Name: **Key**

ECON/ACCT/BUSA 222—Bethany College

**Exam 03**

* There are 110 possible points on this exam. The test is out of 100.
* You have two hours 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.
* You are allowed a calculator for the exam. This calculator cannot be capable of storing equations. This calculator cannot double as a cell phone.
* Be sure to read all instructions and questions carefully.
* Remember to show all your work.
* Recall basic logic. “Water is wet” is a true statement. “Water is wet and leopards have stripes” is a false statement.
* You are allowed one 3” by 5” note card with the exam. You are allowed any information you deem important on it. You can also interpret this as having 30 square inches of paper (two 15 square inches, one for each side of the note card).
* *Please print clearly and neatly.*

**Part I: Multiple Choice.** *Choose the best answer to the following.*

4 points each.

1. Suppose you run a regression with an earner’s income predicting his/her children’s level of education. If you add the earner’s level of education as an explanatory variable, what problem, if any, will you get?
   1. **Multicollinearity**
   2. Heteroscedasticity
   3. Autocorrelation
   4. B & C
   5. None of the above / There is no problem

*The earner’s income and his/her level of education would be highly correlated. If two independent variables are highly correlated, that’s a problem with multicollinearity.*

1. What does R2 represent?
   1. The fraction of the variation the regression explains
   2. Explained Sum of Squares divided by Residual Sum of Squares
   3. Explained Sum of Squares divided by Total Sum of Squares
   4. **A & C**
   5. None of the above

*Option (C) is true by definition. Option (A) is what that means: the proportion of the variance explained by the model.*

1. Which of the following would best be represented with a dummy variable (or series of dummy variables)?
   1. Age
   2. Gender
   3. Color of shirt
   4. **B & C**
   5. None of the above

*Age is numerical; it doesn’t need a dummy variable. Gender and shirt color aren’t numerical so you’d have to term them into one variable (for gender) and several variables (for color of shirt).*

1. Suppose you ran a regression of AGE predicting INCOME (thousands of dollars). If your estimated line was INCOME = -1.5 + 1.2\*AGE + ε, how much more income do you expect to make for every *two* years you age?
2. -0.3
3. 0.9
4. 1.2
5. **2.4**
6. It is impossible to tell given the information provided

*Two times the coefficient—1.2—is 2.4. Since this concerns a change (every two additional years), we don’t consider the intercept.*

1. Suppose you ran a regression and found heteroscedasticity. What should you do first?
2. Start over with new variables
3. Drop as few variables as possible to remove the heteroscedasticity
4. **Check how significant your explanatory variables are**
5. A & B
6. None of the above

*Heteroscedasticity doesn’t influence the significance too much. If your variables are strongly significant, at this level we can assume you’ll still be fine. (At a more advanced level, we could employ robust standard errors to fix that…but Excel isn’t capable of that correction to my knowledge.)*

1. As you add explanatory variables to a regression, what always happens?
   1. Your F-stat falls.
   2. The p-values of the variable(s) you started with decrease.
   3. **R2 increases.**
   4. A & B
   5. None of the above

*The only thing we know for sure is the R2 will increase. As you add explanatory variables, you will explain more. That’s why we have adjusted R2, which may fall as you add explanations.*

1. 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?

* 1. **He has too many variables.**
  2. He will get heteroscedasticity.
  3. He has the same player having more than one favored strategy.
  4. A & C
  5. 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 as it is would introduce multicollinearity. And yes, there are Rock-Paper-Scissors tournaments.*

1. Suppose you ran a regression with the natural log (ln) of income (LN(INCOME)) predicting the square footage of that person’s apartment (SQUARE FT) and the result is: SQUARE FT = 600 + 130(LN(INCOME)) + ε. According to your model if a person’s income increased by 1%, how much larger do you predict their house to be?
2. **131.3**
3. 731.3
4. (1.01)130 – 1
5. 600 + (1.01)130 – 1
6. It is impossible to tell given the information provided

*To find this answer, simply remember the equation from lecture 9 for when only X is logarithmic: ΔY = (1 + %ΔX)β. So we take (1.01)130 = 131.3. Note we do not need the intercept.*

1. When testing if several different groups have the same mean, why don’t we just take the t-statistic for each pair?
   1. You always use the z-test, not the t-test.
   2. It can take too long.
   3. We might have Type I error.
   4. **B & C**
   5. None of the above

*If we have lots of pairs of variables to test, not only is it time consuming but we’re likely to get two means that are different simply by chance. This is sometimes called the multiple comparisons problem; ANOVA fixes it.*

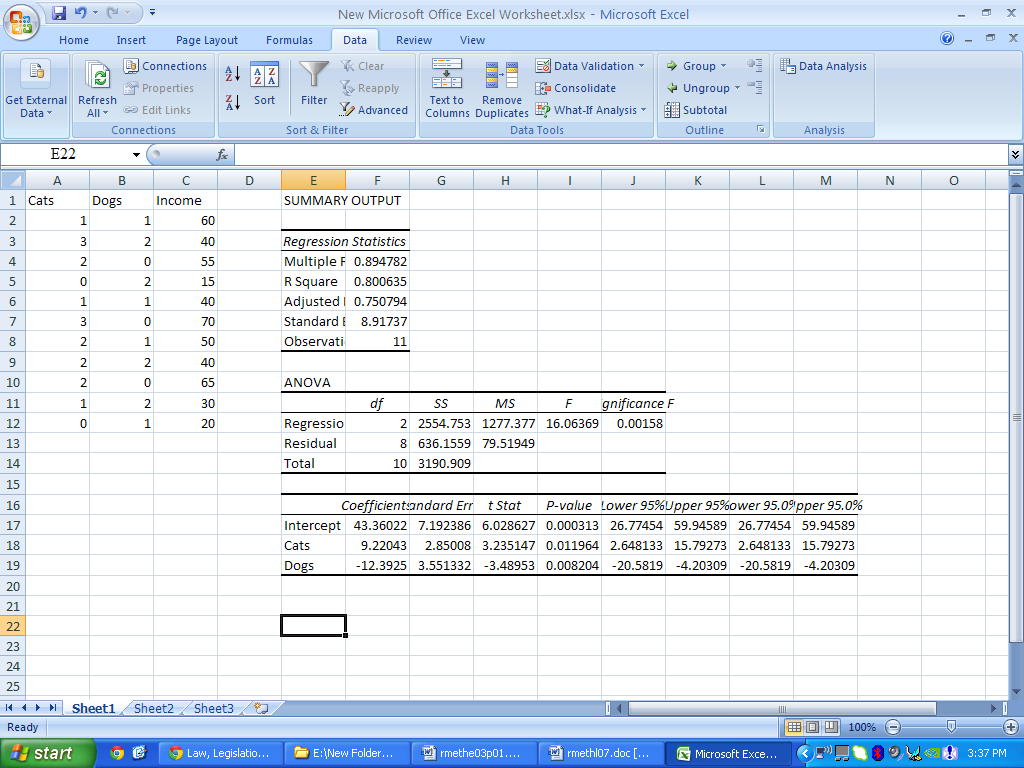
1. Suppose the coefficient for an independent variable—number of cars per person—is 1.5. Suppose you changed the independent variable so it considers the number of cars per 100 people. What is the coefficient now?
   1. **0.015**
   2. 0.15
   3. 15
   4. 150
   5. It is impossible to tell with the information 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.*

1. Which of the following is an example of a scalar?
2. A country’s population
3. The number, pi.
4. 10 million
5. **B & C**
6. None of the above

*Both options are constants (even if pi is irrational), which is what you need for a scalar. Only option A is not a constant since its value changes depending on the observation you are considering.*

1. Consider the following output from Excel, with cats and dogs predicting income in thousands of dollars (the data here are made up):



According this output, what is the estimated line (all values are rounded to three decimal places)?

1. INCOME = 0.801 + 0.751 \*CATS + 16.064\*DOGS + ε
2. INCOME = 0.002 + 0.012\*CATS + 0.008\*DOGS + ε
3. INCOME = 16.064 + 9.220\*CATS – 12.393\*DOGS + ε
4. **INCOME = 43.360 + 9.220\*CATS – 12.393\*DOGS + ε**
5. None of the above

*Here, β0 is the intercept coefficient, β1 is the CATS coefficient, and β2 is the DOGS coefficient. Considering previous lectures you might also be suspect if this data were real—why would cats make you wealthier and dogs make you poorer? Reverse causation is a possible answer. I suspect (which is why I make the data turn out this way) that wealthier people tend to own cats since they are lower maintenance and wealthier people are probably busier. Dogs, while harder to take care of, act as a relatively low-cost security system which makes them more attractive to lower-income people. This is armchair reasoning, of course, and it’s possible both of these effects are not important enough to influence the relationship. It could go the other way, for all I know.*

**Part II: True/False.** *Answer true or false, and justify your answer.*

10 points each.

1. A linear regression with an “error term” (a.k.a. residual) indicates someone made a mistake.

*False. The “error term” is better thought of as the explained part of the regression. All regressions have them; if you don’t have one—if you can explain all variation in your dependent variable—that’s when you probably did something wrong.*

1. If your estimated line is Y = 0.5 – 5X1 + 2.2X2 + ε, and X is statistically significant, then that means for every additional unit of X1 you get 5 additional units of Y.

*True. The level that Y changes equals to the coefficient of the independent variable in question, even more multivariable analysis. One can see this by subtracting the difference of two Ys and the two equations and simplifying such that: ΔY = βΔX, where β is the coefficient. If ΔX is 1, then ΔY is β.*

1. If Harry Kim is interested if ten different airlines have the same portions of flights take off late, he’d definitely want to use a two-factor ANOVA test.

*False. While he’d definitely want to use ANOVA, it’s not clear if he’d prefer two-factor or one-factor. Indeed, since we don’t have any source to consider the variation with his samples classes of different airlines, we can’t tell if we’d like to use that information (e.g. the days they took off) to do a two-factor test.*

**Part III: Short Answer.** *Answer the following.*

16 points each.

1. By now you should have outlined your regression model, even if it is a first draft. Using all your variables, write your regression model in the general form we discussed in class. (HINT: The general form used betas.)

*The answer depends on your model but the basic formula is:*

*Y = β0 + β1X1 + β2X2 +…+ βkXk + ε*

*You’d just need to replace the Xs and the Y with your independent and dependent variables, respectively.*

1. Francis runs a regression with a sample of 33 and with 17 explanatory variables (including the intercept, which is included in *k*). His R2 is 0.70. What is his adjusted R2?

*Recall the equation for adjusted R2:*

*Or, 1 – (1 – 0.70)(33 – 1)/(33 – 17) = 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.*