

Take-home Graded Practice Opportunity 4

Due date: November 8, 2011, 5:00 p.m.

1. Are the Gauss-Markov assumptions required to perform a t -test that determines whether an estimated OLS coefficient is significantly different than zero? Why or why not?

No. The assumption of normally distributed errors is not required, because asymptotically, we have shown the OLS estimator to be normally distributed. Therefore, the normality assumption is only necessary when dealing with small sample sizes in order to argue that the t statistic has a t distribution. Therefore, using a t distribution or standard normal distribution to compare with the t statistic yields approximately valid results. As the sample size increases, the approximation becomes more accurate.

2. Suppose that you estimate an OLS regression of \mathbf{Y} on \mathbf{X} and retrieve the associated residuals, $\hat{\boldsymbol{\epsilon}}$. What is the R^2 of an OLS regression of the residuals $\hat{\boldsymbol{\epsilon}}$ on \mathbf{X} ? Prove and explain the intuition.

Intuitively, when we regress Y on \mathbf{X} , all of the information in \mathbf{X} is used to explain variation in Y . Therefore, the regression residuals represent the variation in Y that cannot be explained by \mathbf{X} . Regressing these residuals on \mathbf{X} will yield parameter estimates that are equal to zero and an R^2 that is zero, because \mathbf{X} cannot explain variation in the Y that hasn't been explained by \mathbf{X} in the first regression. Therefore, the squared correlation between information in Y not explained by \mathbf{X} and information in \mathbf{X} is zero.

3. True or false: if an estimated coefficient is significant at the 10% level, then it is also significant at the 5% level. Explain.

False. The statement is reversed. Therefore, if we find significance at a 5% level, then there is also significance at the 10% level. Not the other way around.

4. Suppose that you have data on middle schoolers' age, weight, and parents' income (in thousands).

Age	Weight	Income
14	110	45
12	95	28
13	105	40
12	100	40
14	105	42

Consider the model: $Weight = f(Age, Income) + \varepsilon$. Compute the following by hand:

- (a) The estimated coefficient vector $\hat{\beta}$, predicted weights vector, the residuals vector.

The following is a summary of the results:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 5 & 65 & 195 \\ 65 & 849 & 2554 \\ 195 & 2554 & 7773 \end{bmatrix} \quad \mathbf{X}'\mathbf{X}^{-1} = \begin{bmatrix} 49.11 & -4.64 & 0.29 \\ -4.64 & 0.54 & -0.06 \\ 0.29 & -0.06 & 0.01 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} 49.7 \\ 2.56 \\ 0.51 \end{bmatrix}$$

$$\hat{Y} = \begin{bmatrix} 108.64 \\ 94.78 \\ 103.51 \\ 100.96 \\ 107.10 \end{bmatrix}$$

$$\hat{\varepsilon} = \begin{bmatrix} 1.35 \\ 0.22 \\ 1.49 \\ -0.96 \\ -2.10 \end{bmatrix}$$

- (b) The vector of standard errors. Using just the standard errors, determine whether each estimated coefficient is statistically different from zero and explain how you are reaching your conclusion.

First, determine the vector standard errors by taking the square root of the main diagonal of the estimated variance-covariance matrix: $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$.

$$\widehat{Var}(\boldsymbol{\beta}|\mathbf{X}) = \begin{bmatrix} 231.32 & -21.86 & 1.38 \\ -21.86 & 2.54 & -0.29 \\ 1.38 & -0.29 & 0.06 \end{bmatrix}$$

$$SE(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 15.21 \\ 1.60 \\ 0.25 \end{bmatrix}$$

The rule-of-thumb is that if the estimated coefficient is roughly twice the size of the associated standard error, then it is statistically different from zero at the 95% confidence level. However, this is only true if you have a large dataset. In our case, we only have five observations and more importantly, 2 degrees of freedom. Therefore, we must use look up the critical value from the t-table: 4.303 at the 5% significance level, and 2.92 at the 10% significance level. Therefore, the estimated coefficient must be four times as large as the standard error to conclude statistical significance at the 95% confidence level, and roughly three times as large to conclude statistical significance at the 90% confidence level.

The intercept term, $\hat{\beta}_0$, is the only statistically significant variable, and it is so at the 90% confidence level. The remaining variables are not statistically different from zero.

(c) The adjusted R^2 , AIC, and BIC.

$$\begin{aligned} R^2 &= 1 - \frac{\frac{1}{n-k} \sum_{i=1}^n \hat{\epsilon}_i^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \\ &= 1 - \frac{\frac{1}{5-3} \cdot 9.42}{\frac{1}{5-1} \cdot 130} \\ &= 0.86 \end{aligned}$$

$$\begin{aligned}
AIC &= \log\left(\frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2\right) + \frac{2k}{n} \\
&= \log\left(\frac{1}{5} \cdot 9.42\right) + \frac{2 \cdot 3}{5} \\
&= 1.83
\end{aligned}$$

$$\begin{aligned}
BIC &= \log\left(\frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2\right) + \frac{k}{n} \log(n) \\
&= \log\left(\frac{1}{5} \cdot 9.42\right) + \frac{3}{5} \log(5) \\
&= 1.60
\end{aligned}$$

- (d) If the true model $Weight = f(Age, Income, Parents_Education_Level) + \varepsilon$, do you think the estimated coefficient associated with the income variable is upward or downward biased? Explain the economic intuition behind your answer.

The coefficient associated with the income variable is likely to be biased upward, because there is typically a positive correlation between household income and the education level of parents.

5. Suppose that the random variable Y has the following pdf: $f_Y(y_i) = \theta e^{(-\theta y_i)}$, for $y_i > 0$. Perform the following:
- (a) Solve for the maximum likelihood estimator (MLE). That is, set up the likelihood and log-likelihood function, and determine the parameter $\hat{\theta}$ that maximizes this function.

The likelihood function is as follows:

$$\begin{aligned}
L(\theta|Y) &= f(Y_1|\theta) \cdot f(Y_2|\theta) \cdot f(Y_3|\theta) \cdots f(Y_n|\theta) \\
&= \theta e^{(-\theta y_1)} \cdot \theta e^{(-\theta y_2)} \cdot \theta e^{(-\theta y_3)} \cdots \theta e^{(-\theta y_n)} \\
&= \prod_{i=1}^n \theta e^{(-\theta y_i)}
\end{aligned}$$

Taking the log of the likelihood function yields the following log-likelihood function:

$$\begin{aligned}
LL(\theta|Y) &= \ln \left(\prod_{i=1}^n \theta e^{(-\theta y_i)} \right) \\
&= \sum_{i=1}^n \ln (\theta e^{(-\theta y_i)}) \\
&= \sum_{i=1}^n (\ln \theta - \theta y_i)
\end{aligned}$$

First-order conditions with respect to θ yield:

$$\begin{aligned}
\frac{\partial LL(\theta|Y)}{\partial \theta} &= \left(\sum_{i=1}^n \frac{1}{\theta} - \sum_{i=1}^n y_i = 0 \right) \Big|_{\hat{\theta}} \\
\frac{n}{\hat{\theta}} &= \sum_{i=1}^n y_i \\
\hat{\theta} &= n / \left(\sum_{i=1}^n y_i \right)
\end{aligned}$$

- (b) Suppose you have a sample for $Y : y_i = \{1, 0.8, 1.3, 1.6, 2\}$. What is the MLE estimator given this sample?

Using the estimator from part (a):

$$\begin{aligned}\hat{\theta} &= n / \left(\sum_{i=1}^n y_i \right) \\ &= 5 / 6.7 \\ &= 0.746\end{aligned}$$

6. You have just been hired (to a hypothetical but very high paying consulting position). Your first project is to analyze the economic impacts of pet ownership policies on property values. In the city of Bozeman, if households wish to own more than two dogs, they must purchase a kennel license and abide by the regulations of those licenses. Suppose that the city's mayor (a pet lover) is considering removing this requirement. In a paragraph, explain your strategy for setting up an empirical evaluation of this policy. In completing this task:

- (a) Discuss potential dependent variables and explain why they may be good choices.

Examples include property values of homes in Bozeman, prices of recently sold homes, number of days a house is on the market, etc.

- (b) Outline the potential explanatory variables and why you would want to include them in the model.

Examples include number of kennel licenses issues, number of citations given by animal control, number of schools, age of homes, size of homes, number of parks, demographic characteristics of individuals, etc.

- (c) For each explanatory variable, hypothesize the effect that the variable could have on the dependent variable.
- (d) Explain using economic intuition what the economic effects of the law removal may have on property values.

The effects are likely dependent on individuals' attitude toward dogs. If individuals prefer to have more dogs, then property values would likely rise, as the demand for homes in a pet-friendly city would increase. Alternatively, if an increased number of dogs is perceived as a nuisance and detriment to the community, then the demand for homes and property values would drop.