# **Additional Study Questions**

# **Includes Omitted Variable Bias, Squared Terms, Interaction Terms (different slopes)**

# **The Value of Advertising**

***Tangerine***sells a variety of consumer electronics, including the U-Phone and U-Pad. Their marketing department is analyzing their adverting data. Their dataset includes:

* Product Name
* Total Sales (in dollars)
* Dollar Amount Spent on Advertising (“***Ads”***).

Each observation is monthly data on a single product. In the past, ***Tangerine*** determined the amount to spend on advertising (***Ads***) based on the results of focus group ratings (“***Focus Ratings***”). Tangerine believes ***Focus Ratings*** are a good measure of how useful the product is, since previously they had measured that the higher the ***Focus Ratings***, the higher the ***Total Sales***.

However, Tangerine , has stopped collecting information on ***Focus Group Ratings*** because it is too costly. Instead, Tangerine wants to choose the amount to spend on advertising based on data analysis of how ***Ads*** affect ***Total Sales***.

The marketing department proposes the following model of sales:

1. ***Total Sales*** =a0 + a1\****Ads***

where a0 and a1 are parameters to be measured by regression.

They then estimate the regression, which gives:

1. ***Total Sales=* 600 + 8.90*\*Ads***

The engineering division disagrees, and says that Total Sales also depends on how useful the good is to consumers which is measured by focus-group ***ratings***. They argue that a better model for total sales is:

1. ***Total Sales***= b0 + b1\****Ads*** +b2\****Focus Ratings***

where b0 ,b1 and b2 are parameters to be measured by regression.

Unfortunately, the marketing department does not have the ***Focus Ratings*** data they used previously that would allow them to run that regression. But Tangerine really wants to know the value of B1 in order to optimally choose ***Ads***. The marketing department does know the relationship they had previously used to decide ads based on Focus Ratings:

1. ***Ads*** = -25 + 5\****Focus Ratings***

Which can be rearranged and rewritten as:

1. ***Focus Ratings*** ***=*** 5.0 + .2**\**Ads***
2. Which of these equations (1, 2, …) is the limited model? **1 (Total Sales =a0 + a1\*Ads) or 2 -- the same model with coefficient values : Total Sales =600+8.9\*Ads**

Which of these equations (1, 2, …) is the full model?  **3** **Total Sales =B0+B1\*Ads+B2\*Ratings**

Which of these equations (1, 2, …) is the background model? **5 Ratings=5+0.2\*Ads** **(or 4, which is the same equation written differently)**

1. Based on all of this evidence, do you think that if *Tangerine* increases ***Ads*** by $1, their ***Total Sales*** would go up by $8.90? If not, do you think total sales would go up by more or less than $8.90? Can you explain why intuitively?

**Ads are positively correlated with ratings. We also think ratings should have a positive effect on sales. Thus, the effect of Ads in the limited model (8.9) also captures some of the positive effect of ratings. If you increase spending on Ads, the original ratings won’t change, and so the total sales will go up by less than $8.90.**

**The coefficient B1 in the full model is a better estimate for the causal effect of advertising than the coefficient in the limited model. We are not given B1.**

**Recall our equation C1 = B1+B2\*A1 , where the Cs are from the limited model, Bs from the full model, and As from the background model. Here we have**

**8.9= B 1+ B2\*0.2**

**And since B2 is positive (ratings increase sales), then B1 must be below 8.9.**

1. Assume that from other studies, you know b2=40 (where b2 is the effect of ***Focus Ratings*** on ***Total Sales*** holding constant ***Ads***). Then, holding constant Focus Ratings, what effect do Ads have on Total Sales? In other words, what is b1? **Show your calculations (handwritten).**

**8.9= b1 + b2\*0.2**

**We are told b2=40, so substituting in, we have 8.90 = B1 + 40\*0.2 => B1 = 0.90**

**Holding Ratings constant, if Tangerine increases Ads by $1, their Total Sales would go up by $0.90.**

**The coefficients in the full model are (based on what has given): Total Sales = B0 + 0.90\*Ads + 40\*Ratings**

1. Based on the value of b1 that you calculated, will increasing Ads by $1 raise or lower *Tangerine’s* profits? Explain.

**Answer: NO. Our best estimate is that spending $1 more on ads increases revenue by only $0.90, holding constant product ratings. So the extra revenues don’t even cover the cost of the ads! (Plus, we have to subtract away the cost of making the extra product.)**

**2. Predicting Weight**

You are hired by the Department of Health and Human Services to help understand the determinants of the obesity epidemic in the US. You are given data on more than 20,000 individuals, aged 22-60. You have the following information:

* Weight in pounds
* Height in inches
* Gender
* Age
* Immigrant status
* Marital status

With this data in hand you start by running several regression models where the dependent variable is weight. The results are reported in the table below: (See Classnotes Chapter 22 on how to read this table.) **Standard errors in parentheses**

|  |  |
| --- | --- |
|  | Dependent Variable |
|  | I. Weight in pounds |  | II. Dummy for Overweight |
|  | (A) | (B) | ( C) | (D) |  | ( E) | (F) |
| Intercept | 179.49 | -183.43 | -140.46 | -135.62 |  | 0.47 | 0.25 |
|  | (0.30) | (3.97) | (6.43) | (6.43) |  | (0.004) | (0.02) |
|  |  |  |  |  |  |  |  |
| Immigrant Dummy | -16.71 | -6.35 | -7.93 | -7.87 |  |  |  |
|  | (0.76) | (0.66) | (0.67) | (0.67) |  |  |  |
|  |  |  |  |  |  |  |  |
| Height in Inches |  | 5.39 | 4.3 | 4.29 |  |  |  |
|  |  | (0.06) | (0.08) | (0.08) |  |  |  |
|  |  |  |  |  |  |  |  |
| Male Dummy |  |  | 12.98 | 6.32 |  | 0.15 | 0.2 |
|  |  |  | (0.66) | (0.89) |  | (0.006) | (0.03) |
|  |  |  |  |  |  |  |  |
| Age |  |  | 0.95 | 0.92 |  |  |  |
|  |  |  | (0.19) | (0.19) |  |  |  |
|  |  |  |  |  |  |  |  |
| Age Squared |  |  | -0.008 | -0.007 |  |  |  |
|  |  |  | (0.002) | (0.002) |  |  |  |
|  |  |  |  |  |  |  |  |
| Married Dummy |  |  | -2.73 |  |  |  |  |
|  |  |  | (0.49) |  |  |  |  |
|  |  |  |  |  |  |  |  |
| Male x Married  |  |  |  | 2.89 |  |  |  |
|  |  |  |  | (0.73) |  |  |  |
|  |  |  |  |  |  |  |  |
| Female x Married  |  |  |  | -7.81 |  |  |  |
|  |  |  |  | (0.66) |  |  |  |
|  |  |  |  |  |  |  |  |
| Female x Age |  |  |  |  |  |  | 0.005 |
|  |  |  |  |  |  |  | (0.0004) |
|  |  |  |  |  |  |  |  |
| Male x Age |  |  |  |  |  |  | 0.004 |
|  |  |  |  |  |  |  | (0.0005) |
|  |  |  |  |  |  |  |  |
| Adjusted R2 | 0.0195 | 0.2697 | 0.2882 | 0.2919 |  | 0.0221 | 0.0312 |

1. Explain ***exactly*** why the coefficient of immigrant goes from more negative to less negative from column (A) to column (B).

**Given that it goes from more negative to less negative, this means that the limited model coefficient, c1, was underestimated and thus there is a negative bias.**

**(The exact bias is the difference in the coefficients or -16.71 –(-6.35) = -10.36**

**Because b2, the coefficient for height, is positive in the fuller model (column B), this means that a1, or the relationship between immigrant status and height is negative. Thus, immigrants are, on average, shorter than non-immigrants, and some of this effect was captured by the limited model.**

**Mathematically, the bias a1 b2 =<0 and b2 >0, then a1 <0**

**In pictures:**

 **Immigrant -6.35 (neg)**

 **Weight**

 **Bias neg**

**Must be neg**

 **height 5.39 (pos)**

1. In ONE SENTENCE interpret the coefficient of Male (6.32) in regression (D).

**Unmarried males, on average, weigh 6.32 pounds more than unmarried females controlling for other right side explanatory variables (immigration status, height, and age).**

1. What is the predicted difference in weight between a married female and an unmarried male, who have the same age, immigrant status, and height. ***Show your handwritten calculations:***

**Married female minus unmarried male:**

**12.98\*0 - 2.73\*1 minus 12.98\*1 - 2.73\*0=-2.73 - 12.98 = -15.71**

1. Based on your models, and assuming no differences in other variables besides age, whom do you predict to have a larger increase in their weight (in pounds) from this year to the next on average?
	1. QM222 students
	2. Their professors
	3. Same for both groups
	4. Cannot tell

Explain why this is your answer, showing any calculations used (handwritten):

1. **QM222 students**

**Explain why this is your answer, showing any calculations used (handwritten):**

**QM222 students are younger…. Lets say on average 20 while professors are on average 45.**

**The change in weight using calculus is dweight/dage = .95 - .016 age**

**The higher the age, the lower the slope. Or, plug in numbers to this waution and get:**

 **For students aged 20, dweight/age= 0.63. For faculty aged 40 dweight/age= .31**

**So students gain more weight each year.**

**You can also do this by isolating the age terms and plugging in numbers, first 20 then 21. And first 40 then 41**.

1. Using regression C, would you say that the relationship between weight and age is linear or nonlinear? If non-linear, sketch what it looks.

**Nonlinear since the coefficient on Age squared is significant. It looks like this. Extra: I can calculate the slope .95 + 2\* (-.008) \* Age and set the slope equal to zero which is where this curve peaks. This turns out to be at the age of 59.4**

1. (4 points) Being overweight increases the risk of complications during pregnancy. The department of health wants to know if more than 50% of women are overweight. Based on model (E), what would you tell them?

**95% CI: Female average ± 2\*SE = .47 ± 2\*.004: (.462, .478)**

**Because .50 falls outside the confidence interval, we reject the null hypothesis that 50% of woman are overweight. In other words, we are very confident that fewer than 50% of women are overweight.**

1. (4 points) In ONE SENTENCE interpret the coefficient on Male x Age in regression (F).

**On average, for each additional year of age, the probability that a male is overweight increases by .4 percentage points.**

1. (4 points) One of the drawbacks of the Linear Probability Model is that it sometimes predicts probabilities lower than 0 or higher than 1 for the relevant population (those in the sample). Regression (F):

a. Predicts a negative probability for some in the sample

b. Predicts a probability greater than 1 for some in the sample

c. Both (a) and (b) are true

**d. None of the above (Note that there are no negative coefficients to push the probability negative, and at the oldest age in the sample of 60 the two age terms each cannot be more than 30, so that the the prediction (including all variables) can’t go anywhere near 1.**

3. (5 points) Teaching hospitals are associated with medical schools, and both medical professors and doctors-in-training (who have already completed the 4 years of medical school instruction) work there. For instance, in Boston, the Boston Medical Center is the main teaching hospital for BU, the Tufts Medical Center is the main teaching hospital for Tufts, and Mass General, Beth Israel-Deaconess, Brigham’s and Boston Children’s Hospital are all teaching hospitals for the Harvard Medical School. These hospitals are known for having the most advanced technology and cutting edge treatments for many rare diseases and cancers.

Researchers have found that the likelihood of people admitted to a teaching hospital dying while at the hospital is significantly higher than the likelihood of dying for people admitted to other hospitals. Some people conclude from this that they should avoid going to teaching hospitals. Why is this likely to be a wrong conclusion that could lead to more people dying? Answer in 1 or 2 sentences.

**The most gravely ill people with a lower chance of survival are going to go to the hospitals with the best resources (teaching hospitals) out of necessity in order to improve the chances of survival. This difference in selection will bias the statistics.**

**We are talking about the likelihood of dying among all who were admitted to each type of hospital. That’s the right thing to compare.**

**Part II: Education, Siblings and Criminal Behavior**

The National Longitudinal Study of Adolescents to Adult Health (AddHealth) is a nationally representative survey that followed a group of people from when they were adolescents to when they were adults. The following analysis is from the 2008 AddHealth survey when the sample was aged 25 to 34.

In the attached regressions, we use the following variables:

Education: Years of education (e.g. 12 is high school, 16 is college, 18 is masters, 20 is PhD/other doctorate)

Siblings: Number of siblings the person had. (Siblings refer to both brothers & sisters)

SiblingsSq: The square of Siblings

Male: An indicator/dummy variable for gender. (If male, male=1; if female, male=0)

Arrested: An indicator/dummy variable for whether the person was ever arrested.

Jailed: An indicator/dummy variable for whether the person was ever jailed. (No one was jailed who

 wasn’t also arrested.)

Using this data, we have run regressions where Education is the dependent variable. The regressions are listed in the Part II Table at the end of this test. Use these regressions to answer the following questions:

1. (5 points) Use Regressions 1 and 2 to answer this question. Which of the following statements is true? CIRCLE ONE:

**On average, men have fewer siblings than women do.**

Explain how you arrived at your conclusion, showing any calculations that you used to answer this question. If we cannot tell, say what information you would need to figure it out.)

**-0.55= -0.586+ bias, which implies that the bias is positive. Positive bias – a1\*b2**

**The bias is the product of two components: 1) relationship between siblings and years of education (a1) and 2) relationship between siblings and male (a1). Since 1) is negative, we know**

1. (4 points) In ***common sense words***, can you explain why the coefficient on arrested is a much less negative number in regression 6 than in regression 4? (1-2 sentences)

**Regression 4 is missing jailed. What bias does this create?**

**An individual cannot be jailed without having been arrested first , and jail time served is also negatively associated fewer years of education. Thus by omitting “jailed” in Regression 4, we have overestimated the negative relationship between arrest and years of education.**

1. (4 points) Expert A looks at these regressions and claims that being jailed is very bad for youth since – if they get jailed – they end up getting less education and therefore have fewer opportunities to succeed in life. Expert A believes that it would be good policy if fewer arrested youth who were still in school (high school or college) were not jailed but instead got probation (i.e. not put in jail but followed carefully, with a lot of supervision by both police and the school). What evidence in these regressions might support his opinion? (1-2 sentences)

**The coefficient on jailed (-.860) is negative and significant in Regression 6. . This means one can be 95% confident that holding gender, siblings and arrested constant, being jailed is negatively associated with years of education. In other words, people who were arrested – of the same gender and with the same # of siblings – get .860 fewer years of education (|t|>8). Expert A is interpreting this negative coefficient as causal.**

Which regression(s) did you use to answer this question? CIRCLE ONE OR MORE:

Regression 1 Regression 2 Regression 3 Regression 4 Regression 5 **Regression 6**

1. (5 points) Expert B disagrees. She believes that there is an important missing (omitted) variable in regression 6 and argues that if you added this variable, the coefficient on jailed would fall dramatically in absolute value and become insignificant. Can you think of a variable that is missing from regressions 5 and 6, is likely to have a causal effect on education, and would drastically lower the coefficient on jailed in absolute value?

Omitted Variable:  **drug use THIS IS JUST AN EXAMPLE OF A REASONABLE ANSWER**

Explain why you think that adding this variable will lower the coefficient on jailed. (1-2 sentences)

**We are looking for a factor that would have a negative bias when it is missing from the education equation.**

**People who use a lot of drugs tend to get less education (b2). People who use a log of drugs are more likely to get jailed (a1). This causes a negative bias in regression 6.**

**Part II Table of Regressions. Dependent Variable: Years of Education**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|   | (1) | (2) | (3) | (4) | (5) | (6) |
| male | -0.550 | -0.586 | -0.591 | -0.311 | -0.375 | -0.292 |
|   | (0.060) | (0.059) | (0.059) | (0.059) | (0.058) | (0.059) |
| siblings |   | -0.178 | -0.321 | -0.291 | -0.289 | -0.283 |
|   |   | (0.012) | (0.027) | (0.027) | (0.027) | (0.027) |
| siblings sq |   |   | 0.013 | 0.012 | 0.012 | 0.012 |
|   |   |   | (0.002) | (0.002) | (0.002) | (0.002) |
| arrested |   |   |   | -1.086 |   | -0.633 |
|   |   |   |   | (0.066) |   | (0.086) |
| jailed |   |   |   |   | -1.366 | -0.860 |
|   |   |   |   |   | (0.081) | (0.105) |
| intercept | 14.462 | 14.992 | 15.223 | 15.336 | 15.259 | 15.312 |
|   | (0.040) | (0.054) | (0.067) | (0.066) | (0.065) | (0.065) |
| R-squared | 0.0164 | 0.0553 | 0.0616 | 0.1099 | 0.1119 | 0.1214 |
| Adj. R-Squared | 0.0162 | 0.0550 | 0.0610 | 0.1092 | 0.1112 | 0.1206 |
| # Observations | 5067 | 5067 | 5067 | 5067 | 5067 | 5067 |
| standard errors in parentheses |  |  |  |  |

 **Question 6:**

Executives at a major financial company are trying to model which households own stocks. They collected data for a national sample of households from around the country. The data they collected includes:

own\_stock: whether or not the household owns any stocks.

college: whether the most educated person in the household completed college

highschool: whether the most educated person in the household completed high school but not college

wealth: the total amount of wealth in millions of dollars (includes value of real assets

 like houses; of financial assets like bank accounts, bonds excluding stocks; and

 subtracts out amount the household owes including loans, mortgages etc.)

house: whether or not the household owns their own house

number of adults: the number of adults in the household.

They ask you to model who owns stocks, so you run a set of regressions with “own\_stock” as the left hand side (Y) variable.

1. You first run Regression 1 at the end of this test. Without any statistics terms or jargon, what does the coefficient 0.3322 tell us? The best answers will be concise, exact, intuitive and in commonplace words.

**On average, the probability of a household of which the most educated person completed college but not high school owns stock is 0.3322 higher than the probability of a household of which the most educated person did not complete college.**

1. What is the difference in the probability that someone with only a high school degree owns stock compared to someone with a college degree? For full credit, show the calculations used to get this answer.

Difference: \_\_\_\_\_\_\_\_0.20\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Calculations:

**Using regression one, difference = 0.3322 - 0.1296**

1. You know that people working in the financial industry tend to own more stock and you know that people working in the financial industry all have at least a college education. If you collected data and made a dummy of whether the person worked in the financial industry and added that dummy to the variables in Regression 1, what would happen to the coefficient on college? (CIRCLE ONE)

COEF. INCREASES COEF. INCREASES **COEF. DECREASES**  COEF. DOESN’T CHANGE WE CAN’T TELL

Explain exactly how you know:

**Bias: College increases 🡪 financial industry increases 🡪 stock increases**

**Bias is positive.**

**So when you add the financial industry variable, the coefficient on college decreases because it takes away some positive bias.**

1. Would it be possible for Regression #1 to predict a probability greater than 1?

 CIRCLE CORRECT ANSWER:

YES **NO**

If you answer NO, why? If you answer YES, give an example that predicts a probability greater than 1.

**NO: The highest value would be if they completed college = 0.1086 + .3322 = .4408**

1. The R-squared in regression 1 is 0.11. Does this means that the education level is not an important factor in determining stock ownership? Explain your answer in 1 sentence.

**No. R-square only states the percentage of variance that could be explained by the variables, and t-stat shows whether the variable is important which in this case is significant.**

1. You then run the additional regressions #2 through 4 as well. Which of these four regressions should you use to predict stock ownership? (CHOOSE ONLY BETWEEN #1, #2, #3, #4)

Regression #\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**4**\_\_\_\_\_\_\_\_\_

Why? (1 correct reason is sufficient)

**Highest adj Rsq, Lowest SEE**

1. Consider 2 households with the same levels of education and wealth. If one also owns a house, on average how much more (or less) likely is it that that household also owns stocks? Give the 68% confidence interval for this estimate. (CHOOSE ONLY BETWEEN regressions #1, #2, #3 and #4.) For full credit, show how you get your answers.

How much more likely? \_\_\_\_\_\_\_\_\_\_**16.3 percentage points**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

68% confidence interval: \_\_\_\_\_\_\_\_\_**14.62 - 18.62 percentage points**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Show how you got your answers:

**Use regression three because we only know education level, house ownership, and wealth. The coefficient on house ownership is 16.3%. The 68% CI of the coefficient is:**

**= 0.1634 + (0.017169)\*1 = 0.180569 = 18.06 percentage points**

**= 0.1634 – (0.017169)\*1 = 0.146231 = 14.62 percentage points**

Which regression did you use to answer this question and why?

Regression # 3 Why?

**We only know education level, house ownership, and wealth; so regression 4 cannot be used. Among regression 1 – 3, regression 3 is the best one as it has higher adjusted R-square.**

1. You believe that owning your own house has an impact on owning stock. Based on the regression chose in part f, what is the probability that you are wrong?

Probability: **basically 0 (1.07\*10-18 )**

1. Based ONLY on regressions #1, #2, #3 and #4 (and not intuition), which of the following is true? Circle the correct statement and explain exactly how you know. Which regression(s) did you use to answer this question?

CIRCLE ONE:

**Students who graduated from college own more net wealth than those who didn’t finish high school.**

* Students who graduated from college own less net wealth than those who didn’t finish high school
* We lack the information to tell which of the above statements is correct.

Explanation:

**The coefficient of college in regression 2 (representing direct effect) is smaller than coefficient of college in regression 1 (representing combined effect). That means the bias is positive. Also, we know that the direct effect of wealth on stock ownership is positive (from the coefficient in regression 2), so the correlation between college and wealth must be positive.**

**Bias due to College increases 🡪 net wealth increases 🡪 stock ownership increases**

Regression(s) used in answer: **1 & 2**\_

1. You next run Regression #5 which takes Regression #3 and replaces wealth with two interaction terms made by multiplying the dummy for owning a house times wealth and a dummy for not owning a house times wealth. The coefficient on the house\*wealth is .01152 and on nohouse\*wealth is .56037. Explain in regular, non-statistical words exactly what we learn from each of these and from the difference between them, and WHY you think there is a difference between them. The best answers will be concise, exact, intuitive in commonplace words.

**If you have no house, the more wealth you have, the more likely you will put it into stocks.**

**If you have a house, more wealth will not make it more likely (t=.06) to put this extra wealth in stocks, probably because you’ll put it in your house.**

**(Note however that people who have a house *are* more likely to also have stocks, as you can see from the coefficient of house.)**

**SEE QUESTIONS FOR THE REGRESSIONS**