Final Exam
ECNS 561
Fall 2017
120 total points possible

Name___________________________

1.) (25 points total) In matrix notation, suppose we have the following population model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{X}$ is an n by k matrix of k independent variables for n observations, $\boldsymbol{y}$ is an n by 1 vector of observations on the dependent variable, $\boldsymbol{\beta}$ is a k by 1 vector of unknown population parameters that we wish to estimate, and $\boldsymbol{\varepsilon}$ is an n by 1 vector of errors.

a.) (8 points) Using matrix notation, solve for the OLS estimator (i.e., the $\widehat{\boldsymbol{\beta}}$ that minimizes the sum of squared residuals).

b.) (6 points) Show that $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

c.) (8 points) Now, instead of the model above, suppose the true population model is given as

(1) $$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

However, suppose you only have data to estimate the following model

(2) $$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u} \qquad \text{where } \mathbf{u} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

Using matrix notation, show that the OLS estimator for equation (2) is biased.

d.) (3 points) Consider the OLS estimator for equation (2) in part b.) above. Under what two scenarios will the bias in this estimator be eliminated? Be short and brief in your answer.

2.) (10 points total) Consider the regression model under the first five Gauss-Markov assumptions:

$$y_i = \beta x_i + \varepsilon_i.$$

Let $\bar{\beta}$ denote an estimator of β that is constructed as $\bar{\beta} = \frac{\bar{y}}{\bar{x}}$, where $\bar{y}$ and $\bar{x}$ are the sample means of y_i and x_i, respectively.

a.) (4 points) Show that $\bar{\beta}$ is a linear function of $y_{1,}\ y_{2,\ \ldots,}\ y_{n.}$

b.) (6 points) Show that $\bar{\beta}$ is unbiased.

3.) (10 points) Consider the following regression model under the first five Gauss-Markov assumptions

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Using summation notation, show that the OLS estimator $\widehat{\beta_1}$ is a consistent estimator of β_1. Make sure to state any necessary assumptions required in your proof.

4.) (20 points total) Suppose you are interested in estimating the relationship between education and crime at the individual level. For your research, you also collect data on an individual's income and cognitive ability (measured by an IQ-type test). Let $\widetilde{\beta_1}$ be the estimate from the following regression

$$Crime_i = \beta_0 + \beta_1 Education_i + \varepsilon_i.$$

And, let $\widehat{\beta_1}$ = be the estimate from the following regression

$$Crime_i = \beta_0 + \beta_1 Education_i + \beta_2 Income_i + \beta_3 Cognitive_i + \varepsilon_i.$$

a.) (5 points) If *Education* is highly correlated with *Income* and *Cognitive* in the sample, and *Income* and *Cognitive* have large partial effects on *Crime,* would you expect $\widetilde{\beta_1}$ and $\widehat{\beta_1}$ to be similar or different? Explain.

b.) (5 points) If *Education* is almost uncorrelated with *Income* and *Cognitive*, but *Income* and *Cognitive* are highly correlated, will $\widetilde{\beta_1}$ and $\widehat{\beta_1}$ tend to be similar or different? Explain.

c.) (5 points) If *Education* is highly correlated with *Income* and *Cognitive,* and *Income* and *Cognitive* have small partial effects on *Crime,* would you expect se($\widetilde{\beta_1}$) or se($\widehat{\beta_1}$) to be smaller? Explain.

d.) (5 points) If *Education* is almost uncorrelated with *Income* and *Cognitive, Income* and *Cognitive* have large partial effects on *Crime,* and *Income* and *Cognitive* are highly correlated, would you expect se($\widetilde{\beta_1}$) or se($\widehat{\beta_1}$) to be smaller? Explain.

5.) (10 points total) Instead of the individual-level data considered in the problem above, suppose you have county-level data and estimate the following regression:

$$Crime_c = \beta_0 + \beta_1 Education_c + \varepsilon_c,$$

where c index counties. The dependent variable *Crime* is the crime rate per 1,000 population and the independent variable *Education* is the average number of years of education in the county.

Suppose you use OLS to estimate the above equation and you obtain the following coefficient estimates (with standard errors in parentheses):

$$\widehat{\beta_0} = 30.0\ (12.0)$$
$$\widehat{\beta_1} = -1.50\ (0.25)$$

a.) (5 points) Given these coefficient estimates, plot the OLS regression line below. Make sure to label your graph and axis intercept points.

b.) (5 points) Suppose raw data points for three MT counties are as follows:

County	*Crime*	*Education*
Fergus	8.0	12.0
Rosebud	20.0	14.0
Yellowstone	25.0	8.0

In the graph above, indicate the residuals for these counties and provide the values for each.

6.) (35 points total) Levitt (2008) used data from the Fatality Analysis Reporting System (FARS) to study the effectiveness of child car safety seats relative to standard seat belts. He found that child safety seats, in actual practice, do not provide discernable improvement over adult seat belts in reducing traffic fatalities among children ages 2 through 6. Building on Levitt (2008), Anderson and Sandholt (forthcoming) used data from the FARS for the period 2008-2016 to study the effectiveness of booster seats relative to child safety seats and adult seat belts for children ages 2 through 9. In particular, they estimated the following OLS regression:

$$(1) \qquad \mathit{Fatality}_{ivct} = \beta_0 + \beta_1 \mathit{Booster\ Seat}_{ivct} + \beta_2 \mathit{Child\ Seat}_{ivct} + \beta_3 \mathit{Seat\ Belt\ Only}_{ivct} + \boldsymbol{X1}_{ivct}\boldsymbol{\beta}_5 + \boldsymbol{X2}_{vct}\boldsymbol{\beta}_6 + \boldsymbol{X3}_{ct}\boldsymbol{\beta}_7 + \varepsilon_{ivct},$$

where $\mathit{Fatality}_{ivct}$ is equal to one if child i in vehicle v, crash c, and year t died, and is equal to zero otherwise. The variables $\mathit{Booster\ Seat}_{ivct}$, $\mathit{Child\ Seat}_{ivct}$, and $\mathit{Seat\ Belt\ Only}_{ivct}$ are mutually exclusive dummy variables for the type of restraint device used. The vectors $\boldsymbol{X1}_{ivct}$, $\boldsymbol{X2}_{vct}$, and $\boldsymbol{X3}_{ct}$ include controls for individual-, vehicle-, and crash-level characteristics, respectively. The table below lists the controls included in $\boldsymbol{X1}_{ivct}$, $\boldsymbol{X2}_{vct}$, and $\boldsymbol{X3}_{ct}$.

Definitions of controls included in $\boldsymbol{X1}_{ivct}$, $\boldsymbol{X2}_{vct}$, and $\boldsymbol{X3}_{ct}$

Individual-level characteristics ($X1_{ivct}$)	
Child seat position	
Front	= 1 if child was sitting in a front seat of the vehicle, = 0 otherwise
Back	= 1 if child was sitting in a back seat of the vehicle, = 0 otherwise
Male	= 1 if male, = 0 otherwise
Age	Age of child in years
Vehicle-level characteristics ($X2_{vct}$)	
Car	= 1 if vehicle was a car, = 0 otherwise
Light Truck	= 1 if vehicle was a light truck, = 0 otherwise
Model Year ≤ 1990	= 1 if vehicle model year was pre 1991, = 0 otherwise
1990 < Model Year ≤ 2000	= 1 if vehicle model year was between 1991 and 2000, = 0 otherwise
Model Year > 2000	= 1 if vehicle model year was post 2000, = 0 otherwise
Vehicle Weight (1,000s lbs.)	Vehicle weight in thousands of pounds
Point of impact	
Non-Collision	= 1 if initial point of contact was classified as "non-collision", = 0 otherwise
Front	= 1 if initial point of contact was at the front of the vehicle, = 0 otherwise
Rear	= 1 if initial point of contact was at the rear of the vehicle, = 0 otherwise
Other Contact Point	= 1 if initial point of contact was at an "other" contact point, = 0 otherwise
Driver Unbelted	= 1 if driver was unbelted, = 0 otherwise
Driver Uninjured	= 1 if driver was uninjured, = 0 otherwise
Driver Minor Injury	= 1 if driver suffered a minor injury, = 0 otherwise
Driver Major/Fatal Injury	= 1 if driver suffered a major injury or died, = 0 otherwise
Driver Past Accident	= 1 if driver was in a previous accident in the past 3 years, = 0 otherwise
Driver Past Violation	= 1 if driver was charged with a driving violation in the past 3 years, = 0 otherwise
Crash-level characteristics ($X3_{ct}$)	
Persons in Crash	Number of persons involved in the crash
Cars in Crash	Number of cars involved in the crash
Speed Limit < 55 MPH	= 1 if speed limit was less than 55 mph, = 0 otherwise
Rural Road	= 1 if crash was on a rural road, = 0 otherwise
Early Morning	= 1 if crash occurred during the early morning hours (1:00 a.m. to 5:59 a.m.), = 0 otherwise
Daytime	= 1 if crash occurred during the daytime hours (6:00 a.m. to 7:59 p.m.), = 0 otherwise
Evening	= 1 if crash occurred during the evening hours (8:00 p.m. to 12:59 a.m.), = 0 otherwise
Weekend	= 1 if crash occurred during the weekend (Friday, 6:00 p.m. to Monday, 5:59 a.m.), = 0 otherwise

a.) (5 points) Given that the choice of restraint type is not random and determined by parents, which of the controls listed above do you think serve as the best proxies for unobserved parental preferences and attitudes? Briefly justify your answer.

b.) (5 points) In addition to the covariates listed above, the authors also considered specifications where they controlled for state fixed effects (i.e., a vector of state dummy variables). What do state fixed effects accomplish? That is, what source of omitted variable bias is controlled for when including state fixed effects in the regression above?

c.) (5 points) Given that booster seats, child safety seats, and seat belts exhaust the potential restraint type choices, how do we interpret β_1, β_2, and β_3 in equation (1)?

d.) (5 points) The below table represents estimates based on equation (1). The first column illustrates results for the entire sample, while the second and third columns illustrate results for children ages 2-5 and 6-9, respectively.

Booster Seats and the Probability of Fatality

	Full sample (Ages 2-9)	Ages 2-5	Ages 6-9
Booster Seat	-.068***	-.035	-.090***
	(.020)	(.026)	(.026)
Child Seat	-.073***	-.053**	-.080***
	(.019)	(.024)	(.030)
Seat Belt Only	-.081***	-.060***	-.094***
	(.018)	(.022)	(.025)
Mean rate of death for unrestrained children	.117	.112	.121
Hypothesis tests (p-values)			
Booster Seat = Child Seat	.651	.247	.626
Booster Seat = Seat Belt Only	.171	.104	.713
Child Seat = Seat Belt Only	.337	.502	.429
N	4,484	1,814	2,670

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

Notes: Each column represents results from a separate OLS regression based on data from the Fatality Analysis Reporting System (2008-2016). The dependent variable is equal to one if the child died in the accident, and is equal to zero otherwise. The models also control for $\boldsymbol{X1}_{ivct}$, $\boldsymbol{X2}_{vct}$, and $\boldsymbol{X3}_{ct}$. Standard errors are in parentheses.

Are these results, in general, economically meaningful? Provide a brief explanation as to why you think these estimates are economically meaningful or not.

e.) (5 points) Based on the regression results, which restraint device appears to be the most effective at reducing traffic fatalities among children? Make sure to briefly justify your answer.

f.) (5 points) Previous research suggests that sitting in a back seat may reduce the risk of death among child passengers. How would you modify equation (1) to explore whether booster seats, child safety seats, or seat belts are more effective for child passengers riding in a back seat as opposed to the front?

g.) (5 points) Because the FARS data represent information provided by the reporting police officers at the scene of the crash, it is possible that measurement error in the type of child restraint device used exists. That is, it is possible that reporting officers at the scene of the crash do not always correctly record the type of restraint device used by child passengers (e.g., an officer may incorrectly record a booster seat as a child safety seat).

i.) What assumption do we require for this type of measurement error to lead to conservative estimates of our coefficients of interest?

ii.) How might one modify equation (1) to allow for this type of measurement error?

References

Anderson, D. Mark and Sina Sandholt. "Are Booster Seats More Effective than Child Safety Seats or Seat Belts at Reducing Traffic Fatalities among Children?" Forthcoming at *American Journal of Health Economics*

Levitt, Steven. 2008. "Evidence that Seat Belts are As Effective as Child Safety Seats in Preventing Death for Children Ages Two and Up." *Review of Economics and Statistics,* 90(1): 158-163.

7.) (10 points) In Dr. Urban's lecture, using county-level panel data, she presented a body of work that looks at the effect of the local unemployment rate on voter turnout. Specifically, consider the following model, where V_{ct} is county-level voter turnout for each year, u_{ct} is the unemployment rate in county *c* for year *t*, and δ_c and γ_t are county and year fixed effects, respectively:

$$V_c = \beta_0 + \beta_1 u_{ct} + \delta_c + \gamma_t + \epsilon_{ct}$$

She discussed several threats to internal validity in determining causality. Does measurement error in the independent variable affect the estimates of β_1? If so, explain how it could bias the estimate.