Name: _____
BSAD 210—Montgomery College
David Youngberg

EXAM 3

Practice B

- There are 110 possible points on this exam. The test is out of 100.
- You have one class period to complete this exam, but you should be able to complete it in less than that
- Please turn off all cell phones and other electronic equipment.
- Be sure to read all instructions and questions carefully.
- Remember to show all your work. Writing down what you put into Excel is sufficient to show your work.
- To access Data Analysis on Excel, select File (top left), then Options, then Add-ins, then Go… (for Excel Add-ins), then select Analysis ToolPak.
- Try all questions! You get zero points for questions that are not attempted.
- Note the last sheet lists all the equations you will need for this exam.
- *Please print clearly and neatly.*

Part I: Matching. *Write the letter from the column on the right which best matches each word or phrase in the column on the left. You will not use all the options on the right and you cannot use the same option more than once.*

2 points each.

1. ___ Average
2. ___ Dummy variable
3. ___ ε
4. ___ Omitted variable bias
5. ___ Observed value
6. ___ Percentage points
7. ___ Predicted value

A. A missing control that's pivotal to the analysis
B. Best guess with a linear regression
C. Best guess without a linear regression.
D. Example: Average intelligence
E. Example: Employment status
F. Example: Years of employment
G. For every observation, there is one of these.
H. If you want to say the difference between 8% and 10% is 2, describe 2 as this.
I. This minus the residual is the estimated value.

Part II: Multiple Choice. *Circle the best answer to the following.*

4 points each.

8. If the dependent variable of a regression is NEWSPAPER? (if a person gets a newspaper delivered or not), and a predicted value is 0.79, what does that mean?
    a. It means the person gets 79% of a newspaper delivered.
    b. It means this person definitely gets a newspaper delivered.
    c. It is impossible to meaningfully answer this question without knowing what the independent variables are.
    d. It means nothing; it would only mean something if it was 0 or 1.
    e. It means something, but nothing listed here.

9. How should one fix a regression with multicollinearity?
    a. Add additional independent variables until the multicollinearity goes away.
    b. Change your dependent variable until multicollinearity is gone.
    c. Remove both variables which are causing multicollinearity.
    d. Remove one of the variables which are causing multicollinearity.
    e. None of the above

10. What's the difference between the correlation coefficient of two variables and a regression line with two variables?
    a. Only correlation determines if the correlation is positive or negative.
    b. Only regression considers causation.
    c. Only regression shows the magnitude of change (i.e. the slope of the best fit line).
    d. A & C
    e. B & C

11. Autocorrelation is a bigger concern for a time series than for a cross-sectional. Why?
    a. Because it's unlikely two variables will be highly correlated with each other.
    b. Because over a long period of time, the standard deviation is naturally higher compared to instances that happen all at once.
    c. Because the data have a natural order, thus a pattern in the residuals is more likely.
    d. Because the residuals are more likely to form a normal distribution if they happen at different times.
    e. Because there's always a way to organize the data to generate a pattern.

12. One of the assumptions of linear regression is that the residuals follow a normal distribution. What does that mean?
    a. Most observations are within one standard deviation of the mean.
    b. Most of the "unexplained" parts of the regression are small, relatively speaking.
    c. Each large variable has a small variable paired with it.
    d. The residuals are never less than zero.
    e. None of the above

13. Use Practice Exam 3 Data Set 2 for this question. Which pair of variables is least correlated?
    a. Salary and Experience
    b. Experience and Mistakes
    c. Female? and Experience
    d. There is not enough information to determine the answer.
    e. There is enough information but none of the above are correct.

14. Consider Practice Exam 3 Data Set 2 for this question. Imagine you wanted to tell a story about the connection between Mistakes and Experience. Which makes more sense: Mistakes cause Experience or Experience causes Mistakes?
    a. Experience causes Mistakes, because you're less like to screw up if you what you're doing.
    b. Experience causes Mistakes, because more experience means a higher salary and a higher salary means you'll be more careful to not lose the job.
    c. Mistakes cause Experience, because if you make mistakes, you get fired and can't get experience.
    d. Mistakes cause Experience, because you learn from mistakes.
    e. None of the above make sense

15. What is homoscedasticity?
    a. An assumption of linear regressions.
    b. When the observed mean has a consistent difference from the predicted mean.
    c. When the number of explanatory variables is less than the number of observations.
    d. A & B
    e. A & C

16. Can adjusted R^2 ever be negative?
    a. No, because R^2 captures the portion of the variation that's explained and a portion can never be less than zero.
    b. No, because adjusted R^2 uses the number of explanatory variables and that's never negative.
    c. Yes, because the original R^2 and the number of explanatory variables are completely different inputs into the adjusted R^2 equation.
    d. Yes, but only if the number of observations is less than the number of explanatory variables.
    e. None of the above

17. When is a variable in a regression is statistically significant at the 99% level?
    a. If the p-value is more than 0.01
    b. If the p-value is less than 0.01
    c. If the t-statistic is more than 2.576
    d. If the t-statistic is less than 2.576
    e. More than one of these is true.

18. Imagine a regression with YEAR predicting SALES, based on sales data from 1990 to 2010:

$$SALES = 100{,}000 + 6500 * YEAR + \varepsilon$$

    What should SALES be for the year 2100?
    a. 685,000
    b. 13,650,000
    c. 13,750,000
    d. You can't use year as an independent variable.
    e. You shouldn't predict so far out of the original range.

19. What does the ε at the end of a regression mean?
    a. The amount you'd have to add to (or subtract from) the predicted value to get the observed value.
    b. The degree the regression matters, holding the number of explanatory variables constant.
    c. The value that's added to the regression to make it more accurate.
    d. B & C
    e. None of the above

Part III: Short Answer. *Answer the following.*
16 points each.

20. Use Practice Exam 3 Data Set 2 to answer this question. It includes hypothetical data on a hypothetical company, Gibson Industries. Run a regression with Mistakes as the dependent variable and the other variables (Age, Female?, Experience, and Salary) as the independent variables. Answer the following:
    a. Which variables are statistically significant at the 99.9% level? Which ones are significant at the 99% level? Which ones are significant at the 95% level? How do you know?
    b. Is this model good as a whole? How do you know?

a. _____

b. _____

21. Which your regression from the previous question in mind, answer the following:
    a. Does this model have any multicollinearity? How do you know? Provide (print and include) the evidence.
    b. If the model has multicollinearity, how should you change this model? If it doesn't have multicollinearity, what's the general strategy to fix the model?

a. _____

b. _____

22. Consider the following hypothetical regression, with Views is the number of views a YouTube video has, Length is how long the video is, in minutes, Cats is the number of cats that appeared in the video, HD? is if the video is in HD (1) or not (0), and Time is the time of day the video was uploaded (as in, the hour number after midnight; 5 AM is 5, 2 PM is 14). All coefficients, except the one for Time, are statistically significant.

$$VIEWS = 5{,}000 - 75 * LENGTH + 90 * CATS + 700 * HD? + 50 * TIME$$

Answer the following:

a. Estimate how many Views a video would have if it's twenty minutes long, has two cats, is in HD, and was uploaded at 6 PM.
b. If you add ten cats to a video, how should the expected number of views change?
c. Give the "punchline" interpretation of the Length variable: "For every additional minute the video is…"
d. This regression has some causation issues concerning a mistaken linear relationship. What variable is probably not linearly related and why? (There may be multiple correct answers to this question.)

a.

b.

c. _____

_____

_____

d. _____

_____

_____

_____

_____

Exam 3 Equation and Information Reference

Function	*Output*
ABS	The absolute value of an input
AVERAGE	Arithmetic mean of a dataset
CONFIDENCE.NORM	Determines the margin of error to make a confidence interval (known σ)
CONFIDENCE.T	Determines the margin of error to make a confidence interval (unknown σ)
CORREL	Correlation coefficient of two variables
CTRL + `	Show formulas
CTRL + F	Find
CTRL + P	Print
CTRL + X	Cut highlighted area
CTRL + C	Copy highlighted area
CTRL + V	Paste highlighted area
CTRL + Z	Undo
F4	Makes cell reference absolute
GEOMEAN	Geometric mean of a dataset (adjustments must be added manually)
LARGE	Larger values of a dataset (k=1 is largest, k=2 is second largest, k=3 is third largest…)
MAX	Maximum value of a dataset
MEDIAN	Median of a dataset
MIN	Minimum value of a dataset
MODE	Mode of a dataset
NORM.DIST	Returns the normal distribution for a specified mean and standard deviation.
NORM.INV	Returns the inverse of the normal cumulative distribution for a specified mean and standard deviation.
NORM.S.DIST	Returns the standard normal distribution.
NORM.S.INV	Returns the inverse of the standard normal cumulative distribution. Useful for finding critical z scores.
QUARTILE	The 0^{th} to 4^{th} quartile of a dataset
SQRT	Finds the square root of the value in question.
SMALL	Smaller values of a dataset (k=1 is smallest, k=2 is second smallest, k=3 is third smallest…)
STDEV.S	Standard deviation of a sample
T.INV	Finds area under a t distribution; useful for finding one-tailed critical t scores.
T.INV.2T	Finds area under a t distribution; useful for finding two-tailed critical t scores.
T.TEST	Various two population tests which use a t score.

Geometric Mean

$$Geometric\ Mean = \sqrt[n]{\prod_{i=1}^{n}(1 + x_i)} - 1$$

Weighted Average

$$Weighted\ Average = \frac{\sum_i^n (w_i x_i)}{\sum_i^n w_i}$$

Coefficient of Variation

$$CV = \frac{s}{\bar{x}}$$

Confidence interval for proportion

$$\widehat{CI}_{\bar{p}} = \bar{p} \mp z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Adjusted R^2

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Hypothesis testing

z-test

$$z_{\bar{x}} = \left|\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right|$$

t-test

$$t_{\bar{x}} = \left|\frac{\bar{x} - \mu}{s/\sqrt{n}}\right|$$

z-test (proportion)

$$z_p = \left|\frac{\bar{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right|$$

Critical z scores

Use =NORM.S.INV command

Confidence	*α*	$z_{\alpha/2}$	z_{α}
95%	0.05	1.960	1.645
99%	0.01	2.576	2.326
99.9%	0.001	3.291	3.090

Critical t scores

Use T.INV or T.INV.2T commands or see the table on the last page

p-values

Make your calculated value negative and then use one of the following (make sure cumulative is turned on):

	1 tail	*2 tails*
z	NORM.S.DIST	Multiply 1 tail result by 2
t	T.DIST	