Name: ____________________________
BSAD 210—Montgomery College
David Youngberg

EXAM 3

Practice A

- 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. ___ Adjusted R^2
2. ___ β
3. ___ Confounding variable
4. ___ Dummy variable
5. ___ Homoscedasticity
6. ___ Multicollinearity
7. ___ R^2

A. Adding explanatory variables always increases this.
B. Adding explanatory variables may lower this.
C. Can only be two different values
D. Describes the percent explained.
E. Example: Child’s education level causing parents’ education level when it should be parent’s education level causing child’s education level.
F. Example: Amazon.com visits causes Wikipedia visits when it should be the proliferation of the internet causing both.
G. For any variable, the null hypothesis is that this is zero.
H. When the distribution of residuals does not change as the independent variables change.
I. When two or more explanatory variables are highly correlated with each other.

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

8. Often in multivariable regression analysis, you’re most interested in one of the independent variables. What do we call the independent variables you’re not as interested in?
   a. Confounding variables
   b. Control variables
   c. Explanatory variables
   d. Dummy variables
   e. None of the above

9. Francis runs a regression with a sample of 33 and with 16 explanatory variables (excluding the intercept). His R^2 is 0.70. What is his adjusted R^2? Remember to show your work.
   a. 0.36
   b. 0.40
   c. 0.60
   d. 0.70
   e. None of the above

10. Suppose you run the following regression using county data:

$$SPENDING = 800 + 15{,}000 * \%D + 0.08 * INCOME$$

Where SPENDING is the K-12 public school funding per pupil, in $; %D is the percent of a registered voters who are Democrats; and INCOME is the average household income. Suppose only INCOME is statistically significant. If %D increases by 1 percentage point, how does SPENDING change?

   a. It increases by $150
   b. It increases by $950
   c. It increases by $15,000
   d. It increases by $15,800
   e. It doesn't change.

11. Suppose you ran a regression of STEPSPERDAY predicting BMI (Body Mass Index; higher values implies fatter people). Also suppose your estimated line was (STEPSPERDAY is statistically significant):

$$\text{BMI} = 40 - 0.002 * \text{STEPSPERDAY}$$

How would BMI change if someone started walking 300 more steps per day?

   a. BMI would increase by 0.6
   b. BMI would decrease by 0.6
   c. BMI would increase by 0.002
   d. BMI would decrease by 0.002
   e. None of these

12. Tyron is interested if the strategy a player prefers in a Rock-Paper-Scissors tournament can be used to predict the player's age. Assume all players have one and only one favored strategy. Tyron gathers and records his data (the first variable asks if the player prefers Rock, etc.) and a section of the output is indicated below.

Age	**Strategy**		
	Rock?	*Paper?*	*Scissors?*
14	1	0	0
18	0	0	1
12	1	0	0
31	0	1	0
…	…	…	…

What, if anything, is wrong with how Tyrone recorded his data?

   a. He has too many variables.
   b. He will get heteroscedasticity.
   c. He has the same player having more than one favored strategy.
   d. A & C
   e. None of the above / Nothing is wrong with it.

13. Use the Practice Exam 3 Data Set 1 for this question. Which pair of variables is the most highly correlated?
    a. Number of competing stores in district & Annual profit, in thousands
    b. Number of families in the sales area, in thousands & Annual profit, in thousands
    c. Number of competing stores in district & Advertising spent, in thousands
    d. Square feet, in thousands & Annual profit, in thousands
    e. None of the above

14. The percent of families who own their own home is positively correlated with average income across different U.S. states. Suppose a governor attempts to increase average income by subsidizing homeownership. If this is a mistake, what is the most likely reason?
    a. Reverse causation: home ownership is the result of high incomes.
    b. Reverse causation: something else is causing both variables.
    c. Confounding variable: home ownership is the result of high incomes.
    d. Confounding variable: something else is causing both variables.
    e. None of the above / The governor has the correct interpretation

15. Which of the following pair of variables is positively correlated?
    a. Value of a car and the owner's income
    b. Time spent taking a shower and the number of forks that person owns
    c. Frequency of rain and how often people play outside
    d. A & C
    e. None of the above

16. Suppose time spent playing video games and non-violent criminal activity are negatively correlated. While there might be a causation story (video games offer a safe outlet for criminal urges), a confounding variable could also be an explanation. Which of the following is the most likely reasonable confounding variable?
    a. Frequency of police patrols
    b. Frequency of new video game releases
    c. Weather
    d. A & B
    e. None of the above

17. As you add explanatory variables to a regression, what always happens?
    a. Your F-stat falls.
    b. The p-values of the variable(s) you started with decrease.
    c. R^2 increases.
    d. A & B
    e. None of the above

18. Safara is using city-level data to predict the level of a city's average income. Which of the following explanatory variables should be adjusted for population?
    a. If the city is on a river or not.
    b. The number of schools in the city.
    c. The percent of the city is Asian American.

d. B & C
e. All of the above

.

19. Which of the following would best be represented with a dummy variable (or series of dummy variables)?
    a. Type of pet a household has
    b. Household income
    c. Number of children in a household
    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 1 to answer this question. It includes hypothetical data on a hypothetical grocery store chain called The Happy Spud. Run a regression with Annual Profit being predicted by Square Feet, Inventory Value, Advertising Spent, Number of Competing Stores, and Number of Families. Answer the following:
    a. Which variables are statistically significant at the 99.9% level? How do you know?
    b. Is this model good as a whole? How do you know?
    c. If you were to add Region to the model, what would you have to do first?

a. ____

____

____

b. ____

____

____

c. ____

____

____

21. In the previous question, you used Happy Spud franchise data to run a regression predicting store profitability.
    a. Does this model have any multicollinearity? How do you know? Provide (print and include) the evidence.
    b. If it does have multicollinearity, what's the best way to fix this model and why?

a. ____

____

____

b. ____

____

____

22. Imagine you’re running a regression predicting recidivism (if a convicted felon reoffends) based on the inmate’s age (AGE), time spent in prison (PRISON), and good behavior (GOOD?). Age and time spent in prison are measured in years. RECIDIVST? and GOOD? are both dummy variables: 1 means the felon reoffended or had good behavior while in prison and 0 means the felon did not reoffend or did not have good behavior in prison. Suppose this is your estimated regression line (all coefficients are statistically significant):

$$RECIDIVIST? = 0.45 - 0.01 * AGE + 0.04 * PRISON - 0.15 * GOOD?$$

Answer the following:

a. For a 17-year-old felon who spent three years in prison with good behavior, what is the predicted value of RECIDVIST?
b. In normal language, what does the value in (A) mean?
c. Give the “punchline” interpretation of the AGE variable: “For every additional year of age the convict is…”
d. This regression has some causation issues concerning a confounding variable. What is that confounding variable 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 0th to 4th 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

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

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	