Questions and Answers for Applied Statistics and Multilevel Modeling
(Columbia University POLS 4726 / STAT 5293, Spring, 2024, Prof. Andrew Gelman)

During the semester, every student was required to contribute before each class to a shared Google doc by putting in a question about the reading or the homework, or by answering another student's question. The material on this document helped us guide discussion during the class.

At the end of the semester, the students were required to add one more question, which then I responded to in the document itself. Perhaps these questions (with students' names removed) and answers will be of interest to others.

Student: For most of the assignments in this class, we have dropped missing data in datasets. I am wondering how I can use Stan to handle missing data. I understand the generated quantities block can be used to model missing data, yet for data for which we might have weak priors or that is MNAR, what is the value is using Stan over simple regression imputation/MICE? In such cases, is imputation inappropriate? Is there a certain amount of missingness above which stan or imputation is futile?

I am also curious how we can determine priors for Stan. In our final project, my partner and I used weak priors since there was little literature on our topic. This made me think (1) about how to think about issues of publication bias in choosing priors from the literature and (2) how to think about priors for a field with mixed/limited evidence. Specifically, I have been thinking about the use of hair cortisol in psychology research. There's limited understanding of what hair cortisol actually is (it doesn't correlate with perceived stress measures in many studies) or how it should relate to other variables. If trying to understand the expected association between, for example, maternal hair cortisol and infant amygdala size (where we can expect small samples), how could we use Stan to determine an "expected" association when the published studies are likely underpowered and present varied results and effect sizes?

AG: Regarding missing data: Analysis of missing data is discussed in chapter 17 of Regression and Other Stories. In Stan we can impute missing data, although it requires some annoying programming steps; see here: https://mc-stan.org/docs/stan-users-guide/missing-data.html and here: https://discourse.mc-stan.org/t/guidelines-for-practical-imputation-with-stan/32024/

AG: Regarding priors: We discuss some of this in section 3 of this paper: http://stat.columbia.edu/~gelman/research/unpublished/causal_quartets.pdf
It should be possible to set a somewhat vague prior that still keeps the parameter values within a reasonable range, given the applied context.

> **Student:** As for me, we can use stan to build an imputation model first. We can model the relationship between observed variables and missing variables to estimate plausible values for missing data points. We can treat the imputed values as random variables with associated uncertainty. And by doing so, we can increase the certainty of the missing value or make the imputed value more accountable. Then, within stan, we can specify a joint model incorporating the imputation process and the analysis model to account for the observed values while accounting for the imputed values. And for the last question in the first paragraph, I think if the amount of missingness is very high, it may be difficult to obtain reliable estimates from the data, regardless of the method used.

Student: When we were going through the types of hierarchical models, I was (and I still am) a little confused on how we should go about choosing what we allow to vary and what we allow to correlate or what we allow to be a global parameter. There are ultimately a lot of choices we can make in terms of

how we fit the model and that is without getting into priors. Some of this boils down to intuition, but if I lack intuition what are the best ways to determine which model fits the data and which model specification should be used.
To use a real world example, I TA for a class where the professor has a fundamentals day on monday and a case study day on wednesday. As the TA, I take attendance everyday and mark each individual as present (1) or absent (0). I noticed just from taking attendance early on that Wednesdays seemed to have less people attending. So, I began messing around with the data and using the principles of our class on hierarchical models. I had three variables: date, student, present, and whether the date was a wednesday. The outcome was whether a student was present and the variables that I allowed to vary either slope wise or intercept wise were the date and student variables since the observations were nested in both. Still, I had no prior on which model specification was the correct to use to determine whether Wednesdays matter and I am still not entirely sure. What would have been the best way to determine which model specification best fit the data?

AG: Our general recommendation with regression is to first include main effects, then include interactions of the most important predictors, where importance refers to your applied goals and also to the size of the effects. For multilevel models, this implies first including varying intercepts, then varying slopes for the most important predictors (notably, the treatment effect in a causal study).

AG: For your real-world example, I guess I'd start with a sequence of models something like this:
present ~ date + wednesday
present ~ date + wednesday + date:wednesday
present ~ (1 | student) + date + wednesday + date:wednesday
present ~ (1 + date + wednesday | student) + date + wednesday + date:wednesday
and I would code "date" as a continuous predictor equal to 0 for the first day of class and 1 for the last day of class (i.e., date = proportion of the semester that has gone by so far).

Student: When fitting multilevel models in this course and my final project, I struggled with simplification of high-dimensionality in datasets and the time for fitting large datasets. We discussed through assignments and in class different ways to simplify models and to sample subsets of large datasets to make the fitting process more efficient. In my line of work in climate science, we often deal with large datasets including geospatial datasets which are often high dimensional. In applications, large data may pose an issue, particularly in spatio-temporal datasets and large time series. My questions are 1a) how can multilevel models be extended to handle complex data types such as spatial data, network data, high-dimensional data, or large geospatial time series? And 1b) what strategies or algorithms may be most effective to increase efficiency for large/high-dimensional datasets? We discussed that Stan uses Hamiltonian Monte Carlo, would other adjacent algorithms such as Gibbs sampling or variational inference be useful in this context? Lastly, 1c) if there are good examples of multilevel modeling for spatio-temporal or high dimensional problems, I would be very interested in reading more, so any relevant literature would be amazing!

A second tangentially related curiosity that I have been wondering is regarding causal inference with complex multilevel models. Specifically, in complex systems, isolating causality can be challenging due to strong nonlinear interdependencies and interconnections and high-dimensional causal discovery networks. I am curious: 2) what are robust strategies to effectively combine multilevel modeling and causal inference to address questions where hierarchical data structures also require causal interpretation, particularly for large, high-dimensional datasets? Thank you!

AG: Here's an example of Stan being used with spatial models:
http://stat.columbia.edu/~gelman/research/published/ward_et_al_2023.pdf

And here's an example of a hierarchical time series model:
http://stat.columbia.edu/~gelman/research/published/Harvard_Data_Science_Review.pdf
There are various spatio-temporal examples in the literature, but I have not worked on any spatio-temporal models myself. Regarding computing, Stan has HMC and it also has two variational methods–ADVI and Pathfinder–which can be used for big data.

> **Student:** This semester I also joined a research group focusing on the data science application in climate science. I can fully understand it is very hard to deal with high-dimensional data. What I learnt is a machine learning algorithm PCA (Principal Component Analysis) that can help reduce dimensions by projecting data onto a direction v $\in R^d$ (the original data may have higher dimension D, D>>d) such that the variance of the projected data is maximized. It helps to keep the most important information of the data, at the same time, reduces dimensions. But this is an unsupervised approach, I am not sure whether it is suitable for your case or not. Besides, I am also wondering how to apply multilevel models to address this kind of questions. Is there any easy but efficient way?
> **Student:** For (1a), I believe multi-level extensions are applicable anywhere. In time-series (my interest), we might have autocorrelated data with a pre-defined hierarchical structure (e.g. categories, locations, etc.), where explicitly modeling co-movements might make sense. This connects to (1b); to the best of my knowledge, there are many "general-purpose" strategies to scale Bayesian computation. There is the Black-Box Variational Inference (e.g. using automatic differentiation, I believe also implemented in Stan! https://mc-stan.org/docs/2_19/reference-manual/vi-algorithms-chapter.html) that approximates the posterior distribution via an optimization approach. The optimization approach can be advantageous over sampling using HMC in terms of speed w/ large/high-dimensional datasets. The approach is non-model-specific, which is also advantageous in that we're able to explore many many models, which is important to evaluate and actually use our models. A caveat is that the approach is not (asymptotically) exact, so I think it makes sense to rely on some diagnostic measures to assess whether VI worked or not (e.g. https://arxiv.org/abs/1709.02536, https://proceedings.mlr.press/v80/yao18a.html). We can also think about "model-specific" strategies, where Gibbs sampling (often faster than HMC) is specific to the model. Or, VI with judicious model-specific choice of the approximation may come into play.

Student: During my final project, I was unsure how to fit a model in the Bayesian framework. The small amount of data and the lack of dimensions makes the estimation of effects difficult. In that sense, the data can be said to be not very great. In that situation, is it possible to still investigate the question of interest? We decided that the lack of evidence itself is worthwhile, but during presentation this was called into question.
If given another opportunity, should we have gathered more information? Or perhaps we should have asked a different question based on what is available? There is also a slight time series aspect for the data since some of the entries are before the others. Therefore there are some other features that might be engineered. Would love to know what we can do better.

AG: There are three issues here. First, is the programming fitting the model correctly? You should be able to check that by simulating fake data from the model and checking that the fitting procedure can roughly recover the assumed parameter values. Second, does the model fit the data? You should be able to check that by simulating replicated data from the model using the posterior predictive distribution, then graphing the replicated data and comparing to graphs of the observed data. Third, would it help to

have more data? You should be able to check that by simulating fake data of a larger sample size and seeing how accurately you can recover the assumed parameter values.

Student: I fully understand that it can be difficult to work with little data that has few dimensions in the Bayesian framework. Nonetheless, I believe one of the main benefits of Bayesian causal inference is that prior information can be included into your model. You could also use models which are less sensitive to outliers or small sample sizes. As ROS explains, “Bayesian inference is generally recommended for regression. If prior information is available, you can use it, and if not, Bayesian regression with weakly informative default priors still has the advantage of yielding stable estimates and producing simulations that enable you to express inferential and predictive uncertainty.” (p. 16 ROS).

AG: And, to be fair, we don't have lots of good examples in our books of using strong priors for regression coefficients. We need to do more of this, both in applied work and in textbooks and tutorial materials.

Student: I find that sometimes interpreting the results of the model is more difficult than the modeling itself, and making the model meaningful and easy to understand may require a lot of preparation.
I am curious about dealing with the 'time' variable when building a model. Taking my project topic as an example, we were exploring salary in the data science community. The dataset included records from 2020 to 2024, but this is not a repeated experiment, thus maybe some people complete surveys many times while some only take once time. What we did was fitting a varying intercept model given different kinds of jobs involving ’ time‘ variable. But in fact I am considering here the 'time' variable causing bias for the model. Different years have different economic environments. Compared with given different kinds of jobs, maybe we can try to group different years. If there is a repeated survey, I expect the results will be more interpretable.
What’s more, I am curious about the difference between setting priors and pre-test variables. If we have access to a strong prior, will this compromise the lack of pre-test variables? When considering the influence of a strong prior, does it mitigate the necessity for pre-test variables? In other words, if we possess robust prior information, does this compensate for the absence or limited availability of pre-test variables in our analysis, and to what extent?

AG: You can include pre-test variables in your regression model. The prior distribution will be on the coefficients of the regression model. If you fit a model separately to several years of data, you can plot the estimate +/- standard error vs. time to get a sense of the variation of the coefficients over time.

Student:

1. During our final project about the Indian election turnout rate difference, we originally thought about using an IRT model for our dataset. Our first thought was that we would use the local election turnout rate as our “ability” and the assembly election as our “difficulty”. However, we also want to add other variables such as urbanization rate, sex of the candidate and party identification and more as our “difficulty” so that maybe a larger urbanization rate will be “easier” for the turnout rate to be large. But we weren’t really sure how to have so many “difficulties” for such a model. I was wondering how or if it’s even possible to have such IRT model that would take multiple difficulty inputs.
2. Also more of a general question about methods of deciding error rate for multilevel model: If I have a large dataset with a lot of potential predictor variables and unsure of which one to choose for our model, what’s the best way of selecting them if they are all valid? for example, if I have x1 to x5 predictors and in the first iteration, I have model such as y ~ x1 + x2 + (1 | x1). And if it turns

out that x1 to be an insignificant predictor, should I take it out before adding other predictors? I would like to know more about the intuition and methods of identifying and solving potential overfitting problems of the multilevel model.

AG: Regarding your question #1: In an IRT model, ability and difficulty are latent parameters. In your example, turnout rates and election outcomes are data. It would not make sense to map them into ability and difficulty parameters. It is better to figure out what your research question is, and then fit a regression model predicting your outcome of interest from your predictors.

Regarding your question #2: No, I do not think you should include or exclude parameters based on statistical significance. Take a look at chapter 12 of Regression and Other Stories for discussions of model building. The same principles apply for multilevel models.

Student: For Question 2, I think techniques like forward selection and backward elimination are effective in identifying key predictors in multilevel models. Forward Selection starts with a model containing only the intercept, possibly including random effects. Predictors are added sequentially, each assessed using AIC, BIC, or p-values from likelihood ratio tests, and retained if they significantly improve the model. The process continues until no further improvements are noted. Backward Elimination begins with a full model, removing predictors one by one based on the same criteria, continuing until only significant contributors remain. These methods help manage the complexity of multilevel models, striving for a balance between simplicity and explanatory power. In multilevel models, these methods might be computationally demanding. Changes in fixed effects may affect the appropriateness of the random effects structure, necessitating a reassessment of the random effects whenever fixed effects are modified. Both forward selection and backward elimination aim to balance model simplicity with explanatory power, essential for building effective and interpretable multilevel models. Additionally, regularization techniques like Ridge and Lasso offer further assistance. In multilevel models, where predictors at different levels might be highly correlated, Ridge addresses this by penalizing the size of the coefficients, thus mitigating overfitting risks. Lasso, by applying a penalty proportional to the absolute value of the coefficients, can reduce some coefficients to zero, effectively performing variable selection and simplifying the model.

AG: No. Never do forward or backward selection. It's a good idea to build up your ultimate model by starting simple, but the way to do this is to keep adding things that make sense, if necessary combining predictors if estimates are noisy. Do not use statistical significance as a criterion. Also, there should be no need to balance model simplicity with explanatory power. In general, more complicated models are better, using priors if necessary to stabilize inferences.

Student:

- This is just my opinion, but having many dimensions of "difficulty" would actually be a good motivation to use IRT-type models; we're using the model to abstract the complex structure that is present in the available data, and nothing is stopping us from assuming the "difficulty" to be multi-dimensional (apart from maybe computational constraints).
- Also, I slightly disagree with the above on the use of forward- and backward-selection w/ AIC/BIC/p-values. I'm assuming Novak's context involves a Bayesian IRT-type model. Both AIC and BIC is based on frequentist theory involving

KL-divergence from "the true data" versus predictions (where AIC look at the "forecast error" when forecasting one period ahead after looking at data of size n, and BIC looks at the "cumulative forecast error" for n periods from data n=0). Using p-values (or something like stepwise regression) is also a bad idea (see https://www.stata.com/support/faqs/statistics/stepwise-regression-problems/).

- One way I know is to compare the posterior model probability between different models.
- There are also Bayesian variants of regularization (e.g. Bayesian LASSO: https://www.tandfonline.com/doi/abs/10.1198/016214508000000337), which strictly speaking does not induce sparse estimates if we are simply running MCMC, which is something to be mindful of.

AG: Regarding predictive model evaluation, I recommend chapter 7 of BDA3 and also this 2017 article from Statistics and Computing:
http://stat.columbia.edu/~gelman/research/published/loo_stan.pdf

Student: The frameworks and models we've learned this semester have been really great at getting me to think more deeply about the assumptions and variation underlying any dataset. In particular, I've been trying to move towards a Bayesian workflow (laid out very well in this paper by yourself and others) when I approach data analysis. The approach makes a lot of sense for helping a researcher/analyst probe their model assumptions, check fit and computation, as well as modify and expand the model in an iterative way. What I'm trying to reconcile now is how much of this workflow should be communicated to an outside audience, in particular research papers that one may want to eventually submit for some publication. I think there is a lot of inertia in the standard format of Intro-Lit Review-Model-Results-Conclusion and most papers that I read move very quickly through the Model and Results sections, sometimes just presenting a table with coefficients (highlighting stat significance mostly). My guess is that most researchers spent a lot of behind-the-scenes time wrangling their data and building their model, but that process (and assumptions in the decisions made) don't get communicated. So, should we be adjusting the format of communication to include more info on model construction, evaluation, and modification? And if so, will that face pushback from editors and audiences that may adhere to a standard structure for communication?

AG: I do think it is better to communicate more of these steps. Here is an example: the published paper is here: http://stat.columbia.edu/~gelman/research/published/millennium_final.pdf and the earlier work is here: http://stat.columbia.edu/~gelman/research/unpublished/1507.02739.pdf and here:
http://stat.columbia.edu/~gelman/research/unpublished/MVP_paper_technical_JRSSA_short.pdf
Unfortunately, we were not able to publish that earlier work! It can be easier to demonstrate this sort of workflow in books (such as our forthcoming Bayesian Workflow book), case studies (such as these: https://mc-stan.org/users/documentation/case-studies) and informal outlets (as here: https://statmodeling.stat.columbia.edu/2022/02/18/hierarchical-model-golf-putting-struggle/ or, for a simple example, here: https://statmodeling.stat.columbia.edu/2024/05/04/you/).

Student: This class was the first time I got to use Stan.

- I was quite surprised at how the model building process was much smoother with Stan: coming from a Bayesian econometrics background, where it's common practice to meticulously derive every single full conditional posterior when building and estimating a single Bayesian model.
- I was also surprised to see how various numerical tricks were implemented in Stan to improve convergence (e.g. orthogonalization w/ QR, non centered parameterization, etc.) of HMC. (If you have experience coding HMC, you know the hyperparameter tuning is quite tedious, which was unrequired AND worked well in Stan!)
- At the same time, the generality was at the cost of poor convergence or speed for a certain class of models (e.g. we discussed in class the stationary AR).
- My natural question/thought was, whether we are able to implement a HMC-within-Gibbs sampler (or if there are any plans to do so) that combines Stan's efficient HMC implementation, alongside model-specific conditional sampling strategies that is known to work well. Taking AR for instance, the forward-filtering backward-sampling algorithm (aka FFBS) is known to be extremely fast and efficient to jointly (and not single-moves) sample latent states in dynamic models, which can be useful for models with latent correlated year effects. Implementing FFBS is quite easy as well while requiring no tuning nor rejection. It would be nice to know if there are Stan implementations of these conditional sampling strategies.

AG: I think that with a good reparameterization you should be able to resolve some of these time series computation problems, for example by parameterizing using independent or approximately independent error terms rather than highly correlated parameters. I'm not sure, though. You could try with an example and see how it goes. You can also ask on the Stan Forums (again, supplying an example) and someone might have a useful reply.

Student: One issue my partner and I ran into during our final project was finding the best way to include time in our multilevel models. Our project explored the differences in voter turnout rates between Indian local and national elections. As a reminder, the years for national elections were consistent across states, but the elections within states occurred on distinct schedules. One of our biggest struggles was figuring out how to match each of the schedules, which ultimately created two datasets, one of which removed some potentially redundant data. We spoke about this in our second presentation, and I continue to be a bit confused by Professor Gelman's feedback. He suggested we effectively reformat our dataset from "wide" to "long" and create indicator variables if the election is at the local or national level. That would certainly help us avoid losing/having to remove data. However, I think the outcome variable would have to change directly to turnout percentage (from difference), and, more importantly, I'm curious if we would then lose some of the multilevel/hierarchical structure? There were, naturally, variables that were specific to the state level, for example our urbanization data, that I'm unsure how we would include for the local election rows in the "long" dataset. Continuing with the thread of time, I've looked in ROS and Data Analysis Using Regression and Multilevel/Hierarchical Models at the sections on panel data, and am also trying to connect this to our class material from weeks 6 and 7 on measuring CD4 percentages of children with HIV over time. In the textbooks, panel data is mostly covered within the context of causal inference, which I don't think is applicable in this context. I'm wondering if there is a better way to run the multilevel model from our data (we set year as a fixed effect and random effect), as well as if there are suggestions for additional reading on how to handle panel data? Or if anyone has experience with the MCMCpack package, which the textbook notes is good to use with multilevel modeling of panel data.

AG: You should be modeling voter turnout in percentage, not absolute votes. It would not make sense to model absolute votes, in that any effects you might expect to find should be on the percentage scale.

Regarding your other question: in multilevel modeling, panel-data models are called nonnested models. We have examples in my book with Hill. Don't think about so-called fixed or random effects; instead, just allow intercepts and possibly slopes to vary by country and by year.
Also, I do not recommend MCMCpack; it will be easier to use rstanarm, brms, or just write the model directly in Stan.

Student: For the final project my partner and I conducted, we mainly relied on the tutorial provided for the homework due class 10a to conduct multilevel regression and post-stratification. As shown during our final results, it seems like the MRP estimates by state are too similar to each other. In other words, the small variability between each state's opinion on affirmative action did not appear to be plausible since as affirmative action has long been a divisive issue among the states. I revisited our code and realized that we made a rookie mistake of not including `state` as a categorial variable for the model used to generate the estimated poststratification cells. Thus, in an effort of obtaining more accurate estimates, I readjusted the hiearchiacal model to make sure the state variable is included in the fit as a categorical variable (i.e., varying intercept), regenerated the prediction matrix, and replotted the comparison plot. However, the results did not seem to improve (i.e., state estimates are still very close to each other). After reading the paper non-representative polls done on Xbox, of which Prof. Gelman is an author, it appears that the overall framework seem quite similar, with some flaws in our approach. Since we obtained our poststratification table via the `ccesMRPprep` library, I later realized that the number of poststratification cells is not equal to the product of all the levels included in our paper. In the Xbox paper, it was explicitly stated, "*`Specifically, we generate the cells by considering all possible combinations of sex (2 categories), race (4 categories), age (4 categories), education (4 categories), state (51 categories), party ID (3 categories), ideology (3 categories) and 2008 vote (3 categories), thus partitioning the data into 176,256 cells.`*" In our poststratification table, there are two levels of gender, six categories for ethnicity, three levels of education, and 50 categories of state, totalling to 2*6*3*50 = 1800 poststratification cells (1740 in our case, since some combinations has an *n* = 0). However, when fitting the hierarchical model, there are two levels of gender, eight categories for ethnicity, six levels of education, and 50 categories of state, totalling to 2*8*6*50 = 4800 possible combinations of demographic variables. Is the reason why MRP failed to provide meaningful results for our paper due to the mismatch between the levels presented in our multilevel regression model (4800) and the number of poststratification cells (1800)? In other words, is it a requisite to ensure that mismatch does not exist for the poststratification table? If so, what is the best way to overcome these inconsistencies? For instance, it is known that the US Census did not include "Middle East" as a category for race and ethnicity until recently. In that case, what is the best practice to obtain poststratification tables? An example from the Xbox paper is utilizing the Current Population Survey (CPS) in junction with the exit poll data from the 2008 presidential election. What are some other approaches to ensure the poststratification cells are correctly derived?

AG: Yes, when poststratifying you should include all interactions of the factors as poststratification cells. If you do not have all the relevant information, you can either reduce the complexity of your model (for example, just using 4 or 5 ethnicity categories instead of 8) or else you can use some estimate of the poststratification table. In your above example, it sounds like you might have a coding error. You can check by simulating a large fake data set from your model, then fitting your model to the simulated data and checking that you can approximately recover the assumed parameter values.

Student: As for me, during my final project about the world happiness report, my first issue is about interpreting the model. Firstly, how the Bayesian model can be further refined to account for potential confounders and reverse causality. In my model, I take health and gdp into account, there is a strong correlation between health and happiness. Health can lead to happiness, and happiness can also lead to

health. Are there any indicators for their relationship? Do I need to make a reverse regression to see how different variables affect other variables? And how we can interpret the coefficient in this way. What's more, I build a time series multilevel models, and I give a strong prior for the autocorrelation, but it seems that the autocorrelation tends to approach 1, when there are more years, which seems make the model meaningless in interpreting the coefficients. If I directly limit the range of autocorrelation or just set it as a small constant, such as 0.1, will it be better? If not, how to make the model more interpretable when I try to choose the potential autocorrelations.
My Second question toward the multilevel model is related to the group data. If the distribution for different groups varies a lot, or some groups have a strong prior but other groups have weak prior, will it have influence on the varying intercept or slope? For instance, in my dataset, during the 10 year period, we have full data for most europe countries, but it does not contain the full data for some turmoil countries or areas, which means that they probably have different distribution, so how to solve such problems? Or we just suppose that they are independent from the same distribution? Can I set a value to split different situations, if there is full data, they will follow one distribution, if the data is limited, they should go to another distribution?

AG: For your first question, read Appendix B of Regression and Other Stories, in particular B.4 (interpret regression coefficients as comparisons) and B.9 (do causal inference in a targeted way). Each causal question requires its own analysis. So, rather than fitting a model and then trying to interpret it causally, go the other direction: frame the causal question and then do as much modeling as possible to adjust for differences between treated and control groups. It's not so helpful to talk about the effect of "health" or of "happiness." Instead, think of possible treatments (interventions) and how you could estimate their effects. Section 21.5 of Regression and Other Stories (on causes of effects and effects of causes) is relevant here too.
For your next question: If you're getting autocorrelations near 1, I recommend you include some time-varying predictors in your model. For a simple example, if you have a linear trend in your data but you don't include it in your model, this will look like nonstationarity and it will induce high autocorrelations. But if you include a linear time trend along with the autocorrelation, the model will fit better and make more sense.
For your final question: If you have more data for some countries than others, that's fine: the Bayesian inference should give you stronger inferences for the countries with more data and weaker inferences for the countries with less data.

Student:

1. In the final project that my partner and I worked on (using MRP), the variable of interest, attitude toward affirmative action, was measured by a census question with the possible choices of 1 = favor, 2 = oppose, 3 = neither favor nor oppose, and 4 = not sure. To make meaningful interpretations, we reordered the choices as 1 = favor, 2 = neither favor nor oppose, 3 = oppose, and we dropped the "not sure" option, because it had much fewer responses than the other ones, and it did not fit into the favor-to-oppose spectrum. Including or dropping these responses also did not visibly change the multilevel regression results. However, in similar cases where the levels of a categorical variable requires some manipulation to be interpreted ordinally or numerically, what might be a better option for manipulating the data to make statistical interpretation make sense while minimizing the risk of biasing the data? Or would an alternative analytical method be more appropriate?
2. A general question regarding the multilevel models that we have constructed and discussed in class: I noticed that in the multilevel models, we would sometimes include a variable as both a varying slope and a fixed slope (e.g., y ~ (1 + x | z) + x). I wonder how this method differ from including a variable only as a varying slope or a fixed slope (e.g., y ~ (1 + x | z) or y ~ (1 | z) + x).

> How would the interpretation of the two models be different? In what cases might each model work best?

AG: I would recommend coding "not sure" as the same as "neither favor nor oppose" and then just doing a check to see if that is consistent with your data. In general with this sort of response with several possible levels, I recommend coding as a continuous variable. It's also possible to fit a model such as ordered logit that has a latent continuous outcome, but typically this is not worth the trouble.

For your second question: if you include x as a varying slope, you should also include it in the main model. So, y ~ x + (1 + x | z) is fine. But y ~ (1 + x | z) is bad (with rare exceptions). You can think of a varying slope as an interaction between x and the indicators for z. When including an interaction you also should include all the main effects.

> **Student:** For your question 2, I think when building multilevel models, including a variable as both varying and a fixed slope, y ~ (1 + x | z) + x, allows us to capture the overall effect of that variable across all groups while also modeling group-specific deviations through varying slopes. While including the variable only as a varying slope y ~ (1 + x | z) provides information of group-specific deviations but without estimating a global effect. And including only as a fixed slope assumes the effect is consistent across groups and allows group-level intercepts to vary. It depends on the specific research questions – when the average effect of x is important but varies across groups, we can choose the first one, and when only group-specific deviations of x are of interest, we can use varying slope models. When we are considering the effect of x is consistent across groups, we can use fixed slopes models.

AG: I just recommend avoid the terms fixed and random because of their ambiguity; for example see this explanation: https://statmodeling.stat.columbia.edu/2015/03/22/no-fixed-random/

Student: Being an undergraduate student with only a concentration in statistics, this class was a huge step forward from anything I have learned in my previous coursework. Consequently, I really enjoyed getting to see what Professor Gelman's workflow looked like for a lot of the examples we did in class/homework. Many of the models we looked at in this class were completely foreign to me, but they really provided me with a great new insight of ways to approach statistical questions. Due to the knowledge gap I feel I came into this class with, there are several questions I have had in regards to things we learned this semester.

1. I have been having a hard time figuring out exactly when it is appropriate to use the lmer function in R vs. when to write a stan model. From working on the final project, my partner and I, initially used just the lmer function and then in a later iteration of the project, we decided to include a Stan model. It was clear that by using the Stan model we could specify more information, which certainly improved the fit of the model to some degree, but I feel like I am not entirely clear on why. It is obviously noticeable that the lmer function runs much faster than a Stan model, and I assume this is because it is actually including less information. So, all this is to say what is really unclear to me is exactly what information Stan includes and what information lmer includes? Then, based on the answer to this question, the next question that follows is, when is it appropriate to use each of these modeling techniques? Is Stan always the better option information wise, but lmer runs faster thereby making it more practical in certain

situations? Or should something about the data or question we are trying to answer with the data determine when we opt to use lmer vs. Stan?

2. Some of the lines of information that get included in a Stan model confuse me as to their purpose. For example, what does the array[] do that is sometimes included in the data section of the stan model? Why do we sometimes include array and other times not? I know this has to do with the data but I am unsure how to tell when array is necessary and when it is not. Another example is the matrix part of data that is also only sometimes included. Why do we only sometimes need a predictor matrix? Is there any reading that anyone can recommend that does a good job of explaining when each line in Stan does? I have tried to look online without much success. Any advice on this would be greatly appreciated because I see the power of what Stan can do so I would like to understand it better to feel more comfortable working with it.

Just a final thought, I really appreciated all of Professor Gelman's thoughts on how to present data based on our final projects. The advice given on graphs, explaining models, etc. is something I will definitely hold with me well beyond this class.

AG: lme4 does marginal maximum likelihood. It gives a point estimate of the variance parameters, unlike Stan which averages over their posterior distributions. You can also fit Stan models using rstanarm and brms, which are no faster than Stan but are convenient because they use the lme4 syntax. There's also the blme package in R, which is a slight modification of lme4 that allows informative priors.
Regarding Stan itself, I recommend the documentation in the Stan User's Guide and other documentation here: https://mc-stan.org/users/documentation/ and also the Stan case studies here: https://mc-stan.org/users/documentation/case-studies.html

> **Student:**
> On the question of what the difference is between lmer and stan: LMER does maximum likely optimization and is equivalent to the optimize function in Stan. We use Stan to sample from the posterior (using Hamiltonian Monte Carlo simulation to get draws), which is much more computationally challenging. The confidence intervals we see with stan are from the draws from the posterior, whereas with LMER, we see the reported standard errors of the estimates. I believe if you have less data, your posterior would look less "nice" (smooth, maybe), and so you would want to use stan for that. If you have a lot of data, stan (and MCMC) might become computationally taxing, so lmer might be preferable.
>
> Array[] is a type that indicates a collection of some other type. For example, an "array[] int" indicates a collection of ints. You can read more on it here. We use arrays frequently when we want to pass in a bunch of observations (or predictors) that are of the same type. For example, if we have 100 scores that are integers we'd use an "array[100] int", which would tell stan to expect 100 integers in an array. You need to tell stan how many elements are in the collection. If you want a "array[100] real", you can use "vector[100]" instead. A matrix is an "array of arrays" (or maybe a better explanation is a "vector of vectors"). When we have multiple predictors, we might want to use a matrix instead of explicitly passing in multiple arrays.

Student: In the application of Multilevel Regression with Poststratification (MRP) for electoral prediction using polling data, one critical challenge is the selection and creation of poststratification

cells that accurately reflect the demographic and geographic variations in voter behavior. The traditional approach often involves somewhat arbitrary definitions of poststratification cells based on common demographic variables such as age, income, and education. However, this method may overlook subtler, yet significant, demographic interactions that can impact voting behavior. Considering the limitations of traditional methods in capturing complex interactions and nonlinear relationships, how can machine learning models be effectively integrated within the MRP framework to optimize the creation of these cells? Specifically, could unsupervised learning techniques such as cluster analysis be used to identify naturally occurring groupings within populations based on a broader range of demographic, behavioral, and social variables, which might not be initially apparent? Furthermore, how might supervised learning algorithms help in dynamically adjusting the weights or influence of different levels in the hierarchical model based on real-time data inputs? This integration poses several challenges, including ensuring the interpretability of the machine learning-enhanced MRP model and managing increased computational complexity. Moreover, in what ways can Bayesian methods be employed to assess and communicate the uncertainty inherent in these more complex models, especially when used to inform decisions in high-stakes scenarios like predicting election outcomes?

AG: I don't think it would make sense to use cluster analysis to set up poststratification cells. Often the choice of poststratification cells depends on what is the population information available. It is also possible to include poststratification variables for which you don't have population information; in that case some modeling is necessary. Here's an example where we used religious attendance as one of our poststratification variables:
http://stat.columbia.edu/~gelman/research/published/socsci-12-00430.pdf

Student:

1. Multilevel models are no doubt useful for prediction problems, especially when there are a small number of samples. They can also leverage the hierarchical nature of blocked experimental design for calculating causal treatment effects. I have some confusions, however, for when multilevel models can successfully enable us to draw causal inferences in observational data. Observational causal inference with single-level regression seems more standardized: control for all potential confounders (especially pre-treatment outcome) and avoid controlling for any potential mediators or colliders. But how do we deal with group-level variables (like varying intercepts) that are correlated with main effects? My impression is that this breaks the exogeneity assumption needed in regression to make causal inferences. This question was addressed in chapter 23 part 2 of Data Analysis Using Regression and MLMs, where the correlation between varying intercepts and a treatment was modeled explicitly by having an aggregated treatment partially determine the varying intercepts. Does this resolve the issue specifically by removing correlation between main effects and varying effects? If not, what's happening? Further, even when we model treatment effects as varying such that we no longer have a correlation between main effects and varying effects, couldn't the varying intercepts/treatment effects still be correlated with other lurking main effects? As shown in causal quartets, heterogeneous treatment effects are common, so it seems we would always need to worry about how varying treatments correlate with main effects. Are these types of broken assumptions particularly pernicious?
2. Additional point: can the comparison between frequentist factors and MLM-esq varying intercepts be used to better help us better answer the above question? During my final project, I originally included a categorical variable on cuisine as a factor rather than a varying intercept. Consequently, the cuisine variable pulled explainability away from restaurant-level effects. When I

included cuisine as a varying intercept, the restaurant-level variables absorbed all of the explainability and cuisine became null. If we solely care about prediction, it doesn't matter apart from numerical stability. If we're making causal inferences, how can we know which specification offers the right estimates?

AG: You can think of a multilevel model as a regression model where the coefficients are estimated using regularization. From that perspective, multilevel modeling allows you to adjust for more things when using regression for causal inference (and please say "adjust for," not "control for"; see here: https://statmodeling.stat.columbia.edu/2019/01/25/regression-matching-weighting-whatever-dont-say-control-say-adjust/) and that is a good thing, but it does not resolve all problems of causal inference any more than any other regression method resolves all problems of causal inference. So just think of multilevel modeling as a regularization tool that allows you to adjust for more things.
Also, if you allow the treatment effect to vary by group, this can be useful–it's something that multilevel modeling lets you do.
Regarding your second question: you should include the cuisine variable and also include varying intercepts for restaurants. This should not reduce the varying intercepts for restaurants to zero, but it should reduce them, which makes sense, as your restaurant-level predictor is explaining some of this variation. The answer to this question doesn't depend on if you're doing causal inference or just prediction; indeed, causal inference is just a special case of prediction where you are specifying the causal predictor, thus defining potential outcomes.

Student:

1. We used the glmer function in class to fit generalised linear mixed-effects model. In reading about these, I found that one of the assumptions behind this is that the random effects(by which I mean the varying intercepts or varying slopes) have a normal distribution. To what extent can we expect this assumption to hold, especially with smaller samples as is often the case when we use multi-level models? For casese where the data is clustered into a small number of groups, how does this assumption play out?
2. I was faced with the problem of missing data in my final project, and I used a model similar to a zero-inflated model where I first modeled the probability of missingness and then created a similar model for non-missing data. However, I wished to use stan to impute the missing data as well. The bayesian approach to this involves estimating multiple samples for those missing data, incorporating various covariates. However, in this case the data was not missing at random. Missingness of the outcome variable (number of people affected in natural disaster) was correlated with the predictors (income level of country, type of disaster). I read about data imputation on stan, and saw it uses predictive mean matching, which may not work in this instance because it is missing not at random. What are different ways to tackle this problem? Should I try simulating the data anyways?
   Student: The assumption that random effects follow a normal distribution is helpful for coming up with computationally tractable estimation.I agree that this may not hold, especially for small sample sizes, but I think this is where you need to do posterior diagnostic checks. Maybe you need to use a more robust approach or maybe a different distribution of random effects.

AG: If you want to model the varying coefficients as having a non-normal distribution, what I recommend you do is to include group-level predictors in your model. The point is that the varying coefficients represent unexplained group-level variation, and once you've included some group-level predictors, it's typically more reasonable to assume a normal distribution for that error term.

For your second question: yes, imputation can be difficult and the most important thing is to be transparent. I don't have any great advice here except for what's in chapter 17 of Regression and Other Stories; also you can do some graphical checks of the imputations: http://stat.columbia.edu/~gelman/research/published/paper77notblind.pdf
In response to the student who responded to the above question: it seems natural to consider nonnormal distributions of these varying coefficients, and sometimes this can make sense, but usually I think we benefit much more by including additional predictors rather than focusing on the distribution of the error terms.

Student:
In our final project, we applied an ideal point model to estimate the ability scores of basketball teams. While our model did not accurately predict the score differences of each game, our primary objective was to accurately ascertain the abilities of the teams rather than their game outcomes. We put in a considerable amount of effort towards validating the accuracy of the ability scores we derived. Thankfully we had the season's final standings, which helped us to verify that our ability scores were sensible.

In typical predictive modeling, performance evaluation is straightforward, often based on clear-cut objective functions. However, in the context of our Bayesian ideal point model, our evaluation was predominantly conducted through posterior predictive checks, which sometimes seemed overly qualitative. These checks included comparing posterior distributions of specific parameters or ability scores, looking at likely intervals of parameters. Although we considered using the Kullback-Leibler divergence to compare the actual outcome distribution with our generated outcome distribution, we couldn't lean on this as we usually do with objective functions. Minimizing the KL divergence isn't exactly what we're trying to do either. The Bayesian model development loop appears to lack the straightforwardness of the frequentist approach, which typically involves iterative model modification and evaluation based on specific metrics.

How can we more quantitatively assess the accuracy of the ability scores estimated by a Bayesian ideal point model? What are effective strategies or metrics that can be employed within the Bayesian framework to create a more structured and quantitative model development loop similar to that used in frequentist approaches? How can we enhance the rigor of our model evaluations to ensure that the Bayesian model not only fits the data well but also truly captures the underlying processes governing team performance?

AG: I recommend embedding your problem into a decision model. That is, consider some hypothetical decisions that could be made, set up costs and benefits, and then work out the optimal decision that maximizes expected value. Then you can evaluate your procedure using simulation and see if the resulting decisions make sense, see how your decisions do in comparison to the assumed truth from your simulations, etc. To set up the hypothetical decision problem takes some work (some decisions!) but I think the effort is worth it.

> **Student:** While I agree that model evaluations are straightforward in the typical iid setting predictive setting, I think the difficulty with evaluating Bayesian ideal point models are rather due to the difficulty of evaluating ideal point models in general, and not due to Bayesian/frequentist dichotomy. In typical settings, it's conceptually straightforward to perform out-of-sample predictive evaluations based on e.g. cross validation, whether the model is Bayesian or not. If the model is frequentist, you could evaluate point predictions. If you are Bayesian, you would simply use the Monte Carlo average as your point estimate, approximate upto MC approximation error. In

settings where we use ideal point models, however, "out-of-sample" could be framed in many different ways, and some of them might be nonsensical depending on the context.

Student: Since the data is a time series, one way might be to consider fitting consecutive models causally on portions of the data and seeing how the model performs on the predictions for subsequent matches. For example, we could have fit a model for the first 40 games and then seen how likely those ideal points captured the winners of the following 10 games. For an acausal estimate, we could have tried training models on random subsets of the data and seeing how those predicted the winner of unseen games.

AG: I agree that cross validation can make sense. Here's a paper on cross validation for time series: https://arxiv.org/abs/1902.06281

Student:
- When doing item response models, in class we evaluated the fit by graphing response curves per item and seeing where the trend does not match our expectation. For example, we graphed the percentage to get the question right for a given function as a function of student ability to identify questions that were incorrectly graded or questions that were “ineffective”. Is there a standard way to evaluate the fit for such IRT models, especially when you consider adding predictors in (i.e. what if we added what students had for breakfast as a predictor)? Additionally, we found that with our item response models, we got huge confidence intervals for the outcomes of matches. It was suggested that if we had higher quality predictors for the teams, that might have made the fit better and resulted in tighter confidence bounds. When not present, I would assume these predictors are “absorbed” into the ability scores for teams. Would adding them help here, and if so how?
- In this class, we spent a lot of time dealing with cases where we have a reasonable amount of data (say <100,000 observations) and we can look at the plot for an individual group pretty well. What diagnostic tools can we use when we have huge amounts of groups (i.e. thousands or hundreds of thousands)? Are there other parts of the workflow that would change? What if we have millions of data points? How can we do diagnostics when it’s very difficult to plot that much data?
- When does it make more sense to display data in a table versus a plot? It seemed like from class that a chart was almost always better, but at the same time a plot can sometimes display too much information or make it harder to focus on distinctions than maybe a single number could. In a related question: when does it make sense to display a single number aggregate (i.e. average of a group, or maybe average +/- sd) versus a more complicated graph displaying a distribution (like a qq plot or histogram). Additionally, one advantage you have of a table versus a plot is that you can see the exact numbers. Are there any clever ways to provide the precision of a table with the pattern ease of a line chart?

AG: For quantitative evaluation of fit, you can use leave-one-out cross validation as discussed in chapter 7 of BDA3 and this article: http://stat.columbia.edu/~gelman/research/published/loo_stan.pdf and you can use the loo package in R.

If you have hundreds of thousands of observations, you can plot a grid of graphs on a single page. No need to try to cram all your datapoints into a single plot. It's amazing how much you can see with a careful grid. Also, for binary data you can plot binned averages.

I recommend never using tables; see discussion here: http://stat.columbia.edu/~gelman/research/published/dodhia.pdf If people might want the exact numbers, you can post the raw data online.

Finally, I usually would prefer to display scatterplots rather than histograms. I rarely am interested in univariate distributional summaries. And if the data being plotted have roughly unimodal and symmetric distributions, I'd just summarize by mean and sd (or median and mad sd, as discussed in Regression and Other Stories).

Student:
I have been trying to gain a better understanding of how priors are selected and the implications of selecting strong/weak/flat priors. Thank you for your insight in advance!

(i) Is there ever an acceptable case for including a non-informative/flat prior in a study?
We are generally building on findings discovered from past studies or some sort of assumption we have about the world that would necessitate the inclusion of some prior. And in the case that there is no knowledge on the phenomenon being analyzed, couldn't you simply take a portion of the data, perform an analysis, and use the posterior from that analysis as your prior in your study assessing the remaining data. (ii) In the cases where all values are equally likely, wouldn't that also be considered a prior, and a strong one at that? I would imagine that that would also lead to some distortion of the posterior. In subject-areas where studies are predominantly frequentist, or bayesian containing non-informative priors, I would assume that results would be a lot noisier/contradictory. With regards to the replication crisis in psychology and many of the other social sciences, could this partially be attributed to the lack of inclusion of prior information in published studies?

(iii) In studies with high-dimensional data, would the inclusion of informative priors create distorted results?
When setting priors, we assume independence between parameters. If we are dealing with highly dimensional data, where it's difficult to assume total independence, adding informative priors for parameters with collinearity, I would imagine, creates results that are distorted and overconfident.

(iv) As a follow up question, how do you handle discrepancies in prior information across parameters? Let's say there's a lot of prior information on parameter a, so you add a highly informative prior, and very little information on parameter b, so you add a flat prior. How would you make sense of those results which may be skewed in a direction favoring a? One suggestion I can think of would be to develop several models with different prior values (both flat, both weak, one strong/one weak) and compare posterior distributions across all the fits.

(v) On a different note, for one of the homework assignments we were asked to predict opinion on gun control in a region, given republican vote count and the rural population. The model we developed was a non-nested model with both rep percentage and region/rural population as main effects. With non-nested models, I am slightly confused on how to interpret the results. I understand how to interpret them in silo, "for every percentage increase in … results in a change of …", but how would I factor in the other parameter in this interpretation? For example, in interpreting the impact of a region being rural on gun opinion, shouldn't I also mention the republican vote count parameter's impact on my findings? Would this require that I include an interaction between region and republican vote count in my model?

AG: (i) See our page on prior choice recommendations: https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations I do think that flat or weak priors can be useful. We think of them as placeholders: you use a flat or weak prior, get your inferences, and if your inferences are too wide and you realize you have additional information available that would help constrain the inferences, you go back in and include that as a prior.

(ii) Yes, I think the use of flat priors has contributed to the replication crisis; see discussions here: http://stat.columbia.edu/~gelman/research/published/bayes_management.pdf and here: http://stat.columbia.edu/~gelman/research/published/failure_of.pdf
(iii) In high-dimensional problems (high dimensionality in the sense of a large number of parameters, not high dimensionality in data), any prior will be strong. In such problems, the prior is really a key part of the model, not just something that's added in. For an example of a high-dimensional problem (actually, just 30 dimensions, but the data are sparse, so 30 parameters is a lot in this example) where a seemingly noninformative prior turns out to be way too strong, see section 3 of this paper: http://stat.columbia.edu/~gelman/research/published/deep.pdf
(iv) If you want to include more prior information on one parameter than another, that should be fine; the Bayesian math should be able to handle this.
(v) When we say "non-nested," we are talking about non-nested varying coefficients, for example a model with intercepts that vary by country and by year. In the gun control example, the state-level predictors for R vote and %rural are not multilevel factors–they're just simple regression predictors–so we would not refer to this as a non-nested model. Regarding your particular question: I do not think it makes sense to give these coefficients causal interpretations. Recall Section B.4 in appendix B of Regression and Other Stories: interpret regression coefficients as comparisons. So don't say, "for every percentage increase in … results in a change of …"; instead say, "comparing two states that . . . corresponds to an average difference of . . ."

Student:
This is much more of a philosophical question, but maybe it's interesting to some people. I've been thinking a lot this semester about the notion of iterative model fitting and its relationship to scientific (or social scientific) discovery. A motif running through the class, in my mind, has been the general idea of fitting simple models, seeing how they break or what "doesn't make sense" about them, and trying more complex models or making adjustments to find better ones. If we're fitting a model with the aim of "discovery," though—maybe we want to discover a candidate's county-specific approval ratings—this step of examining a model seems potentially problematic to me. If I adjust a model, perhaps strengthening the partial pooling effect, based on my prior intuition for what I think "should" be happening, don't I limit my ability to detect when something novel or surprising really is the case?

Through steps and steps of iterative model development, maybe adding or removing interaction terms, transforming covariates or responses, strengthening or changing priors, my result seems more and more "reasonable." I think I'm doing a great job, but what if the choices I'm making through this iterative process—crucially, informed by the fact that I can see the results of a model—implicitly nudge my results further and further away from what is *truly* going on?

We all, of course, have prior beliefs and expectations (Alabama is very Republican, urban areas are more wealthy, etc) and want to leverage them to fit better models. But how can we make sure that we don't accidentally *create* the results we expect to see through this model fitting process? How do we balance these expectations with the knowledge that something surprising might actually be going on, in a principled way?

This felt especially relevant to me when it came to our discussion of partial pooling. A great case for partial pooling is that *we would expect* a certain amount of variability across our categories but due to noise and small sample sizes, we see far more variability in practice. So we use our intuition, and pull the estimates in towards some conditional mean to stabilize them. But what if the true value in one category really is somehow different than the rest?

I know it's just a computational tradeoff to determine how much pooling is applied, but it feels unsettling. Is there a kind of sensitivity analysis we can do to understand just how different a given value would have to be for it to not receive as much pooling? Is that even helpful?

AG: There is indeed a contradiction between workflow (iterative model building) and inference (which is conditional on a model). Ultimately you can check things by external validation on new data. Also there are analytic ways to adjust for overfitting (see chapter 7 of BDA3 and this article: http://stat.columbia.edu/~gelman/research/published/loo_stan.pdf).
Regarding partial pooling: you ask, "what if the true value in one category really is somehow different than the rest?" In real life they will all be different from each other. But if you are concerned, what if the true value in one category really is _much_ different from the rest, then it would make sense to fit a model with a group-level predictor capturing what you think would make this category different. Or you could fit a long-tailed or mixture model, which is mathematically equivalent to the regression with group-level predictor, if this group-level predictor is not observed. (A latent-variable regression, when integrating over the latent variable, corresponds to a mixture model.)

Student: This class was the first time I used Stan and focused on a more frequentist approach to econometrics. In the beginning, I was still unsure how to integrate Stan to run my models, but once I understood the intuition behind it, I was able to run my models more smoothly.

Prior to this course, I have gotten very comfortable in the application of R and the application of different *classical inference regression models* to receive estimates and predictions that have well understood statistical properties, low bias, and low variance. However, this course made me realize that my data often does not have these properties, and the application of Bayesian inference has the great advantage of taking this into account.

With regards to the fitting of my models, it was a novel approach for me as a statistics student to reflect more critically on the *prior information* I receive from my data and how to incorporate it into inferences, rather than simply summarizing my data straight away, which I have done in the traditional approach of classical inference so far.

In return, I question the traditional approach to causal inference more critically, as the pure summary of my data can have limited values as predictions. Instead, Bayesian inference has the great strength to make valid predictions, even if there is little data, or the priors are weak and noninformative. Moreover, even though similar results can be obtained to the classical estimates, I realized that there is a great strength in conducting simulation draws to grasp the predictive uncertainty of the model.

Throughout this class, I found great joy in the "park" example discussed in class, as I have encountered ideal point estimation models in the past, yet never had the chance to understand them in more depth. Not only did this example show me how to rank the respondents' preferences on a scale depending on their attitude, but also showed me how to cleverly graph these points along the axis of "agreeability" with the park rule that has been established. In the future, I can see myself applying this knowledge to my own research interest to analyze voting decisions of members of parliament, or members of the cabinet.

(I am also still wondering to what extent the order of the questions asked in the park survey has an influence on their answers.)

AG: One concern about Bayesian inference is the strong dependence on the model. So it's important to check model fit. One cool thing about Bayesian inference is that it's fully probabilistic, so you can simulate

replicated data from the posterior predictive distribution and compare to the observed data, as discussed in chapter 6 of BDA3.

Student: For another class of mine (Frontiers of Justice), I just read a public health review paper on racial residential segregation and its impacts on health outcomes (Williams and Collins, 2001, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1497358/pdf/12042604.pdf). Towards the end of the paper, the authors write that "theoretically driven multilevel analytic models are needed" to understand how physical and social environments relate and interact with each other and individual factors to determine health (pg 412). This got me thinking...

I've been mulling over this notion of the relationship between theory and quantitative work a lot recently, because I like to think of myself as a quantitative person with interdisciplinary interests in various social sciences. So I'm curious as to your thoughts on the ways in which theory can inform, help, or hurt quantitative empirical research? What has been the role of theory in your own research over the years?

AG: I do think theory is important, although often it comes up in the background, more to motivate the questions than to form our models. Here is an example of a model from pharmacology that is heavily theory-based: http://stat.columbia.edu/~gelman/research/published/bois2.pdf and here is an example of research from political science that has a lot of implicit theory but without any formal model: http://stat.columbia.edu/~gelman/research/published/bjps1993.pdf Here is an example from political science which is almost entirely theoretical: http://stat.columbia.edu/~gelman/research/published/gelmanaxelrod_001.pdf and here is an example where we use data to criticize certain theories: http://stat.columbia.edu/~gelman/research/published/gelmankatzbafumi.pdf