STA 371H Midterm Grading Guide, Spring 2017

Question 1: Short answers (30 points)

A) Define the term "p-value."

A p-value is the probability of observing a test statistic as extreme as, or more extreme than, the test statistic actually observed, assuming that the null hypothesis is true.

(Could also say "at least as extreme as" for the phrase "as extreme as, or more extreme than." The answer should refer to a test statistic for full credit. Using a term like "result" or "data" in lieu of "test statistic" should receive most of the credit, but not all, assuming everything else is right.

B) Concisely describe the stepwise-selection algorithm. (It's OK to just list the steps.)

Start from a baseline model. Consider all possible one-variable additions to or deletions from the model. Choose the single addition or deletion that yields the best improvement to the prediction error (you could also use terms like generalization error, forecasting error, mean-squared error, MPSE, etc). Then repeat the process, using the new model as the baseline model, until no further improvement is possible.

C) Under what general circumstances should we expect to need an interaction term in a statistical model?

Something like this: if the effect of X on Y is modulated/changed by some third variable Z, then we need an interaction between X and Z in the model.

Or: if the effect of X on Y is context-specific (dependent on circumstances, etc), then we need an interaction between X and whatever variable specifies the context.

(If you didn't specify that the interaction was between X and the context/modulating variable Z, you didn't get full credit.)

Question 2: Essay (30 points)

We have met the concept of statistical adjustment in at least two different contexts. First, we met it in ordinary linear regression with a single predictor variable. Here, we wished to look at the y (response) variable after "controlling for" or "adjusting for" the x (predictor) variable. Second, we met this concept in multiple regression models, with more than one numerical predictor.

Briefly describe what is meant by "statistical adjustment" in each context, making sure to address/explain the claim we encountered in class that "statistical adjustment is just subtraction." Give one example of statistical adjustment, either from class/reading or your own imagination. Though it's not necessary to draw pictures, feel free to do so if you find that it helps you convey an idea.

Below is one version of a good answer, recognizing that the details varied from good answer to the next.

In one-variable linear regression, "statistical adjustment" involves looking at the residuals, and interpreting them as y adjusted for x. The logic here is that our model for y is $y_i = \beta_0 + \beta_1 x_i + e_i$, which splits y_i into two parts: a fitted value $\hat{y}_i$ that's predictable by x, and a residual that isn't. If we want to adjust for or subtract the part of y predictable by x, we just take $y_i - \hat{y}_i = e_i$.

In multiple regression, "statistical adjustment" is about estimating a partial slope for y versus some variable x_1, holding other variables constant. This is also subtraction. For example, in a two-variable model we have

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$$

If we want to adjust for the effect of x_2 on y, we move it to the left-hand side, by subtraction:

$$y_i - \beta_2 x_{i2} = \beta_0 + \beta_1 x_{i1} + e_i$$

The left-hand side is y, adjusted for the effect of x_2, and the right-hand side is a regression on x_1. This is why β_1 is a partial slope.

A good example of statistical adjustment was the Austin critics data, in which we adjusted the price of a meal at a restaurant (y) for the quality of food there (x). The residuals from this model told us how cheap or expensive a meal was, relative to the price predicted by its quality.

Question 3: Interpreting data analysis (40 points)

In this question, you will look at data on traffic fatalities and seatbelt usage collected from all 50 US States (plus DC) from 1983–1997, a period in which many states were implementing laws requiring that all passengers in a car wear seatbelts.

The variables in the data set are:
- state: a categorical variable for the state in which the observation was taken.
- year: the year of the observation.
- YearsSince1983: same as year, but centered so that 1983 starts at 0, 1984 is 1, 1985 is 2, etc.
- fatalities: number of traffic fatalities per billion miles logged by drivers in that state in that year.
- seatbelt: seat-belt usage rate by drivers in that state in that year, as self-reported by respondents to a survey in the state. This is expressed as percent out of 100, so 100 means everyone uses their seatbelts.
- speed65: whether there is a 65 MPH (or lower) speed limit on state highways. States with "no" for this variable had highway speed limits of higher than 65 MPH.
- drinkage: whether there was a minimum drinking age of 21 in the state during that year.

Over the next several pages, the results of several statistical plots and analyses are shown. Use these results to decide whether the following statements are true, false, or undecidable/ambiguous in light of the evidence provided. If true, cite supporting evidence. If false, propose a correction and cite supporting evidence. If undecidable/ambiguous, *make your best guess in light of what you do know* and explain what evidence you'd like to see in order to decide the question to your satisfaction.

The statements: true or false?

8 points per question, with partial credit assigned for the quality of your explanations.

The answers to most questions turned on your decision of which model to use, and your reasoning/evidence in support of that choice. As long as you made this reasoning/evidence clear somewhere, you didn't have to repeat it for every single Part. E.g. it was fine to argue in the answer to A (or any part) why you needed to use Model 3, and then for every subsequent part, freely draw conclusions based on Model 3 without re-stating your reasons for preferring Model 3. There are reasonable arguments to be made for Models 3 or 4 (or even Model 2 – this is pretty tenuous, but it was acceptable if you were super explicit in your reasoning). See details below.

A) Holding other relevant factors equal, increasing a state's seatbelt-usage rate by 1% seems to decrease traffic fatalities by about 0.11 deaths per billion miles driven (on average across all states), with a 95% confidence interval of about (-0.14, -0.09) for the partial slope on seatbelt usage rate.

False. These numbers come from Model 1, but Model 1 doesn't adjust for state or year. The correct answer should use Model 3 to get the average effect of seatbelt usage rate across all states. A better confidence interval is (-0.08, -0.03), rounding to the nearest tenth (although your answer might not have rounded, which is OK too).

A full-credit answer would look something like this: "False, a better estimate is a decrease of 0.05, with a confidence interval of (-0.08, -0.03), from Model 3. Model 1 doesn't account for state and year as confounders, which we know are important from the output of Model 3 (either the ANOVA table or the confidence intervals of model coefficients)."

To get full credit, you must show somewhere on the test that you appreciated Model 3 was the better Model to use than Model 1. (You might have been explicit about this evidence in your answer on another part, and that's OK, as long as it's there somewhere.) Good reasons for prefering Model 3 include:
- the ANOVA table for Model 3, which shows that both state and year have large contributions to R^2
- the confidence interval on the "YearsSince1983" variable.
- the pictures that show fatality rates differ substantially across states and over time (although if this was your *only* reason, you would get only 7/8 points, since the two pictures don't actually prove that these two variables are confounders for seatbelts, only that they predict y.)

Any "true" answer receives zero credit.

It is conceivable that a student may have answered false and proposed the confidence interval from Model 2 rather than Model 3, which would be (-0.7, 0). This is a bit of a stretch, since the addition of state in Model 3 improves R^2 by an enormous margin (a 60% boost). But there is no permutation test or formal measure of uncertainty/statistical significance supporting the inclusion of state, even if it's pretty obvious that it has an effect. So if you went with Model 2 rather than Model 3 and were very careful about explaining why – or even said that the answer was ambiguous and you'd want to see a formal test of the "state" variable in Model 3 – then you would still receive full credit. If you went with the model 2 confidence interval on the grounds that it adjusted for year as a confounder, but didn't explain carefully why you were not using Model 3 instead, then you got 5/8 points.

An answer that referred to the "seatbelts" term in Model 4, which had a confidence interval of (-0.045, 0.095), also received zero credit. That seatbelts slope definitely does not refer to the average effect of seatbelts across all states – it is only the slope for the baseline state in the model that assumes an interaction between seatbelts and state.

B) Holding other relevant factors equal, the average fatality rate across all states seemed to change by about -0.65 fatalities per billion miles per year from 1983-1997, with a 95% confidence interval of roughly (-0.8, -0.5).

This confidence interval comes from Model 2. Probably the best answer was something like the following:

"False, it decreased by -0.5 per year, CI (-0.6, -0.4), from Model 3. Model 2 omits state as a confounder, which looks very important in Model 3 (boosts R^2 by 60%)."

You could also quote the confidence interval on YearsSince1983 from Model 4 if you argued that the interaction term looked significant in light of the permutation test.

You could also say "ambiguous" here, on a couple of different grounds.
- You could say it was ambiguous between the Model 2 estimate and the Model 3 estimate, and that you wanted to see a formal test of significance for the "state" variable in Model 3 to decide the question. Saying this got 7/8 points; saying this plus "But I believe it's Model 3 with what I know," for any of the reasons cited above, got full credit.
- You could also say that it was ambiguous between the Model 3 estimate and the Model 4 estimate; see reasoning below in answer to Part D.

C) Even once we adjust for other relevant confounders, there are statistically significant differences among the states in their overall fatality rates.

Any "false" answer received zero credit. There is certainly no evidence that the claim is false.

Good answers here are either "true" or "ambiguous." If you said true, you had to argue using the ANOVA table for Model 3 or 4, which shows that the "state" variable improves R2 by a whopping 60%. This doesn't

formally establish statistical significance. For partial credit, you could have pointed to any of the state-level coefficients whose confidence intervals didn't contain zero – but this didn't receive full credit, because the answer pointing to the 60% improvement in R^2 is much more convincing here, since most state-level coefficients were omitted from the output.

"Ambiguous" was also a fine answer, as long as you did two things: (1) said that you wanted to see a formal test for the state variable in Model 3/4, or look at *all* of the state-level coefficients; and (2) made your best guess for true or false and supported your answer. If you said "ambiguous" and did (1), you got 6/8. To get 8/8 with "ambiguous" as an answer, you had to do (1) and (2).

Bottom line is that in order to receive full credit, you could not simply ignore or not address the fact that state-level dummy variables improved R^2 by 60%.

D) The relationship between seatbelt usage and fatality rates differs from state to state.

I personally think that the best answer here is that it's ambiguous, for the reason that it's genuinely ambiguous whether the interaction term between state and seatbelts is significant! In comparing Model 4 vs Model 3, R^2 for Model 4 is 0.91319. This is near the right of the histogram from the permutation test under the assumption that the interaction term is useless, which isn't in the middle, but not way out in the tails, either. (If you actually calculated the p value, it would be 0.07ish, but I didn't give that information to you.)

However, either a true or a false answer would be acceptable, as long as you clearly understood that to answer the question, you needed to assess the importance of the interaction term in Model 4, versus no interaction term in Model 3. Reasonable people could look at R^2 for Model 4, compare it to the histogram from the permutation test, and disagree about the better answer.

Grading guidelines:
- If you didn't talk about the interaction term in Model 4 at all, you received zero credit. It's the only evidence in what you were given that could, in principle, allow you to decide the question.
- If you talked about the interaction term but provided no evidence that we either needed it or didn't need it, then you got 4/8 points.
- If you cited R2_improve or sd_improve from the ANOVA table, but didn't refer to the permutation test, you got 5/8 points.
- If you referred to the permutation test, but used some value of R2 other than 0.913 as a test statistic, you got 4/8 points.
- To decide the question, I personally think that you really just need more data, but no one was penalized for not coming up with a concrete suggestion for what would actually decide the question. It was all about knowing to look at the interaction term.

E) States with higher drinking ages (21) had lower traffic fatality rates, holding other relevant factors equal.

This looks false. In Models 3 and 4, the coefficient on drinkage=Yes actually has a *positive* coefficient, and the confidence interval spans either side of zero (no effect). You could also have said "ambiguous" and gotten full credit, if you argued that the confidence interval in either Model 3 or 4 was too wide to support a definitive conclusion about whether drinking-age laws decreased fatality rates.

You could also say "ambiguous" if you argued that you needed to see a formal test of the "state" variable in Model 3 to decide whether to use Model 2 or 3 to answer the question. We also accepted "true" on the basis of the confidence interval in Model 2 – but again, ONLY if you were super explicit about why you were using Model 2, as described above. If you used Model 2 to support a "true" answer without the careful argument about wanting to see statistical significance for the "state" variable, then you got zero credit.

If you used Model 1 to support any answer, you got zero credit. If you used the boxplots to support a "true" answer, you got zero credit, without some other counterbalancing answer involving a model. These boxplots show an overall relationship, and the question is asking about a partial relationship!