**Applications: Answers to Questions**

**CRITICAL ANALYSIS OF DATA-DRIVEN CONCLUSIONS**

**Case #1:**

* 1. The active predictions are: 1) Improving first serve percentage from less than or equal to 75% to above 75% will change the winning percentage from 76% to 93%; 2) Increasing the number of times at net from less than or equal to 30 to above 30 will change the winning percentage from 78% to 91%.
  2. The treatments are: first serve percentage above 75% and number of times at the net above 30. The outcome associated with both treatments is the probability of winning the match.
  3. WinMatchi = α1 + β1Serve75i + U1i & WinMatchi = α2 + β2Net30i + U2i. For the former equation WinMatch is dichotomous, equaling one if Serena wins the match, and Serve75 is dichotomous, equaling one if Serena’s first serve percentage > 75%. For the latter equation, Net30 is dichotomous, equaling one if Serena goes to the net more than 30 times.

1. 1. The key assumptions are: 1) The prior matches from which these statistics were generated are effectively a random sample of matches Serena typically plays and 2) Serena randomly chooses whether to serve above 75% and whether to come to the net more than 30 times.
   2. For serve percentage, Serena may serve more aggressively on her first serve against stronger opponents, hoping to get some needed easy points; hence she non-randomly chooses her serving strategy based on opponent quality. For times at the net, Serena may come to the net more often against weaker opponents. Such opponents will likely have a harder time hitting the ball past her, and this strategy can help end the match faster. Hence, Serena non-randomly chooses her net approaches based on opponent quality again.
   3. Yes, it would change, in that the non-randomness of the treatment could not be due to opponent quality. However, each of the treatments could still be non-random even with the same opponent, since any given opponent may be playing well or poorly on any given day. To illustrate, let’s call the opponent Martina. Serena may serve more aggressively on her first serve if Martina is playing well and less aggressively if Martina is not playing well. Similarly, Serena may come to the net more often if Martina is not playing well, and less often if Martina is playing well. Here, Serena non-randomly chooses her serving strategy and net approaches based on how well that particular opponent is playing at that time.
2. Yes, then they are being used for passive prediction. If as observers, we see Serena having a first serve percentage above 75% or that she approaches the net more than 30 times – even if we don’t think those strategies are randomly chosen – we can use the figures in Table A1 to predict the likelihood of Serena winning (still assuming the prior matches from which the statistics were generated are effectively a random sample of matches Serena typically plays).
3. Perhaps the most plausible example would involve Serena’s approaches to the net. Suppose Serena decided to randomize her strategy for each match (e.g., via a coin toss), where a realization of heads means she plays serve and volley (meaning many approaches to the net) and a realization of tails means she plays at the baseline (meaning few approaches to the net). In this case, the randomness of Serena’s net strategy means there cannot be any feature of a match (e.g., opponent quality or how well an opponent is playing) that affects whether Serena wins that is systematically varying with her net strategy. Consequently, we can conclude (with a large enough “random sample”) that the difference in win likelihood is, on average, the effect of her net strategy.

**Case #2:**

1. The treatment is switching and the outcome is money saved on car insurance.
2. The people who receive the treatment (i.e., people who switch) *chose* to switch; they did not get randomly selected and forced to switch.
   1. No. The effect of the treatment on the treated (ETT) is the effect of switching on savings for those who switched. Here, the people who chose to switch are likely those people who found switching to be most financially beneficial; that is, they are the people for whom switching resulting in the most savings. Hence, this group experiences a larger treatment effect compared to the average, meaning ETT > ATE.
   2. Yes. Even though there is non-random treatment assignment, the specifics of this problem imply that Selection Bias is still zero. To see this, first note what Selection Bias means in this context. We need to compare the expected outcome from not receiving the treatment for those who received the treatment to the expected outcome from not receiving the treatment for those who did not receive the treatment. In this context, not receiving the treatment means that a person did not switch insurance. And if switching doesn’t occur, there are no savings. Hence, these two values both equal 0 – both switchers and non-switchers could expect to save nothing if they did not switch.
   3. Savingsi = α + β\*Switchi + Ui. Here, Savings is the amount of money saved by person i, and Switch is a binary variable equal to one if person i switches insurance and 0 otherwise.
   4. The intercept (α) must be zero, following the logic of 3b above. Since Switch is a binary treatment, α is the population mean of Savings when Switch = 0 (i.e., when a person does not switch). If a person does not switch, that person cannot save. Therefore, we have α = 0.
   5. The slope is the difference in the mean of Savings when Switch equals 1 and the mean of Savings when Switch equals 0 (i.e., ). The first piece is $582 and the second piece is $0, so the difference is $582. Therefore, the estimated slope will be .
   6. Since people freely choose to switch insurance, it is very likely that unobservables that affect a person’s savings (U) are correlated with the decision to switch (Switch). Hence, U and Switch are correlated, meaning we violate one of the key assumptions (i.e., U and Switch are uncorrelated) needed for us to interpret our regression results as causal. As an example, an insurance company might need more young people to balance its risk portfolio, and so offers people under 35 a good price. Here, age is then in the U term, and those who are young will be more likely to switch, creating a positive correlation between U and Switch.
   7. Consider an unobserved variable, say V, that is in the error term. If V is positively related to savings, then V and Switch are likely positively related as well (people will tend to switch when they get a high value for V and hence a high amount of savings); consequently, the bias will be positive in this case (product of two positive correlations). If V is negatively related to savings, then V and Switch are likely negatively related as well (people will tend not to switch when they get a high value for V and hence a low amount of savings); consequently, the bias will be positive in this case as well (product of two negative correlations). Therefore, it is likely the bias is positive in general. That is, the savings experienced by the switchers overstates the savings the average person should expect from switching.

**Case #3:**

1. The treatment is the price of cereal and the outcome is cereal revenue.
2. To simplify this discussion, we can consider low prices as “treatment” and high prices as “no treatment”…
   1. Revenues in Texas may be higher than Ohio even when both charge high prices (i.e., “no treatment”) due to Texas residents having stronger preference for cereal compared to Ohio residents.
   2. Lowering price in Texas (i.e., “getting treated”) may have a bigger impact than lowering price in Ohio because of more elastic demand in Texas. That is, residents in Texas may be more price sensitive when it comes to cereal compared to residents in Ohio.
3. To simplify this discussion, we can again consider low prices as “treatment” and high prices as “no treatment”…
   1. In a given state, consider the times when price was high (“no treatment”) vs. times when price was low (“treatment”). It may be the case that, were price high during both times (i.e., there is no treatment at any time), revenues would be higher during the times when price was actually low compared to the times when price was actually high. For example, it could be that price is low during times of high cereal demand (perhaps Winter) and high during times of low cereal demand (perhaps Summer).
   2. In a given state, consider again the times when price was high (“no treatment”) vs. times when price was low (“treatment”). It may be the case that, consumers were more responsive to price during the times when price was actually low compared to the times when price was actually high. For example, it could be that price is low during times when there was an influx of competing products (e.g., frozen breakfast sandwiches).
   3. Assuming we have a random sample of prices and revenues from Texas and Ohio in our data, when we observe a price decline of $1 in Texas or Ohio, on average we expect to observe an increase in revenue of $8,705.
   4. If we want an estimate of the average effect of a change in price on revenues across Texas and Ohio, we likely need to consider an instrumental variables approach. This is because it will be difficult for us to produce an assumed data-generating process that accounts for selective pricing across time according to various demand-influencing factors. Consequently, we may try to find a cost-side instrument, perhaps costs for ingredients or transportation costs, that is correlated with price but not unobservables in the revenues equation. Here, we would use 2SLS to get an estimated effect of price with which we could make active predictions.

**WRITTEN EXPLANATIONS OF DATA ANALYSIS AND ACTIVE PREDICTIONS**

**Case #1:**

* 1. Perhaps the most notable feature of variables in Table A4 is substantial variation. In particular, the outcome variable (Claims per 100) and treatment variable (Deductible) both vary considerably. Hence, there is potential return on identifying an effect (since there is substantial variation that may be “controllable”) and there is potential for finding an effect (since there is significant treatment variation in the data).
  2. As noted in part a, variation in Deductible is important, because it is necessary in order to find a treatment effect. In contrast, if Deductible doesn’t vary, or only varies a little bit, there is no way for us to tell how the outcome (Claims per 100) changes with treatment (Deductible) changes. We need changes in treatment to measure treatment effects.
  3. The point estimate means that, holding % between 25-65, % married, traffic index and wealth index fixed, a $1 increase in deductible is associated with a decline in claims per 100 of about 0.0323.
  4. The p-value means that, if the coefficient on Deductible in the population regression equation is zero, the probability of observing a coefficient at least as large (in absolute value) as -0.0323 in our sample is approximately nil.
  5. These bounds provide a range that, with 95% probability, contains the coefficient on Deductible in the population regression equation.

1. If we only cared about co-movement of Claims per 100 and Deductible, they allow us to learn about partial correlation (controlling for each of the additional variables). If we care about causality, they can help us build an assumed data-generating process for which the necessary assumptions for causality can be credibly defended.
2. Conceptually, a control is a variable that, if excluded, would be a confounding factor. Hence, a control is a variable that is believed to be both correlated with the treatment and has an effect on the outcome. In contrast, a proxy variable is one that moves with a confounding factor, but besides this co-movement with a confounding factor, has no influence on the outcome. For the four variables besides Deductible, one could make an argument as to why they might impact the outcome, and so they are likely all playing the role of controls.
3. If the deductible goes up by $200, this will reduce claims per 100 by 6.46 on average.
4. To get to the conclusion in problem #5, we are assuming the data-generating process as described in the case, our data are a random sample, and the error in the data-generating process has mean zero and no correlation with the included variables. If these are true, then we can use sample results to tell us about the population (random sample) and view variation in the treatment (and controls) as analogous to random treatment assignment (i.e., there aren’t unobserved factors that affect the outcome systematically moving with the X’s). The latter claim allows us to infer causality in the same way a scientific experiment would – differences in the outcome corresponding to differences in the treatment are the causal effect of the treatment, since other factors affecting the outcome do not systematically co-move with the treatment (and therefore cannot confound our measurement of its effect).
5. We may worry that local managers are choosing the deductible based on expected claim rates. Further, these managers may be able to predict claim rates better than information on age, marital status, traffic, and wealth can. If this is the case, they may, for example, set a higher deductible when they expect more claims and vice versa. In such a scenario, it would not be the case that Deductible is uncorrelated with the error term (they would instead be positively correlated), breaking down the reasoning presented for problem #6. We can even sign the bias in this case; the unobserved propensity for a claim is positively correlated with claims and positively correlated with the deductible, leading to a positive bias. So, the estimated coefficient on Deductible is less negative than it is in the true data-generating process.

**Case #2:**

* 1. Yes, it is crucial. Given price is one of the treatments whose effect we’d like to measure, we need to see that treatment vary. Without treatment variation, we cannot compare outcomes across different treatment levels.
  2. This pricing rule creates a perfect linear relationship between price and battery life. Hence, it creates perfect multicollinearity, meaning we cannot separately identify the coefficient on price from the coefficient on battery life. In Table A7, we would be forced to fix one of these two coefficients (usually at zero), and then estimate the other given this fixed value.
  3. The point estimate means that, holding Price fixed, a one-day increase in battery life is associated with an increase in unit sales per 100 of about 3.01.
  4. The p-value means that, if the coefficient on Battery Life in the population regression equation is zero, the probability of observing a coefficient at least as large (in absolute value) as 3.01 in our sample is approximately 0.001.
  5. These bounds provide a range that, with 95% probability, contains the coefficient on Battery Life in the population regression equation.
  6. The point estimate means that, holding Battery Life fixed, a $1 increase in Price is associated with an decrease in unit sales per 1,000 of about 0.36.
  7. The p-value means that, if the coefficient on Price in the population regression equation is zero, the probability of observing a coefficient at least as large (in absolute value) as -0.36 in our sample is approximately 0.002.
  8. These bounds provide a range that, with 95% probability, contains the coefficient on Price in the population regression equation.

1. Unit Sales per 1,000 will increase by 9.03 (3 x 3.01).
2. To get to the conclusion in problem #4, we are assuming the data-generating process as described in the case, our data are a random sample, and the error in the data-generating process has mean zero and no correlation with the included variables. If these are true, then we can use sample results to tell us about the population (random sample) and view variation in the treatments as analogous to random treatment assignment (i.e., there aren’t unobserved factors that affect the outcome systematically moving with either of them). The latter claim allows us to infer causality in the same way a scientific experiment would – differences in the outcome corresponding to differences in the treatment(s) are the causal effect of the treatment(s), since other factors affecting the outcome do not systematically co-move with the treatment(s) (and therefore cannot confound our measurement of its effect).
3. We can calculate this by taking the following ratio (3.01 / 0.36), which equals $8.36. This figure comes from the fact that increasing battery life by one day increases unit sales per 1,000 by 3.01 – and if we raise price by $8.36, unit sales per 1,000 would go back to the starting level (i.e., go back down by 3.01). Hence, we have solved: -0.36\*X = -3.01, and X = 8.36.
   1. If price/battery life offerings were done in this way, we would worry that price and/or battery life is correlated with the error term, i.e., unobservables that affect unit sales. This would break down our ability to infer causality from the regression estimates.
   2. We may try to collect panel data, allowing for fixed effects regression analysis. This added data would be particularly useful if we believe that within-market price/battery life changes are not in response to unobservables affecting demand (e.g., they are instead due to market-level changes in costs).

**Case #3:**

* 1. The point estimate means that, holding time of day fixed, in increase in ad duration from 15 seconds to one minute is associated with an increase in the likelihood of a click of 0.2.
  2. The p-value means that, if the effect of Duration60 in the population regression equation is zero, the probability of observing a coefficient at least as large (in absolute value) as 0.201 in our sample is approximately nil. (Notice we are using causal terminology right away, since the use of an instrument already implies we are analyzing for causality).
  3. These bounds provide a range that, with 95% probability, contains the effect of Duration60 in the data-generating process.

1. For Morning, going from Afternoon to Morning decreases the likelihood of a click by 0.032. For Evening, going from Afternoon to Evening increases the likelihood of a click by 0.014. For Night, going from Afternoon to Night decreases the likelihood of a click by 0.025.
2. The first stage indicates it is relevant, as its p-value in the first stage is essentially nil. As to it being exogenous, we must make a theoretical argument. Specifically, we have to believe that unobservables affecting the likelihood of a click on this ad are not correlated with the level of congestion on the network when the ad ran.
3. To make a causal interpretation for the coefficient on Duration60, we are assuming the data-generating process as described in the case, our data are a random sample, and the instrument we use is valid. If these are true, then we can use sample results to tell us about the population (random sample) and view variation in the treatment (stemming from variation in the instrument) as analogous to random treatment assignment (i.e., there aren’t unobserved factors that affect the outcome systematically moving with either of them). The latter claim allows us to infer causality in the same way a scientific experiment would – differences in the outcome corresponding to differences in the treatment(s) are the causal effect of the treatment(s), since other factors affecting the outcome do not systematically co-move with the treatment(s) (and therefore cannot confound our measurement of its effect).
4. Since longer ads were being shown to people more prone to buy, we have a positive correlation between the treatment and a part of the error (proneness to buy) and a positive correlation between proneness to buy (in the error) and the likelihood of a click. Hence, the bias is likely to be positive (positive x positive).

**PROJECTS: COMBINING ANALYSIS WITH REASON-BASED COMMUNICATION**

Answers to associated questions will depend on approaches taken by the students. Answers from the prior six cases provide a good template for the types of answers one should expect to these questions. Also, the project data provided to instructors has examples of regressions to run (in STATA do files), along with the formulas used to generate the original data.