Econometrics Midterm (Solutions)

Ryan Safner

ECON 480 - Fall 2018

Note that these suggested answers are longer than answers I expect you to provide and answers that would be sufficient for full points. I have simply tried to give you the most complete explanation to understand why the answer is correct, as well as to diagnose and address some of the most common mistakes I saw on these questions.

1. [5 points] Why is OLS called Ordinary Least Squares? No need to calculate anything, but how are the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ chosen?

Solution: OLS estimators are chosen to *minimize* the sum of the squared errors or residuals, $\hat{u}$. In terms of a line between the data points, the parameters of the line (intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$) are chosen to minimize the distance between any observed data point and the line itself. Remember, that distance between the line (predicted value) and the data point (actual value) is the residual or error, $\hat{u} = Y_i - \hat{Y}_i$. We take the square of these residuals (to always ensure positive distance), and try to minimize this number to get a line that best "fits" the data. Formally:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2$$

2. [5 points] There are two sources of randomness in OLS estimates. What is the difference between sampling randomness and modeled randomness?

Solution: Sampling randomness comes from the fact that each OLS regression is estimated from a specific sample of data. If we were to take another sample and run OLS, we would get different estimates (i.e. different $\hat{\beta}_0$ and $\hat{\beta}_1$).

Modeled randomness is our error term, ϵ, which includes all factors affecting Y other than X. We assume the error term is random, by definition. Each sample may have different values for the other unknown factors constituting the error term. Perhaps everyone who was asked in a survey happened to be hungry at the time. This would

lead to different OLS estimates than another sample who happened to not be hungry when asked.

3. [5 points] In your own words, explain what the regression R^2 means. Explain verbally two different ways that it can be calculated. What does it mean to have a low R^2?

Solution: The R^2 is a measure of how well the OLS regression line "fits" our observed data points. It is a measure of the share of the total variation in Y (TSS) that is explained by the variation from our model (ESS), where $R^2 = \frac{ESS}{TSS}$ and $ESS = \sum(\hat{Y}_i - \bar{Y})^2$ and $TSS = \sum(Y_i - \bar{Y})^2$.

Another way to calculate R^2 is by subtracting the fraction of variation in Y that is *not* explained by our model (i.e. the error of our model, SSR). Here, $R^2 = 1 - \frac{SSR}{TSS}$ where $SSR = \sum {\hat{u}_i}^2$

Yet another way is by squaring $corr(X, Y)$.

The closer R^2 is to 1, the better the fit, the closer to 0, the poorer the fit. Low R^2 tells us that there are better models, including more variables, that explain the variation in Y.

4. [5 points] Explain, in your own words, what a p-value is, and how it is used to establish statistical significance.

Solution: The p-value is the probability that, under the null hypothesis (i.e. assuming the null hypothesis were true), we would get a value for our parameter at least as extreme (as far from the mean) as the value our sample found. Another way to interpret this is that the p-value is the probability we commit a Type I error: the probability that, if the null hypothesis were true, we falsely reject it from our sample evidence.

Using our notation, $p = P(T > t)$: that is, the p-value is the probability (big P) that, if the null hypothesis were true, we would obtain a test statistic t at least as extreme (large) as the one we found with our sample.

Be careful, the p-value is *not* the probability that our alternative hypothesis is true given our findings (commonly believed)! In fact it is basically the opposite, the probability of our findings being valid given the null hypothesis!

5. [10 points] Both using formulas and your own words, describe what exogeneity and endogeneity mean, and how they are related to bias. What can we learn about the bias?

Solution:

The OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates of the true population parameters β_0 and β_1 if and only if X is *exogenous*. That is to say, if $corr(X, \epsilon) = 0$ (i.e. there is no correlation between X and any unobserved variable that affects Y), then $E[\hat{\beta}_1] = \beta_1$.

If X *is* correlated with the error term, then X is *endogenous*. The true expected value of the OLS estimator is

$$E[\hat{\beta}_1] = \beta_1 + corr(X, \epsilon)\frac{\sigma_\epsilon}{\sigma_X}$$

The bias is $\big(E[\hat{\beta}_1] - \beta_1\big)$, i.e. the difference between average estimated sample slope and the 'true' population slope, so we can determine first the *size* of the bias based on how large $corr(X, \epsilon)$ is. The stronger the correlation, the larger the bias.

Second, we can determine the *direction* of the bias depending on the sign of $corr(X, \epsilon)$.

- If X and ϵ are positively correlated (move in the same direction), we know that we have *overstated* the true effect of ΔX on ΔY, since a change in Y is picking up both a change in X and a further change (in the same direction as X) in the unobserved ϵ.
- If the correlation is negative (move in opposite directions), we know that we have *understated* the true effect of ΔX on ΔY, since a change in Y is picking up both a change in X that is dampened by a change in the opposite direction of ϵ.

6. [10 points] Suppose a professor wants to estimate the effect of class attendance on grades. The professor randomly selects students and collects data on the number of classes that student attends ($Attendanec$) and that student's final grade ($Grade$). The professor then estimates an OLS model of the form

$$Grade_i = \beta_0 + \beta_1 Attendance_i + \epsilon_i$$

and finds a large positive coefficient on the estimated parameter $\hat{\beta}_1$. Is this an unbiased estimate of the impact of attendance on grades? Why or why not? Do you expect the estimate to overstate or understate the true relationship between $Grade_i$ and $Attendance_i$?

Solution: Knowing something about a student's attendance tells us something about other unobserved characteristics about a student that likely matter for their grades. This implies that the $E(\epsilon|Attendance) \neq 0$ and $corr(Grades, \epsilon) \neq 0$. Students who

attend class are more likely to study more (which increases grade), and be more diligent and conscientious (which increases grade). We might also make arguments for correlations between attendance and ability. In these cases, $corr(Attendance, \epsilon) > 0$. Thus, we are likely to overstate the effect of Attendance on Grades, because all of those other factors also matter. The professor risks a greater probability of a Type I error.

7. [10 points] A discrete random variable X has the following pdf:

x_i	$P(x_i)$
10	0.1
20	0.2
30	0.3
40	0.4

Calculate the standard deviation of X.

Solution: First we must calculate the mean. The mean is:

$$E(X) = \mu_X = \sum X_i p_i$$

$$E(X) = 0.1(10) + 0.2(20) + 0.3(30) + 0.4(40)$$

$$E(X) = 1 + 4 + 9 + 16 = 30$$

Now we must calculate the variance, the average squared distance from the mean of 30 weighted by the probability of each value:

$$var(X) = E[(X_i - \mu_X)^2] = \sum (X_i - \mu_X)^2 p_i$$

$$var(X) = 0.1(10 - 30)^2 + 0.2(20 - 30)^2 + 0.3(30 - 30)^2 + 0.4(40 - 30)^2$$

$$var(X) = 0.1(400) + 0.2(100) + 0.3(0) + 0.4(100)$$

$$var(X) = 40 + 20 + 40 = 100$$

$$sd(X) = \sqrt{var(X)} = \sqrt{100} = 10$$

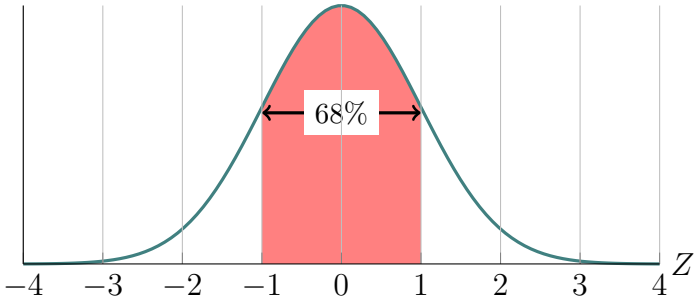
8. [10 points] Suppose exam grades are normally distributed with a mean of 75 and standard deviation of 10. Convert to the standard normal distribution and estimate the following probabilities:

   (a) [4 points] What is the probability of a student's exam grade being between a 65 and an 85?

Solution: Standardize to Z-score:

$$P(65 < Y < 85) = P(\frac{65-75}{10} < \frac{Y-75}{10} < \frac{85-75}{10})$$

$$P(65 < Y < 85) = P(-1 < Z < 1)$$



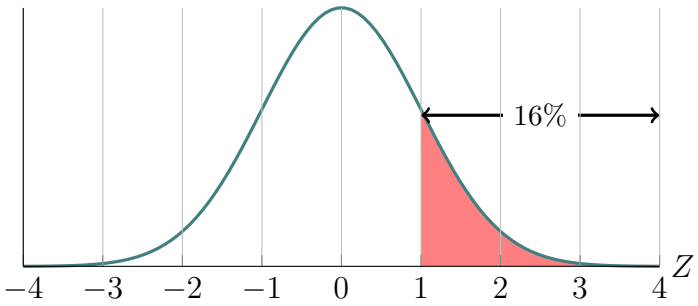
From the 68-95-99.7% rule, **68%** of the probability will fall within 1 standard deviation (1Z) of the mean.

(b) [3 points] What is the probability of a student's exam grade being above an 85?

Solution: Standardize to Z-score:

$$P(Y > 85) = P(\frac{Y-75}{10} > \frac{85-75}{10})$$

$$P(Y > 85) = P(Z > 1)$$



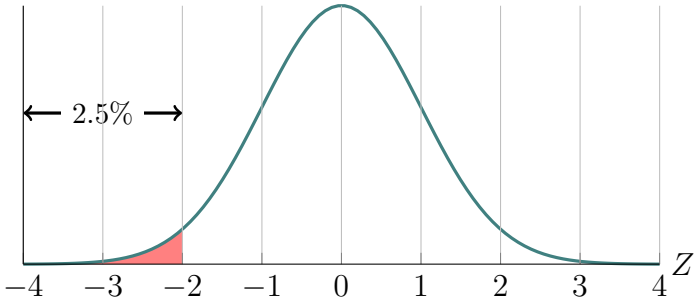
From the 68-95-99.7% rule, 68% of the probability will fall within 1 standard deviation ($\pm 1Z$) of the mean. This means that 32% will fall beyond 1 standard deviation ($\pm 1Z$) of the mean. We only want the right tail, above the Z-score of 1, so this would be half of this, or **16%**.

(c) [3 points] What is the probability of a student's exam grade being lower than a 55?

Solution: Standardize to Z-score:

$$P(55 < Y) = P(\frac{55-75}{10} < \frac{Y-75}{10})$$

$$P(55 < Y) = P(-2 < Z)$$



From the 68-95-99.7% rule, 95% of the probability will fall within 2 standard deviations ($\pm 2Z$) of the mean. Thus, 5% of probability is beyond this area. We want the left-tail, the area to the left of a Z-score of 1, so half of 5% is **2.5%**.

9. [20 points] The results of a sample regression of the number of violent crimes (per 1,000 people) on the number of police officers in a city are reported as follows:

	(1) Crimes
Police	-2.00
	(4.00)
Constant	45***
	(10)
n	100
R^2	0.150
SER	7

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

(a) [1 point] Write out the regression equation (i.e. using the estimates of the coefficients).

Solution:

$$\widehat{\text{Crimes}} = 45 - 2 \times \text{Police}$$

(b) [2 points] What does the estimate for $\hat{\beta}_0$ mean in terms of the regression line? Interpret what the estimate means in context.

Solution: $\hat{\beta}_0$ is the intercept of the regression line, where it crosses the vertical axis. Literally, it means that a town with 0 police officers will have a predicted number of violent crimes equal to $\hat{\beta}_0$: 45.

(c) [2 points] What does the estimate for $\hat{\beta}_1$ mean in terms of the regression line? Interpret what the estimate means in context.

Solution: $\hat{\beta}_1$ is the slope of the regression line. Literally, it means if a town has 1 additional police officer, the predicted number of violent crimes in the town changes by $\hat{\beta}_1$: it decreases by 4 (and vice versa).

(d) [2 points] What does the SER mean in terms of the regression line? What does it mean in context?

Solution: The standard error of the regression is the average size of the residual error, or the average distance of any data point from the regression line. Literally, it means there will be on average a difference in a town's violent crimes of 7 from our model's predicted value of violent crimes for the town.

(e) [2 points] Suppose a small town has 10 police officers. What is the predicted number of violent crimes?

Solution:

$$\widehat{ViolentCrimes}_{smalltown} = 45 - 2(10) = 25$$

(f) [3 points] Suppose that same small town is actually in our data and has had 20 crimes occur. What is the residual for this town? Is this a reasonably good prediction for the current model (and how do you know)?

Solution:

$$\begin{aligned} \hat{\epsilon}_i &= Y_i - \hat{Y}_i \\ \hat{\epsilon_{smalltown}} &= 20 - 25 \\ &= -5 \end{aligned}$$

This is smaller than the average error (measured by the SER) of 7, so it is a better than average prediction.

(g) [2 points] Write down the null and alternate hypotheses for testing whether Police has *any* effect on Crime.

Solution:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

(h) [3 points] Calculate the test-statistic.

Solution:

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sigma_{\hat{\beta}_1}} \\ &= \frac{-2 - 0}{4} \\ &= -\frac{1}{2} \end{aligned}$$

(i) [3 points] If the p-value is 0.618, what can we conclude at the $\alpha = 0.05$ level?

Solution:

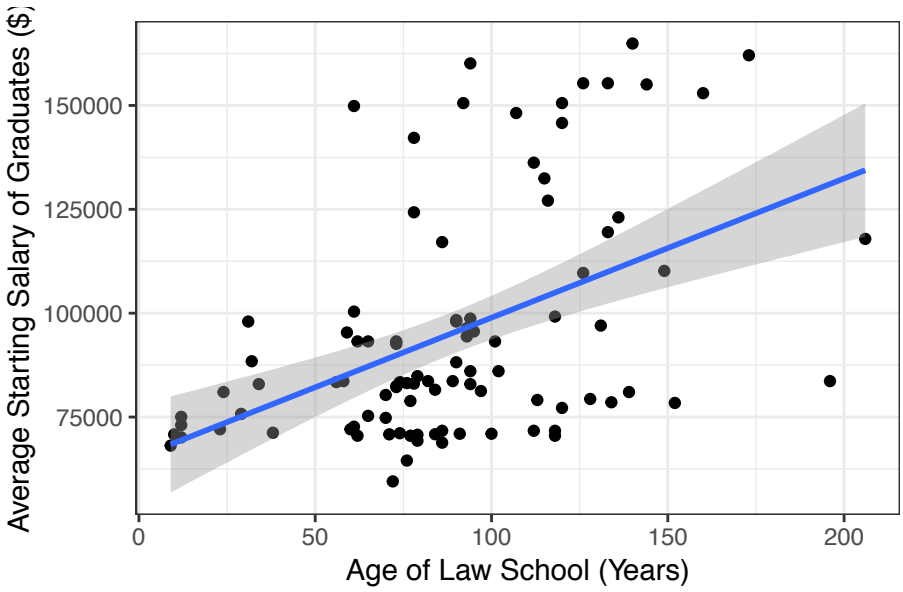
$$2[P(T > t)] = 2[\texttt{tcdf(0.5,1E99, 99)}] = 2[0.309] = 0.618$$

We have insufficient evidence to reject the null hypothesis. According to our sample data, there appears to be no statistically significant relationship between the number of police officers and the amount of violent crimes.

10. [20 points] U.S. News and World Report ranks law schools. Suppose their researchers want to estimate estimate the effect of how old a law school is (in years) on the average starting salary of graduates class in their first legal job (in \$1,000s). They randomly select 95 law schools and create the following scatterplot and graph of their OLS regression line:

(a) [3 points] Are the errors likely to be homoskedastic or heteroskedastic? Explain your answer both in terms of the graph, and the economic intuition why the errors might be this way.

Solution: The errors are likely to be heteroskedastic. The size of the residual errors, the distance from the data points to the regression line, changes for different ages of law schools. For example, the very young schools in the lower left are all quite clustered well around the line, but much older schools are more widely distributed from the line.



We would expect something like this, especially that *all* newer schools would all probably have low earnings potential for their graduates, since the schools have not been around long enough to create a reputation that creates strong demand for their graduates.

Older schools, however, are more likely to vary once their reputation becomes established – Stanford and Golden Gate University School of Law were established about 10 years apart around the turn of the 20th century: Stanford a top 10 national law school, Golden Gate is in the top 10 *worst* law schools in the country.

(b) [2 points] The output for the regression from `R` is below:

```
Call:
lm(formula = salary ~ age, data = lawschool)

Residuals:
   Min     1Q Median     3Q    Max
-47433 -16117  -5761   8550  64019

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 65405.75    6353.98  10.294  < 2e-16 ***
age           335.09      65.51   5.115 1.68e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 24890 on 93 degrees of freedom
```

```
Multiple R-squared:  0.2196,Adjusted R-squared:  0.2112
F-statistic: 26.17 on 1 and 93 DF,  p-value: 1.675e-06
```

Write an equation for the estimated regression line, and write the standard errors of the estimates beneath them in parentheses.

Solution:

$$\widehat{\text{Salary}} = \underset{(6354)}{65406} + \underset{(66)}{335} \times Age$$

(c) [2 points] Interpret the R^2 of this regression.

Solution: 21.96% of the variation in salary is explained by our model using age. This is quite low, indicating other variables likely affect the salary.

(d) [2 points] Interpret the standard error of the regression in this context.

Solution: The average residual error in our model is \$24,894, meaning on average, the actual salaries of graduates is \$23,304 larger or smaller than our predicted salary: which is a pretty big issue.

(e) [2 points] For a law school that is relatively new, 10 years old, what is the average graduate's predicted starting salary?

Solution:

$$\widehat{\text{Salary}} = 65406 + 335(10) = \$68,756$$

(f) [3 points] Suppose that this new law school is in our data, and has an *actual* average starting salary of \$90K. Calculate the residual for this law school. Is this a reasonably good prediction for the current model (and how do you know)?

Solution:

$$\begin{aligned}\hat{\epsilon}_i &= Y_i - \hat{Y}_i \\ \hat{\epsilon_{newschool}} &= 90000 - 68756 \\ &= -21,244\end{aligned}$$

The prediction error is \$21,244 below the actual value. This is just smaller than the average error (the SER) of 24890, so it is a better than average prediction.

(g) [5 points] Is Age exogenous or endogenous? Would we have reasons to believe that we have overestimated or underestimated the effect of age on salary?

Solution: It is likely that age is endogenous, that $corr(\epsilon, Age) \neq 0$ and $E[\epsilon|Age] \neq 0$. Knowing something about the age of the school likely tells you other things about the school that may affect salary – such as the location of the school (the west was settled after the east), whether the law school is independent (more new schools are) or part of a larger university, whether the law school is for-profit (more new schools are) or non-profit, etc. Age is negatively correlated with law schools in the West, independent, and for-profit (all are more recent), and these things also probably negatively correlate with average salary (it is lower for these types of schools). Hence, we actually *under*estimate the effect of age on starting salary!

Again, you can come up with different arguments, possibly even over the bias going in different directions. But at the very least, we can agree Age is probably endogenous.

11. [5 points] **Bonus**: The OLS assumption of exogeneity is sometimes referred to as the "Zero Conditional Mean" assumption. Explain what this means and give an example.

Solution: In general, the standard deviation is a descriptive statistic that measures the average deviation of a variable from its mean:

$$s.d.[X] = \sigma_X = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X - \mu_X)^2}$$

(Note this is for a population standard deviation, for sample standard deviation:

$$s_X = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(X - \bar{X})^2}$$

However, it's when we talk about *sampling distributions*, that is, a distribution of the descriptive statistic itself, like the sample mean, $\bar{X}$. The central limit theorem tells us that for large enough n, the distribution of the sample mean is approximately normal, with mean μ_X and the *standard error*:

$$s.e.[\bar{X}] = \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

Note the distinction between $\sigma_{\bar{X}}$ ("sigma of X bar") and σ_X ("sigma of X"). You need to know the population standard deviation σ_X to find the standard error of the sample mean, $\sigma_{\bar{X}}$. The standard error of the sample statistic ($\bar{X}$ or $\hat{\beta}_1$) describes the average deviation of that sample statistic from its expected value (i.e. the expected

value of the sample statistic if we were to run many samples and calculate the sample statistic each time).

12. [5 points] **Bonus**: A peer of yours, who is a major in another social science (we won't say which!), says he is not interested in linear regression models. Instead he says he gets all he needs to know from estimating correlations. For example, in assessing the effect of Class Size on Test Scores, he claims knowing the correlation coefficient between the two, -0.226, is sufficient to understand the relationship. What response might you have for your peer?

Solution: First of all, the regression R^2 is calculated by squaring the correlation coefficient, so correlation can tell you something about the strength of the relationship (e.g. how much variation in Y is explained by the model).

However, while the correlation coefficient tells you something about the direction and strength of the relationship between two variables, it does *not* inform you about the marginal effect of a one unit increase in the explanatory variable. Hence it cannot answer the question whether or not the relationship is important (although even with the knowledge of the slope coefficient, this requires further information). Correlation also does not allow you to make predictions based on the data, this requires knowing the equation of the regression line. Your friend would not be able to answer the question which policy makers and researchers are typically interested in, such as, what would be the effect on test scores of a reduction in the class size by 1 (or 2, or 10, etc)?