be less powerful than the BG test when the null hypothesis is false.

### 7.4.6 Dealing with Autocorrelation

There are several ways to deal with autocorrelation in the data. If the correlation structure and autocorrelation factors are known, then we could transform the OLS regression to find autocorrelation-robust estimators. In large samples, we may be able to use OLS parameter estimates and estimate an autocorrelation-robust variance matrix. And in small samples, we will discuss when we may be able use OLS estimators directly. In all cases, we will be dealing with an AR(1) autocorrelation process.

#### Feasible GLS

As with heteroskedasticity, we can derive a robust estimator directly if the parameters of the  $\Omega$  matrix are known. This estimator is as follows:

$$\hat{\beta} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}(\mathbf{X}'\Omega^{-1}\mathbf{Y})$$

The associated estimated variance structure is:

$$\widehat{\text{Var}}(\hat{\beta}|\mathbf{X}) = \frac{1}{T}(\mathbf{Y} - \mathbf{X}\hat{\beta})'\Omega^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta})[\mathbf{X}'\Omega^{-1}\mathbf{X}]^{-1}$$

Let's consider how the OLS regression model can be transformed to yield this estimator and variance. Recall the AR(1) process model:

$$\mathbf{Y}_t = \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t$$

where

$$\boldsymbol{\varepsilon}_t = \rho\boldsymbol{\varepsilon}_{t-1} + e_t$$

Solving for the reduced form yields:  $\mathbf{Y}_t = \mathbf{X}_t\boldsymbol{\beta} + \rho\boldsymbol{\varepsilon}_{t-1} + e_t$ . Now, we'd like to substitute for  $\rho\boldsymbol{\varepsilon}_{t-1}$  in order to remove any of the remaining autocorrelated terms.

$$\mathbf{Y}_{t-1} = \mathbf{X}_{t-1}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{t-1}$$

$$\boldsymbol{\varepsilon}_{t-1} = \mathbf{Y}_{t-1} - \mathbf{X}_{t-1}\boldsymbol{\beta}$$

$$\rho\boldsymbol{\varepsilon}_{t-1} = \rho\mathbf{Y}_{t-1} - \rho\mathbf{X}_{t-1}\boldsymbol{\beta}$$

Substituting into the reduced form yields:

$$\mathbf{Y}_t = \mathbf{X}_t\boldsymbol{\beta} + \rho\mathbf{Y}_{t-1} - \rho\mathbf{X}_{t-1}\boldsymbol{\beta} + \mathbf{e}_t$$

$$(\mathbf{Y}_t - \rho\mathbf{Y}_{t-1}) = (\mathbf{X}_t - \rho\mathbf{X}_{t-1})\boldsymbol{\beta} + \mathbf{e}_t$$

$$\mathbf{Y}_t^* = \mathbf{X}_t^*\boldsymbol{\beta} + \mathbf{e}_t$$

You can see that the transformed equation now has an error term that is white noise. It should be necessary to note that doing this transformation will cause you to lose the first observation (because we are differencing the variables). To avoid the loss, we can transform the first observation only by:  $\sqrt{(1-\rho)}$ . This is the generalized least squares application for dealing with autocorrelation.

What's the problem with this method? Well, nothing if we know  $\rho$ . But almost always, we don't know its value *a priori*. There have been several proposed methods for estimating  $\rho$ .

*Durbin-Watson  $\rho$*

The first method is proposed by Theil and Nagar (1971), which is based on the Durbin-Watson statistic:

$$\hat{\rho} = \frac{n^2(1 - d/2) + k^2}{n^2 - k^2} = \frac{n^2 \cdot \left( \frac{\sum_{t=2}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\sum_{t=1}^T \hat{\varepsilon}_t^2} \right) + k^2}{n^2 - k^2}$$

The term  $\hat{\varepsilon}_t$  is the residual from an initial OLS regression. The estimation  $\hat{\rho}$  is then used to transform the variables. The transformed variables are regressed, and the resulting parameters and variance matrix are autocorrelation-robust.

*Iterative methods*

Iterative methods proposed by Cochran and Orcutt (1949) and Hildreth and Lu (1960) use successive approximation techniques to determine the best estimator of  $\hat{\rho}$ . These methods can be used to estimate  $\rho$  for AR(p) processes. The procedure is as following:

1. Use the Durbin-Watson approach to get the initial estimate  $\hat{\rho}_1$ .
2. Using  $\hat{\rho}_1$ , transform and estimate the autocorrelation-robust OLS equation. The resulting estimator is  $\hat{\beta}^*$ .
3. Using the estimator  $\hat{\beta}^*$  and the transformed  $\mathbf{Y}^*$  and  $\mathbf{X}^*$ , retrieve the transformed regression residuals,  $\hat{e}_t$ .
4. Regress  $\hat{e}_t = \rho_2 \hat{e}_{t-1} + \nu$ . The estimated coefficient is  $\hat{\rho}_2$ .
5. Repeat steps 2-4 with the value of  $\hat{\rho}$  in step 3. Continue this process until  $\hat{\beta}^*$  converges.

**Newey-West autocorrelation-robust standard errors**

Similar to the motivation of using White’s heteroskedasticity-robust standard errors (see section 7.3.3), Newey and West (1987) proposed using OLS estimation, but correcting the variance-covariance matrix such that it is robust to autocorrelation. It is necessary to note, however, that this procedure is only appropriate with a large sample size. The variance matrix is calculated as follows:

$$\hat{\Sigma} = \frac{1}{T} \mathbf{X}' \text{vecdiag}(\hat{\epsilon}\hat{\epsilon}') \mathbf{X} + \frac{1}{T} \sum_{j=1}^L \sum_{t=j+1}^T \left(1 - \frac{j}{L+1}\right) \hat{e}_t \hat{e}_{t-j} [\mathbf{x}_t \mathbf{x}'_{t-j} + \mathbf{x}_{t-j} \mathbf{x}'_t]$$

where  $L \approx T^{1/4}$  (it is the maximum lag of autocorrelation). As with White’s heteroskedasticity-robust standard errors, you can compare these robust standard to the ones produced by OLS to understand the degree of problems that autocorrelation may be causing.

**OLS approach**

There are instances when the OLS method provides better standard error estimators than FGLS or the autocorrelation-robust variance matrix. Griliches and Rao (1969) find that in small samples with  $\rho < 0.3$ , OLS outperforms the other two methods. What is a small enough sample size? Griliches and Rao suggest that samples containing less than 20 observations indicate higher efficiency using OLS methods.

# Econometric References and Resources

- Ashenfelter, O., P. Levine, and D. Zimmerman. 2006. *Statistics and Econometrics: Methods and Applications*. Hoboken, NJ: John Wiley & Sons, Inc.
- Belsley, D., E. Kuh, and R. Welsh. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley and Sons.
- Breusch, T., and A. Pagan. 1979. "A Simple test for Heteroskedasticity and Random Coefficient Variation." *Econometrica* 47:1287–1294.
- Cameron, A., and P. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge, MA: Cambridge University Press.
- Casella, G., and R. Berger. 2002. *Statistical Inference*, 2nd ed. Pacific Grove, CA: Duxbury.
- Davidson, R., and J. MacKinnon. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.
- Durbin, J., and G. Watson. 1950. "Testing for Serial Correlation in Least Squares Regression - I." *Biometrika* 37:409–428.
- Greene, W. 2003. *Econometric Analysis*, 5th ed. Pearson Education, Inc.
- Hamilton, J. 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press.
- Koenker, R. 1981. "A Note on Studentizing a Test for Heteroskedasticity." *Journal of Econometrics* 17:107–112.
- Koenker, R., and G. Bassett. 1982. "Robust Tests for Heteroskedasticity Based on Regression Quantiles." *Econometrica* 50:43–61.
- Shumway, R., and D. Stoffer. 2005. *Time Series Analysis and Its Applications*, 2nd ed. New York, NY: Springer Science & Business Media, LLC.

- Stock, J., and M. Watson. 2006. *Introduction to Econometrics*, 2nd ed. Boston, MA: Pearson Education, Inc.
- Verbeek, M. 2008. *A Guide to Modern Econometrics*, 3rd ed. West Sussex, England: John Wiley & Sons, Ltd.
- White, H. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48:817–838.
- Wooldridge, J. 2002. *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.
- . 2009. *Introductory Econometrics*, 4th ed. Mason, OH: South-Western Cengage Learning.

# Appendix 1: Statistical Tables

Table 1: Cumulative Areas Under the Standard Normal Distribution

<b>z</b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
<b>-3.0</b>	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
<b>-2.9</b>	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
<b>-2.8</b>	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
<b>-2.7</b>	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
<b>-2.6</b>	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
<b>-2.5</b>	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
<b>-2.4</b>	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
<b>-2.3</b>	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
<b>-2.2</b>	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
<b>-2.1</b>	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
<b>-2.0</b>	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
<b>-1.9</b>	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
<b>-1.8</b>	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
<b>-1.7</b>	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
<b>-1.6</b>	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
<b>-1.5</b>	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
<b>-1.4</b>	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
<b>-1.3</b>	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
<b>-1.2</b>	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
<b>-1.1</b>	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
<b>-1.0</b>	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
<b>-0.9</b>	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
<b>-0.8</b>	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
<b>-0.7</b>	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
<b>-0.6</b>	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
<b>-0.5</b>	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
<b>-0.4</b>	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
<b>-0.3</b>	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
<b>-0.2</b>	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859

(continued on next page...)

*Econometric References and Resources*

<b>z</b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
<b>-0.1</b>	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
<b>0.0</b>	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
<b>0.1</b>	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
<b>0.2</b>	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
<b>0.3</b>	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
<b>0.4</b>	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
<b>0.5</b>	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
<b>0.6</b>	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
<b>0.7</b>	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
<b>0.8</b>	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
<b>0.9</b>	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
<b>1.0</b>	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
<b>1.1</b>	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
<b>1.2</b>	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
<b>1.3</b>	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
<b>1.4</b>	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
<b>1.5</b>	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
<b>1.6</b>	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
<b>1.7</b>	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
<b>1.8</b>	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
<b>1.9</b>	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
<b>2.0</b>	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
<b>2.1</b>	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
<b>2.2</b>	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
<b>2.3</b>	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
<b>2.4</b>	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
<b>2.5</b>	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
<b>2.6</b>	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
<b>2.7</b>	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
<b>2.8</b>	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
<b>2.9</b>	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
<b>3.0</b>	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

*Example:* If  $z \sim N(0, 1)$ , then  $P[Z \leq -1.45] = 0.0735$ . Similarly,  $P[Z \leq 2.38] = 0.9913$ .

Table 2: Critical Values of the  $t$  Distribution

	<i>Critical value, <math>\alpha</math></i>					
	<b>1-tailed</b>	<b>0.10</b>	<b>0.05</b>	<b>0.025</b>	<b>0.01</b>	<b>0.005</b>
<b>Degrees of Freedom</b>	<b>2-tailed</b>	<b>0.20</b>	<b>0.10</b>	<b>0.05</b>	<b>0.02</b>	<b>0.01</b>
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250
10		1.372	1.812	2.228	2.764	3.169
11		1.363	1.796	2.201	2.718	3.106
12		1.356	1.782	2.179	2.681	3.055
13		1.350	1.771	2.160	2.650	3.012
14		1.345	1.761	2.145	2.624	2.977
15		1.341	1.753	2.131	2.602	2.947
16		1.337	1.746	2.120	2.583	2.921
17		1.333	1.740	2.110	2.567	2.898
18		1.330	1.734	2.101	2.552	2.878
19		1.328	1.729	2.093	2.539	2.861
20		1.325	1.725	2.086	2.528	2.845
21		1.323	1.721	2.080	2.518	2.831
22		1.321	1.717	2.074	2.508	2.819
23		1.319	1.714	2.069	2.500	2.807
24		1.318	1.711	2.064	2.492	2.797
25		1.316	1.708	2.060	2.485	2.787
26		1.315	1.706	2.056	2.479	2.779
27		1.314	1.703	2.052	2.473	2.771
28		1.313	1.701	2.048	2.467	2.763
29		1.311	1.699	2.045	2.462	2.756
30		1.310	1.697	2.042	2.457	2.750
40		1.303	1.684	2.021	2.423	2.704
60		1.296	1.671	2.000	2.390	2.660
90		1.291	1.662	1.987	2.368	2.632
120		1.289	1.658	1.980	2.358	2.617
$\infty$		1.282	1.646	1.962	2.330	2.581

*Example:* for a 5% significance value two-tailed test with 20  $df$ , the critical value is 2.086.

Table 3: 5% Critical Values of the  $F$  Distribution

		Numerator Degrees of Freedom									
		1	2	3	4	5	6	7	8	9	10
Denominator Degrees of Freedom	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	
$\infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	

*Example:* for a 5% significance value with numerator  $df = 8$  and denominator  $df = 40$ , the critical value is 2.34.

# Appendix 2: SAS Basics

## Overview

SAS is a power statistical analysis tool that is used in many different industries and by many companies. SAS enables you to organize and analyze data using both graphical and analytical methods. The software is flexible, dynamic, and is frequently updated to include statistical and econometric features that are emerging in the professional fields.

## Starting SAS (Windows OS)

Use the Windows **Start** menu to navigate to the SAS executable as follows:

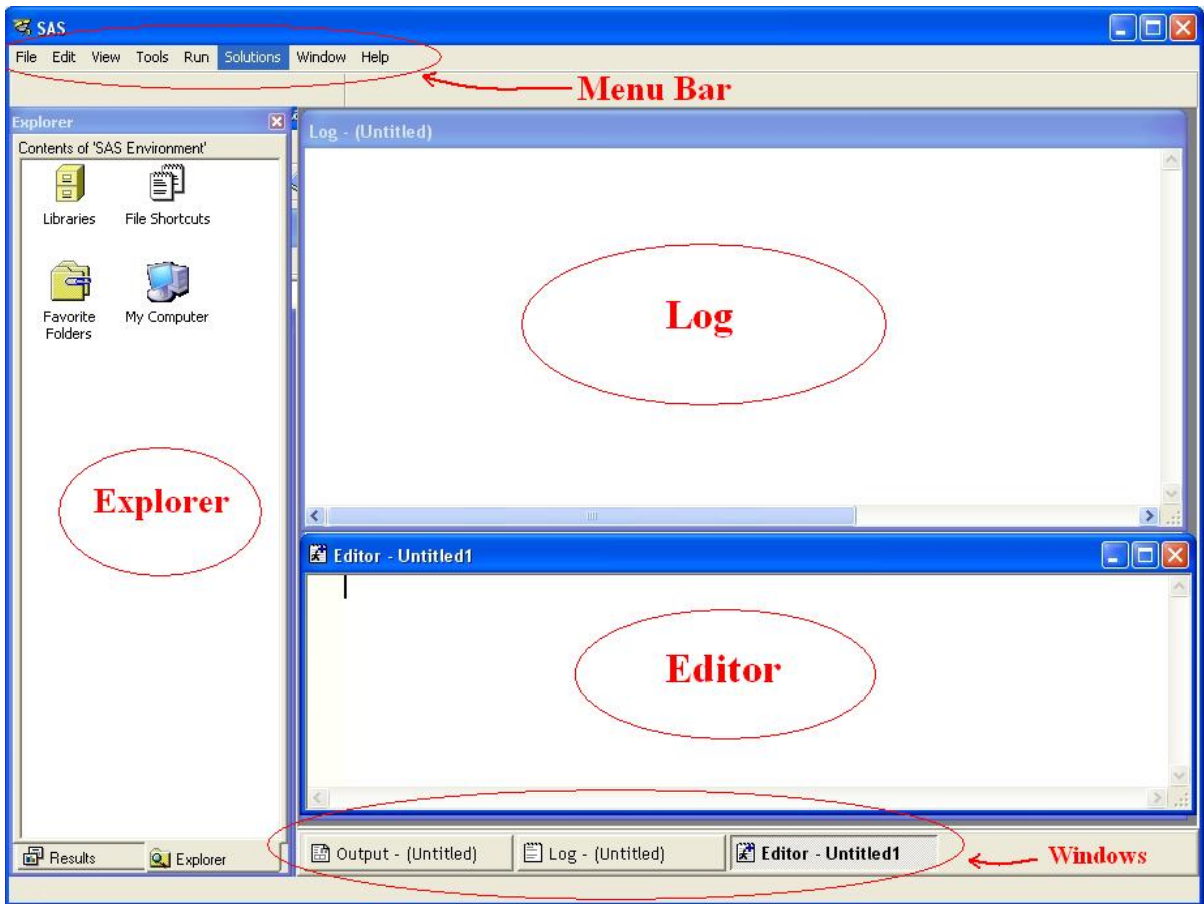
Start → All Programs → SAS → SAS 9.2 (English)

## Initial View

When you launch SAS, you will see the initial view, shown in figure A2.1.

- **Log:** Used to examine system messages. The log shows errors in and successful command execution.
- **Editor:** Used to enter programming commands, which tell SAS how the user wants to analyze the data.
- **Explorer:** Used to navigate to active data sets and results. Click the appropriate tab at the bottom of the Explorer panel.

Figure A2.1: Initial SAS View



- **Output (on Windows bar):** Used to view the analytical results.
- **Menu Bar:** Used for purposes such as opening and saving files, opening various panels, and setting preferences.

**NOTE:** If you close any of the panels that are open in the initial view, you can do the following to re-open them:

1. Click **View** in the Menu Bar.
2. Select the appropriate panel:

Log	→	Opens the log panel
Enhanced Editor	→	Opens/focuses on the command editor
Output	→	Opens the output panel
Contents Only	→	Opens the Explorer panel

## Importing Data

To import data, you can use the SAS Import Wizard as follows:

1. **IMPORTANT:** If you opened the data file in Excel, you need to close Excel before beginning the Import Wizard. Otherwise, you will receive a SAS error and the data will not be imported.
2. In SAS, click **File** in the Menu Bar and select **Import Data**. The Import Wizard window opens (as shown in figure A2.2).
3. In the “Select a data source from the list below” field, select **Microsoft Excel 97, 2000 or 2002 Workbook** (this is the default). Then, click **Next**. The Connect to MS Excel window opens.
4. In the Connect to MS Excel window, click **Browse**. Navigate to the data file, select it, and then click **OK**.
5. In the “What table do you want to import” field, select the appropriate sheet. Then, click **Next**. **NOTE:** The data files that I provide will only have a single sheet.
6. In the “Library” field, select **WORK** (default).
7. In the “Member” field, enter a name for the data set. Make it something intuitive. For example, if analyzing a data set of wheat prices, name the data set *wheat\_data*.
8. Click **Finish** to complete the Import Wizard. The Import Wizard window closes.

### Checking Data Import

There are two ways to check whether the data was successfully imported and can be used for statistical analyses. These are as follows:

1. Check the Log panel. A successful import will result in the following log message:  
*NOTE: WORK.WHEAT\_DATA was successfully created.*
2. Check the Explorer panel.
  - (a) Click the **Explorer** tab at the bottom of the Explorer panel.
  - (b) Double-click the **Libraries** icon.
  - (c) Double-click the **Work** icon.




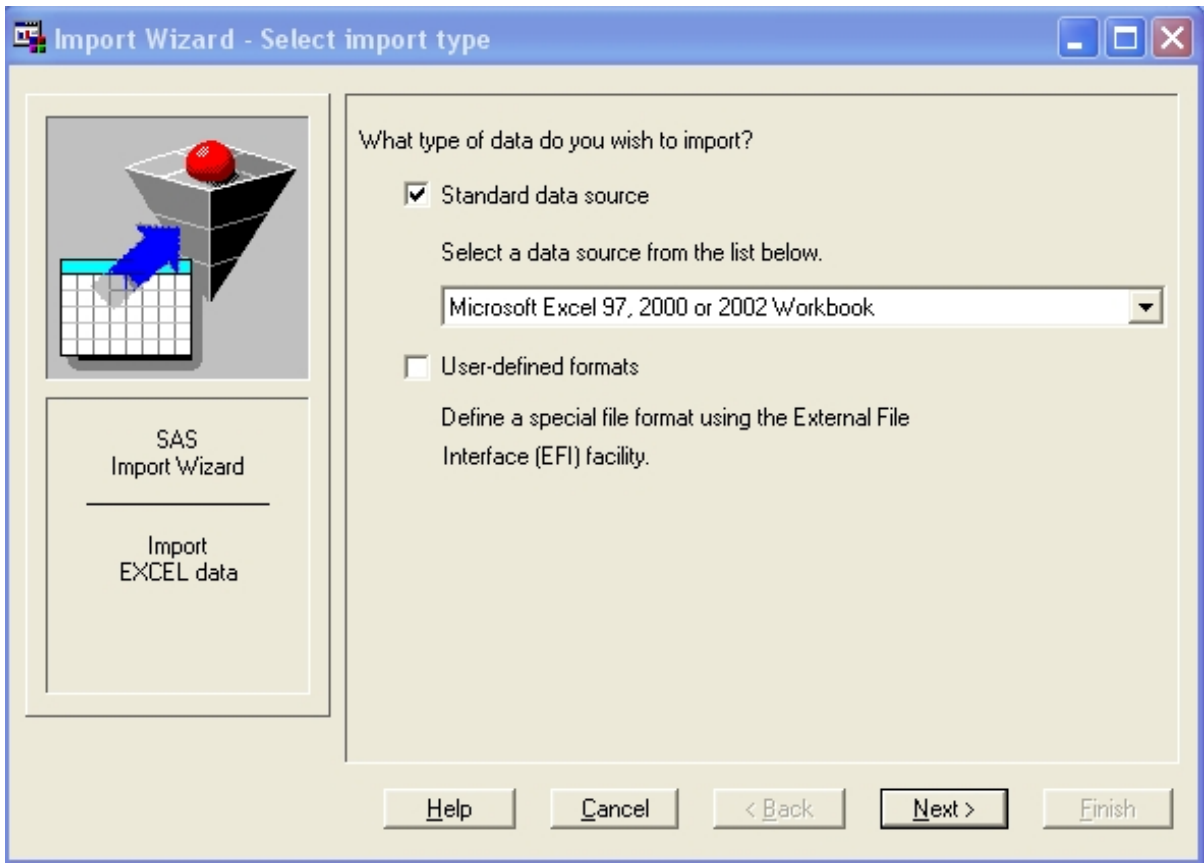
You should now see a data table icon  that is named with the designated data name. In this example, the icon would be named **Wheat\_data**.

Figure A2.2: SAS Data Import Wizard



## Coding in SAS

To successfully write code within SAS, you can follow these guidelines:

- All code is written in the Editor panel.
- The code is often color coded:

BLUE → Syntactically correct  
RED → Incorrectly coded

- **Every** line of code must end with a semicolon (;) (Most errors will stem from users forgetting to end a line with a semicolon).
- Specify the data set that you are using by including the `data = your_dataset` statement following any `proc ...` statement. For example,

```
proc means data=wheat_data;
```

- After finishing writing a snippet of code, you must end the snippet with the **run;** statement. For example,

```
proc means data=wheat_data;  
run;
```

- To execute the written code, click the **Run** icon  found in the Icon bar, which is directly below the Menu bar.

**NOTE:** You can execute only a certain part of the written code. To do so, highlight the code that you would like to execute, and then click the **Run** icon.

### Troubleshooting Tips

As a general rule, you should make a habit of checking the Log panel after you executed the code. In most cases, a successfully executed code will result in blue messages in the Log panel, as well as a message that indicates that the analysis was successfully completed.

- The Log panel has the following color scheme:

BLUE	→	Successfully executed analyses
RED/BURGUNDY	→	Errors and/or Warnings
GREEN	→	Warnings

- Check if you have placed a semicolon at the end of *every* line of code.
- Check for incorrectly spelled code (e.g. options, dataset name, variable names).
- Check that you have included a snippet of code with the **run;** statement.

## Calculating Summary Statistics

In almost all cases, one of the first (and very useful) steps that you want to take when dealing with data analysis is to generate summary statistics. These often include the mean (average), minimum, maximum, number of observations, number of missing observations, standard deviation, and others. With just the summary statistics, you are able to gain a general overview of the data's characteristics.


## Summary Statistics Using SAS

To calculate summary statistics using SAS, you will use the MEANS procedure. In the Editor panel, the MEANS procedure can be coded as follows:

```
proc means data=dataset_name options ;  
run;
```

The *options* designate which summary statistics to produce. To produce the mean (average), minimum, and maximum for the wheat prices, you can use the following code:

```
proc means data=wheat_data mean min max ;  
var wheat_p;  
run;
```

After you have written the above code in the Editor panel, click the **Run** icon . The Output panel will automatically be displayed with the resulting summary statistics. In addition to the mean, minimum, and maximum, you can produce other summary statistics by specifying additional options. The following options can be specified:

### Options for PROC MEANS

---

---

Command	Resulting Summary Statistic
mean	Mean (average)
min	Minimum
max	Maximum
n	Number of observations
std	Standard deviation

---

## Implementing Graphical Analysis

Although analytical summary statistics are extremely useful for providing an overview of the data, such an overview may not reveal some additional important information, which may be relevant. A visual representation of the dataset can provide another perspective on examining the data. By plotting and examining the graphical representation, you can observe important factors such as trends, outliers, and patterns.


## Graphical Analysis Using SAS

To produce a plot of the data series, you will use the SGPLOT procedure. In the Editor panel, the SGPLOT procedure can be coded as follows:

```
ods graphics on ;
proc sgplot data=dataset_name ;
series y=price_variable x=time_variable ;
run;
ods graphics off ;
```

To generate the plot of the wheat prices over time, you can use the following code:

```
ods graphics on ;
proc sgplot data=wheat_data ;
series y=wheat_p x=date ;
run;
ods graphics off ;
```

After you have written the above code in the Editor panel, click the **Run** icon .

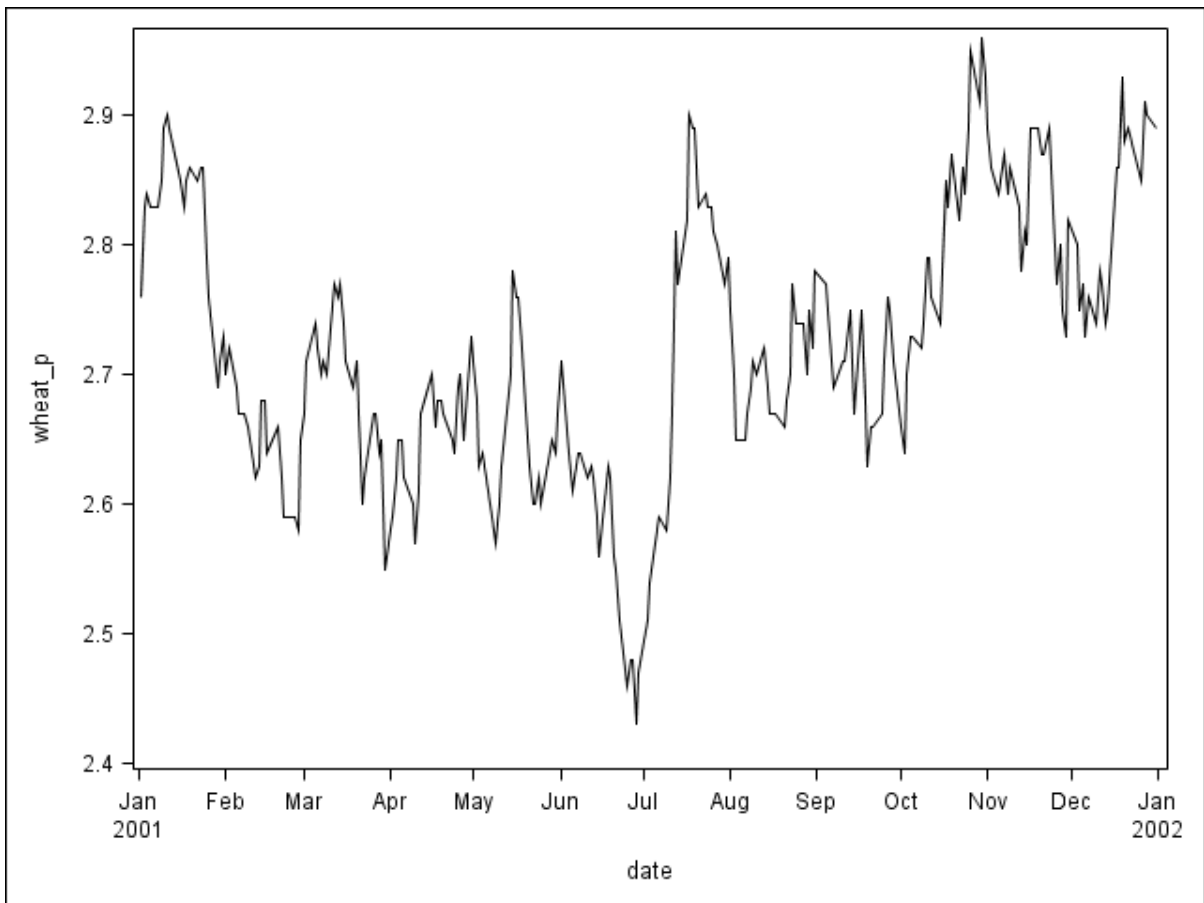
### Retrieving the Resulting Plot

To retrieve the resulting plot, you should do the following:

1. In the Explorer panel, click the **Results** tab at the bottom of the Explorer panel.
2. Double-click the bottom-most icon “Sgplot: The SAS System.” The “Sgplot: The SAS System” folder expands.
3. Double-click the icon “The Sgplot Procedure.” “The Sgplot Procedure” folder expands.
4. Double-click the picture icon “The Sgplot Procedure.” A picture viewer opens with the plot.

For the wheat data example, the plot of the wheat price is shown in figure A2.3.

Figure A2.3: SAS SGPlot Procedure Graphic



## SAS Code Samples

```
/* I. Specifying a library name */
libname myLib "C:\Users\anton.bekkerman\Documents\";

/* II. Importing data */

/* Method 1: Use the SAS wizard to import Excel, CSV, Access,
   DBA, etc. data files */

/* Method 2: Use code to import Excel, CSV, Access, etc. data files */
/* Example 1: Importing Excel files */
PROC IMPORT OUT= MYLIB.DT1
            DATAFILE= "C:\Users\anton.bekkerman\Documents\Classes\
            ECNS561\SAS Lab\dt1.xls"
            DBMS=XLS REPLACE;
            GETNAMES=YES;
GUESSINGROWS=50;
RUN;
/* Example 2: Importing CSV files */
PROC IMPORT OUT= myLib.dt1
            DATAFILE= "C:\Users\anton.bekkerman\Documents\Classes\
            ECNS561\SAS Lab\dt1.csv"
            DBMS=CSV REPLACE;
            GETNAMES=YES;
            DATAROW=2;
            GUESSINGROWS=50;
RUN;

/* III. Creating a data set by entering values */

data myLib.dt1;
input record year farm $ acres chickens $2.;
label record="Farm Number"
       year="Year"
       farm="Farm Name"
       acres="Planted Acres"
       chickens="Number of chickens"
;
datalines;
```

```
2 1999 HappyFam 554 44
1 1999 MmDonuts 334 32
3 1999 NoRecord . .
5 1999 TinyCow 1234 56
8 1999 Wheater 442 12
14 1999 Deere 994 72
1 2000 MmDonuts 324 41
2 2000 HappyFam 604 35
3 2000 NoRecord . .
5 2000 TinyCow 1532 76
8 2000 Wheater 204 9
14 2000 Deere 763 63
```

```
;
run;
```

```
/* IV. Basic data manipulation */
```

```
/* a. Sorting data */
```

```
/* 1. Sort data by a variable in ascending order (lowest is first) */
```

```
/* Example: Sort farms such that those with lowest acreage are first */
```

```
proc sort data=myLib.dt1;
by acres;
run;
```

```
/* 2. Sort data by a variable in descending order (highest is first) */
```

```
/* Example: Sort farms such that those with highest acreage are first */
```

```
proc sort data=myLib.dt1;
by descending acres;
run;
```

```
/* 3. Sort data by more than one variables. That is, sort within subgroups */
```

```
/* Example: Sort farms such that those with lowest acreage in each year
are first */
```

```
proc sort data=myLib.dt1;
by year acres;
run;
```

```
/* b. Removing/keeping certain observations */
```

```
/* 1. Remove specific observations */
```

```
/* Example: Remove farms that are named "Wheater" */
```

```
data myLib.dt2;  
set myLib.dt1;  
if farm="Wheater" then delete;  
run;
```

```
/* 2. Remove observations according to a criteria */
```

```
/* Example: Remove all farms that have acreage above or equal to 1,000 */
```

```
data myLib.dt2;  
set myLib.dt1;  
if acres >= 1000 then delete;  
run;
```

```
/* 3. Keep observations according to a criteria */
```

```
/* Example: Keep only farms that have acreage between 300 and 900 acres */
```

```
data myLib.dt2;  
set myLib.dt1;  
if 300 < acres <= 900;  
run;
```

```
/* 4. Remove missing observations */
```

```
/* Example: Remove farms that have a missing observation for acres */
```

```
data myLib.dt2;  
set myLib.dt1;  
if acres = . then delete;  
run;
```

```
/* c. Manipulating variables (columns) */
```

```
/* 1. Keep only certain variables from an original data set */
```

```
/* Example: Keep only the year and farm name variables */
```

```
data myLib.dt2;
set myLib.dt1(keep=year farm);
run;

/* 2. Delete certain variables from the original data set */

/* Example: Drop the farm number and acreage variables */
data myLib.dt2;
set myLib.dt1(drop=record acres);
run;

/* 3. Rename variable names */

/* Example: Rename the farm number variable from "record" to "number" */
data myLib.dt2;
set myLib.dt1(rename=(record=number));
run;

/* 4. Convert a character variable into a numeric */

/* Example: Convert the character variable describing the number of
chickens on a farm into a numeric format */
data myLib.dt2;
set myLib.dt1;
chickN = input(chickens, 8.);
drop chickens;
label chickN="Number of chickens";
rename chickN=chickens;
run;

/* V. Advanced data manipulation */

/* First, let's create four data sets for us to use */

data myLib.dta1;
input record year farm $ acres chickens $2.;
label record="Farm Number"
      year="Year"
      farm="Farm Name"
      acres="Planted Acres"
```

```
    chickens="Number of chickens"
;
datalines;
2 1999 HappyFam 554 44
1 1999 MmDonuts 334 32
3 1999 NoRecord . .
5 1999 TinyCow 1234 56
8 1999 Wheater 442 12
14 1999 Deere 994 72
;
run;

data myLib.dta2;
input record year farm $ acres chickens $2.;
label record="Farm Number"
    year="Year"
    farm="Farm Name"
    acres="Planted Acres"
    chickens="Number of chickens"
;
datalines;
2 2000 HappyFam 554 35
1 2000 MmDonuts 324 41
3 2000 NoRecord . .
5 2000 TinyCow 1532 76
8 2000 Wheater 442 9
14 2000 Deere 763 63
;
run;

data myLib.dta3;
input record year farm $;
label record="Farm Number"
    year="Year"
    farm="Farm Name"
;
datalines;
2 1999 HappyFam
1 1999 MmDonuts
3 1999 NoRecord
```

```
5 1999 TinyCow
8 1999 Wheater
14 1999 Deere
1 2000 MmDonuts
2 2000 HappyFam
3 2000 NoRecord
5 2000 TinyCow
8 2000 Wheater
14 2000 Deere
;
run;

data myLib.dta4;
input farm $ acres chickens $2.;
label farm="Farm Name"
      acres="Planted Acres"
      chickens="Number of chickens"
;
datalines;
HappyFam 554 44
MmDonuts 334 32
NoRecord . .
TinyCow 1234 56
Wheater 442 12
Deere 994 72
MmDonuts 324 41
HappyFam 604 35
NoRecord . .
TinyCow 1532 76
Wheater 204 9
Deere 763 63
;
run;

/* a. Setting one data set into another */

/* Example: Set the dataset "dta2" onto "dta1" */
data myLib.dta5;
set myLib.dta1 myLib.dta2;
run;
```

```
/* b. Merging data sets together */
```

```
/* Example: Merge data sets "dta3" and "dta4" together by farm name */
```

```
proc sort data=myLib.dta3;  
by farm;  
run;  
proc sort data=myLib.dta4;  
by farm;  
run;
```

```
data myLib.dta5;  
merge myLib.dta3 myLib.dta4;  
by farm;  
run;
```

```
/* c. Append data sets */
```

```
/* Example: Append data set "dta2" to "dta1" */
```

```
proc append base=myLib.dta1  
data=myLib.dta2  
force;  
run;
```

```
/* d. Exporting a SAS data set */
```

```
/* Method 1: Use the SAS wizard to export a data set */
```

```
/* Method 2: Code the export of a data set */
```

```
PROC EXPORT DATA= WORK.WEIGHT_IML  
OUTFILE= "C:\Users\anton.bekkerman\Documents\Classes\  
ECNS561\SAS Lab\dt1_exp.csv"  
DBMS=CSV REPLACE;  
PUTNAMES=YES;  
RUN;
```

```
/* VI. Exploring and summarizing the data */
```

```
/* a. Producing basic summary statistics about the data set */
```

```
/* Example 1: Default summary statistics about all variables in a data set */
```

```
proc means data=myLib.dt1;
run;

/* Example 2: Default summary statistics about only specific variables */
proc means data=myLib.dt1;
var acres;
run;

/* Example 3: Selected summary statistics */
proc means data=myLib.dt1 n nmiss mean median std;
run;

/* b. Frequency statistics */

/* 1. Basic frequency and additional statistics */
proc univariate data=myLib.dt1;
run;

/* 2. Descriptive frequency statistics */
proc freq data=myLib.dt1;
tables farm acres;
run;

/* c. Visualizing the data -- graphics */

/* 1. Producing a basic scatter plot of the data */

/* Example: Seeing acreage as a function of farm number */
ods graphics on;
proc sgplot data=myLib.dt1;
scatter x=record y=acres;
run;
ods graphics off;

/* 2. Producing a series plot (i.e. connecting the dots) */

/* Example: Seeing the number of holidays in Spain taken by US residents */
ods graphics on;
proc sgplot data=sashelp.tourism;
series x=year y=vsp;
```

```
run;
ods graphics off;

/* 3. Producing a bar graph */
/* Example: Bar chart of acres on farms in 1999 */
ods graphics on;
proc sgplot data=myLib.dt1;
where year=1999;
vbar farm / response=acres;
run;
ods graphics off;

/* Example: Bar chart of comparing acres in 1999 and 2000*/
data myLib.dt1_graph;
set myLib.dt1;
if year = 1999 then acres99 = acres;
if year = 2000 then acres00 = acres;
label acres99 = "Acres in 1999"
      acres00 = "Acres in 2000";
run;
ods graphics on;
proc sgplot data=myLib.dt1_graph;
yaxis label = "Acreage";
vbar farm / response=acres99;
vbar farm / response=acres00
barwidth=0.5
transparency=0.2;
run;
ods graphics off;

/* 4. Producing a histogram and density curve */

/* Example: Distribution of acreage */
ods graphics on;
proc sgplot data=myLib.dt1;
histogram acres;
density acres / type=normal;
density acres / type=kernel;
run;
ods graphics off;
```

```
/* VII. Introduction to SAS matrix language */

/* Entering into IML (interactive matrix language) mode */
proc iml;

/* Set option to display row and column numbers */
reset autoname;

/*Construct a vector; columns separated by space, rows separated by "," */
v1 = {1 2 3 4};
print v1;

/*Construct a matrix; columns separated by space, rows separated by "," */
m1 = {1 2, 3 4};
print m1;

/* Change a value in a vector or a matrix */
/* Example: Change the value of the second column in the vector v1 */
v1[,2] = 5;
print v1;

/* Example: Change the value of the second row, first column in matrix m1 */
m1[2,1] = 5;
print m1;

/* Construct an identity matrix; I(num columns) */
i1 = i(2);
print i1;

/* Construct a matrix of all ones; J(rows, cols, value) */
ones1 = j(2,2,1);
print ones1;

/* Transpose a vector or matrix; t(vector name) or (vector name)' */
v1t = t(v1);
m1t = m1';
print v1t, m1t;
```

```
/* Inverse of square matrix; inv(matrix name) */
m1inv = inv(m1);
print m1inv;

/* Adding or subtracting a scalar */
v2 = v1 - 1;
m2 = m1 - 1;
print v2, m2;

/* Multiplying/dividing by scalars */
v3 = v1 # 2;
m3 = m1 # 2;
print v3, m3;

/* Raising to a power */
v4 = v1 ## 2;
m4 = m1 ## (0.5);
print v4, m4;

/* Vector / matrix addition/substraction (must be the same dimensions) */
v5 = v1 + v1;
m5 = m1 + m1;
print v5, m5;

/* Vector/matrix multiplication (must be appopriate dimensions) */
v6 = v1 * v1';
m6 = m1 * m1';
print v6, m6;

/* Determine the rank of a matrix */
m1rank=round(trace(ginv(m1)*m1));
print m1rank;

/* Read a SAS data set into a matrix */
/* 1. Specify which data set you wish to import into IML */
use sashelp.bweight;

/* 2. Specify that you want to read all of the variables into the matrix */
read all into bweight;
print bweight;
```

```
/* 3. Read only specific variables and only a range of observations  
into a matrix */
```

```
/* Example: Read only the first 100 observations of the variables  
"weight" "black" and "married" */
```

```
read point (1:100) var {weight black married} into bweight;  
print bweight;
```

```
/* Summary statistics */
```

```
summary var {weight black married} stat{mean std min max} opt{save};  
print weight, black, married;
```

```
/* Export matrices into SAS data sets */
```

```
cols = {"weight" "black" "married"};  
create weight_ims from bweight[colname=cols] ;  
append from bweight;
```

```
/* Exit from IML */
```

```
quit;
```