STA 371H Midterm, Spring 2018: Grading Guide

Part 1 (25 points)

I give some explanations below that weren't necessary to get credit. They're there to help you understand why the answer was what it was.

A) *Define the term "standard error." (10 points)*

The standard error is the standard deviation of an estimator's sampling distribution.

Grading note: if you defined this as the standard deviation of the residuals in a regression model, we gave you 5 out of 10 points. Some references call this the "residual standard error," although this is not a term we have used in this course. You get partial credit here because, while a correct definition of the term "residual standard error," this is not the correct definition of the term "standard error" as we have used it repeatedly throughout the course. (The potential confusion between these concepts is why I have never used the term "residual standard error" to refer to the standard deviation of the residuals in a regression model.)

B) *Suppose we are running a study whose goal is to assess the partial relationship between consumption of vegetables (X) and blood pressure (Y). However, we believe that frequency of exercise (Z) is a confounder for the relationship between X and Y. True or false: to account for this confounder, we must build a regression model that incorporates an interaction term between X (vegetable consumption) and Z (exercise). (5 points)*

False. Explanation: you do need to add exercise to a regression model, but you don't necessarily need to add an interaction. You would only add an interaction if you thought that the effect of vegetable consumption on blood pressure was different for exercisers than for non-exercisers.

C) *Suppose you have a sample of N=60 houses on the Austin real-estate market, and you find that the relationship between size of the house in square feet (x) and sales price (y) has a slope of $\beta = \$225$ per square foot. You want to provide a standard error for this number as an estimator of the size-price relationship for all 355,000 houses in the city of Austin. You decide to use bootstrapping to do this: that is, by taking B different bootstrap samples from your original sample, each of size M. How large should M, the size of each bootstrap sample, be? (5 points)*

M = 60. Explanation: to calculate a standard error by bootstrapping, we resample with replacement from the original sample. These samples must be the same size as the original sample, which here is 60.

D) _Which of the following is the best definition of the term p-value? Just write the correct number in your blue book; no need to copy out the statement. (5 points)

1. The probability that the null hypothesis is false, given the observed test statistic.

2. The probability of observing the test statistic that we actually observed, given that the null hypothesis is true.

3. The probability of falsely rejecting the null hypothesis, given the observed test statistic.

4. The probability of observing a test statistic as extreme as or more extreme than the test statistic that we actually observed, given that the null hypothesis is true._

Answer: 4.

Part 2 (30 points)

Suppose we have two candidate regression models for forecasting the value of some outcome y (e.g. the price of a house), given some features x (e.g. size, location, and so on). Let's call them models 1A and 2A, each with a different set of candidate variables that might be important for predicting y. We've used a previously collected data set to fit both models, and now our goal is to assess whether model 1A or model 2B is the better predictive model. That is, we want to know which one is better at forecasting future values of y for "out-of-sample" data points that weren't used to fit the original models.

A) *One obvious thought here is to look at each model's in-sample fit, by comparing their residual standard deviations σ_e on the past data (i.e. the original data that was used to fit both models). Briefly explain why this won't necessarily give you a sensible answer about whether Model 1A or Model 2A will make better predictions on future data. (10 points)*

The residual standard deviation σ_e measures how well our model predicted *past* data. But a model with a small σ_e might be overfit to the past data. In particular, if one model is large (lots of variables) and the other model is small (few variables), the large model might easily have a lower in-sample error but a higher out-of-sample error on new data. A good example was the house-price model from the course packet, where a model with 11 variables outperformed a model with 130 variables on new data, despite the fact that the 130-variable model had a smaller σ_e.

Grading note: you didn't need to give an example to receive full credit, but that example from the course packet certainly helped to communicate that you knew what you were talking about. For full credit you had to explain the idea of overfitting somehow—not necessarily mentioning the word, but clearly exhibiting that you understood to the concept.

B) *Suppose now that you decide to refine each of Model 1A and 2A by running the* stepwise selection *algorithm, using each of these two models as a different starting point. (To be super explicit, this means running stepwise selection twice: once starting from Model 1A, and once from Model 2A.) Briefly describe how the stepwise selection algorithm works. It's OK to just list the steps involved.* (10 points)

1) Start from a baseline model. 2) Consider all possible one-variable additions to or deletions from that model. 3) Choose the single addition or deletion that most improves the out-of-sample prediction error (RMSPE) versus the current model. 4) Repeat steps 1-3 until no further improvements are possible.

Technical note: you didn't have to point this out to receive full credit, but stepwise selection actually uses an *estimate* of out-of-sample RMSPE, rather than the real thing.

C) *Your application of stepwise selection from Part B will give you two refined models (call them 1B and 2B, corresponding to your original models 1A and 2A, respectively). Now describe how we might compare the generalization error of these two models without actually collecting any new data. That is, how could we assess whether Model 1B or 2B is likely to do better at forecasting future values of y for data that wasn't used to fit the original models? (10 points)*

We could split our data into a training and testing set. We could then fit Models 2A and 2B to the training set only, and use the testing set to assess each model's prediction error (RMSPE) on the held-out ("future") data. Ideally we would average our estimates of out-of-sample RMSPE over multiple train/test splits in order to minimize the effect of which particular random train/test split we happened to select.

8 points for explaining train/test splits correctly, 2 points for mentioning that we should average over multiple splits.

Part 3 (45 points)

In this question, you will look at data on dengue fever in Latin America. This data was compiled by a group of STA 371H students in spring 2017 for their course project. I'll let them describe the data and the problem in their own words:

> As one of the most prolific diseases in the world, dengue fever is a mosquito-borne infection that affects almost half a billion people every year. Mainly endangering the tropics, the fever at best causes symptoms like vomiting and at worst can turn life-threatening. Thousands die every year while researchers attempt to find the best way to fight the pandemic. The disease first started to become prevalent after World War II as mosquitoes were able to travel around the world much easier than before. Since the disease can't spread directly between people, the spread of mosquitoes led to the spread of dengue fever. Specifically, the Aegypti mosquito is the main vector for the virus. Researchers understand this link and have turned their attention to efforts like eliminating still water where mosquitoes breed and encouraging residents to wear clothes that cover more of their skin. If you want to stop dengue fever, you have to stop the mosquitoes.
>
> This dataset includes weekly information from two Latin American cities: San Juan, Puerto Rico and Iquitos, Peru. Along with the number of dengue fever cases each city encountered in the week, included are various environmental measures that describe precipitation, temperature, vegetation, and more. The variables are all intended to be in some way related to how mosquitoes breed and spread. Studies have shown that mosquitoes breed in hot, warm, and green areas. The dataset includes complete data on over 1100 weeks from these two cities in years between 1990 and 2010.

Each row in the data set corresponds to a single week in a single city. The variables in the data set are as follows:
- total_cases: Total recorded number of dengue fever cases that week. This is the outcome variable of interest in all regression models.
- city: City in whivh the data was recorded (sj = San Juan, Puerto Rico; iq = Iquitos, Peru)
- season: Season the data was recorded (spring, summer, fall, winter)
- specific_humidity: Average specific humidity in grams of water per kilogram of air for the week. This is a raw measure of humidity based purely on how much water is in the air.
- tdtr_k: Average Diurnal Temperature Range (DTR) for the week. DTR is the difference between the maximum and minimum temperature for a single day.
- precipitation_amt: Rainfall for the week in millimeters

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 a guess if you think there's at least partial evidence one way or another,* and explain what evidence you'd like to see in order to decide the question to your satisfaction. (Note: all quoted numbers are rounded off a bit; I'm not trying to trick you here by making subtle rounding errors that invalidate an otherwise true statement.)

The statements: true or false?

A) *Suppose we look at the overall relationship between dengue fever cases and city (i.e. not adjusting for other variables like humidity and temperature). It appears as though San Juan (city = sj) has about 1.4 more cases of dengue fever per week than Iquitos, with a 95% confidence interval of 1.29 to 1.51 cases.*

False. Model 1 (main effect only for city) has log(total_cases) as the response. San Juan therefore had exp(1.4) times as many cases as Iquitos, with a 95% confidence interval of exp(1.29) to exp(1.51).

Grading note: the key thing for you to recognize was that the city=sj coefficient in Model 1 encoded a multiplicative change, not an additive change. Answers that implied an *additive* change of exp(1.4) did not receive full credit.

B) *On average across both cities, fall is the worst season for dengue fever, and winter is the second worst, with about* $e^{0.545} \approx 1.72$ *times as many cases per week during fall as during winter.*

True. This statement seems consistent with Figure 1, where fall is highest and winter is second-highest across both cities. We can also see from Model 2, where winter is the baseline season, that the fall coefficient is 0.545; this gives us the average difference in log cases from winter to fall, on average across both cities. Therefore on average across both cities, cases increase by a factor of exp(0.545) in fall, compared to winter.

Grading note: the question doesn't say anything about "holding other variables constant." It just asks straight up whether cases are higher in some seasons versus others. So the season coefficients in models 3 and 4 are really not as helpful as the ones in Model 2. (What would it even mean, substantively speaking, to "hold humidity constant" between winter and summer?) The fall coefficient in Model 4 is especially misleading as an answer to this question: it is the intercept in the humidity-log(cases) relationship during the fall season. It is not telling you what's happening "on average" across both cities during the fall, but rather what's happening specifically when humidity is 0 during the fall.

We would also accept an "ambiguous" answer here, if you said something like as follows: "I can't tell whether the question is asking for an overall or a partial relationship. If overall, then true (citing numbers from Model 2), but if partial, then false (citing numbers from Model 3)."

C) *There is clear evidence that higher daily temperature range is associated with a negative effect on the number of dengue fever cases in both cities.*

False. tdtr_k is a variable in both Models 3 and 4, and in both models its coefficient's confidence interval contains both positive and negative values. There is not clear evidence that the effect goes in a particular direction.

D) *Higher specific humidity yields the same increase in dengue fever across all seasons, with a one-unit change in humidity producing an increase in dengue cases by a factor of roughly* $e^{0.09} \approx 1.09$.

False. It seems like the humidity effect isn't the same across all seasons; it actually changes from season to season. We know this because the permutation test of Model 4 versus Model 3 shows that the interaction between season and humidity is a useful addition to the model ($R^2 = 0.45$ is very far to the right of the histogram under the null hypothesis of no interaction).

Grading note: if you said false and cited Figure 3, claiming that the relationship looked steeper in some seasons (i.e. summer) than others, but didn't mention the permutation test, you received partial credit. You could have also received full credit by citing the confidence interval on the interaction term between summer and humidity. If you said true and cited the coefficient on humidity in Model 3 without addressing the permutation test, then you received 2 points: you're drawing a sensible conclusion from Model 3, but you're missing the big picture provided by the test. If you said false and cited the interaction term in Model 4, without referring to the permutation test or confidence intervals for those terms, you received partial credit. You got the answer right, but didn't sufficiently explain why it was right.

E) *Humidity itself is actually irrelevant in predicting the number of dengue cases. What matters most is the amount of rainfall in a given week, and once we adjust for rainfall, there is not a strong partial relationship between humidity and dengue cases.*

Undecidable. We have no model with rainfall in it, and so using the information given, there's no way to assess the partial relationship of humidity and dengue cases, holding rainfall constant.