Name: **KEY**
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. 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. **C** Average
2. **E** Dummy variable
3. **G** ε
4. **A** Omitted variable bias
5. **I** Observed value
6. **H** Percentage points
7. **B** 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.

1. *If you don't have a regression, your best guess of a variable's value for any observation is the average of that variable. For example, if I told you the average age of the class is 22.4 and then I asked you what Fernando's age is, your best guess would be 22.4.*
2. *Dummy variables are categorical: a zero or one (or a series of dummy variables if there are more than two categories). Employment status (unemployed or employed) is categorical so it's the only dummy variable there. (Economics students should note that there's a third category—not in the labor force—in which case employment status would have to be indicated with two dummy variables instead of one.)*
3. *The residual (ε) is what's "leftover" from the regression. It's the difference between what the regression predicts and what the value actually is. Since each observation should have a deviation from prediction (even if that prediction is zero), the number of residuals equals the number of observations. Try it yourself: next time you run a regression on Excel, ask it to include the residuals and you'll see a number of residuals equal to the number of observations.*
4. *A regression suffering from omitted variable bias doesn't have an important control. The lack of that control messes up the results, causing statistical significance where there shouldn't be and a lack of significance where there should.*
5. *If the residual is what's left over after prediction, then subtracting it from the observed value is the predicted value.*

6. *A percentage point is the difference between two percents; the difference between 8% and 10% is two percentage points, as is 50% and 52%. In contrast, percent relies on the starting value. Going from 8% to 10% is a 25% increase; going from 50% to 52% is a 4% increase.*
7. *If you have a regression, you can use traits (variables) you know about any particular observation to predict the dependent variable. For example, suppose I told you the average age of the class is 22.4* ***and*** *that the estimated line to predict age is AGE = 18 + 0.6*SEMESTERS (where SEMESTERS is the number of college semesters the student completed). If I asked you what Fernando's age was, given that he's completed two semesters of college, your best guess wouldn't be 22.4; it would be 19.2.*

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.

Observations with dummy variables can only be 0 or 1 but predicted variables are a different matter. Predicted values are the result of many different observations. Just like the average of a dummy variable is not limited to 0 and 1, a predicted dummy variable isn't either.

A predicted dummy variable is best interpreted as a percent chance. A value of 0.79 means there is a 79% chance that this person gets a newspaper delivered. Note this is the not the same thing as option A; you either get the paper delivered or you don't. You don't get 79% of a paper delivered.

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

Multicollinearity occurs when two independent variables are highly correlated, creating a redundancy. Since it takes two variables to have a high correlation, you only need to remove one of them to fix the problem.

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

The correlation coefficient does not incorporate causation; it does not claim that one variable is causing the other. It also does not indicate how much one changes if the other changes; a perfectly correlated set of observations with a positive gradual slope has the same correlation coefficient as a perfectly correlated set of observations with a positive steep slope: +1. The correlation coefficient only indicates the degree that two variables "move together."

A regression line using two variables can tell us if two variables are positively or negatively correlated because that would be revealed by the slope of the line. A negative coefficient would indicate negative correlation.

11. Imagine a regression with income (in dollars) predicting house size (in square feet). Suppose the coefficient of income is 0.052. If you changed the regression so that income is in thousands of dollars, what would the new coefficient bet?
    a. 0.000052
    b. 0.00052
    c. 5.2
    d. **52**
    e. It is impossible to tell without re-running the regression.

If you change income into "in thousands of dollars" then the observations would get a thousand times smaller. $50,000 would become $50, for example. That means the coefficient has to be a thousand times larger to compensate.

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

A residual is the difference between what the regression predicts and what observation actually is; it's the unexplained part of the regression. If the residuals form a normal distribution, then that means the bell curve centers at the average size of the residual, the average of which is always zero (the sum of all residuals in a linear regression also always equals zero). Some residuals are far from zero (either negative or positive) but most are nearby, forming a normal distribution.

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.

Female? and Salary, with a correlation coefficient of 0.0074, are least correlated because the correlation coefficient is closest to zero.

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

The negative correlation between the two implies that more mistakes is paired with less experience and vice versa. While it's possible that if you make a lot of mistakes you get fired and thus can't learn (option C), it seems odd to claim that there's no learning process going on to explain the correlation.

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

It is an assumption of linear regression but it occurs when the variance of the observed observations from the predicted line is consistent.

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

It's entirely possible to have an R^2 of 0.1 and five explanatory variables and 11 observations.

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1} = 1 - (1 - 0.1)\frac{11-1}{11-5-1} = 1 - (0.9)\frac{10}{5} = 1 - 1.8 = -0.8$$

In other words, these are completely different inputs into the equation. While R^2 can't be less than zero or greater than one, adjusted R^2 is not subject to those bounds.

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.

P-values indicate the lowest alpha you can cite and claim statistical significance. Since the alpha for 99% is 0.01, B is the answer.

Note that C is not correct. While 2.576 is the critical value for 99% significance, regressions use the t distribution, not the z distribution. If the sample size is low enough, a t-statistic of 3.000 may still not be statistically significant because there will be so few degrees of freedom. This is why it's best to look at the p-value; the p-value incorporates the degrees of freedom and does not require referencing a critical value.

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.

This regression is based on 20 years of sales data but the question refers to estimating 90 years after the observations end. It's unreasonable that the coefficient would be the same all that time, given how much change an industry can endure over the course of a generation, let alone four generations.

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

The ε at the end of the equation is the residual; it's not always there because it represents the difference between a predicted value and an observed value. A regression line that reflects the predicted value has no ε and a regression line that reflects the observed value has ε.

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?

When you run the regression, you get should get something like this:

Regression Statistics	
Multiple F	0.958371115
R Square	0.918475193
Adjusted I	0.902170232
Standard E	2.847420775
Observati	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressio	4	1826.883899	456.7209746	56.33102556	1.32083E-10
Residual	20	162.1561014	8.107805071		
Total	24	1989.04			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	39.33650517	5.381219269	7.309961405	4.56195E-07	28.1114785
Age	0.19640797	0.107669301	1.824177993	0.083103424	-0.028186256
Female?	-6.597941282	1.199835446	-5.499038474	2.20709E-05	-9.100754159
Experienc	-0.817071194	0.276255073	-2.957669465	0.007782035	-1.393329178
Salary	8.8159E-05	0.000175831	0.501385489	0.621573959	-0.000278618

a) *Because Female? has a p-value of just 0.000022, it's statistically significant at the 99.9% level (because it's less than 0.001). No other variables are significant at that level.*

   Experience is significant at the 99% level; its p-value is 0.0078, which is less than 0.01.

   Besides the variables that are already significant, nothing else is significance at the 95% level. Age gets close, but it's p-value is 0.0831 and it needs to be less than 0.05 for it to be significant at this level.

b) *The model is good as a whole because the F stat is large (note "large" is objective here as the significance F is less than even 0.001; F is significant at the 99.9% level). You could also note that R^2 is very high, but the F stat should be the focus of the answer as that's a better measure due to its objective standards of what's "good" or not.*

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) Yes, it has multicollinearity; Experience and Salary are highly correlated.

	Age	*Female?*	*Experience*	*Salary*
Age	1			
Female?	0.266105749	1		
Experience	0.415695676	0.052584673	1	
Salary	0.348830304	0.007388825	0.981878192	1

Note this correlation table does not include the dependent variable, Mistakes, because the dependent variable is not relevant in determining multicollinearity.

b) Either Experience or Salary, as these variables are redundant with one another, should be removed from the regression.

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) First, note that because Time is not statistically significant, it's coefficient is indistinguishable from zero. This doesn't mean it's definitely zero—we've failed to reject the null hypothesis, not accepted the null—but we have no evidence it's anything but zero. Thus, we should ignore the Time coefficient as so:

$$VIEWS = 5{,}000 - 75 * (20) + 90 * (2) + 700 * (1) = 4{,}380$$

b) *Adding ten cats should add (90*10 = 900) 900 views.*
c) *The punchline simply reports the coefficient: For every additional minute the video is, the number of views decrease by 75.*
d) *The most obvious answer to me is to note that the Cats variable is likely not linear. It's strange to think that the first cat would add the same number of viewers as the 100th cat. There are likely diminishing returns to cats: each additional cat added will be less and less valuable. You probably couldn't even tell the difference between a video with 99 cats and a video with 100 cats.*

   You could tell a similar nonlinear story about Length. Going from one minute to two minutes probably won't lose you any viewers, but going from 30 minutes to 31 minutes probably be quite noticeable.

   Note that the HD dummy variable can't be a good answer here; there's only two values the variable can be so you can't claim there's a nonlinear relationship.

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

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

Confidence interval for proportion

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

z-test

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

t-test

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

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%	5%	1.960	1.645
99%	1%	2.576	2.326
99.9%	0.1%	3.291	3.090

Critical t scores

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

Adjusted R^2

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