Econometrics Final Exam

ECON 480 – Safner

Fall 2018

Name: $\underbrace{\hspace{6cm}}_{\text{I pledge that I have neither given nor received unauthorized aid on this exam}}$

Please read all of the following information and wait until authorized to start the exam:

- You have **2 hours** to complete this exam.
- This test consists of **9** required questions worth **100** points (25% of final course grade), and up to 2 bonus questions, worth additional points.
- Please read all instructions and question prompts carefully.
- Describing your thought process will give you a better chance to earn partial credit for wrong answers.
- You can ace this test without use of a calculator, but you can use a calculator. **Do not use your phone**.
- Good luck!

1. [5 points] What are the conditions that would cause a regression to suffer from omitted variable bias?

2. [5 points] Suppose we wish to determine if movies that make more money at the Box Office tend to be illegally downloaded more. Imagine we were able to track (weekly) the number of illegal downloads and the box office performance of all movies released in 2018, such that each individual observation (mw) is movie m in week w. We estimate the following two-way fixed effects model:

$$\text{Illegal downloads}_{mw} = \beta_0 + \beta_1 \text{Box Office Performance}_{mw} + \alpha_m + \theta_w + \epsilon_{mw}$$

Explain (i) what sorts of factors are absorbed in α_m, (ii) what sorts of factors are absorbed in θ_w, and (iii) if we were to include another X_{mw} variable in the regression, what must be true about it.

3. [5 points] Why are we as econometricians not necessarily trying to maximize R^2 (or even $\bar{R^2}$) for our regressions?

4. [5 points] When would we run an F-test? How does the F-test tell us what we want to know? No need to write down a formula, but explain intuitively what exactly F is measuring.

5. [10 points] For *each* of the following scenarios, under the following model

$$\hat{Y}_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

explain in a sentence what happens to:

i) the *unbiasedness* of OLS estimates and

ii) the *variance* (or standard error) of OLS estimates:

a. [2 points] A variable Z is omitted that is correlated with both X_{1i}, X_{2i}, and Y_i

b. [2 points] X_{1i} and X_{2i} are multicollinear (but not perfectly so)

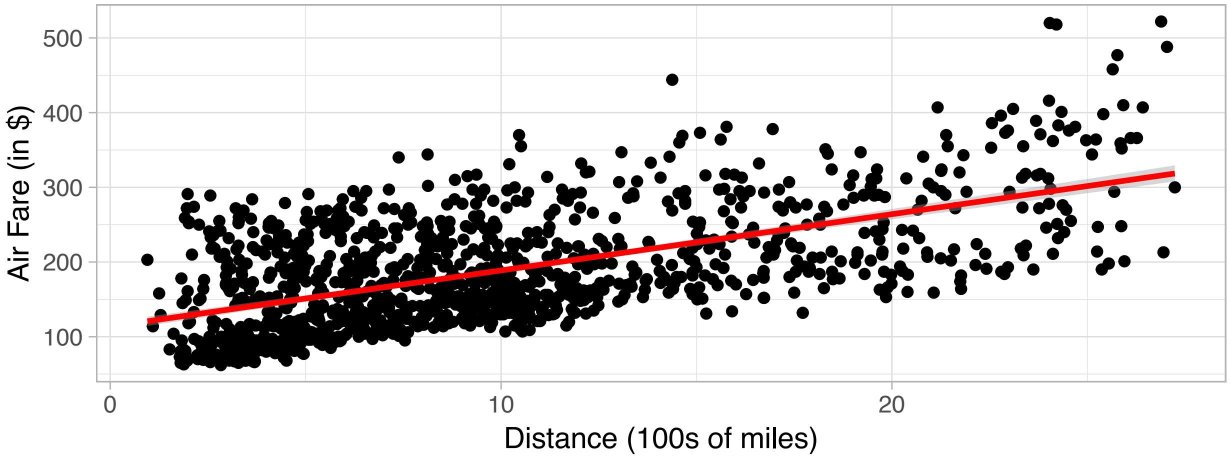
c. [2 points] We drop 50% our observations

d. [2 points] There are significant outliers of X_{1i} and X_{2i}

e. [2 points] Residuals are heteroskedastic.

6. [10 points] Suppose you wanted to estimate the impact of legalizing marijuana on crime rates. You collect panel data on crime rates in Colorado and neighboring Nebraska each year for the 2010-2018 period. Since 2014, marijuana is legal in Colorado and remains illegal Nebraska. Describe how you could isolate the effect of legalizing marijuana on crime rates. Be specific: fully describe the model estimation strategy, intuition, and write out the OLS regression equation to estimate.

7. [20 points] A researcher for a travel blog wants to estimate how the distance of a particular flight route affects its price. She collects data on 1,149 flights, and plots a scatterplot below showing the relationship between the the distance of the route (in hundreds of miles) and the airfare (in $).



a. [4 points] As a baseline regression, she runs a regression of airfare `fare` on distance in hundreds of miles (`dist.100s`).

```
##
## Call:
## lm(formula = fare ~ dist.100s, data = fares)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -114.89  -48.23  -16.64   42.86  225.56
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 113.5473     3.4530   32.88   <2e-16 ***
## dist.100s     7.5248     0.2968   25.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.54 on 1147 degrees of freedom
## Multiple R-squared:  0.3592, Adjusted R-squared:  0.3586
## F-statistic: 642.9 on 1 and 1147 DF,  p-value: < 2.2e-16
```

(i) Interpret the coefficient on distance. (ii) How much would we expect a ticket for a flight going 1,000 miles to cost?

b. [3 points] She then wants to check if there are any nonlinearities in the relationship. She generates a variable `dist.sq` which is the distance^2, and includes it in the regression reported below. Is there evidence that she should include `dist.sq`? Why or why not?

```
##
## Call:
## lm(formula = fare ~ dist.100s + dist.sq, data = fares)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -126.60  -48.84  -16.19   40.85  226.58
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 125.43103    6.13612  20.441  < 2e-16 ***
## dist.100s     4.90708    1.15689   4.242  2.4e-05 ***
## dist.sq       0.10359    0.04426   2.341   0.0194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.42 on 1146 degrees of freedom
## Multiple R-squared:  0.3622, Adjusted R-squared:  0.3611
## F-statistic: 325.5 on 2 and 1146 DF,  p-value: < 2.2e-16
```

c. [4 points] Use the regression in (b) to find the marginal effect of distance on fare. What is the marginal effect for a distance of 500 miles? 50,000 miles?

d. [3 points] She further generates a variable `dist.cube` which is the distance^3, and includes it in the regression reported below. Is there evidence that she should use a cubic model? Why or why not?

```
##
## Call:
## lm(formula = fare ~ dist.100s + dist.sq + dist.cube, data = fares)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -145.17  -48.36  -15.00   40.12  230.78
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 103.153608  10.092538  10.221  < 2e-16 ***
## dist.100s    13.035271   3.147591   4.141 3.71e-05 ***
## dist.sq      -0.650588   0.275293  -2.363   0.0183 *
## dist.cube     0.019221   0.006925   2.775   0.0056 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.24 on 1145 degrees of freedom
## Multiple R-squared:  0.3665, Adjusted R-squared:  0.3648
## F-statistic: 220.8 on 3 and 1145 DF,  p-value: < 2.2e-16
```

e. [3 points] She then runs an F-test on `dist.sq` and `dist.cube`. What do the results of the F-test tell us?

```
## Linear hypothesis test
##
## Hypothesis:
## dist.sq = 0
## dist.cube = 0
##
## Model 1: restricted model
## Model 2: fare ~ dist + dist.sq + dist.cube
##
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1   1147 4343764
## 2   1145 4294206  2     49558 6.6071 0.001403 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

f. [3 points] She then tries another specification, and converts *both* fare and distance into natural logarithms. Interpret the coefficient on `l.dist`.

```
##
## Call:
## lm(formula = l.fare ~ l.dist, data = fares)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71756 -0.26507 -0.04334  0.26349  1.03965
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.68465    0.10208   26.30   <2e-16 ***
## l.dist       0.36856    0.01517   24.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.339 on 1147 degrees of freedom
## Multiple R-squared:  0.3397, Adjusted R-squared:  0.3391
## F-statistic: 590.2 on 1 and 1147 DF,  p-value: < 2.2e-16
```

8. [20 points] Does taking econometrics help your career prospects? Below is a regression table from a sample of 50 students and a description of the variables included in the various regressions.

Variable	Description
Salary	Annual starting salary at first job (in $)
ln(salary)	The natural logarithm of Salary
Metrics	Dummy variable (=1 if student took econometrics, =0 if student did not)
GPA	Grade point average (4.0 scale)
Female	Dummy variable (=1 if student is female, =0 if male)

	(1)	(2)	(3)	(4)
	Salary	Salary	Salary	ln(Salary)
Metrics	4,736***	5,033***	5,024***	0.162***
	(466.3)	(413.5)	(414.8)	(0.0132)
GPA		1,643***	1,658***	0.0553***
		(364.8)	(368.1)	(0.0121)
Female			-204.8	-0.00649
			(400.4)	(0.0134)
Constant	29,105***	24,200***	24,242***	10.11***
	(308.6)	(1,024)	(1,027)	(0.0341)
n	50	50	50	50
R^2	0.615	0.737	0.738	0.728

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

a. [3 points] Interpret the estimate on Metrics for the baseline (1^{st}) regression.

b. [3 points] For regression (2), the estimate on Metrics has changed. Do you think GPA the absence of GPA caused significant omitted variable bias in regression (1)? Note that the correlation between Metrics and GPA is -0.1395.

c. [3 points] For regression (3), interpret the coefficient on Female. Is this statistically significant?

d. [2 points] Using regression (3), what is the predicted starting salary for a female with a GPA of 3.5 who has taken econometrics?

e. [2 points] Using regression (3), interpret $\hat{\beta}_0$ in the context of a particular student's starting salary.

f. [2 points] Regression (4) is a log-linear model. Interpret the estimate on Metrics, and the estimate on GPA.

g. [3 points] Why not include another dummy variable,"No Metrics" $= \begin{cases} 1 \text{ if student } i \text{ has NOT taken econometrics} \\ 0 \text{ if person } i \text{ has taken econometrics} \end{cases}$

into the regression?

9. [20 points] The results of a regression of $\texttt{Happiness}_i$ (self-reported on a scale of 1 (least happy) to 10 (most happy) on $\texttt{Salary}$ (in \$10,000s) and $\texttt{Married}_i = \begin{cases} 1 \text{ if person } i \text{ is married} \\ 0 \text{ if person } i \text{ is unmarried} \end{cases}$ surveying 500 random individuals are reported as follows:

$$\widehat{\text{Happiness}}_i = \underset{(0.20)}{1.50} + \underset{(0.10)}{0.50} \text{ Salary}_i + \underset{(0.50)}{2.00} \text{ Married}_i + \underset{(0.20)}{1.00} \left(\text{Salary}_i * \text{Married}_i\right)$$

a. [4 points] Explain what each coefficient (i.e. $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$) means in the context of this regression.

b. [2 points] Write out (and simplify if applicable) the estimated regression equation for *unmarried* individuals.

c. [2 points] Write out (and simplify if applicable) the estimated regression equation for *married* individuals.

d. [6 points] Do these two regressions from parts (c) and (d) have the same *intercepts*? Do they have the same *slopes*? Test the answer to each of these questions statistically.

e. [6 points] *Roughly sketch* the two regression lines from (c) and (d) on the same graph below. Label slopes, intercepts, and equations of each line by using the coefficients from part (a). You can use just the abstract "betas" rather than the exact numbers for labels, and the lines need not be fully accurate or drawn to scale.