Midterm Exam

Instructions

INSTRUCTIONS:

- There are 8 multiple-choice questions worth 3 points each and 2 short answer questions worth 76 points for a total of 100 points. Budget your time in each section accordingly.
- The midterm exam is closed-book, but you can use a simple calculator, one letter size sheet with formulas on one side, and R for very simple computations.
- For the test, you may assume that the samples are large enough to satisfy the Central Limit Theorem, so a normal approximation to the t-statistic is appropriate.
- The duration of the exam is 95 minutes.

Question 1

3 / 3 pts

A researcher analyzing the determinants of earnings and she has data on 16 occupation categories that exhausts all possibilities. If the researcher runs a regression of earnings on a binary (dummy variable) for all 16 categories, which least-square assumption is violated?

Correct!

- (●) No perfect multicollinearity
- ( ) Conditional mean of errors is zero
- ( ) i.i.d. Sample
- ( ) Large outliers unlikely

Question 2

0 / 3 pts

A hypothesis test on a regression coefficient fails to reject the null hypothesis at the 5% significance level. This means:

Correct Answer

- ( ) If the null hypothesis were true, the chance of obtaining an estimate as far from the null as the one we obtained would be more than 5%.

The null hypothesis is true.

The null hypothesis is false.

You Answered

If the null hypothesis were true, the chance of obtaining an estimate as far from the null as the one we obtained would be less than 5%.

Question 3

3 / 3 pts

The R-squared for a regression is found to be 0.45. This means that:

Correct!

55% of the variation in Y around its mean is due to factors not included in the regression.

The regression provides evidence of a causal relationship between Y and the regressors

The regression coefficients suffer from omitted variable bias.

45% of the variation in Y around it mean is due to factors no included in the regression.

Question 4

3 / 3 pts

Question 4-8: Suppose you run a simple regression of house price (P) measured in US dollars on a dummy variable, G, that takes 1 if the neighborhood has at a park or green space area, and 0, otherwise.

$$\widehat{P}_i = 114{,}000 + 236{,}000\, G_i,\ i=1,\ldots,n=400,\ R^2 = 0.34$$
$$(3{,}000.5)\quad (98{,}500)$$

4. According to the estimate on G:

Correct!

- (●) Houses in neighborhoods with parks are, on average, 236,000 US dollars more expensive than house in neighborhoods without parks.
- ( ) An increase in the number of parks by 1, will increase, on average, the house price by 236,000 US dollars
- ( ) The mean house price in neighborhoods with parks is 114,000 US dollars.
- ( ) The mean house price in neighborhoods with parks is 236,000 US dollars.

Question 5

3 / 3 pts

The 95% confidence interval on G:

Correct!

- (●) has the upper bound higher than 400,000
- ( ) contains 0.

- has all values below 0
- has the upper bound lower than 400,000

Question 6

3 / 3 pts

Suppose you do not like working with large numbers and you re-scale the house prices to thousands and run the above regression again. You expect:

Correct!

- (•) The estimated coefficient on G to decrease and the intercept to decrease.
- The estimated coefficient on G to decrease and the intercept to increase.
- The estimated coefficient on G to decrease and the intercept to remain the same.
- The estimated coefficient on G to remain the same and the intercept to decrease.

Question 7

3 / 3 pts

Poor people are less likely to live in neighborhoods with parks and green spaces than rich people. The housing of poor people is typically less valuable. Because of this, in a simple regression of P on G, the coefficient on G will probably be:

Correct!

- (●) Upward biased as an estimate of the causal impact G on P.
- ( ) Downward biased as an estimate of the causal impact of G on P.
- ( ) Unbiased
- ( ) Exactly zero.

Question 8

3 / 3 pts

Suppose that you obtain more data on house prices and neighborhood characteristics increasing the sample size to 800. When you estimate the same regression, other things equal, you expect:

Correct!

- (●) The standard error of the slope to decrease
- ( ) The coefficient on G to be unbiased.
- ( ) The standard error of the slope to increase.
- ( ) None of the above.

Question 9

10 / 10 pts

Short-Answer Question 1 (Part a):

A real estate researcher wants to understand the main determinants of house prices in Boston area. He collects data from the real estate pages of the Boston Globe, during 1990s, on homes sold in Boston during this time. Here are the variables available for the analysis and their descriptive statistics:

Variable	Description
House price	House Price in $1,000s
Lot size	Size of the lot in square feet
House size	Size of the house in square feet
Number of bedrooms	Number of bedrooms
Colonial	1 if home is colonial style, 0 otherwise

Determinants of House Prices

Statistic	House Price	Lot Size	House Size (sq feet)	Number of bedrooms
N	88	88	88	88
Mean	293.55	9,019.86	2,013.69	3.57
St. Dev.	102.71	10,174.15	577.19	0.84

First, the researcher focuses on the effect of the lot size on the house prices. The regression results in the following output:

$$\widehat{House\ Price}_i = 261.937 + 0.004^* Lot\ Size_i,\ R^2 = 0.12,\ SER=96.89$$
$$(21.9) \quad (0.001)$$

(a)Interpret the regression results carefully (make sure to interpret both the intercept and slope estimate, R2 and SER

Your Answer:

A one square foot increase in Lot size is associated with an increase in House Price by $0.004(in 1000s)

When the lot size is 0 square foot, the sample mean of the House price is $261.937 (in 1000s). However, having Lot size=0 does not make sense.

R^2 = 0.12 represents that 12% of the variation in House Prices (in $1000s) is explained by Lot Size.

SER = 96.89 represents that, on average the deviation (or spread) of the actual house price around the regression line is $96.89 (in 1000s).

Question 10

5 / 5 pts

Short-Answer Question 1 (Part b):

(b) Calculate the t-statistic to determine whether the slope coefficient is statistically different from zero. Justify the use of of a one-sided or a two-sided test.

Your Answer:

H0 : beta1 = 0

Ha : beta1 is not 0

t-stat = (0.004-0)/(0.001) = 4

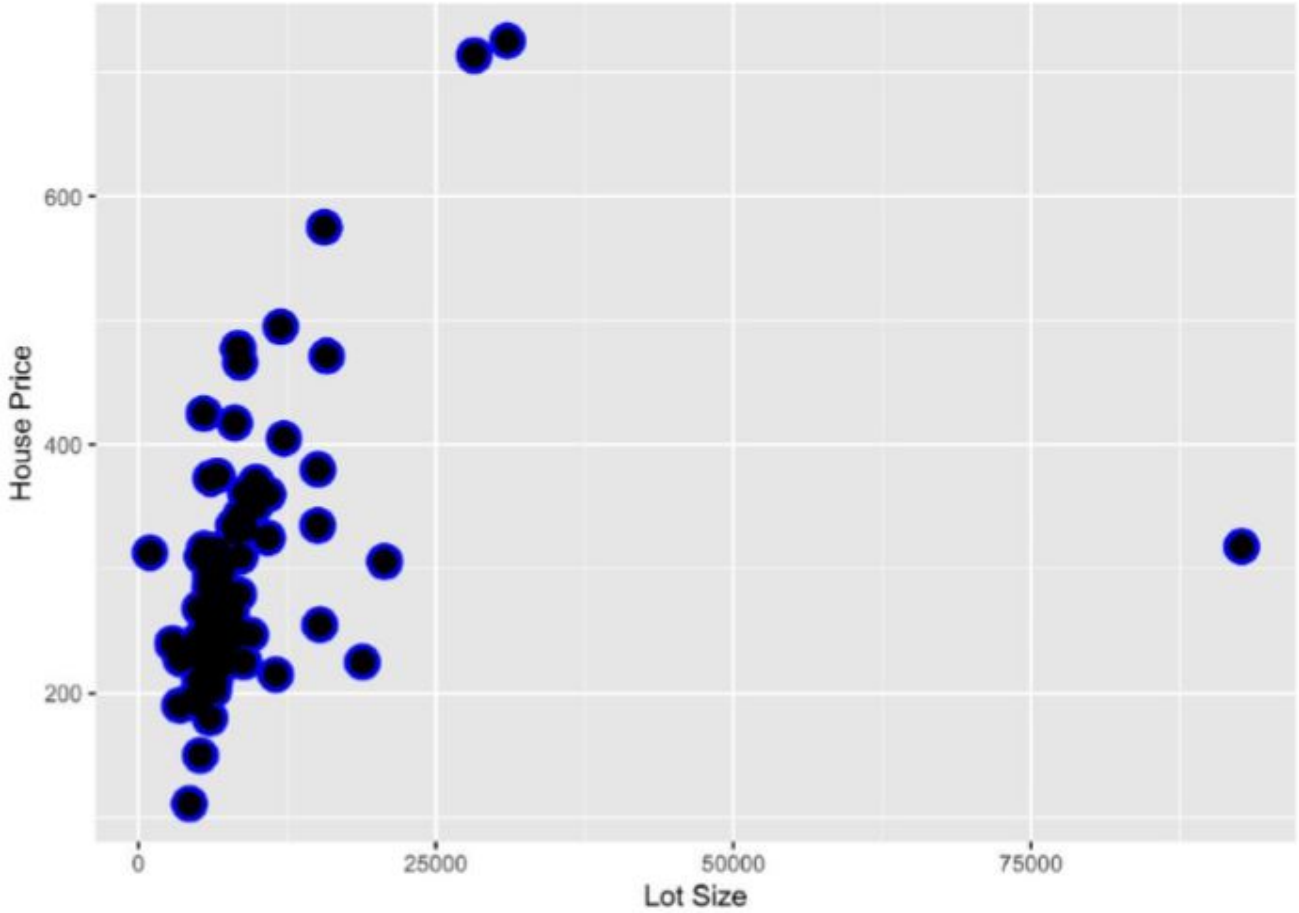
The t - statistics is 4.0 for the slope. Since we expect a positive sign on the slope (as lot size increases, we expect the house prices to increase), we should conduct a one-sided test. We can reject the null hypothesis that the slope estimates are equal to zero at less than 5% significance level. We can reject the null hypothesis at even 1% as critical value is 1.645 for one-sided test.

Question 11

5 / 5 pts

Short-Answer Question 1 (Part c):

(c) You accidentally forgot to use the heteroskedasticity-robust standard errors option in your regression package and estimate the equation using homoskedasticity-only standard errors. You quickly plot the data to see if you have to worry about your mistake:



Should you re-run your regression with heteroskedasticity-robust standard errors. If you do, do you expect the estimate and the standard error of the estimate to increase, decrease or remain unchanged. Explain carefully.

Your Answer:

It is difficult to say if the regression suffers from Heteroskedasticity based on the graph. We should plot a graph of residuals vs Lot Size to see if the scatter plot has any patterns or if it is randomly scattered.

To be on the safe side, let's assume the model does suffer from Heteroskedasticity and re-run the model with heteroskedastic-robust standard errors. If there is a heteroskedasticity issue, when we re-run the

regression, then we expect to see an increase in the standard error of slope estimate and no change in slope estimate. Even if the model has homoskedastic errors, re-running the model with robust errors will not change the estimates.

Question 12

5 / 5 pts

Short-Answer Question 1 (Part d):

(d) You take a second closer look at your plot. Is the relationship linear? Are there any outliers? Should you exclude the outliers from the analysis? Why or why not?

Your Answer:

Yes, the relationship between House price and Lot Size looks linear.

Yes, we can see that there is one outlier at extreme right of the scatter plot.

Excluding the outlier depends on the context. If the outlier came into existence because of some data entry error, then we can safely exclude it. However, if there is some additional context to that outlier, then that point should not be excluded.

For this question, assuming the outlier is an erroneous entry, we should exclude it. Because this point is a large outlier, it will contradict a least-square assumption (NO LARGE OUTLIERS), hence to get unbiased estimates of slope, we should remove it.

Question 13

5 / 5 pts

Short-Answer Question 1 (Part e):

(e) Using the regression results and descriptive statistics, what is the correlation coefficient between house price and lot size? Explain.

Your Answer:

Correlation Coefficient between House price and lot size is 0.346.

Since this is a regression with only one regressor, the correlation coefficient r^2 is equal to R^2 of the regression.

R^2= r^2 = 0.12

therefore, r = root(0.12) = 0.346

It suggests that House Price and Lot Size are positively correlated.

Question 14

20 / 20 pts

Short-Answer Question 2 (Part a):

On the same dataset with some modifications, the researcher runs a set of 4 regressions. The table below displays the output of these regressions:

Determinats of House Prices

	Dependent variable:			
	price			
	(1)	(2)	(3)	(4)
Lot Size	0.015	0.009	0.009	0.010
	(0.003)	(0.002)	(0.002)	(0.002)
House Size		0.104	0.086	0.087
		(0.012)	(0.012)	(0.012)
No. of Bedrooms			20.482	15.733
			(7.113)	(7.558)
Colonial				24.443
				(13.750)
Constant	169.882	12.195	-28.301	-32.822
	(19.076)	(22.926)	(25.103)	(26.265)
Observations	87	87	87	87
R2	0.492	0.745	0.765	0.775
Adjusted R2	0.486	0.739	0.756	0.764
Residual Std. Error	74.020	52.776	51.000	50.133

(a) Write out the formula for the fitted regression line in column (2) and (3). Interpret each of the estimated slope coefficients in words, citing specific numbers and units.

Your Answer:

Column 2:

House Price = 12.195 + 0.009 LotSize + 0.104 HouseSize

Keeping house size constant, a one square foot increase in Lot size is associated with an increase in House Price by $0.009(in 1000s) (or Keeping house size constant, a one square foot increase in Lot size is associated with an increase in House Price by $9).

Keeping lot size constant, a one square foot increase in House size is associated with an increase in House Price by $0.104(in 1000s) (or Keeping lot size constant, a one square foot increase in House size is associated with an increase in House Price by $104).

Column 3:

House Price = -28.301 + 0.009 LotSize + 0.086 HouseSize + 20.482 NoOfBedrooms

Keeping all other things constant, a one square foot increase in Lot size is associated with an increase in House Price by $0.009(in 1000s) (or Keeping all other things constant, a one square foot increase in Lot size is associated with an increase in House Price by $9).

Keeping all other things constant, a one square foot increase in House size is associated with an increase in House Price by $0.086(in 1000s) (or Keeping all other things constant, a one square foot increase in House size is associated with an increase in House Price by $86).

Keeping all other things constant, an increase in the number of bedrooms by 1, is associated with an increase in House Price by $20.482 (1000s) (or Keeping all other things constant, an increase in the number of bedrooms by 1, is associated with an increase in House Price by $20482).

Question 15

10 / 10 pts

Short-Answer Question 2 (Part b):

(b) Does the table suggests that the regression in column (1) suffers from omitted variable bias? Use the numbers in the table to support your answer. Is it consistent with what you would have expected? What LS square assumption is likely violated in column (1)?

Your Answer:

Yes, there is an omitted variable bias in regression 1. As house size affects House Price positively, and house size is positively correlated with Lot size, regression 1 suffers from an Upward Bias. As the omitted variable is positively correlated with both, we would expect upward omitted variable bias. This is also confirmed with the numbers in regression output as after adding the house size, the slope coefficient decreased from 0.015 to 0.009.

The least-square assumption violation in regression 1 is that E(UI | X) does not equal to 0. Therefore, we have a biased and inconsistent estimator.

Question 16

5 / 5 pts

Short-Answer Question 2 (Part c):

(c) Which of the four regressions has the better overall fit, in your view? Explain carefully, citing specific numbers.

Your Answer:

We can find the best model by comparing the adjusted R^2. We see that the fourth regression explains 76.4% of the variation in house prices. The third regression, explains 75.6% of the variation in house-prices around its mean compared to 48.6% in the first regression and 73.9% in the second regression.

Based on these numbers we can say that Fourth regression is the best and has a better overall fit than all other models presented.

Question 17

6 / 6 pts

Short-Answer Question (Part d):

(d) Using the model in column (2), calculate the predicted house price for a 3,000 square feet house on a 10,000 square feet lot . If the house price is $400,000, what is the residual?

Your Answer:

Column 2:

House Price = 12.195 + 0.009 LotSize + 0.104 HouseSize

House Price = 12.195 + 0.009 (10000) + 0.104 (3000) = $414.195 (in 1000s) = $414195

If the house price is $400,000

Residual = Actual_Y - Predicted_Y

Residual = $400000 - $414195 = -$14195

Question 18

5 / 5 pts

Short-Answer Question (Part e):

(e) In column (5), interpret carefully the coefficient estimate on Colonial using numbers?

Your Answer:

24.443

Holding all other variables constant, houses with colonial-style (Colonial = 1) have, on average, House Price $24.443 (in 1000) more than houses with no colonial-style(Colonial =0).

or Holding all other variables constant, houses with colonial-style (Colonial = 1) have, on average, House Price $24443 more than houses with no colonial-style(Colonial =0).

Quiz Score: **97** out of 100