ECN 102, Spring 2020

Final Exam Review
Multiple Choice

Multiple Choice Problem 1

In linear OLS regression, a major problem arises if

(a) important regressors are omitted

(b) unnecessary (or irrelevant) regressors are included

(c) neither a. nor b

(d) both a. and b.

Answer: a. A relevant regressor is one that is correlated with y. If an omitted relevant regressor is correlated with any of the included regressors, then it is a *confounding variable*: estimates suffer from the **omitted variable bias** because $E[u|x] = 0$ is violated, and they are also inconsistent.

If an omitted relevant regressor is not correlated with any of the included regressors, then estimates are unbiased and consistent and all is fine.

An irrelevant regressor is one that is not correlated with y. In this case, including it (or failing to include it) will still give unbiased and consistent estimates. But adding an irrelevant regressor will increase standard errors and make inference less precise. Researchers tend to err on the side of adding an irrelevant regressor than omitting a relevant one.

Multiple Choice Problem 2

Variable x increased by 10 percent. It follows that $\ln(x)$ increased by approximately

(a) $\exp(10)$

(b) $\exp(0.1)$

(c) 10

(d) 0.1

(e) none of the above

Answer: d. Changes in logs are approximately proportional changes. If x_1 increases by 10% to x_2, then

$$\frac{x_2 - x_1}{x_1} = 0.10 \approx \log(x_2) - \log(x_1)$$

Multiple Choice Problem 3

Prices doubled over seven years. It follows that the annual inflation rate is

(a) less than 10 percent

(b) between 10 and 12 percent

(c) between 12 and 14 percent

(d) between 14 and 16 percent

(e) more than 16 percent

Answer: b. The rule of 72 says

$$\text{years to double (7)} \;\approx \frac{72}{\text{percentage growth rate}},$$

which gives us

$$\text{percentage growth rate} = \frac{72}{7} = 10.29\%$$

Multiple Choice Problem 4

Consider $\widehat{\log(y)} = 7 + 5x$. Then the predicted value for y is

(a) $\hat{y} = e^{7+5x}$

(b) $\hat{y} = \log(7 + 5x)$

(c) $\hat{y} = e^{s_e^2/2} e^{7+5x}$

(d) $\hat{y} = \log(s_e^2/2) \log(7 + 5x)$

(e) none of the above

Answer: c. It's tempting to exponentiate both sides, $e^{\widehat{\log(y)}} = e^{7+5x}$. Problem is, $\log(\hat{y}) \neq \widehat{\log(y)}$. So

$$e^{\log(\hat{y})} \neq e^{\widehat{\log(y)}} = e^{7+5x}.$$

Simply taking $\hat{y} = e^{7+5x}$ will lead to a **retransformation bias** in the value of $\hat{y}$.

Have to adjust by $e^{s_e^2/2}$ to correct for the bias: $\hat{y} = e^{s_e^2/2} e^{\widehat{\log(y)}}$.

Bias correction only valid if errors are homoskedastic and normally distributed.

Multiple Choice Problem 5

Let d be an indicator variable for whether female. The regression model $y = \beta_1 + \beta_2 x + \beta_3 d + \beta_4 dx + u$ is one with

(a) different intercept coefficient by gender

(b) different slope coefficient by gender

(c) both a and b

(d) neither a nor b

Answer: c. Suppose we're looking at only men. Then $d = 0$ and the model is $y = \beta_1 + \beta_2 x + u$, so β_1 is the intercept and β_2 is the slope for men.

Now suppose we're looking at only women. Then $d = 1$ and the model is $y = \beta_1 + \beta_2 x + \beta_3 + \beta_4 x + u$. Or better yet,

$$y = (\beta_1 + \beta_3) + (\beta_2 + \beta_4)x + u$$

So $\beta_1 + \beta_3$ is the intercept and $\beta_2 + \beta_4$ is the slope for women.

Multiple Choice Problem 6

We regress $y = \beta_1 + u$. Then β_1 is

(a) 0

(b) 1

(c) the mean of y

(d) none of the above

Answer: c. Regressing y on only a constant gives the mean of y. Intuitively, what is the best guess for y if you don't have any other information? Guess its average value.

Multiple Choice Problem 7

We obtain OLS estimate $\hat{y} = 2 + 5d$ where d is an indicator variable taking values 0 or 1. Then

(a) the mean is $y = 7$ for those observations with $d = 0$

(b) the mean is $y = 2$ for those observations with $d = 1$

(c) both a and b

(d) neither a nor b

Answer: d. When $d = 0$, the regression gives $\hat{y} = 2$, so (a) is no good. When $d = 1$, the regression gives $\hat{y} = 7$, so (b) is no good.

Multiple Choice Problem 8

Let indicator *female* $= 1$ if female and *female* $= 0$ if male; and let indicator *male* $= 0$ if female and *male* $= 1$ if male. You try to run regression $wage = \beta_1 + \beta_2 female + \beta_3 male + u$ to see if there is a difference in average wage of men and women, but OLS fails catastrophically because

(a) there are important omitted variables

(b) male and female form clusters that require cluster-robust standard errors

(c) there is perfect multicollinearity because of the dummy variable trap

(d) none of the above, OLS is totally fine

Answer: c. If a person is female, then $female + male = 1 + 0 = 1$. If a person is male, then $female + male = 0 + 1 = 1$. In all cases, $female + male = 1$. It follows that, say, $female = 1 - male$, or in words, female is a perfect linear function of male. This is a source of **perfect multicollinearity** called the **dummy variable trap**. The solution is to drop one of the dummy variables, say $wage = \beta_1 + \beta_2 female + u$ (or drop the intercept; the former is more common).

Multiple Choice Problem 9

Let indicator *female* $= 1$ if female and *female* $= 0$ if male; and let indicator *male* $= 0$ if female and *male* $= 1$ if male. Consider $wage = \beta_1 + \beta_2 female + u$. The coefficient β_2 captures

(a) the average wage of women

(b) the average wage of men

(c) how much higher of a wage women earn than men, on average

(d) how much higher of a wage men earn than women, on average

(e) none of the above

Answer: c. The regression does not contain *male*, so men are the **reference category** and therefore the included dummy variables are interpreted relative to men. If male, we have $wage = \beta_1 + u$. If female, we have $wage = \beta_1 + \beta_2 + u$. So β_2 is how much more women earn, on average, than men.

Multiple Choice Problem 10

Let indicator $\textit{female} = 1$ if female and $\textit{female} = 0$ if male; and let indicator $\textit{male} = 0$ if female and $\textit{male} = 1$ if male. Consider $\textit{wage} = \beta_1 + \beta_2 \textit{male} + u$. The coefficient β_2 captures

(a) the average wage of women

(b) the average wage of men

(c) how much higher of a wage women earn than men, on average

(d) how much higher of a wage men earn than women, on average

(e) none of the above

Answer: d. Now the regression does not contain *female*, so women are the reference category and therefore the included dummy variables are interpreted relative to women. If male, we have $\textit{wage} = \beta_1 + \beta_2 + u$. If female, we have $\textit{wage} = \beta_1 + u$. So β_2 is how much more men earn, on average, than women. *Interpretation depends on which dummy variable is omitted!*

Multiple Choice Problem 11

Let indicator *female* $= 1$ if female and *female* $= 0$ if male; and let indicator *male* $= 0$ if female and *male* $= 1$ if male. Consider $\textit{wage} = 20 + 2 \times \textit{male} + e$. Then it's also true that

(a) $\textit{wage} = (20 \times \textit{female}) + (22 \times \textit{male}) + e$

(b) $\textit{wage} = (22 \times \textit{female}) + (20 \times \textit{male}) + e$

(c) $\textit{wage} = (2 \times \textit{female}) + (20 \times \textit{male}) + e$

(d) $\textit{wage} = (20 \times \textit{female}) + (2 \times \textit{male}) + e$

(e) none of the above

Answer: a. The regression given has female as the reference category. Therefore it says *wage* $= 20$ on average for women and *wage* $= 22$ on average for men. That's exactly what the regression in (a) says:

$$\begin{aligned} \textit{female} = 1, \textit{male} = 0 \quad &\Longrightarrow \quad \widehat{\textit{wage}} = 20 \\ \textit{female} = 0, \textit{male} = 1 \quad &\Longrightarrow \quad \widehat{\textit{wage}} = 22 \end{aligned}$$

Multiple Choice Problem 12

Suppose we estimate a model with nonlinear regressors and find $\hat{y} = 3 + 2x + x^2 - 5\log(z)$. The marginal effect of a change in x equals

(a) $2x + x^2$

(b) $2 + 2x$

(c) 2

(d) none of the above

Answer: b. The term *marginal effect* is just a fancy way of saying the partial derivative: by how much do we expect y to change when we consider marginally different x?

$$\frac{\partial \hat{y}}{\partial x} = 2 + 2x$$

Multiple Choice Problem 13

Consider $\hat{y} = 3 + 3x^2 + 5x + \sin(z)$ with data points $(1, 1)$, $(2, 4)$, $(4, 6)$. The average marginal effect (AME) of x on y is

(a) $\frac{1}{3}\sum_{i=1}^{3} 6x_i$

(b) $\frac{1}{3}\sum_{i=1}^{3}(3 + 3x_i^2 + 5x_i)$

(c) $\frac{1}{3}\sum_{i=1}^{3}(3 + 3\overline{x}^2 + 5\overline{x})$

(d) $\frac{1}{3}\sum_{i=1}^{3}(6x_i + 5)$

Answer: d. The marginal effect is $\partial\hat{y}/\partial x = 6x + 5$. Sum this up over all $n = 3$ observations and we get

$$\underbrace{(6 \times 1 + 5)}_{11} + \underbrace{(6 \times 2 + 5)}_{17} + \underbrace{(6 \times 4 + 5)}_{29} = 57.$$

Therefore the average marginal effect is $57/3 = 19$.

Multiple Choice Problem 14

Consider $\hat{y} = 3 + 3x^2 + 5x$ with data points $(1, 1)$, $(2, 4)$, $(4, 6)$. The marginal effect at the mean (MEM) is

(a) $6\overline{x} + 5$

(b) $\frac{1}{3}\sum_{i=1}^{3}(6x_i + 5)$

(c) $\frac{1}{3}\sum_{i=1}^{3}(3 + 3x_i^2 + 5x_i)$

(d) $3 + 3\overline{x}^2 + 5\overline{x}$

Answer: definitely a, and also b by coincidence. The marginal effect is $\partial\hat{y}/\partial x = 6x + 5$. The mean of x is $(1 + 2 + 4)/3 = 7/3$. Therefore the marginal effect (evaluated) at the mean is

$$\text{MEM} = 6 \times \frac{7}{3} + 5 = 19.$$

When quadratic, AME = MEM. But not always the case in generality.

Multiple Choice Problem 15

You are regressing cross-sectional data and suspect there is heteroskedasticity, but errors are still independent. Then you use

(a) default standard errors

(b) heteroskedasticity-robust standard errors

(c) cluster-robust standard errors

(d) HAC standard errors

Answer: b. In other words, use heteroskedasticity-robust standard errors when assumption 3 alone fails. In practice, heteroskedasticity is the rule, homoskedasticity the exception.

Multiple Choice Problem 16

You are regressing time series data and you suspect autocorrelation. You should use

(a) default standard errors

(b) heteroskedasticity-robust standard errors

(c) cluster-robust standard errors

(d) heteroskedasticity and autocorrelation-consistent (HAC) standard errors

Answer: d. Autocorrelation is a fancy way of saying that a variable is correlated with its past values, and therefore error terms are also correlated with their past values. (Think GDP being persistently above trend during an expansion and persistently below trend during a recession.) HAC standard errors accounts for the presence of error dependence as well as heteroskedasticity.

Multiple Choice Problem 17

You are regressing panel data consisting of individuals from different cities as panel members. You should use

(a) default standard errors

(b) heteroskedasticity-robust standard errors

(c) cluster-robust standard errors

(d) HAC standard errors

Answer: c. Errors for individuals in the same city are likely correlated, so each city forms its own cluster. Using cluster-robust standard errors accounts for the presence of error dependence within a cluster.

Multiple Choice Problem 18

The F-statistic distribution for test of overall significance,

$$F \equiv \frac{\text{ESS}/(k-1)}{\text{RSS}/(n-k)} \sim F(k-1, n-k),$$

is valid when

(a) Assumptions 1-2 hold and n is sufficiently large

(b) Assumptions 1-3 hold and n is sufficiently large

(c) Assumptions 1-4 hold and n is sufficiently large

(d) None of the above

Answer: c. The distribution is approximate unless errors are normally distributed, in which case the distribution is exact. In practice, assumption 3 breaks down and we have to use a heteroskedasticity-robust F-statistic instead. It is not easy to calculate, so we simply use the Stata command `test x z` to test for joint significance of x and z, which gives a p-value.

Multiple Choice Problem 19

The main lesson from regression analysis of school scores on the California Academic Performance Index is that

(a) by far the biggest determinant is teacher quality

(b) by far the biggest determinant is educational attainment of parents

(c) by far the biggest determinant is student disadvantage (English learner, free meals)

(d) all of a, b, and c are substantial determinants

Answer: b. Regressing the Academic Performance Index (API) on parent educational attainment alone gives an adjusted R-squared of 0.834. Doing a multiple regression that accounts for whether the student is poor (i.e. qualifies for free meals); speaks English fluently; attends a year-round school; learns from fully-credentialed teachers; or learns from emergency-credentials teachers (i.e. substitute teachers made full time) increases adjusted R-squared to 0.852. The additional regressors are jointly statistically significant ($p = 0.000$), so they add something. But not a whole lot.

Multiple Choice Problem 20

You test $H_0 : \beta_2 \leq 0$ against $H_a : \beta_2 > 0$ and find a p-value of 0.03. You conclude that

(a) β_2 is statistically significant at 5%

(b) β_2 is statistically insignificant at 5%

(c) there is not enough information to make any conclusion about whether β_2 is statistically significant or not

Answer: b. Okay, so we're comparing two different tests:

$$H_0 : \beta_2 \leq 0 \text{ against } H_a : \beta_2 > 0$$
$$H_0 : \beta_2 = 0 \text{ against } H_a : \beta_2 \neq 0 \qquad \text{(test of statistical significance)}$$

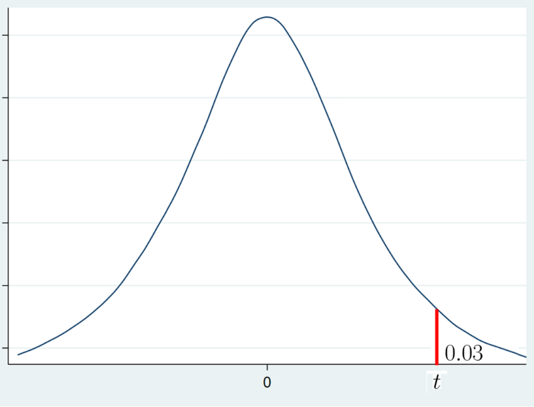
The first thing to note is that you get the same t-statistic for both tests,

$$t = \frac{b_2 - 0}{\text{se}(b_2)}$$

Given p-value says: if null is true, then there is a 3% probability of observing a t-statistic at least as large in the positive direction as the one we've observed.

Multiple Choice Problem 20

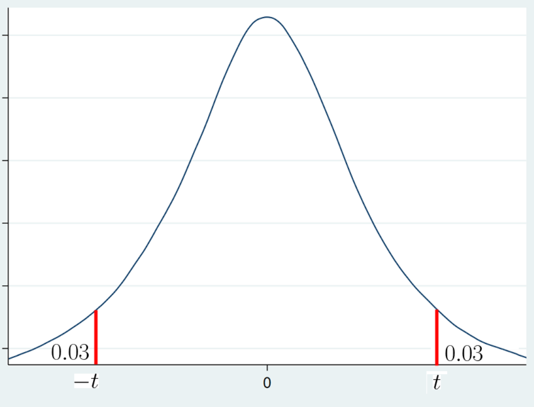
Answer: b. p-value here says: if the null is true, then there is a 3% probability of observing a t-statistic at least as large in the positive direction as the one we've observed.



Two-sided p-value says: if the null is true, then there is a 3% probability of observing a t-statistic at least as large in both positive and negative direction as the one we've observed.

Multiple Choice Problem 20

Answer: b. Two-sided *p*-value says: if the null is true, then there is a 3% probability of observing a *t*-statistic at least as large in both positive and negative direction as the one we've observed.



Therefore the *p*-value for the two-sided test is 0.06 and we fail to reject $H_0 : \beta_2 = 0$ at 5% significance. Conclude β_2 is statistically insignificant at 5%.

Multiple Choice Problem 21

You regress y on x with $n = 1000$ observations and get an F-statistic of $F(1, 998) = 36$. You conclude that

(a) $H_0 : \beta_2 = 0$ should be rejected at any conventional significance level

(b) $H_0 : \beta_2 = 0$ should not be rejected at any conventional significance level

(c) there is not enough information to make any conclusion about $H_0 : \beta_2 = 0$

Answer: a. When testing a single regressor, the $F(1, n - k)$ statistic is exactly the t-statistic squared. Therefore

$$t^2 = F(1, 998) = 36 \quad \Longrightarrow \quad t = \sqrt{36}$$

We don't know whether t is positive or negative six, but by now you should recognize that ± 6 is a very large t-statistic that will warrant rejection for any conventional significance level when doing a two-sided test.

Multiple Choice Problem 22

Using a regression to predict the actual outcome of an individual is

(a) less precise than using the regression to predict the average outcome for all similar individuals

(b) more precise than using the regression to predict the average outcome for all similar individuals

(c) equally precise as using the regression to predict the average outcome for all similar individuals

(d) not enough information

Answer: a. The key is OLS assumption 2, $E[u|x = x^*] = 0$. This says that the prediction is correct *on average* because the regression has zero error *on average* when we consider individuals with $x = x^*$. Therefore the **conditional mean** prediction is relatively precise because the errors average out.

But just because the regression has zero error *on average* doesn't mean it will have zero error for a specific individual. That additional source of uncertainty makes it more difficult to make a **forecast** about an individual.

Homework 5, Problem 5 Correction

I plugged in the wrong number and could have simplified more, so there's been some confusion here. My bad.

- P is the principal (initial level of infected), r is daily the rate of growth of infected, t how many days have passed
- Level is initially 75, so $P = 75$
- After t periods of growth, level is $A = Pe^{rt}$
- Doubles in three days? $150 = 75e^{r \times 3}$
- Gotta find infection rate r

Homework 5, Problem 5 Correction

- Write $150 = 75e^{r\times 3}$ as $2 = e^{3r}$
- Take logs of both sides

$$\log(2) = \log\left(e^{3r}\right) = 3r\log(e) = 3r$$

- Solve: $r = \frac{\log(2)}{3} \approx 0.2310$

Homework 5, Problem 5 Correction

- Therefore we can write either $A = 75e^{0.2310t}$, which I did
- Or to be more specific and avoid some rounding

$$A = 75e^{\frac{\log(2)}{3}t} = 75\left(e^{\log(2)}\right)^{t/3}$$

$$= 75 \times 2^{t/3}$$

which used is in the solutions.

- Here's where I made the mistake. The data set starts with day 1. So that means in day 39, 38 days have passed. So we plug in $t = 38$. I plugged in $t = 39$. Oops.
- $A = 75e^{0.2310 \times 38} = 486,741$
- $A = 75 \times 2^{38/3} = 487,650$