Stat 5100: Questions and Answers

Module 1

1.1

Section 5.1: *In the question identifying linear/non-linear models, the only non-linear model is the last one, correct? It is only nonlinear when the beta is a part of something nonlinear?*

Correct. We only need linear form in the parameters we are trying to estimate. The values of X are observed, not estimated, values.

Module 2

2.1

Section 2 - Estimation of ϵ in Theory: *What does it mean to "constrain" one of the ϵ_i? In other words, what does it mean for ϵ_n to be fully determined if we know $\underline{\epsilon}$ and $\epsilon_1, \ldots, \epsilon_{n-1}$? What is the connection to the "constrained" concept and the number of unconstrained observations/degrees of freedom?*

This is a difficult concept to both understand AND to explain. I will do my best via example:

Suppose I am interested in the length measurements of sea turtles. Suppose now that I have a sample size of 2 turtles. The average length is simply the sum of the two lengths divided by 2. Now suppose that I only know the length of one turtle but also know the average length of the two turtles. I can easily find the length of the second turtle by solving the equation $\frac{(x_1+x_2)}{2} = \underline{x}$ for x_2. This means that if I know x_1, then x_2 is *forced* to be a particular value. While x_1 is "free" to be any value, x_2 is fully dependent on x_1 if the average of the two turtles is already known. This means I have one degree of freedom.

Now suppose I have three turtles in my sample, but I know the length of the average and the first turtle. There are an infinite number of combinations of x_2 and x_3 that would return the same average value given a known value of x_1. Now suppose that I know the length of the first two turtles as well as the average of the three. Now the third turtle length is forced to be a specific value to obtain the known average. Thus, one measurement is "constrained", while two are "free", once I know the average value of the three turtles. Thus, in this example we have two degrees of freedom.

To calculate other statistics (such as regression model parameters), I need to have "free" observations. Consider for example if I had only one turtle in my sample. The average would (obviously) be the length of the one turtle, which means that there would be no free

observations. This is troublesome when I go to calculate the variance, which involves finding the squared distance from each point to its average. I cannot do it because the distance is trivially zero since the average and the observation are one and the same. Now suppose I had two turtles with precisely the same length measurements. Now the variance is equal to zero (rather than undefined), because there were infinite combinations of the two turtle lengths that would achieve the same average. It just so happened that the two measurements were identical.

The main point is that calculating a statistic requires at least one observation, leaving less observations "free". The more parameters/statistics you calculate, the less observations have freedom. Given this rule, we cannot calculate more parameters than we have observations in classical ordinary least squares regression.

2.2

Section 4 - Remedial Measures: *Does the statement "Variable Transformations on X and Y (not e)" imply that we cannot transform error?*

Correct. We can only change the model form, which will in turn change the estimated residuals. We cannot transform errors directly.

2.5.1

Page 5 of code: *I see we divide the alpha = .10 by 3, as we are finding joint confidence intervals for three beta coefficients, resulting in the usage of the .1/3 = 0.0333, but do we have to divide alpha by the number of intervals in any other circumstance? For example, on the next page (page 6), we leave alpha as .10 for the prediction limits/intervals. Is this because we are dealing with y values and not betas?*

In both cases we perform the division based on the number of intervals (i.e. the Bonferroni correction). The difference with the prediction intervals is that we perform the division inside of the tinv() function (see the definition of t on the middle of page 6). The reason for the difference is simply code syntax and there is nothing that is statistically different about the two scenarios. Bonferroni correction always involves dividing α by the number of intervals we wish to consider jointly.

2.6

Section 3 - Problems with Multicollinearity*: How does multicollinearity produce (nearly) non-unique estimates of beta coefficients and is this referring to our beta estimates or our Y estimates?*

It is referring to beta estimates: Suppose we have two *exact* copies of the same variable in our model, labeled x_1 and x_2. If only one of the copies was included in the model, the true model

coefficients would be $\beta_1 = 1$ and $\beta_0 = 0$. If we include *both* copies in the model, the following three scenarios would all produce identical predictions:

- $\hat{Y} = 1x_1 + 0x_2$
- $\hat{Y} = 0.5x_1 + 0.5x_2$
- $\hat{Y} = 10x_1 - 9x_2$

This shows that there are an infinite number of combinations of estimated values for β_1 and β_2 that will produce identical model results when two copies of the same variable are included in the model. Now suppose that x_1 and x_2 are *near* copies of each other. Because they are not exact copies, there is only one unique solution. However, there are many different combinations of b_1 and b_2 that would produce nearly identical results. This is what we mean when we talk about non-unique estimates of the betas and explains why the variance of the beta coefficients is inflated when there is strong multi-collinearity.

For those who have taken linear algebra (skip if you have not taken linear algebra): Perfect multicollinearity implies that the matrix $X'X$ will not be of full rank and thus the inverse, which is required for the coefficient estimates, will not exist. Near perfect multi-collinearity results in an ill-conditioned $X'X$ matrix which leads to unstable (though still technically unique) model coefficient estimates.

Module 3

3.1.1 (SAS example only)

Bottom of page 2: *Is "highercheck: " a command or something else? Is that line corresponding to the last box of output on page 3?*

Highercheck is simply a name for the test. We could name it anything we want. This allows you to perform different variations of the same test, each with different names. The last box of output on page 3 corresponds to the "highercheck" test.

3.2

Variable selection/interactions general question: *If I understand correctly, it is possible to have interaction terms that are significant without their corresponding lower ordered terms being significant (but they are included in the model anyway). So, when we run variable selection (like a stepwise selection process, for example), should we run it with any interaction terms we are considering, and then use whatever terms are recommended (plus any associated lower-ordered terms that may have been eliminated)? Or should we use variable selection simply to narrow down our lower ordered terms and then run individual t-tests for any interactions involving only those few lower ordered terms?*

You should never perform stepwise selection with interaction terms. The possibility of significant interactions terms is a major shortcoming of variable selection techniques. It is possible that two variables express a strong interaction on the response but are not significant when considered separately. If we perform variable selection before testing interactions, we run the risk of never testing this interaction if the variables do not survive the first cut. There is not a good solution to this problem, which is why some of the more modern regression techniques (like random forests) are so useful when you have a lot of explanatory variables.

3.4

Section 3 - Cross Validation: *I know we did not use cross validation in our final projects because it is not easy to do in SAS, but theoretically, is this what we should have done, since we only withheld a section of our original data instead of collecting "new" data? I guess I'm having trouble understanding when to use cross validation.*

When you have lots of data, what you did for your project (withholding a portion of the original dataset as a "test" dataset) might be preferred to cross validation. This is because the cost of excluding data is low when you have a large dataset. Cross validation is an alternative to the test/training set paradigm that is most useful when you have few observations available for model training. The advantage of cross validation is that you get an estimate of error for every observation in your dataset, rather than just a small subset.

3.4.1

General coding question: *I know we use a seed to see the same results each time we run the rest of the code, but does it matter the size of the number used for the seed? Did we need 5 numbers in our seed for any reason or could 1234 have done the job equally well?*

Any number can do the job. I have heard that very large prime numbers are preferred for many of the random number generators but that is by no means a requirement.

Module 4

4.1

Section 2: *From the video, we learn that penalized regression techniques "discourage large values of beta," but how exactly is this working? Is it because we are adding on the specific summation piece to the SSE and we are trying to minimize the SSE alongside this new term?*

Yes. The new loss function (i.e., measure of badness) includes a penalty based on the sum of the absolute value or squared coefficients. This makes it so that a large value of beta will only be tolerated if it brings with it a substantial reduction in the sum of squares error. The key is that beta values with large magnitude add to the measure of "badness" with the inclusion of the penalty term.

Section 2.4: *"In the presence of high multicollinearity, LASSO tends to select only one variable from the group of correlated predictors." Is this a bad thing? Because it sounds beneficial to me.*

It is certainly not a terrible thing, but it is not a good thing because there is no guarantee that LASSO will select the most representative variable of the group. Also, because the collinear variables are not identical, they all potentially explain different portions of the variance of Y. By selecting only one member of the group, we may see a substantial loss in predictive power that comes when the other variables are excluded.

4.1c

Question 1: *"When the number of candidate explanatory variables is large, inflated variance may cause us to throw the 'best' predictor variables out in a stepwise search." Are we saying that having more variables implies inflated variance?*

No, but having more variables makes it more likely for us to observe collinear variables, which inflates the variance of the $\beta_k{}'s$.

(Related Question): *And how would this cause us to throw out the best predictor variables in stepwise selection? Is it because somehow the best variables would have higher variance and be penalized more?*

Stepwise selection makes decisions to keep or discard variables based solely on the p-values. If a useful variable for predicting Y is collinear with some other set of variables, the p-value will be inflated, and stepwise selection may mark it for removal. The ultimate issue is that multicollinearity compromises the integrity of the p-values, so they can no longer be used as an effective metric as to which variables are "best" to keep.

Question 2: *"Variables with a small range of values will be unfairly punished if we do not standardize." I thought it would be the opposite: variables with a larger range of values will be unfairly punished because it would seem like they have higher variance.*

It is important to distinguish between the variance of the variables and the variance of the β coefficients. Our goal is to prevent the variance of the β coefficients from being inflated due to multicollinearity. Penalized regression punishes coefficients that are large in magnitude. Variables with small ranges tend to be associated with β coefficients that are larger in magnitude. Think, for example, if we are measuring the association between diamond size and price. A one dollar increase in price is associated with a small increase in the diamond size (i.e. smaller β). However, if we change the unit of measurement to *thousands of dollars* then one unit increase in price is associated with a much larger increase in diamond size.

4.2

Last page - sand compression example: *"Goal: find μ, α, and M to make f roughly equal to 0, and look at the relationship between these three parameters." Was making f roughly equal to 0 just part of this specific research problem or is there a general reason for doing this?*

Goal is specific to this research problem.

4.3

Section 2.2: *I am having trouble wrapping my head around the role q has in determining the size of d. If q is the proportion of observations nearest to the current X-profile, is q the neighborhood?*

Suppose that q = 0.25. This means that we want 25% of the observations in the dataset to be included in the neighborhood centered around a particular point X_i. q defines the *conditions* of the neighborhood, but the *size* of the neighborhood (as defined by d_q) changes as we move from point to point. The point is that d_q dynamically changes to ensure that 25% of the observations are always in the neighborhood, regardless of which point we are centered at. Thus, q fully determines the value of d based upon the particular X_i point we are centered at.

4.3c

Question 2: *"As the number of dimensions increases, the distance between any two points becomes large." Why is this?*

The way we define traditional distance between two observations X and Y is

$D = \sqrt{\sum_i^q (X_i - Y_i)^2}$ where i represents the q different variables contained in X and Y. Suppose now that X and Y each has only two variables. That would mean that we are only summing two squared distances, one for each dimension. It would be tough to get a very large value for D if I am only adding up squared differences across two dimensions. Now suppose that X and Y have 100 variables, meaning that I am adding up squared differences across 100 dimensions. Even if the individual squared differences are small, adding up 100 of them will give me a large value of D. Thus, the more dimensions we have, the more squared differences we are adding together, which increases the overall distance between the points.

Question 3: *What would be a good example of when a regression tree would work out well? It feels like these potential jumps would almost always be problematic.*

The jumps are one of the reasons why random forests are almost always preferred to a single regression tree. Random Forests have an averaging component that "smooths out" some of the jumps in the data that are often problematic as you have observed. However, if the explanatory variables do not vary smoothly themselves, then the jumps become less of an issue. Included is

a link to a paper that successfully uses regression trees to predict the abundance of species in Coral Reefs: **https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2**

Module 5

5.1

Section 3.1: *I understand the mathematical difference between probability and odds, but the difference between the meaning of the two words in this context is what I've never really known how to explain. In the 5.1c notes, we use the word "chance" with regards to probability, so what exactly are odds?*

A rough definition of "odds" is the chance that an event happens divided by the chance that it does not. When odds go up, the chance that an event happens also goes up. Log-odds have some nice mathematical properties that regular probabilities do not, which is why log-odds are usually what is modeled.

5.2.1 (SAS Example)

Page 5 - Analysis of Maximum Likelihood Estimates table*: I was not following what you were saying in the video about the reason we have two intercepts being that we grouped the responses of "important" with "little importance" vs. grouping "important" with "very important"—why are we grouping them?*

In this output, we are assuming that the categories are *ordinal* instead of nominal. We therefore consider the *cumulative* odds for the categories vs the *individual* odds for a single category. In the case of three distinct categories, we are now considering the odds ratio $\frac{P(x\leq 2)}{P(x>2)}$ instead of $\frac{P(x=2)}{P(x\neq 2)}$, this is what I mean by "grouping" in this context.

(Related Question): *Also do we use the Testing Global Null Hypothesis: BETA=0 table for anything?*

Not in this class. I am almost positive that the output of that table is akin to an overall F test in OLS regression.

Module 6

6.1

Section 1.1: *In the video you mentioned something about "overstating our sample size" due to autocorrelation, what does this mean?*

If $X \sim N(\mu, \sigma)$ then, under the assumption that the observations are independent: $\underline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. However, when the observations X are not independent, then the standard deviation of $\underline{X} > \frac{\sigma}{\sqrt{n}}$. In other words, dividing by $\sqrt{n}$ *underestimates* the standard deviation and we should have divided by something *less than* n. This, in essence, means that we overstated oursample size.

Put another way: independent data means that each observation is a totally new piece of information. Autocorrelated data essentially means that some portion of each observation is explained by the other observations and is therefore not a totally new piece of information. Thus, autocorrelated data has an effective sample size *less than* n, because we do not have n distinct pieces of information.

Section 2 - Point 2: *"There is a difference in a series being a function of time (plus random noise) versus a series that is correlated in time." I am having trouble distinguishing between these two things. In another part of the notes, it says, "If a random variable is autocorrelated over time, then observations closer in time will tend to be more similar than observations far away in time," which sounds exactly what I would imagine a time dependent trend/function of time would be. Is it that data with a time dependent trend must also have autocorrelation but autocorrelated data do not necessarily have a time dependent trend? So, if house prices in Utah are an example of data that we would expect to be autocorrelated over time, does this mean that it potentially has a time-dependent trend too, but we do not know for sure? Are we ultimately trying to remove autocorrelation, but not time-dependent trends?*

Autocorrelation is "self-correlation" i.e., observations that are related to each other due to their proximity in space or time. The key is that autocorrelation is all about the *proximity* of the observed values of the response variable and does not directly consider any explanatory variables such as time. Time dependent trends are *external to* actual observations. Trends are modeled as a function of the external variable only and give no direct regard to "close" observations of the response variable. Put differently, autocorrelation is when we use Y to predict other values of Y. Trends are when we use values of X (in this case, X = time) to predict values of Y.

Utah housing prices have consistently risen year over year since 2012. That is a time dependent trend because I am using Year to predict Price. Once I have removed the time dependent trend, then I can check for autocorrelation. Autocorrelation is when I try to use the time-adjusted home prices from February and March to predict home prices in April. Note that I am using

previously observed home prices to predict future home prices, rather than directly using time to predict home prices.

You can have time dependent trends without autocorrelation, and you can have autocorrelation without time dependent trends.

General Question: *What is the difference between ARMA() and ARIMA()? I see one of them seems to have two inputs and the other has three, but I wondered if they were just the same thing.*

Do you remember in the notes where it talks about first and second order differencing? (i.e., modeling the *difference* between sequential observations rather than modeling the observations themselves)? The middle (and additional) argument in ARIMA allows you to perform the differencing as part of the model fit, rather than having to separately create a differenced variable.

General Question: *Is evidence of non-stationarity indicative of time dependent trends or autocorrelation specifically, or just either one?*

Non-stationarity is indicative of time dependent trends, but NOT autocorrelation. Time dependents trends can make it appear as though there is autocorrelation in the model residuals when there is none.

Module 7

7.1

Section 1: *"The key is that the smoothing occurs over one dimension at a time after accounting for the effects of the other dimensions." Does this mean that we apply the smoothing strategy (or different smoothing strategies…?) to one additive function in the model at a time?*

Yes, but we do it iteratively. This usually starts by using a flat line function for each variable, updating the smoothing function for each variable one at a time, then cycling through *all* variable's multiple times until the smoothing functions stop changing very much on each iteration.

General Question: *What would be some examples of smoothing strategies that are not splines?*

LOESS curves are a form of weighted regression. I guess that would be an alternative to splines, though GAMS almost exclusively use some sort of spline strategy because there are lots of computational advantages to splines vs weighted regression.