ECNS561, Fall 2011

Take-home Graded Practice Opportunity 5

Due date: November 29, 2011, 5:00 p.m.

Practice with testing restrictions

1. You are working with the function: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$. Variables $\{X_1, X_2, X_3\}$ are continuous and X_4 is a categorical variable. After estimating the model, you retrieve the following results:

β_0	β_1	β_2	β_3	β_4
-2.47	1.82	0.57	0.19	-3.32

R^2	$\hat{\varepsilon}'\hat{\varepsilon}$	n
0.91	8.97	9

$$(\boldsymbol{X'X})^{-1} = \begin{bmatrix} 8.82 & -0.65 & -0.55 & -0.38 & -1.21 \\ -0.65 & 0.06 & 0.05 & 0.02 & 0.06 \\ -0.55 & 0.05 & 0.05 & 0.01 & 0.02 \\ -0.38 & 0.02 & 0.01 & 0.05 & 0.12 \\ -1.21 & 0.06 & 0.02 & 0.12 & 0.75 \end{bmatrix}$$

(a) Set up and test the joint hypothesis that $X_1 = 0$ and $X_2 = 0$.

$$R = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \qquad q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$(R\hat{\beta} - q) = \begin{bmatrix} 1.82 \\ 0.57 \end{bmatrix} \qquad (R(X'X)^{-1}R') = \begin{bmatrix} 0.057 & 0.045 \\ 0.045 & 0.051 \end{bmatrix}$$

$$F_{stat} = \frac{(\boldsymbol{R\beta - q})'\{\boldsymbol{R}(\boldsymbol{X'X})^{-1}\boldsymbol{R'}\}^{-1}(\boldsymbol{R\beta - q})}{J\hat{\sigma}^2} =$$

$$\frac{\begin{bmatrix} 1.82 \\ 0.57 \end{bmatrix}' \left\{ \begin{bmatrix} 0.057 & 0.045 \\ 0.045 & 0.051 \end{bmatrix} \right\}^{-1} \begin{bmatrix} 1.82 \\ 0.57 \end{bmatrix}}{2 \cdot 2.24} = 24.725$$

The p-value can be calculated as $1 - cdf(F, F_{stat}, J, n-k)$, where $cdf(\cdot)$ represents the cumulative density function evaluated at the value of the F_{stat} with $J,\ n-k$ degrees of freedom.

$$p_{value} = 0.005$$

Therefore, we reject the joint null hypothesis.

(b) Set up and test the joint hypothesis that $X_2 = X_3$ and $X4 = -3$.

$$R = \begin{bmatrix} 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad q = \begin{bmatrix} 0 \\ -3 \end{bmatrix}$$

$$(R\hat{\beta} - q) = \begin{bmatrix} 0.38 \\ -0.32 \end{bmatrix} \qquad (R(X'X)^{-1}R') = \begin{bmatrix} 0.08 & -0.10 \\ -0.10 & 0.75 \end{bmatrix}$$

$$F_{stat} = \frac{(\boldsymbol{R\beta - q})'\{\boldsymbol{R}(\boldsymbol{X'X})^{-1}\boldsymbol{R'}\}^{-1}(\boldsymbol{R\beta - q})}{J\hat{\sigma}^2} =$$

$$\frac{\begin{bmatrix} 0.38 \\ -0.32 \end{bmatrix}' \left\{ \begin{bmatrix} 0.08 & -0.10 \\ -0.10 & 0.75 \end{bmatrix} \right\}^{-1} \begin{bmatrix} 0.38 \\ -0.32 \end{bmatrix}}{2 \cdot 2.24} = 0.388$$

The p-value can be calculated as $1 - cdf(F, F_{stat}, J, n-k)$, where $cdf(\cdot)$ represents the cumulative density function evaluated at the value of the F_{stat} with $J,\ n-k$ degrees of freedom.

$$p_{value} = 0.702$$

Therefore, we cannot reject the joint null hypothesis.

Putting it all together

You are a consultant working for a major airport. You are tasked with determining manners in which to decrease the average waiting time for travelers in airports. To do so, you decide to perform an empirical analysis of waiting times at three other major airports. The U.S. Customs and Border Protection agency posts these times for a number of airports, and these data can be accessed at: http://apps.cbp.gov/awt/index.asp

You decide to randomly select the Hartsfield-Jackson Atlanta International Airport (ATL), the Chicago O'Hare Airport (ORD), and the George Bush Intercontinental/Houston Airport (IAH). After looking at the data, you decide that instead of dealing with "Time of Day" ranges, you will simply round up to the nearest hour. For example, the range "Midnight to 1 A.M." will be "Hour = 1;" the range "1 A.M. to 2 A.M." will be "Hour = 2." Furthermore, you believe that there are distinct differences in the airports and decide to control for these differences using categorical variables. Your initial model is therefore:

$$Avg.\ wait\ time = \beta_0 + \beta_1 Hour + \beta_2 Passengers + \beta_3 Booths + \sum_i \beta_{ji} Airport + \varepsilon$$

where $Hour$ represents the hour of day, $Passengers$ is the average number of arriving passengers, $Booths$ is the number of open service booths, and $Airport$ is a categorical variable representing the airport.

Your goals are the following:

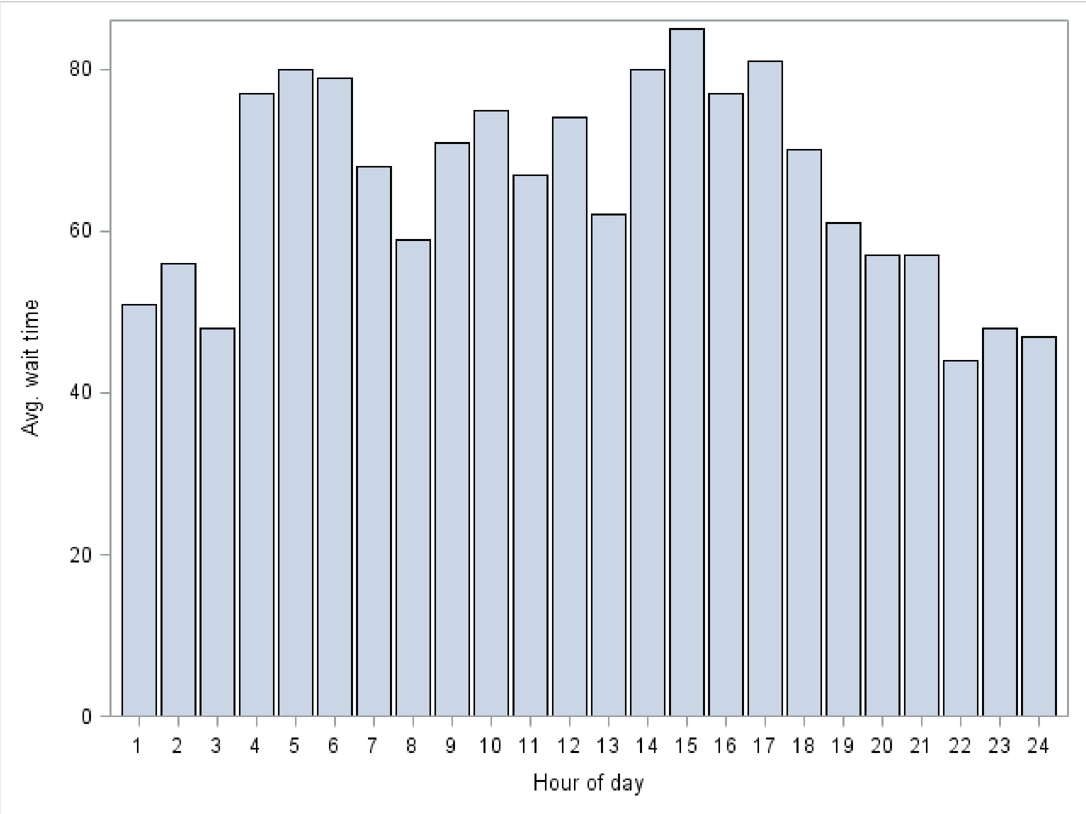
See SAS code on course website

1. Collect, clean, and organize the data.
2. Import the data into your favorite statistical analysis software and explore it using visual and statistical summaries. Provide a brief interpretation of the output and potential economic and econometric implications.

   The summary statistics indicate that on average people wait 22 minutes. Interestingly, passengers must always wait more than 10 minutes. Furthermore, there are on average approximately one booth open for each passenger, and there are at times when there are booths open but no passengers. This may indicate inefficient allocation of resources.

   The figure indicates that during off-peak hours (late night and early morning) there is a substantially lower waiting time then between the primary airport operating hours. This may suggest a structural difference between the data generating process associated with off-peak and peak hours.

Variable	N	Mean	Std Dev	Minimum	Maximum
Avg. wait time	71	22.17	5.57	11	33
Avg. arriving passengers	71	458.83	462.11	0	1,581
Open service booths	71	20.24	11.58	4	43
ATL Dummy	71	0.34	0.48	0	1
ORD Dummy	71	0.34	0.48	0	1
IAH Dummy	71	0.32	0.47	0	1



3. Estimate the model proposed above. Provide economic interpretation of the results.
4. Show that there is multicollinearity in the above model. What are the economic reasons for the multicollinearity? How would you attempt to deal with the multicollinearity?

Parameter Estimates

Variable	**Parameter**	**StdErr**	**tValue**	**VIF**
Intercept	19.998	1.557	12.850	0.000
Hour of day	-0.356	0.073	-4.860	1.118
Avg. arriving passengers	0.002	0.003	0.870	7.016
Open service booths	0.215	0.110	1.960	7.012
ATL Dummy	3.234	1.196	2.700	1.410
ORD Dummy	0.396	1.175	0.340	1.362

Variance inflation factors indicate that the number of arriving passengers and the number of open service booths may be collinear. This is not surprising, because more booths open as more passengers arrive. Because we are interested in making a recommendation to the airport, one way to deal with the problem is to eliminate the variable describing the average number of passengers (since we are not able to control how many passengers arrive, but we are able to the number of open booths).

5. One potential manner in which to reduce waiting times is to increase the number of service booths during peak hours. How would you test whether this would be an effective strategy? Re-estimate the model (after dealing with multicollinearity) that can be used to answer the question. Interpret the results.

To test this hypothesis, we can create an interaction variable: hour $\times$ booths. The estimation results are as follows:

Parameter Estimates

Label	**Parameter**	**StdErr**	**tValue**
Intercept	15.148	2.020	7.500
Hour of day	-0.061	0.129	-0.470
Open service booths	0.653	0.139	4.700
Hour $\times$ Booths	-0.024	0.009	-2.650
ATL Dummy	2.722	1.136	2.400
ORD Dummy	0.743	1.130	0.660

The coefficient associated with the interaction term is negative and statistically significant. This indicates that during the peak hours, additional booths are correlated with less average waiting times.

6. Lastly, you wish to know whether separate models must be used to estimate average waiting times for off-peak and peak hours. That is, you believe that there are structural differences in waiting times occurring between 22:00 (10 p.m.) and 3:00 (3 a.m.), and waiting times occurring between 4:00 (4 a.m.) and 21:00 (9 p.m.). Test this hypothesis and provide an economic interpretation of the findings.

   The $F_{chow} = 1.216$ *and the associated p-value = 0.311. Therefore, we cannot reject the null hypothesis that a single model can be used to describe waiting times during off-peak and peak hours.*

7. What final recommendations can you make to the airport?