Midterm 3

answers

Instructions: You may use a calculator, and one sheet of notes. You will never be penalized for showing work, but if what is asked for can be computed directly, points awarded will depend primarily on the correctness of your numerical answer. Each question and subquestion is worth 4 points. As there are 21 total questions (including subquestions), the exam is out of 84 points, so you get 16 points for free. Good luck!

Problem 1 A regression analysis between sales (in thousands of dollars) and advertising (in hundreds of dollars) resulted in the following least squares line: $Sales = 75 + 6 * Advertising$. This implies that if advertising is \$800, then the predicted amount of sales (in dollars) is:

a. \$4,875

B. \$123,000

c. \$487,500

d. \$12,300

Problem 2 In linear regression analysis, we often perform a two-tail test of the population slope parameter β_1 to determine whether there is sufficient evidence to infer that a linear relationship exists. The null hypothesis is stated as:

A. $H_0 : \beta_1 = 0$

b. $H_0 : \beta_1 = X$

c. $H_0 : \beta_1 \neq 0$

d. None of these choices.

Problem 3 In single-variable regression analysis, if the coefficient of determination (R^2) is .975, then which of the following is true regarding the estimated slope of the regression line, ($\hat{\beta}_1$)?

a. All we can tell is that it must be positive.

b. It must be .975.

c. It must be .987.

D. We cannot tell the sign or the value.

Problem 4 If the coefficient of correlation between X and Y is close to 1.0, this indicates that:

a. Y causes X to happen.

b. X causes Y to happen.

c. Both a and b.

D. There may or may not be a causal relationship between X and Y.

Problem 5 Consider the regression model $Y = \beta_0 + \beta_1 * X + \epsilon$. Standard regression analysis requires that the variance of the error variable ϵ is a constant no matter what the value of X is. When this requirement is violated, the condition is called:

A. heteroscedasticity.

b. homoscedasticity.

c. influential observation.

d. non-independence of ϵ.

Problem 6 You regress Y on X_1, X_2 and X_3 and obtain the following estimates:

Coefficient	estimate
β_0	8
β_1	3
β_2	5
β_3	-4

If X_3 increases by one unit, with X_1 and X_2 held constant, what is the expected change in Y?

a. increase by 1 unit.

b. increase by 12 units.

C. decrease by 4 units.

d. decrease by 16 units.

Problem 7 The coefficient of determination, or R^2, ranges from:

a. 1 to ∞.

B. 0 to 1.

c. 1 to k, where k is the number of independent (X) variables in the model.

d. 1 to n, where n is the number of observations in the dependent variable.

Problem 8 You are interested in the following multiple regression model:

$$Price\_of\_silver = \beta_0 + \beta_1 * Price\_of\_oil + \beta_2 * Price\_of\_gold + \epsilon$$

What is the null hypothesis associated with the F statistic (the "F significance" cell in Excel)?

A. $H_0 : \beta_1 = \beta_2 = 0$

b. $H_0 : \beta_0 = \beta_1 = \beta_2$

c. $H_0 : \beta_1 = \beta_2 = 1$

d. $H_0 : \beta_0 = \beta_1 = \beta_2 \neq 0$

Problem 9 When the independent variables are correlated with one another in a multiple regression analysis, this condition is called:

a. Reverse causality

b. Omitted variable bias

C. Multicollinearity

d. Heteroscedasticity

e. autocorrelation

Problem 10 If a group of independent variables are not significant individually but are significant as a group (low F significance), this is most likely due to:

a. Reverse causality

b. Omitted variable bias

C. Multicollinearity

d. Heteroscedasticity

Problem 11 You are interested in the deterrent effect of police on crime. Using a sample consisting of all cities in the US with a population above 100,000 in 2012, you estimate the following regression model:

$$Crime = \beta_0 + \beta_1 * Police + \epsilon \tag{1}$$

where *Crime* is a city's crime rate and *Police* is the number of police per capita in a city.

Which statistical problem should you be most concerned with in this application?

A. Reverse causality

b. Omitted variable bias

c. Multicollinearity

d. Heteroscedasticity

Problem 12 You are interested in the causal effect of attending a more prestigious university on labor market earnings. Using a random sample of 10,000 40-year old college graduates, you estimate the following regression model:

$$Salary = \beta_0 + \beta_1 * Prestige + \epsilon \tag{2}$$

where *Salary* is annual labor market earnings and *Prestige* is a score from 0 (low prestige) to 100 (high prestige) given by the publication *Inside Higher Ed* to each university. Which of the following statistical problems should you be most concerned with in this application?

a. Reverse causality

B. Omitted variable bias

c. Multicollinearity

d. autocorrelation

Problem 13 Give a plausible example of two variables X and Y such that if the regression $Y = \beta_0 + \beta_1 * X + \epsilon$ were estimated, the following two things would likely be true:

1. X is significantly and positively correlated with Y (the coefficient estimate is positive and its associated p-value is low).

2. X is NOT a causal predictor of Y.

Note: you should give one single example that satisfies both 1-2 above.

Many examples are acceptable. To pick one mentioned in class, drowning deaths regressed on ice cream sales is a particularly stark example of positive correlation without causation.

Problem 14 Referring to your answer in question 13, explain, in detail, in the context of your example why X is significantly correlated with Y if X does not cause Y. Refer to appropriate concepts studied in class.

Again, there are many correct answers. In the drowning deaths example, a good answer would discuss omitted variable bias which could be fixed by including a "heat" variable.

Problem 15 Alice suggests that men who have children earn more than men with no children. To test her theory, she collects a random sample of 34,000 adult males from US census data. Each data point contains annual earnings in thousands of dollars, number of children, and various demographic variables (such as age, race, education, etc.). She estimates the following regression:

$$Earnings = \beta_0 + \beta_1 * Children + \epsilon \tag{3}$$

Her results are as follows:

Regression	**Statistics**
Multiple R	.31279
R Square	.12046
Adjusted R Square	.12046
Standard error	4.533
Observations	34,000

	coefficients	Standard error	t stat	P-value	Lower 95%	Upper 95%
Intercept	44.20	18.10	2.44	8.724	8.724	79.676
Children	12.425	2.489	4.99	5.977 E-007	7.5466	17.3034

a. Interpret the results. What do they predict about the relationship between children and earnings? Is the relationship significant?

By virtue of a low P-value, children is a significant variable for predicting earnings. The coefficient estimate predicts that having one more child will increase earnings by $12,425.

b. How would your answer to part a change were the value of R square to be much higher, say $R^2 = .92$?

While the R^2 is fairly low, this is not a big problem, as it is certain there are many other factors which affect earnings. The only way in which it would change the interpretation of the results is that we would conclude that factors other than children were relatively less important in explaining variation in earnings.

c. If Alice is specifically interested in the causal effect of having one more child on labor market earnings, list (at least) one additional X variable that she could add to the regression that would improve the accuracy of her results.[1] Explain, precisely, what the advantage is to including the variable you suggest.

[1] Do not worry about whether or not the variable you suggest is actually likely to be included in the Census dataset.

There is presumably some amount of omitted variable bias here, and any reasonable answer which offers an omitted variable with explanation is fine. Examples would be age (older people both have more kids and earn more by virtue of having more experience), education (fertility is strongly correlated with education, as is income), marital status (again, both the number of children and earnings are correlated with marital status), or spouse's income (own income and spouse's income are substitutes to some extent, and it's easy to imagine spouse's income being correlated with fertility).

Problem 16 The dean of admissions at Mountain State University wishes to determine what variables predict success at the university. Using a dataset consisting of information on 18,500 former students, he runs the following regression:

$$GPA = \beta_0 + \beta_1 * SAT + \beta_2 * ACT + \beta_3 * HS\_GPA + \beta_4 * HS\_class\_rank + \beta_5 * Number\_of\_Dean's\_List\_appearences + \beta_6 * Number\_extracurricular + \epsilon$$

where GPA is a student's college GPA, on a 4-point scale, SAT is a student's SAT percentile, ACT is a student's ACT percentile, HS_GPA is a student's GPA while in high school on a 4-point scale, HS_class_rank is a student's numerical rank in his high school graduating class (1 is highest possible), $Number\_of\_Dean's\_List$ is the number of semesters in which a student was on the Dean's List during high school, and $Number\_extracurricular$ is the number of extracurricular activities a student participated in.

His results are as follows:

Regression	**Statistics**
Multiple R	.98751
R Square	.85460
Adjusted R Square	.78249
Standard error	1.798.
Observations	18,500

	coefficients	Standard error	t stat	P-value	Lower 95%	Upper 95%
Intercept	1.8	.21	8.57	1.3884	1.3884	2.2116
SAT	.008	.006	1.333	.1825	-.00376	.01976
ACT	-.012	.080	-.1500	.88076	-.1688	.1448
HS_GPA	-.420	.524	-.8015	.4228	-1.447	.607
HS_class_rank	-.017	.287	-.0592	.9528	-.57952	.54552
$Number\_of\_Dean's\_List$	.467	2.31	.2022	.8398	-4.0606	4.9946
$Number\_extracurricular$	.143	.043	3.326	.0008823	.05872	.22728

a. Looking only at the signs (positive or negative) of the coefficient estimates, say for each of $\beta_1, \beta_2, ..., \beta_6$ whether the sign of the coefficient estimate is what we would expect it to be, and why.

SAT: a positive sign means a higher test score leads to a higher GPA. This makes sense.

ACT: a negative sign means a higher test score leads to a lower GPA. This is unexpected.

HS GPA: a negative sign means a higher HS GPA leads to a lower college GPA. This is unexpected.

HS Class rank: a negative sign means a better (lower number) class rank leads to a higher GPA. This makes sense.

Deans list: a positive sign means more HS Dean's list appearances lead to a higher college GPA. This makes sense.

Number extracurricular: a positive sign means the more HS extracurricular activities, the higher one's college GPA. This makes sense, at least in the view of admissions officers.

b. Which of the 6 independent variables are significant predictors of GPA? Which are insignificant?

c. For each of the significant variables you identified in part b, interpret the coefficient estimate, saying what it tells you in plain English.

Only Number extracurricular is significant, with a low P-value of .0008823. It's coefficient tells us that having one additional extracurricular activity is associated with a college GPA that is .143 higher, or about $\frac{1}{7}$ of a grade.

d. In light of your answers to parts a and b, do these results do a good job explaining the determinants of college GPA? If yes, say how you know, and if no, suggest what the problem is and what can be done about it.

We have a high R^2, but 5 out of 6 coefficient estimates are insignificant, meaning that we don't really know much about the determinants of GPA. There is almost certainly a multicollinearity problem here. ACT and SAT scores are highly correlated. HS GPA, HS class rank, and Number of Dean's List are 3 different measures of the same thing (grades), and so are presumably highly correlated. We will get better results by removing at least 3 of the variables: either ACT or SAT, and 2 of the 3 variables measuring HS grades.