Student: ____________________ **Date:** ____________________	**Instructor:** Richeng Piao **Course:** ECON 2560 - Applied Econometrics	**Assignment:** Practice Problem Set 9

1. What is the difference between *internal validity* and *external validity*?

○ **A.** A statistical analysis is said to have *internal validity* if the statistical inferences about causal effects can only be verified by a few res

○ **B.** A statistical analysis is said to have *internal validity* if the statistical inferences about causal effects are valid for the population being

○ **C.** A statistical analysis is said to have *external validity* if the statistical inferences about causal effects are valid for the population bein

○ **D.** *Internal validity* and *external validity* are equivalent.

What is the difference between the *population studied* and the *population of interest*?

○ **A.** The *population of interest* is the population from which the sample was drawn, while the *population studied* is the population to whic

○ **B.** The *population studied* is the population from which the sample was drawn, while the *population of interest* is the population to whic

○ **C.** The *population studied* is mandated by the government, while the *population of interest* is chosen freely.

○ **D.** The *population studied* and *the population of interest* are always equivalent.

Answers B.

A statistical analysis is said to have *internal validity* if the statistical inferences about causal effects are valid for the population being studied. The analysis is said to have *external validity* if conclusions can be generalized to other populations and settings.

B.

The *population studied* is the population from which the sample was drawn, while the *population of interest* is the population to which causal inferences from this study are to be applied.

ID: Review Concept 9.1

---

2. Suppose that you have just read a careful statistical study of the effect of advertising on the demand for cigarettes. Using data from New York during the 1970s, the study concluded that advertising on buses and subways was more effective than print advertising. Use the concept of external validity to determine if these results are likely to apply to Boston in the 1970s; Los Angeles in the 1970s; New York in 2010.

○ **A.** The results are likely to apply to New York in 2010, but not to Boston in the 1970s or to Los Angeles in the 1970s.

○ **B.** The results are likely to apply to Los Angeles in the 1970s, but not to Boston in the 1970s or to New York in 2010.

○ **C.** The results are likely to apply to Boston in the 1970s, but not to Los Angeles in the 1970s or New York in 2010.

○ **D.** The results are likely to apply to any city in the United States regardless of location and year.

Answer: C. The results are likely to apply to Boston in the 1970s, but not to Los Angeles in the 1970s or New York in 2010.

ID: Exercise 9.1

---

3. Are the following statements true or false?

An ordinary least squares regression of Y onto X will be internally inconsistent if X is correlated with the error term.

○ **A.** True.

○ **B.** False.

Each of the five primary threats to internal validity implies that X is correlated with the error term.

○ **A.** True.

○ **B.** False.

Answers A. True.

A. True.

ID: Exercise 9.7

---

4. A statistical analysis is internally valid if:

○ **A.** the population is small, say less than 2,000, and can be observed.

○ **B.** the statistical inferences about causal effects are valid for the population studied.

○ **C.** all t-statistics are greater than $|1.96|$.

○ **D.** the regression $R^2 > 0.05$.

Answer: B. the statistical inferences about causal effects are valid for the population studied.

ID: Test A Ex 9.1.1

---

5. Threats to internal validity lead to:

○ **A.** a false generalization to the population of interest.

○ **B.** the inability to transfer data sets into your statistical package.

○ **C.** failures of one or more of the least squares assumptions.

○ **D.** perfect multicollinearity.

Answer: C. failures of one or more of the least squares assumptions.

ID: Test A Ex 9.1.2

---

6. Comparing the California test scores to test scores in Massachusetts is appropriate for external validity if:

○ **A.** the two income distributions were very similar.

○ **B.** Massachusetts also allowed beach walking to be an appropriate P.E. activity.

○ **C.** the institutional settings in California and Massachusetts, such as organization in classroom instruction and curriculum, were similai

○ **D.** the student-to-teacher ratio did not differ by more than five on average.

Answer: C.
the institutional settings in California and Massachusetts, such as organization in classroom instruction and curriculum, were similar in the two states.

ID: Test B Ex 9.1.1

---

7. The question of reliability/unreliability of a multiple regression depends on:

○ **A.** internal but not external validity.

○ **B.** the quality of your statistical software package.

○ **C.** external but not internal validity.

○ **D.** internal and external validity.

Answer: D. internal and external validity.

ID: Test B Ex 9.1.2

---

8. Internal validity is that:

○ **A.** the estimator of the causal effect should be unbiased and consistent.

○ **B.** inferences and conclusions can be generalized from the population to other populations.

○ **C.** the estimator of the causal effect should be efficient.

○ **D.** OLS estimation has been used in your statistical package.

Answer: A. the estimator of the causal effect should be unbiased and consistent.

ID: Test B Ex 9.1.3

---

9. The true causal effect might not be the same in the population studied and the population of interest because:

○ **A.** of geographical differences.

○ **B.** of differences in characteristics of the population.

○ **C.** the study is out of date.

○ **D.** all of the above.

Answer: D. all of the above.

ID: Test B Ex 9.1.4

---

10. Consider the following population regression function:

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

where β_0 and β_1 denote the intercept and the slope coefficient on X, respectively. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the predicted intercept and the predicted slope coefficient, respectively.

Which of the following are the requirements for ensuring that there is no threat to internal validity in this case? (*Check all that apply.*)

□ **A.** There is no difference between the settings studied and the settings of interest.

□ **B.** There is no difference between the population studied and the population of interest.

□ **C.** The OLS estimator, $\hat{\beta}_1$, is an unbiased and consistent estimator of the population coefficient, β_1.

□ **D.** The confidence interval $\hat{\beta}_1 \pm 1.96 SE(\hat{\beta}_1)$ has a 95% probability of containing the population coefficient β_1.

Results of any campaign to curb traffic violators in California, where the average fine for running a red signal is $395,

(1) ____________ be generalised to an identical population of traffic violators in Nevada, where the fine for running a red signal is $450.

Assume that the difference between $395 and $450 is statistically significant.

If a study can generalize the behavioural statistics of prison gangs in the Allenwood penitentiary in Pennsylvania, to that of prison gangs in Canaan and Lewisburg penitentiaries in Pennsylvania, assuming the populations are identical, the statistical analysis can be said to have (2) ____________.

(1) ○ can
○ cannot

(2) ○ external validity
○ internal validity

Answers C. The OLS estimator, $\hat{\beta}_1$, is an unbiased and consistent estimator of the population coefficient, β_1., D. The confidence interval $\hat{\beta}_1 \pm 1.96 SE(\hat{\beta}_1)$ has a 95% probability of containing the population coefficient β_1.

(1) cannot

(2) external validity

ID: Concept Exercise 9.1.1

11. What is the trade-off when including an extra variable in a regression?

○ **A.** Including an extra variable always makes estimated coefficients more significant, but it makes the results much harder to interpret.

○ **B.** Including an extra variable could make estimated coefficients more significant, but it always decreases the regression R^2.

○ **C.** Including an extra variable always makes estimated coefficients more significant, but it also introduces multicollinearity.

○ **D.** An extra variable could control for omitted variable bias, but it also increases the variance of other estimated coefficients.

Answer: D. An extra variable could control for omitted variable bias, but it also increases the variance of other estimated coefficients.

ID: Review Concept 9.2

12. Consider the following regression model

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

Suppose that Y is measured with random error. Does this mean that regression analysis is unreliable?

○ **A.** Yes.

○ **B.** No.

Now, suppose that X is measured with random error. Does this mean that regression analysis is unreliable?

○ **A.** Yes.

○ **B.** No.

Answers B. No.

A. Yes.

ID: Review Concept 9.3

---

13. Suppose that a state offered voluntary standardized tests to all its third graders and that these data were used in a study of class size on student performance. Which of the following would generate selection bias?

○ **A.** Scores on the standardized test are usually recorded with error.

○ **B.** Schools are unable to control for unobservable student ability.

○ **C.** Class size is simultaneously determined with student performance.

○ **D.** Schools with higher-achieving students could be more likely to volunteer to take the test.

Answer: D. Schools with higher-achieving students could be more likely to volunteer to take the test.

ID: Review Concept 9.4

---

14. A researcher estimates the effect on crime rates of spending on police by using city-level data. Which of the following represents simultaneous causality?

○ **A.** The researcher only has data on cities with both high spending on police and low crime rate.

○ **B.** More spending on police causes the crime rate to decrease. However, there is large measurement error in spending reports, which

○ **C.** There is unobserved officer quality, which is correlated with more spending and lower crime rates.

○ **D.** Cities with high crime rates may need a larger police force, and thus more spending. More police spending, in turn, reduces crime.

Answer: D.

Cities with high crime rates may need a larger police force, and thus more spending. More police spending, in turn, reduces crime.

ID: Review Concept 9.5

---

15. A researcher estimates a regression using two different software packages. The first uses the homoskedasticity-only formula for standard errors. The second uses the heteroskedasticity-robust formula. The standard errors are very different. Which should the researcher use?

○ **A.** The homoskedasticity-only only standard errors should be used.

○ **B.** In this case, both the homoskedasticity-only and the heteroskedasticity-robust are equivalent.

○ **C.** The heteroskedasticity-robust standard errors should be used.

○ **D.** The researcher would need more information to answer this question.

Answer: C. The heteroskedasticity-robust standard errors should be used.

ID: Review Concept 9.6

16. Labor economists studying the determinants of women's earnings discovered a puzzling empirical result. Using randomly selected employed women, they regressed earnings on the women's number of children and a set of control variables (age, education, occupation, and so forth). They found that women with more children had higher wages, controlling for these other factors. What is most likely causing this result?

○ **A.** Measurement error bias.

○ **B.** Simultaneous causality between women's earnings and women's number of children.

○ **C.** Sample selection bias.

○ **D.** Omitted variable bias.

Answer: C. Sample selection bias.

ID: Exercise 9.3

17. Using the OLS Estimates shown in the table below, compare the estimated effects of a 17% increase in average district income on test scores in California and Massachusetts.

Complete the table below. (*Round your response to two decimal places*)

Multiple Regression Estimates of the Student-Teacher Ratio and Test Scores: Data from California and Massachusetts. Dependent variable: average test score in K-8 school districts in California (1); and average combined English, math, and science test score in elementary school districts in Massachusetts (2).

			Estimated Effect of a 17% Increase in Average District Income	
	OLS Estimates	**Standard Deviation of Test Scores Across Districts**	**Points on the Test**	**Standard Deviations**
California (1)				
Average district income (logs)	11.55** (1.83)	19.2		
Massachusetts (2)				
Average district income (logs)	16.87** (3.11)	15.5		

These regressions were estimated using the *n* = 2380 observations in the Boston HMDA data set. The California Testing and Reporting data set uses data from all 420 K-6 and K-8 districts in California with data available for 1999. Individual coefficients are statistically significant at the *5% or **1% level.

Answers 1.96

0.10

2.87

0.19

ID: Exercise 9.4

18. Suppose that $n = 369$ i.i.d. observations for $\left(Y_i, X_i\right)$ yield the following regression results:

$$\begin{array}{l} \hat{Y} = 30.58 + 66.35X,\ SER = 15.88,\ R^2 = 0.76 \\ \quad\ (15.4) \quad\ (12.9) \end{array}$$

Another researcher is interested in the same regression, but he makes an error when he enters the data into the regression: He enters each observation twice, so he has 738 observations (with observation 1 entered twice, observation 2 entered twice, and so forth).

Which of the following estimated parameters change as result? (*Check all that apply*)

□ **A.** The R^2 of the regression.

□ **B.** The standard error of the regression (*SER*).

□ **C.** The estimated intercept and slope.

□ **D.** The standard errors of the estimated coefficients.

Using the 738 observations, what results will be produced by his regression program?

$$\begin{array}{l} \hat{Y} = 30.58 + 66.35X,\ SER = \square,\ R^2 = 0.76 \\ (\square)\ (\square) \end{array}$$

(*Round your responses to two decimal places*)

Which (if any) of the internal validity conditions are violated?

○ **A.** Sample selection.

○ **B.** Simultaneous causality.

○ **C.** Omitted variables.

○ **D.** Measurement error.

Answers B. The standard error of the regression (*SER*)., D. The standard errors of the estimated coefficients.

15.86

10.87

9.11

D. Measurement error.

ID: Exercise 9.6

---

19. A survey of earnings contains an unusually high fraction of individuals who state their weekly earnings in 100s, such as 300, 400, 500, etc.

This is an example of:

○ **A.** errors-in-variables bias.

○ **B.** sample selection bias.

○ **C.** companies that typically bargain with workers in 100s of dollars

○ **D.** simultaneous causality bias.

Answer: A. errors-in-variables bias.

ID: Test A Ex 9.2.3

---

20. In the case of errors-in-variables bias:

○ **A.** the OLS estimator is consistent if the variance in the unobservable variable is relatively large compared to the variance in the mea

○ **B.** binary variables should not be used as independent variables.

○ **C.** the OLS estimator is consistent but no longer unbiased in small samples.

○ **D.** maximum likelihood estimation must be used.

Answer: A.
the OLS estimator is consistent if the variance in the unobservable variable is relatively large compared to the variance in the measurement error.

ID: Test A Ex 9.2.4

21. In the case of errors-in-variables bias, the precise size and direction of the bias depend on:

○ **A.** the sample size in general.

○ **B.** the size of the regression R^2.

○ **C.** the correlation between the measured variable and the measurement error.

○ **D.** whether the good in question is price elastic.

Answer: C. the correlation between the measured variable and the measurement error.

ID: Test A Ex 9.2.5

22. In the case of a simple regression, where the independent variable is measured with i.i.d. error:

○ **A.** $\hat{\beta}_1 \xrightarrow{p} \frac{\sigma_w^2}{\sigma_X^2 + \sigma_w^2} \beta_1.$

○ **B.** $\hat{\beta}_1 \xrightarrow{p} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}.$

○ **C.** $\hat{\beta}_1 \xrightarrow{p} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1.$

○ **D.** $\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}.$

Answer:
C. $\hat{\beta}_1 \xrightarrow{p} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1.$

ID: Test B Ex 9.2.5

23.

Omitted variable bias (1) ____________ a threat to internal validity.

Consider the following population regression equation:

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

where β_0 and β_1 denote the intercept and slope coefficient, respectively. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the estimated intercept and the estimated slope coefficient, respectively.

A measurement error in X_i (2) ____________ result in an errors-in-variables bias.

Suppose that simultaneous equations bias arises in the following set of equations.

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

$$X_i = \gamma_0 + \gamma_1 Y_i + v_i.$$

Which of the following statements are correct? (*Check all that apply.*)

□ **A.** u_i is correlated with v_i

□ **B.** Y_i is correlated with v_i

□ **C.** γ_1 is correlated with β_1

□ **D.** X_i is correlated with u_i

(1) ○ is not
○ is

(2) ○ will not
○ will

Answers (1) is

(2) will

B. Y_i is correlated with v_i, D. X_i is correlated with u_i

ID: Concept Exercise 9.2.1

24. Suppose we have a regression equation of the form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_n X_{ni} + u_i, \ i = 1...n$$

where $\beta_0, \beta_1, ..., \beta_n$ are the population coefficients. Let $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_n$ denote the respective estimated coefficients.

When does missing data pose a threat to internal validity?

○ **A.** Internal validity is threatened when the data are missing because of a selection process that is related to Y_i beyond depending on

○ **B.** Internal validity is threatened when the data are missing because of a selection process that is related to any u_i beyond depending

○ **C.** Internal validity is threatened when the data are missing because of a selection process that is related to Y_i beyond depending on

○ **D.** Internal validity is threatened when data are missing completely at random.

Which of the following statements is not an implication of the regression error being correlated across observations?

○ **A.** The OLS estimators become biased and inconsistent.

○ **B.** The external validity of the analysis is threatened.

○ **C.** The heteroskedasticity-robust OLS standard errors are correct but the homoskedasticity-only OLS standard errors are incorrect.

○ **D.** The regressors are not independently and identically distributed (i.i.d.) draws from their joint distribution.

Answers A.

Internal validity is threatened when the data are missing because of a selection process that is related to Y_i beyond depending on X_i.

A. The OLS estimators become biased and inconsistent.

ID: Concept Exercise 9.2.2

25. A student wants to study whether there is any relationship between the amount of calories (C) consumed by the cattle and their weight (Y). He selects a random sample of 150 cattle and asks their caretakers about their calorie intake.

Consider the following regression function:

$$Y_i = \beta_0 + \beta_1 \widetilde{C}_i + \left[\beta_1 \left(C_i - \widetilde{C}_i\right)\right] + u_i,$$

$$= \beta_0 + \beta_1 \widetilde{C}_i + v_i,$$

where $v_i = \beta_1 \left(C_i - \widetilde{C}_i\right) + u_i$ and β_0 and β_1 denote the intercept and the slope coefficient on $\widetilde{C}_i$, respectively.

The population regression equation written in terms of $\widetilde{C}_i$ has an error term that contains the measurement error, the difference between C_i and $\widetilde{C}_i$. Suppose the student discovers that the measured value, $\widetilde{C}_i$, equals the actual, unmeasured value, C_i, plus a purely random component, w_i.

Suppose the variance of C_i (σ_C^2) and the variance of w_i (σ_w^2) are 0.07 and 0.03, respectively.

Then, under the classical measurement error model, as the sample size increases, the value to which the slope estimator will converge to is □ β_1.

(*Round your answer to two decimal places.*)

In this case, when the value of the slope estimator converges in probability to 0, it means that there is (1) ______ available regarding the relationship between calories consumed by the cattle and their weight.

And, when the value of the slope estimator converges in probability to 1, it means that there is (2) ______ available regarding the relationship between calories consumed by the cattle and their weight.

Consider the following regression function:

$$Y_i = \beta_0 + \beta_1 \widetilde{X}_i + v_i,$$

where $v_i = \beta_1 \left(X_i - \widetilde{X}_i\right) + u_i$ and β_0 and β_1 denote the intercept, and the slope coefficient on $\widetilde{X}_i$, respectively.

Suppose the measured value, $\widetilde{X}_i$, equals the actual, unmeasured value, X_i, plus a purely random component, w_i. Let $\hat{\beta}_1$ and σ_w^2 denote the predicted value of the slope coefficient β_1 and the variance of w_i, respectively.

Which of the following statements regarding the value of the measurement error are true? (*Check all that apply.*)

□ **A.** When the value of the measurement error is very large, $\hat{\beta}_1$ converges in probability to β_1.

□ **B.** When the value of the measurement error is very large, $\hat{\beta}_1$ converges in probability to 0.

□ **C.** When there is no measurement error, $\hat{\beta}_1$ converges in probability to β_1.

□ **D.** When there is no measurement error, $\hat{\beta}_1$ converges in probability to 0.

(1) ○ little useful information
○ no useful information
○ full information

(2) ○ full useful information
○ little information
○ no useful information

Answers 0.70

(1) no useful information

(2) full useful information

B. When the value of the measurement error is very large, $\hat{\beta}_1$ converges in probability to 0., C. When there is no measurement error, $\hat{\beta}_1$ converges in probability to β_1.

ID: Concept Exercise 9.2.3

---

26. A researcher is interested in estimating the relationship between the vital capacity (*VC*, the maximum amount of air a person can expel from the lungs after a maximum inhalation, measured in cm^3) and the age of a person. She collects data from a random sample of 300 individuals (i.i.d) and estimates the following regression equation:

$$\widehat{VC} = \underset{(39.67)}{3{,}210.45} + \underset{(2.56)}{1.87}Age,\ SER = 24.14,\ R^2 = 0.71,$$

where the standard errors are given in parentheses.

Another researcher is interested in the same regression, but she makes an error when she enters the data into her regression program: she enters each observation thrice, so she has 900 observations (with each observation entered thrice).

The standard error of regression for the second researcher's regression equation will be □.

(*Round your answers to two decimal places*.)

In the presence of a measurement error, the values of R^2, the estimated intercept, and the slope coefficient (1) ____________.

(1) ○ triple
○ increase
○ remain constant
○ decrease

Answers 24.09

(1) remain constant

ID: Concept Exercise 9.2.4

---

27. Which of the following statements are accurate for when regression models are used for forecasting? (*Check all that apply.*)

- □ **A.** When regression models are used for forecasting, concerns about unbiased estimation of causal effects are important.
- □ **B.** When regression models are used for forecasting, concerns about external validity are important.
- □ **C.** When regression models are used for forecasting, concerns about unbiased estimation of causal effects are not important.
- □ **D.** When regression models are used for forecasting, concerns about external validity are not important.

Consider the following regression equation:

$$\widehat{\textit{Crime rate}} = \hat{\beta}_0 - \hat{\beta}_1 \times \textit{(No.of police stations in the district)},$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated intercept and slope coefficients, respectively. Assume that the estimated regression is not stable, it has a good explanatory power, and $\hat{\beta}_1$ is not estimated precisely.

Suppose that a married couple wants to assess the safety of a neighbourhood in a particular district they are considering shifting to. The above equation (1) ____________ reliable in forecasting the safety of a neighborhood in the said district.

(1) ○ is not
○ is

Answers B. When regression models are used for forecasting, concerns about external validity are important., C.
When regression models are used for forecasting, concerns about unbiased estimation of causal effects are not important.

(1) is not

ID: Concept Exercise 9.3.1