Final Exam (135 possible points)
ECNS 561
Fall 2015

Name_____

1.) Consider the regression model

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

for $i = 1, \ldots, n$.

a.) **(5 pts)** Specify the least squares function that is minimized by OLS.

b.) **(5 pts)** Compute the partial derivatives of the objection function with respect to β_1 and β_2.

c.) **(10 pts)** Suppose $\sum_{i=1}^{n} X_{1i} X_{2i} = 0$. Show that $\widehat{\beta_1} = \frac{\sum_{i=1}^{n} X_{1i} Y_i}{\sum_{i=1}^{n} X_{1i}^2}$.

d.) **(10 pts)** Suppose $\sum_{i=1}^{n} X_{1i}X_{2i} \neq 0$. Derive an expression for $\widehat{\beta_1}$ as a function of the data (Y_i, X_{1i}, X_{2i}), $i = 1, \ldots, n$.

2.) **(10 pts)** Consider the simple regression model under the first four Gauss-Markov assumptions

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

Using $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$, show that plim$\widehat{\beta_0} = \beta_0$. To show this, you will need to use the consistency of $\widehat{\beta_1}$, invoke the law of large numbers, and use the fact that $\beta_0 = E(y) - \beta_1 E(x_1)$.

3.) Consider the following population regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where $E[\varepsilon|X] = 0$. Suppose we observe a sample of size n: $\{Y_i, X_i\}_{i=1}^{n}$.

Define an estimator of β_1, b_1, as:

$$b_1 = \frac{Y_2 - Y_1}{X_2 - X_1}$$ (i.e., b_1 is simply based on the 1st and 2nd observation of (Y, X)).

a.) **(10 pts)** Is b_1 an unbiased estimator of β_1? Support your answer by mathematically showing whether or not b_1 is unbiased.

b.) **(10 pts)** Is b_1 a consistent estimator of β_1? Support your answer by mathematically showing whether or not b_1 is consistent.

4.) **(5 pts)** You have obtained data on test scores and student-teacher ratios in region A and region B of your state. Region B, on average, has lower student-teacher ratios than region A. Your boss tells you to run the following regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

where X_1 is the class size in region A, X_2 is the difference in class size between region A and B, and X_3 is the class size in region B. What would you tell your boss about this proposed regression equation he has asked you to estimate? Would you have any suggestions on how to improve upon the equation above? (Note: Be short and precise in your answer. Take up no more than the space below.)

5.) **(7 pts)** As we have discussed in class, a commonly run regression is the log-log model,

$$\ln(y) = \beta_0 + \sum_k \beta_k \ln(x_k) + \varepsilon.$$

Using calculus, show that β_k measures the percentage change in y associated with a one percent change in x_k.

6.) **(8 pts)** In some countries (e.g., China), there exist abnormal sex ratios driven by cultural and policy-derived incentives. Prior research suggests a u-shaped relationship between the level of violence in a country and the sex ratio. Suppose you have data on levels of violence and sex ratios for a large sample of countries. To test for a u-shaped relationship, write down the regression you would estimate and the signs on the coefficients you would expect.

7.) This question is adapted from the following journal article:

Clay, Karen, Werner Troesken, and Michael Haines. 2014. "Lead and Mortality." *Review of Economics and Statistics*, 96(3): 458-470.

Lead is a well-known environmental toxin whose adverse effects include infant mortality, morbidity, loss of IQ, and violence. Clay et al. use city-level data from 1900 to examine the effect of waterborne lead exposure on infant mortality. Within their empirical strategy, identification of the effect of lead comes from variation in water chemistry across cities with different types of water service pipes. In particular, they estimate the following regression:

$$(1) \quad mort_i = \alpha_0 + \alpha_1 mixed\_lead_i + \alpha_2 lead\_only_i + \alpha_3 \ln(pH_i - 5.675) + \alpha_4 mixed\_lead_i * \ln(pH_i - 5.675) + \alpha_5 lead\_only_i * \ln(pH_i - 5.675) + \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{S_i}\boldsymbol{\delta} + \varepsilon_i,$$

where $mort_i$ is the infant mortality rate per 100 population in city i in 1900. The indicator variable $mixed\_lead_i$ is equal to one if city i has a mixture of lead and non-lead pipes (and is equal to zero otherwise), while the indicator $lead\_only_i$ is equal to one if city i has lead-only pipes (and is equal to zero otherwise). The omitted category is having no lead pipes. The authors also control for the level of acidity in the water. More acidic water leached lead faster. Because the effects of acidity are expected to differ across cities with no-lead pipes, mixed pipes, and lead pipes, $\ln(pH_i - 5.675)$ is interacted with the lead dummies and is included as a control. The authors use $\ln(pH_i - 5.675)$ to ensure the main effects are evaluated within the range of the sample pH. Specifically, $\ln(pH_i - 5.675)$ is equal to zero at a pH of 6.675, which is the 25th percentile of pH. The authors also control for a vector of city-level characteristics, $\boldsymbol{X_i}$, and a vector of state climatic characteristics, $\boldsymbol{S_i}$. They estimate equation (1) with ordinary least squares.

Note: Just in case you have forgotten high school chemistry, low pH levels are associated with higher acidity. For example, battery acid has a pH = 0 and toothpaste has a pH = 9.

a.) Before estimating equation (1) above, the authors report results from a separate OLS regression where they explore the determinants of a city having lead pipes. Some of their results are as follows:

Determinants of a City Having Lead Pipes in 1900	
	Dependent variable: *lead_only*
Independent variables	
Citypop Q2	.221** (.108)
Citypop Q3	.191* (.110)
Citypop Q4	.410*** (.104)
Precipitation	.082 (.051)
Temperature	-.179*** (.056)
Typhoid	-.011 (.038)
N	172
R^2	.135
* Statistically significant at 10% level; ** at 5% level; *** at 1% level. Notes: *Citypop* are indicator variables for population quartiles, which run from the largest (Q4) to smallest (Q1). *Precipitation* is the state precipitation in inches. *Temperature* is the state temperature in Farenheit. *Typhoid* is the city's typhoid mortality rate per 100 population and serves as a measure of water quality.	

a1.) **(5 pts)** Interpret the coefficient estimate on *Citypop Q2*.

a2.) **(5 pts)** Given these findings, do you have concerns with giving any results based on estimation of equation (1) a causal interpretation? Use only the space below for your answer…be precise and to the point.

b.) The results from the estimation of equation (1) are as follows:

Effects of Lead and pH on Infant Mortality		
	(1)	(2)
	Dependent variable: *mort*	Dependent variable: *mort*
Independent variables		
$mixed\_lead_i$	.043 (.035)	.022 (.031)
$lead\_only_i$	.074*** (.028)	.068*** (.025)
$\ln(pH_i - 5.675)$	-.049 (.033)	-.015 (.026)
$mixed\_lead_i * \ln(pH_i - 5.675)$	-.029 (.039)	.002 (.033)
$lead\_only_i * \ln(pH_i - 5.675)$	-.075** (.037)	-.060** (.027)
Mean infant mortality rate for the sample of cities	.396	.396
N	172	172
R^2	.304	.552
City- and state-level controls	No	Yes

* Statistically significant at 10% level; ** at 5% level; *** at 1% level.

Notes: The city-level controls consist of the city population quartile dummies, fraction foreign born, fraction white, and fraction of women ages 20 to 40. The state-level controls consist of state temperature and precipitation.

b1.) **(5 pts)** Are the statistically significant results of the expected sign? Provide a brief explanation.

b2.) **(5 pts)** At a pH of 6.675 and using the results from column (1), interpret the effect of a city having lead-only pipes. Put the size of this effect into context and describe whether or not you think it is meaningfully large.

b3.) **(5 pts)** The difference between column (1) and column (2) is that the results from the latter are based on a specification that includes city- and state-level controls. Does adding the controls improve the fit of the regression? Are you more or less worried about the potential threat of omitted variable bias when comparing results across these two models?

c.) Towards the end of the paper, the authors propose using non-infant mortality rates (i.e., mortality of all individuals over the age of 1) as a falsification test. That is, they hypothesize that lead pipes should have no effect on non-infant mortality rates.

c1.) **(5 pts)** Suppose they were to find that lead pipes appear to decrease non-infant mortality rates (they actually don't, but suppose they do). Would this call their results based on the estimation of equation (1) into question? Why?

c2.) **(5 pts)** Can you think of another falsification outcome that would be preferred over non-infant mortality rates? Explain why.

8.) This question is based on the paper that I asked you to read entitled "Education and Criminal Behavior: Insights from an Expansion of Upper Secondary School" by Aslund et al. (2015).

a.) **(10 points)** Briefly describe the endogeneity problem the authors address by exploiting the expansion of upper secondary school in Sweden. In doing so, also describe the intuition underlying their natural experiment. How does it help them determine whether or not a causal relationship exists between education and crime?

b.) **(10 points)** Previous studies have exploited compulsory schooling laws to gauge the effect of education on crime. Why is the parameter of interest in Aslund et al. (2015) not necessarily comparable to the parameter of interest in compulsory schooling law studies? To answer this question effectively, you need to think about who exactly is being treated under both types of reform.