Name: **KEY**
BSAD 210—Montgomery College

EXAM 3

Practice #1

- 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. You may print your formulas in Excel using the Show Formulas option in the Formulas tab. Printed versions of your work showing formulas ***and*** showing the results counts as showing your work. But you must include both with your test for "showing your work" to count this way. Write your name on both print outs.
- 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. **B** Adjusted R^2
2. **E** Confounding variable
3. **C** Dummy variable
4. **H** Homoscedasticity
5. **I** Multicollinearity
6. **F** R^2
7. **G** Scalar

A. A constant value you apply to the data, like taking the square root.
B. Adding explanatory variables may lower this.
C. Can only be two different values
D. Example: Child's education level causing parents' education level when it should be parent's education level causing child's education level.
E. Example: Using Amazon.com causes Wikipedia use when it should be the proliferation of the internet causing both.
F. Describes the percent explained.
G. When applied, will not change statistical significance.
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.

1. *By definition, adjusted R^2 adjusts for the number of explanatory variables used. If you add an explanatory variable that doesn't make up for this penalty—that doesn't "pull its weight"—then adjust R^2 may fall. Even though R^2 rises.*
2. *A confounding variable is the underlying reason for a correlation. If you plot yearly Amazon.com users and Wikipedia users, you'll surely get a positive correlation. But that's because of increased accessibility of the internet caused both, not because one caused the other. (D is reverse causation.)*
3. *A dummy variable is either 0 ("no") or 1 ("yes"). It has only two possible values.*
4. *Variance of residuals should be the same across all values of the independent variables; that's the definition of homoscedasticity and an assumption of linear regressions.*
5. *When at least two independent variables are correlated, then at least one variable is made redundant by another variable. Statistical significance falls when it shouldn't, resulting in Type II Error.*

6. *R^2 is the explained sum of squares divided by the total sum of squares; it's the amount of deviation from the average that the regression can explain.*
7. *A scalar is a constant value, but taking the square root wouldn't be a constant transformation. But scalars don't change statistical significance. They change the coefficient (as we've discussed), but statistical significance doesn't change.*

Part II: Multiple Choice. *Choose 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

We call these "controls variables" or simply controls. They hold other important factors constant; they "control" for other factors.

9. Suppose the coefficient for an independent variable—number of cars per person—is 1.5. Suppose you changed the independent variable so it considers the number of cars per 100 people. What is the coefficient now?
    a. 0.015
    b. 0.15
    c. 15
    d. 150
    e. It is impossible to tell with the information provided

To transform the data to "cars per 100 people," you multiply the raw data by 100. This makes sense: if you're looking at 100 times as many people, you'd expect 100 times as many cars. If that value is going up then the coefficient must be proportionally smaller to compensate. It's not as if your predicted value of Y should be much larger just because you're using a trivially different measurement: the coefficient must adjust to keep it the same.

10. Suppose you ran a regression of STEPSPERDAY predicting BMI (Body Mass Index; higher values implies fatter people). Also suppose your estimated line was BMI = 40 – 0.002*STEPSPERDAY + ε. 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

Multiply the coefficient (-0.002) by 300 results in -0.6. Since the coefficient is negative, BMI falls by the indicated amount.

11. 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.

One of the variables is completely redundant—it doesn't matter which. If a player isn't using Rock nor Paper has the primary strategy, Scissors must be their primary strategy. Keeping it as it is would introduce multicollinearity. And yes, there actually are Rock-Paper-Scissors tournaments.

12. Use the Practice Exam 3 Data Set 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

If you make a correlation table with Excel, you'll find this pair has a correlation coefficient of 0.95. Note option D is the second highest (0.89) and option third highest (-0.85). The higher the <u>absolute</u> <u>value</u> of the coefficient, the higher the correlation.

13. 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

Recall the equation for adjusted R^2:

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

Or, 1 – (1 – 0.70)(33 – 1)/(33 – 16 – 1) = 1 – (0.3)(32/16) = 1 – 0.6 = 0.4.

For someone using so many explanatory variables, we should not be surprised by this drop in explanatory power.

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

Options (B) and (C) are interpreting reverse causation and confounding variables incorrectly. While (D) is possible, it is unlikely…what would that underlying variable possibly be? Interesting, E isn't a terrible choice—perhaps homeownership allows people to borrow more because the house is collateral and that might mean they go to college which means, later, they get more income. Maybe, but there's a lot of ifs, maybes, and conditions in that statement. Option (A) is the most likely answer.

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

One would expect cheaper cars would be owned by people with low incomes and expensive cars would be owned by people with high incomes; that is positive correlation. Option C is an example of negative correlation: as it rains more, people will play outside less. Option B is an example of zero correlation. It's hard to imagine the number of forks someone owns would influence or be influenced, directly or indirectly, by how long that person spends in the shower.

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

While you could argue all these of these are confounding variables, only option C is really strong. Both A and B suggest criminals might stop thieving in order to play games—thus the negative correlation—but that relies on the same people switching tasks.

Weather, however, doesn't require that assumption. If the weather is good, fewer people will play video games and more criminals will engage in illegal activity but these don't have to be the same people. The people could be shopping, surfing, going to bars, etc. And because the criminals are outside, they are not sleeping or watching TV and, yes, playing video games. Because C doesn't require that the gamers are the same folk as the criminals, C is the best option.

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

The only thing we know for sure is the R^2 will increase. As you add explanatory variables, you will explain more. That's why we have adjusted R^2, which may fall as you add explanations.

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

Only the number of schools would obviously and inherently increase with population. A single school in a large population has very different implications than a single school in a small population.

The presence of a river (which would be a dummy variable) should not be adjusted for population; the river is either there or it is not. Dividing by the population makes little sense here.

Similarly, the percent of a city that's Asian American is already adjusted for population (number that's Asian Americans/number in the city) so "adjusting" it again would not be helpful.

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

Type of pet is a category—cat, dog, etc—and therefore cannot be expressed as a number. Unless you make it a dummy variable: one variable for cat, one for dog, and, perhaps, one for other (since these categories are not mutually exclusive, you need one dummy variable for each category).

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

20. Use Practice Exam 3 Data Set 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?

Upon running the regression you should output that looks like this:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.986117
R Square	0.972427
Adjusted R Square	0.965862
Standard Error	35.4861
Observations	27

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	5	932635.8	186527.2	148.124	1.23E-15
Residual	21	26444.53	1259.263		
Total	26	959080.4			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	61.57863	41.10265	1.498167	0.148972	-23.899	147.0563	-23.899	147.0563
Square feet, in thousands	27.32292	6.381342	4.281688	0.000331	14.05219	40.59364	14.05219	40.59364
Inventory value, in thousands	-0.08328	0.08069	-1.03213	0.313751	-0.25109	0.084521	-0.25109	0.084521
Advertising spent, in thousands	9.013352	2.651466	3.399384	0.002702	3.499326	14.52738	3.499326	14.52738
Number of competing stores in district	-11.3978	2.781184	-4.09817	0.000514	-17.1815	-5.61397	-17.1815	-5.61397
Number of families in the sales area, in thousands	18.57857	3.888056	4.77837	0.000101	10.49292	26.66423	10.49292	26.66423

a. *At 99.9% confidence, p must be less than 0.001. All variables expect Inventory and Advertising accomplish that.*
b. *This model is quite good. Adjusted R-squared is well above 0.9 and the F stat is large. This model explains a lot of the variation of profitability.*
c. *You would have to change Region to a dummy variable (0 for one region and 1 for the other region), then include it in the regression like a normal explanatory variable. Note you would only have to add one variable even though there are two regions.*

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?

The evidence you would need to demonstrate multicollinearity (or not) is a correlation table. Note profit should NOT be in that table; this is only about the explanatory variables.

	Square feet, in thousands	*Inventory value, in thousands*	*Advertising spent, in thousands*	*Number of competing stores in district*	*Number of families in the sales area, in thousands*

Square feet, in thousands	1				
Inventory value, in thousands	0.695668099	1			
Advertising spent, in thousands	0.392377882	0.472209627	1		
Number of competing stores in district	-0.69970287	-0.58481759	-0.220238621	1	
Number of families in the sales area, in thousands	0.838022882	0.828234129	0.456025334	0.796802911	1

a. *Looking at the correlation table, Number of families is highly correlated with Square feet and Inventory value; the absolute value of each pair's correlation coefficient of greater than 0.8.*
b. *The best thing to do is to drop the Number of Families variable. Since Square feet isn't strongly correlated with Inventory, you can eliminate all multicollinearity by just eliminating that one variable.*

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. *We have information on all the variables; we can use these to estimate recidivism value. 0.45-0.01*(17)+0.04*(3)-0.15*(1) = 0.45-0.17+0.12-0.15=0.25*

b. *Recall that a dummy variable is either zero or one. You can think of observed valued—the data—as percents. This observation has a 100% chance of recidivism (because that person did reoffend) and that observation has a 0% chance of recidivism (because that person did not reoffend). Predicted values will typically fall between zero and one, meaning you can use that to predict the chance they will reoffend. In the case of A, this ex-con has a 25% chance of reoffending.*
c. *This question is asking about the "marginal" effect, or the effect of a one-point change. In this case, a one-year age change results in the dependent variable falling by 0.01. But that doesn't translate into a good punchline; it's not in natural language. So instead we'd say "For every additional year of age the convict is, the chance of recidivism falls by one percentage point."Note this is percentage point and not percent; a 1% decrease would be 1% of a percent, the exact value changing based on how much we're talking about. But that's not what the linear regression tells us. Change is expressed in percentage points.*
d. *For this question, you need to find a confounding variable: something that could cause both the explanatory and dependent variables. For example, this regression claims that being on good behavior causes the chance of reoffending to fall by fifteen percentage points. Does that mean that being on good behavior "teaches" a convict to be more law-abiding? Maybe. Or maybe the kind of convict which gets good behavior is also the kind of convict who is unlikely to reoffend. Personality is our confounding variable.*

   You could also argue that same confounding variable applies to time in prison instead of the GOOD? variable. Since time in prison is correlated with the severity of the crime, it seems reasonable that the kind of people (based on personality or economic circumstances or upbringing) who commit minor crimes are less likely to reoffend. Thus a policy prescription—reducing how long everyone spend in jail—would not make sense. That said, it could make sense if we think of time in prison as normalizing, and thus encouraging, crime. Convicts are socialized into the way criminals operate and also provides a networking opportunity so less time in prison would reduce reoffend rates.

(All this points to how hard statistical analyses can be and remind us that this class is only the beginning. More advanced classes discuss how to isolate these effects and test for these complications.)

As the question mentioned, there are many possible answers. But as long as your answer causes (A) an explanatory variable, (B) the dependent variable, and (C) the causations run in such a way that resulting correlations could be mistaken for a direct connection (as in, if increasing X increase Y and decreases Z, that could be mistaking for increasing Y decreases Z but not for increasing Y increases Z), you have a good answer.

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.

Coefficient of Variation

$$CV_{sample} = \frac{s}{\bar{x}}(100)$$

Confidence interval for proportion

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

Optimal Sample Size

$$n = \left(\frac{z_{\alpha/2}\sigma}{ME}\right)^2$$

$$n = \left(\frac{z_{\alpha/2}}{ME}\right)^2 \bar{p}(1-\bar{p})$$

Critical z scores

Use =NORM.S.INV command

Confidence	α	$z_{\alpha/2}$	z_{α}
90%	0.1	1.645	1.280
95%	0.05	1.960	1.645
99%	0.01	2.576	2.330
99.9%	0.001	3.291	3.090

Critical t scores

Use T.INV or T.INV.2T commands

z-test

$$z_{\bar{x}} = \left|\frac{\bar{x} - \mu_{H_0}}{\sigma/\sqrt{n}}\right|$$

Proportion

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

t-test

$$t_{\bar{x}} = \left|\frac{\bar{x} - \mu_{H_0}}{s/\sqrt{n}}\right|$$

Adjusted R^2

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