Midsemester Course Evaluation report for 36-617, Fall 2022

Brian Junker

10/21/2022

On page 3 below is a summary of your ratings for 36-617, and on the succeeding pages are your text comments and questions for me.

I really appreciate how much you wrote about the course! This is very helpful, not only for future iterations of the course, but also for making decisions about how to pace the second half of the semester, etc. Here are a few comments and reactions to things that caught my eye in your ratings, comments and questions:

Ratings

- I'm pleased to see that the most frequent category is the middle category for "Understanding", "Difficulty" and "Pace", and especially pleased that for "Interest" the next higher category was most frequently endorsed!
- Trouble spots:
  - About 1/3 of the class think the course is a bit too hard; a few more feel the course is way too hard.
  - A slightly larger fraction think the course is going a bit too fast; a few more feel the course is way too fast.

  These are related of course: the material feels hard in part because we are going fast. But there is a bit more going on...

  So far in the course, on the technical side I have essentially been reviewing material that you would have seen in a good undergraduate regression course (not all of you saw that material so this is also a good way to get everyone "on the same page"). Since there is more new material in the second half of the course, the pace will slow somewhat. It may still feel challenging to keep up, so here is a little perspective that may help: I am exposing you to lots of ideas and concepts each week. I do not expect that you will master everything each week, but rather I hope that you are getting enough exposure each week that when you encounter these things in your later professional life you will have some idea what to look for—in textbooks, on the internet, asking colleagues, etc.—to address the data analysis problems that are in front of you. So do the reading and the exercises with the goal of solidifying that exposure to the main ideas, and to some—but not all—of the technical details.

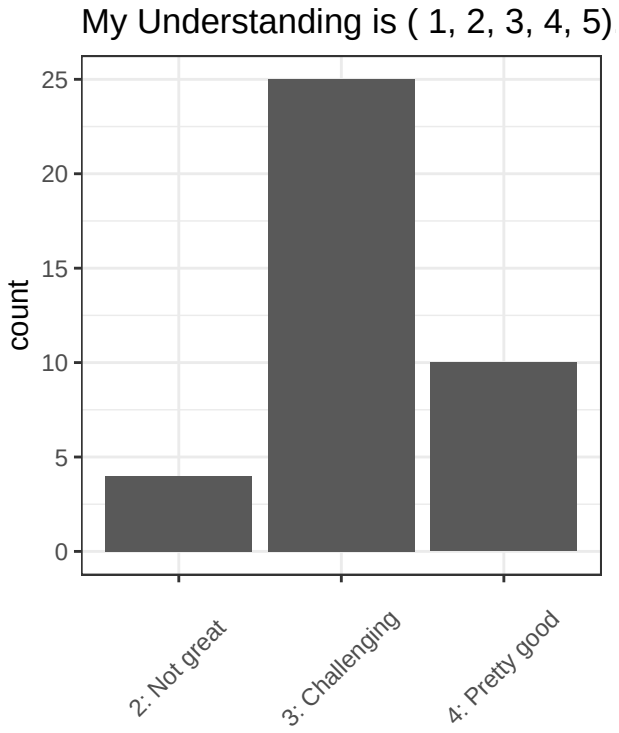
  On the non-technical side, I have been asking you to do a lot of interpretation that some of you may not be used to. I have also emphasized that there can be several "good enough" models, and likely not one single "best" model, which many of you are not at all used to! For me, Statistics is not Computer Science, for which there is often a correct or most efficient algorithm, and it is not Machine Learning, for which prediction or classification is usually the primary goal. In each of those fields, there is often a deterministically correct or best answer. Statistics involves the art of using mathematical data analysis to tell important stories about, and/or to get a better understanding of, the data and the mechanisms that give rise to the data. Statistics also involves grappling with uncertainty, whether that uncertainty is expressed as a confidence interval around a point estimate, or it is expressed in not knowing which is really the "best" model, and making do with a model that is "good enough" to answer the questions that your client or collaborator may have. After all, *"All models are wrong but some are useful!" – G.E.P. Box.*

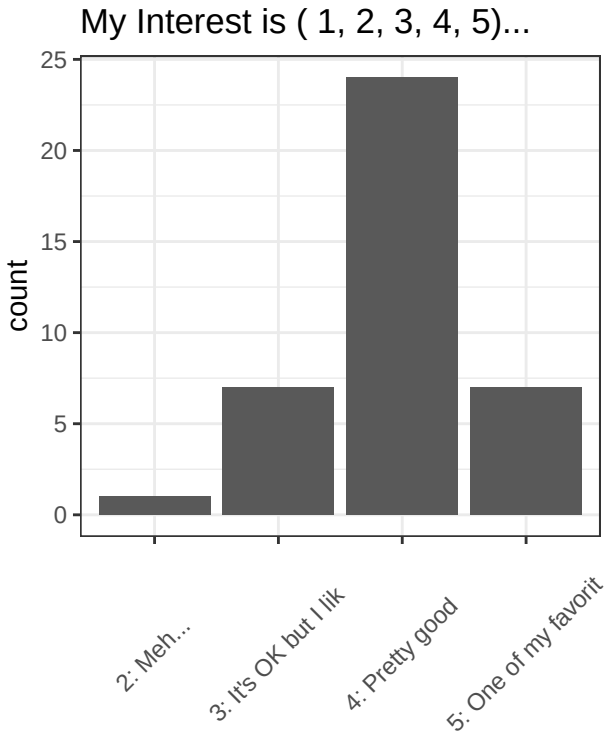
Your Comments

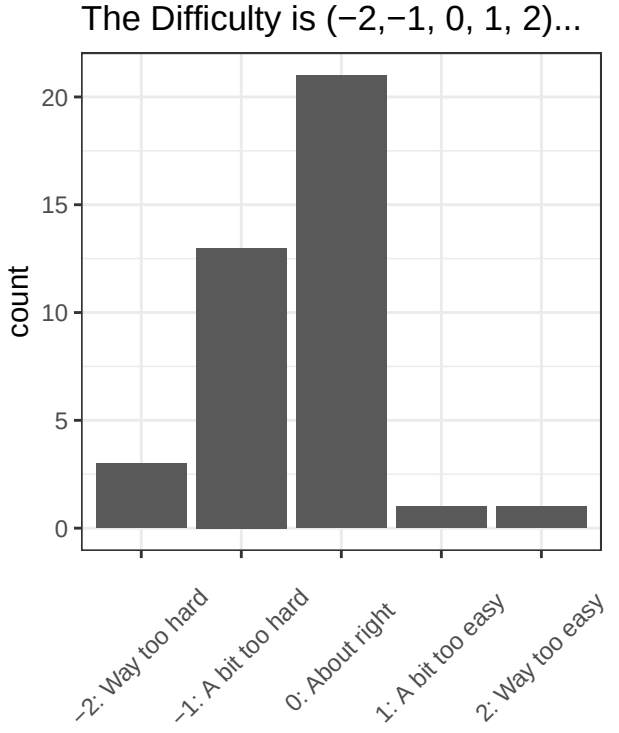
- Nice to see that the zoom recordings are useful!
- Also good to see that you are getting a lot out of Piazza and office hours.
- Several of you seem to prefer Wednesday due dates for homework instead of Monday due dates. I'll ask about that in class.
- It seems like the homeworks might be a bit long. Model fitting, data analysis and interpretation are inherently time-consuming processes, but I will try to tighten up the assignments a bit (also this will give you more time to work on the final report project!).
- I appreciate the feedback (positive, as well as suggestions for improvement!) on the takehome midterm!
- There were some comments in the direction of wanting more coding examples in class. I will try to include coding examples when I can in class, but also please be aware that the hw solutions have complete code to solve the problems, and there are things you can carry forward from old hw solutions to help solve new hw problems. Also, don't forget that your classmates can be great resources, since I allow (credited!) collaboration on hw's!

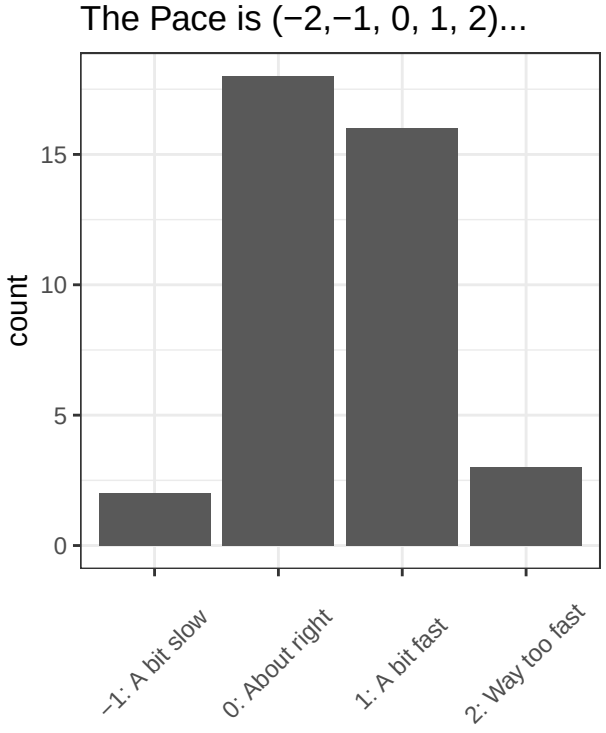
Your Questions & Things You Learned

- Great to see that you are solidifying your knowledge of, and skill in using, linear regression models, diagnostics, methods for model comparison, variable selection, etc.
- There were some questions about whether we will do any machine learning (ML) in this course, and about what would be expected of you in industry jobs.
  - There really is no ML in 36-617, but you can get exposure to ML in 36-662 "Methods of Statistical Learning" (and probably some in 36-615 "Software for large-scale data", 36-616 "Computational methods for statistics", and 36-668 "Text analysis" as well).
  - There definitely is a role for linear regression and its siblings (generalized linear models, GAM's, multilevel models, etc.) in industry. The main advantages of these models is their interpretability, so if you are working with someone who wants to understand and be able to communicate the mechanisms underlying the data you have, these models can be quite useful. If you are working on a problem where the main concern is good prediction, then methods with more of an ML flavor will generally do better, at the cost of interpretability. Sometimes your boss or client will want both, of course, and then you have to manage a tradeoff between predictive accuracy and interpretability. In any case, it is almost always worth trying a linear model or glm first, in part to see if a model with lots of interpretability will suffice to answer your boss's or client's question, and in part to serve as a "baseline" to try to beat with more sophisticated methods that you may be asked to use.
- Some people voiced some rather specific technical questions, that I will not try to answer here. If you'd like to bring them up on Piazza or in office hours, however, I'd be glad to address them there.


My Understanding is (1, 2, 3, 4, 5)
count
25
20
15
10
5
0
2: Not great
3: Challenging
4: Pretty good


My Interest is (1, 2, 3, 4, 5)...
count
25
20
15
10
5
0
2: Meh...
3: It's OK but I lik
4: Pretty good
5: One of my favorit


The Difficulty is (−2,−1, 0, 1, 2)...
count
20
15
10
5
0
−2: Way too hard
−1: A bit too hard
0: About right
1: A bit too easy
2: Way too easy


The Pace is (−2,−1, 0, 1, 2)...
count
15
10
5
0
−1: A bit slow
0: About right
1: A bit fast
2: Way too fast

Comments
The zoom lectures are good because I can use the videos to consolidate what I've learnt and they helped me a lot during my sick leave.
The level of difficult seems about right, a little harder than other courses. The instructor is very helpful and always available to explain concepts outside class and office hours. The zoom lectures are also very helpful. I really enjoy the class and the interaction with the instructor and have no complaints. The only thing would be that the homeworks are a little longer than expected with lesser weightage. The amount of work in homeworks should make them hold higher weightage because most of the work in this course is in homeworks.
Well it is pretty good but for lectures I really hope we can cover more topics, I just felt like the pace is kinda slow.
I think sometimes the instruction in homework is not clear, it is too general.
I think the content is excellent in this course - diving further into these techniques is one of the primary reasons I wanted to do the MSP program. Prof. Junker is clearly knowledgeable and willing to spend the necessary time with students until they understand the content. I think the grading on the homework is on the forgiving side (hence, my rating of a bit too easy above). That said, I think Lorenzo does a good job catching marginal mistakes with -0 rubric items, so I'm not worried that I'm getting bad feedback. It's more that some weeks, I feel like it would be impossible to get below a 90 unless I didn't submit _anything_ for a question. Which is not a terrible problem to have and I'm sure you're aware of it from the grade distribution! But still worth mentioning in my opinion. Overall very happy with the class and I'm looking forward to the second half.
Office hours and Piazza help a lot. It would be better to have more detail explanation on how to interpret the plots on the PowerPoint.
For our lecture PowerPoint, maybe cover more descriptions with full sentences, to illustrate why and when we are using this method.
Somehow the class material is becoming really hard to catch up with, and we may need more help with the coding, maybe more coding solution to sample questions will be helpful
I don't think the material is hard I just think the pace of this class might be a little bit fast such that we don't have enough time to digest the material. I thought the homework is nice but might be a little bit too abstract sometimes I'm not sure what the expectation is. The office hour is super useful.
I've enjoyed the class thus far. I'd prefer if we kept the homework due date on Wednesday nights. It works better with our schedules and allows us to better utilize Monday and Wednesday office hours before its due. I know we are pressed for time with only two lectures per week, but I feel as if I learn best when you go through real data examples and talk through methods/decisions you might make regarding the topic we are covering in the context of the real-world data.
I think the office hours are great and we receive helpful feedback. Piazza is closely monitored, the feedback there is valuable, and the response rate is fantastic. I wish we could apply 1-2 day extensions on maybe 1 or 2 assignments in the semester. There was one quiz that I believe was a bit tricky asking about the multicollinearity based on vif but the rest of the quizzes have done a good job of being brief but accurately assessing our understanding of the material. I believe some of the material gets a lot into the weeds of things (which may very well be the purpose of the course) but I think the main takeaways we learn are most useful. So knowing which graphs to use for diagnostics, why we use them, and how to interpret them is more important than understanding the theory behind them (but it also could be good to know the theory).
Sometimes I feel like I have to search stackoverflow more than I would like for the homeworks but maybe that's not a fair representation of everyone because I have missed a few lectures. Overall though I think everything is very manageable even given my current situation.
I think this course is going at a pretty good pace. I feel like I have learned a lot over the first half of the semester. I like being able to look back at lecture recordings and I think the slides are very useful overall. I think the textbooks can be a little bit dense at times, but not too bad. I am excited to see what the rest of the semester has for us to learn.

(continued)

Comments
The zoom lecture are pretty useful as review materials. The homeworks are a little hard for me.
Sometimes when I come to office hours, I want to ask questions, but I'm not sure if they've already been asked. This happens especially when I can only come to the second half of office hours, at which point I am unsure what has already been discussed. I also really benefit from the discussion that happens during office hours. I think that having an opportunity to attempt to vocalize my own understanding is essential for learning the material. However, office hours are usually solely intended for asking questions (which I think is also a good thing). Essentially, I think that having a time, like a recitation, for the students to discuss their understanding would really benefit our learning. That way, we can try and formalize our current knowledge. Then, if there are gaps in knowledge or it is incomplete, the TA/professor can make a comment, which then facilitates further discussion. Having a set time to do this each week would also eliminate the concern that students would only be present for certain parts of the discussion, but would miss other key points. I hope this feedback helps, and I would be happy to discuss this further!
I feel like the ratio of putting what we learn in practice versus what we talk about in class is pretty good! The assignments can just talk a lot of time when you are inexperienced with model selection.
The due dates for things is suitable for me and the piazza and office hour is useful for me to do the homework.
I think we went through logistic regression really quickly. I am still not 100% clear about intrepreting the plots and what a logit is and how we should intrepret the results. Professor is really helpful and Piazza during office hours which makes the course feel more manageable. Sheather is a little too dense for me to understand. I wish the assignments were due on Wednesday instead of Monday because we don't have enough time to do the assignments if it is due on Monday, and we only get one (professor's) office hour before submitting the assignment. Sometimes, I feel overwhelmed with having to do the quiz and homework all on Monday. The take-home midterm (especially the storytelling part) was really difficult and long for me personally. I don't feel confident that I did well on the midterm.
The piazza and office hours really helped since we might meet same problems as our classmates. I spend lots of time to do the homework. It is stressful but not that bad. Peer pressure sometimes makes me upset.
As for me, I think the difficulty is appropriate, it is not hard but also not easy. I think professor Junker did a great job on explaining the concept and power points. Homework is kind of hard, I usually spend 10h-12h on the home work each week, but it helps us figure out the conceptions, so it is useful. I think the TAs' office hour should be rearranges since it has time conflict with our schedule.
I think I would prefer a due date of Wednesday night over Tuesday night. That's about it
Homework load is a little bit heavy to me. The average time I spent on the homework was about 8 hours each.
The material for this class has been interesting, but also quite dense. When analyzing data, it feels like there are many different ways to approach it, and I appreciate the design of the class finding the best methods, balancing statistical reasoning and practical reasoning. However, I am still adjusting to not having a "definitive" answer in this approach, and with the fast pace of the material, I feel like I lag a week behind in feeling comfortable with the material. The homework assignments do help with that, but it does take time for me to absorb the lecture material. I appreciate the zoom lectures, and being able to rewatch the lectures to gain a better understanding. Some homework assignments are quite long, but I imagine that is simply due to the amount of possibilities for approaches to data. I think with how office hours are structured, homeworks would be better if they were always due on wednesday rather than monday.
The pace of class is a little bit quick.

(continued)

Comments
This course is truly a little bit difficult, but I think I can understand it after spending some of time, especially by finishing the homework. The instructor and TA both do a good job. The whole experience in class is great.
I think the overall class content is very useful, and office hour is also very useful. But sometimes I feel a little bit hard to understand details of function we use and those calculation methods.
* I think the zoom lectures are really useful. I could use them to make up the knowledge I missed or didn't understand in class. The homeworks were challenging, but I think I learned a lot by doing them. The office hours and piazza helped me a lot with my homeworks. * I mentioned the pace of the course was a bit too fast, because there were a lot of math concepts covered in the course, and as a slow learner, I sometimes needed more time to digest the concepts.
I think sometimes the pace is bit fast, since we do not go deep into the theories.
Professor Junker is really great! I benefited a lot from this class. ISLR is great, while Sheather provides no code. Office hours is useful and provoking, but always full of people, maybe a larger room would be better. Homeworks usually consume too much time. And I can hardly spare enough time to finish textbook reading. Piazza is really a good tool to ask questions and learn from others' questions.
I love the lectures, as the instructor shows great enthusiasm. However, I believe the problem sets are the challenging part. It's nice that we have zoom recorded lectures though I rarely used them. The quizzes are nice to have, though it's easy to forget that they exist sometimes. Office hour has been incredibly helpful for both midterm and the homework. Though Latex is a bit difficult and time consuming to work with, I haven't encountered any difficulty submitting things online.
I think that the availability of office hours which promotes a collaborative environment is great, and I think Professor junker does a great job of teaching the class in an engaging way. Everything going through gradescope would be great, in future iterations of the class, you could potentially also have quizzes go through GradeScope.
I really like how the questions in take-home midterm are designed. It allows me to incorporate everything I learned so far to a full linear model building process, based on large amounts of inferences and my understanding to each step. I think what it benefits the most is that it pushes me to know how to analyze the model rather than just memorizing the concepts and theories. Although I learned linear model as a prerequisite course for my undergrad, I can't even say I understand and master to build a linear model step by step until now.
I'm having a good time in the class so far! This class is definitely a challenging one but Professor Junker has been really helpful in assisting us along the way where we struggle, whether that be during lectures, in office hours (especially then), or on Piazza. I like the balance between the theoretical background of the concepts that we cover in class and their tangible applications in R. The lecture slides have been helpful resources to refer back to and the homework assignments (including the take-home midterm) have been appropriate assignments to test mastery of the class content. I do feel, however, that the pace of the lectures is a little fast at times which makes it a little challenging to keep up sometimes. Office hours have been extremely helpful for help relating to the lecture content or homework assignments.

(continued)

Comments
I don't necessarily think the homework or material itself is difficult, but I definitely don't feel confident in my interpretations. It seems like on every homework, there's no one right answer for creating the best models, but this also is confusing because I never know if I'm creating the models correctly. Going to office hours has provided reassurance that other people know the material as well as I do and that we have similar questions, but usually after submitting assignments I feel overwhelmed with how many pages I had and how ambiguous the questions/answers were. I did not like how the take-home midterm used our previous models from HW 4 because I already didn't feel confident on the models that I had created. An improvement to the first question might've been to provide one reduced model for us and then have us compare it to one from our homework, but having to use both of the models from the former homework felt confusing and less than ideal for a midterm that was worth a significant percentage of our grade. The past two assignments I've turned in have also been difficult to submit because of their size. Sometimes Gradescope can't handle large files when you submit the first time. It ends up being fine but I need to keep this in mind the next time I submit so many pages. I've preferred the homeworks being due on Wednesdays opposed to Mondays (before the due date change I usually ended up working all throughout the weekend and not having any time to relax). I feel conflicted about turning in so many pages for assignments because up until now, I had only submitted undergrad homework projects that were a maximum of 20 pages, but I think every assignment has had more than that. I almost feel like the content needs to be more concise because I don't think we'd ever turn in so many pages for a job project. It's also difficult to provide this opinion because I know the homeworks are already only a few questions long (maximum of 4), and I'm not sure how I would improve upon them. Overall I have really enjoyed office hours and I find them very helpful.
The take-home midterm is pretty good which gives me a good overview of how to do linear regression. I hope the homework in the rest of the class could also be similar to that format.
I really like the course in that it allows us to actually use statistical tools to solve problems in the real world context. That being said, I would like more of in-class demonstrations of R coding, more hands on examples (just as we did in one of our classes the other day). Since the class materials are heavily based on the readings, I sometimes find it difficult to connect the textbook reading to the homework assignments. I know that in more recent slides there are more coding examples, but they do not always align exactly to the homework problems, which makes me confused from time to time about the direction I am supposed to take to start off the question. I believe in-class demonstrations will alleviate this kind of confusion.
I believe things are running well so far.
I think the hint on piazza posted by the professor is useful for understanding the course and the homework questions. The midterm is a little bit difficult.
I think you are a great teacher who really gives your all to the course and your students.

Questions
I learnt a lot about how to choose variables when doing regression.
I had never worked with splines and it is very interesting and not so intuitive. Therefore, I would like to understand what models in complex data actually hold meaning to their coefficients and have the most appropriate interpretation.
Will we learn some ML algorithms later in the course?
I know how to use a linear model to solve a practical question (like in the midterm, we need to use every tool we have to solve the problem)
Today, I learned how the smoothness constraint works for cubic splines. I'd like to better understand how the construction of the design matrix enforces that constraint (though I understand that I'll have to look that up on my own).
I've learned to evaluate the diagnostic plots (e.g., marginal model plot, DHARMa residual plot, binned plot etc.) besides residual plots. I do not have any question so far.
Learned: Using ANOVA to test differences between two models (nested). Question: What's the best way to find variable interactions?
one thing I learned is how to dig deeper with the linear model question instead of just summary them. One question I have is how will the linear model be used in real industrial working task?
One thing I learned is that for the analysis of some datasets, there isn't a right or wrong answer, we can interpret it as long as we think it is reasonable. One question I have is that in industry, what do people expect from data scientists? In our data analysis, are we expecting to come up perfect model? Or just show the nature of this dataset? I think based on what I have learned right now, it is reasonable to do it either way.
One thing I learned is Poisson regression. I've only been previously exposed to regular linear regression and logistic regression. One question I have is the interpretation of the binned residual plots. Are you able to determine normality from them with a random scatter? Or are you only able to see the points that have high residuals (crossing the 95% line)? Thank you!
I learned that the breadth of linear models is a lot larger than I originally thought. Although non-parametric models are often times the go-to, it is good to know the foundational validity and theory around linear models.
I feel as if I have gotten more comfortable with allowing my models to sacrifice some predictive power for reliability which is something that has never really been presented to me as an option and I think it's very reasonable. I would like if one day we can go through the whole modeling process as a class. I have difficulties finding interactions or knowing which terms to add a quadratic to. I understand Box-Cox but I feel like sometimes that isn't sufficient for modeling and I don't know how else to fix it.
I think a major strength that I have gained is my ability and level of comfort in interpreting diagnostic plots. Before this class, I understood a residuals vs fitted plot and a QQ plot, the other two were random points to me. I feel like I have a very good understanding of the Leverage Plot as well as what to look for in the scale location plot now. Something I occasionally will get confused over is regarding collinearity. It seems like it is very case by case when we actually care about it (though I understand that is the nature of the subject matter we are examining in this class). So, I guess my question is are there more concrete examples / situations where we really do not care much about collinearity or high VIFs vs when we do care. Would this motivate us into using ridge or lasso to account for this multicollinearity?
I learned how to select models using a lot of methods, like xIc, lasso, anova,etc.

(continued)

Questions
I've learned about Cook's distance, which is great because I have seen that diagnostic plot previously and always wondered to myself what that meant. I am curious about other variance-stabilizing techniques. I feel that on the midterm especially there were a lot of variables with a large range of values, and so I would like to know what other methods of transformation are available to address this.
I learn the using of lasso but the result of the lasso regression is random. And I understand how to find a good transformation. One question is about that, I'm still confuse about the cockroach counts using, maybe I have a misunderstanding of it.
I learned that logistic regression can be used for predicting y's that are 0s and 1s. I am still not clear about how to interpret the results for logistic regression.
Even though I still do not work well, I got the general idea about how to generate a model, how to make the model fits better by using transformation, variable selections.
One thing I have learnt in the course is how to find the better model, we can use AIC, BIC, CAIC, and partial t-test to compare models. The question I have is I don't quite understand what is the difference between 'family = 'quasi" and 'family = 'binomial".
No questions
Model selection is what I think most useful I have learned recently. I haven't get in touch to Lasso and Ridge before.
One thing that I have learned is that there is always a tradeoff with interpretation/practicality and statistical approaches and finding a good balance depends on context. My question is in what fields/settings in the professional world are fine with less practical for better results outside of academia. It seems like a balance is usually preferred.
The method of model selection. How to select model basing on facts.
I learned that in the analysis of linear regression, the first thing we should do is to check whether the Normal Q-Q plot could satisfy the assumption of Normality, it will decide the modelâ€™s fitness. Iâ€™m wondering whether we could talk more about the analysis of the plots, or could we add this kind of contents on the PowerPoint? Thank you so much!
I learned how to use backward, forward selection, BIC, AIC, Ridge and Lasso to make variable selection. Still little confused about the exposure part of Passion regression.
* Learned about which models to apply when predicting continuous variables and categorical variables. * I'm wondering what we could do to master the material of this class. I sometimes feel that even after attending the lectures, reading the textbooks and doing the homeworks, I still couldn't fully understand some of the knowledge covered in this course.
I learned to use how to conduct various forms of regression model and relevant tests in R.
I learned about the whole process of variable screening to build a linear model after getting the data and how to interpret the variable selection including transformations. I still don't know when we should choose poisson regression if given a raw dataset. How can we prove that the data follow the poisson distribution?
I learned how to evaluate goodness-of-fit for Logit and Poisson model using a graphical approach. I still don't understand the linear algebra proof of the hat matrix for generalized linear models

(continued)

Questions
One thing that I learned was how to make splines in LM, I had only used them in GAMs before. One question that I have is, does variable selection in a way tackle overfit?
One thing I have learned: Learning about cross-validation methods and how they could be effective in estimating lasso and ridge regression outputs was really interesting to me. One question I have: I get a little confused between the different types of generalized linear models and which one would be good to use in certain situations.
One thing I've learned from this class is that although I know how to interpret the diagnostic plots and how to decide which models better fit our assumptions of normality based on their plots, it's not always a useful thing to perfect the plots because it can make the models difficult to interpret from a real-world perspective. My question is: What's the boundary between making something interpretable and improving a model? Up until now it seems like that's just a skill you inherently have/are able to build up over time and with experience, but how does a beginner make these decisions?
I learned the procedure of how to do linear regression including variable selection, the transformation of variables, and the explanation of the model to a client. One question I have at this time is in which situation we should use which logistic model. What is the difference between different logistic models like Poisson and Binomial?
One thing Iâ€™ve learned is how to transform variables depending on their distribution. At first I wasnâ€™t sure which transformation I am supposed to use (sqrt, log, etc) but after help from office hours it became much clearer. One question I have is regarding how to interpret transformed variables in words (eg. For sqrt(x), etc.). The explanation I got earlier seemed less interpretable to non-stasticians, and I am wondering whether there is a more plausible interpretation?? (The explanation from homework 4 solution was also a little confusing for me - the interpretation with beauty data). It would be nice if we can talk about this in more detail.
One thing I've learned is that there are many ways to 'evaluate' a model and some methods are more practical than others given the situation. One question I have is "Does model validity really matter in the long run if prediction accuracy is the end goal?"
I learned how to deal with the data to create a fitted model, such as transformation, variable selection and so on. I'm still confused about which method should be used depending on different data.
* I have really learned how to think about solving statistical problems. Today in the Professional Skills and Development Class we had some guest speakers who asked us to think about how to solve a problem as data scientists and I could see the growth in my way of thinking and in how I approached the problem. * I am having trouble analyzing the results in the homeworks.