Sample take-home final exam for Regression and Other Stories

We will send you a topic and a dataset.

You will have 48 hours to do this exam. The exam will be handed out at 8am. Submission deadline is two days later at 8am. You should be able to do it in one day; we're giving you two days so that you can do the exam on day 1, get a good sleep, and then go over your answers on day 2 before submitting.

Select questions that are jointly worth 130 points.

At the top of your submission you should indicate which questions you intend to address with the data. Answer in full sentences. Your submission should be a Rmarkdown file rendered as a pdf. Run `stan_glm` using the `refresh=0` setting to suppress intermediate output. Please also provide your code in a separate R file. The R file should include all commands from the Rmarkdown file and should run independently. It should include library calls and the data should be loaded in the environment from the same folder as the R file is in.

Good luck!

Questions

Answer all questions *which you have chosen* below in the context of this applied problem and the data you have downloaded.

Chapter 1 (10 points):

- Discuss the challenges in this example of generalizing from sample to population, from treatment to control group, and from observed measurements to the underlying construct of interest. Give two sentences for each.

Chapter 2 (10 points):

- Discuss issues of reliability and validity for this example, giving two sentences for each.
- Make a grid of graphs exploring your data, along with a paragraph describing what you have learned. Add another paragraph explaining how you have used the principles of statistical graphics in making your plots, and another paragraph discussing what important aspects of the data you were not able to include in these graphs. Your graphs should communicate key features of the data clearly. Use `par(mfrow())` in base graphics or `facet_wrap/facet_grid` in ggplot2.

Chapter 3 (10 points):

- Consider a deterministic model on the linear or logarithmic scale that would arise here. Graph the model and discuss its relevance to this example.

Chapter 4 (10 points):

- Perform a simple comparison of treated vs. control or exposed vs. unexposed in your data. Compute the standard error and 95% confidence interval for this comparison.
- Discuss a possible source of bias or unmodeled uncertainty and estimate its magnitude, in comparison to the standard error you just calculated. Give a sentence discussing the relative importance of the bias or unmodeled uncertainty, compared to the uncertainty in the comparison from the data.

Chapter 5 (20 points):

- Construct a probability model—a function in R that reflects the process by which (a) the true unobserved data and (b) how the observed data have been generated, e.g. measurement error, selection bias)—that is relevant to your question and use it to simulate some fake data. Graph your simulated data, compare to a graph of real data, and discuss the connections between your model and your larger substantive questions.

Chapter 6 (10 points):

- Take two variables that represent before-and-after measurements of some sort. Make a scatterplot and discuss challenges of "regression to the mean" when interpreting before-after changes here.

Chapter 7 (10 points):

- Fit a linear regression with a single predictor, graph the data along with the fitted line, and interpret the estimated parameters and their uncertainties. Write three sentences, one for each parameter in the model.

Chapter 8 (10 points):

- Use `lm` to fit a regression model with one predictor using least squares. Perform a calculation in R to demonstrate that the estimated slope equals a weighted average of slopes from all pairs of points.

Chapter 9 (10 points):

- Fit a linear regression to your data and use `predict`, `posterior_epred`, and `posterior_predict` to make predictions for a new data point with the predictor set to a reasonable value. In three sentences, summarize the results from these predictions and explain how they differ.

Chapter 10 (10 points):

- Fit a linear regression with multiple predictors including at least one interaction. The model should make sense; that is, there should be a good applied reason for fitting it. Explain each of the estimated parameters and their uncertainties, using one sentence for each parameter.

Chapter 11 (20 points):

- Fit a linear regression with multiple predictors. List all six of the assumptions of the model and explain, in one sentence each, how these are relevant to this example.
- Fit a linear regression with multiple predictors. Use residual plots (at least two) to assess the fit, and describe in two sentences what you found.
- Fit two different linear regressions, each with multiple predictors. Both models should make sense; that is, there should be good applied reasons for fitting them. Compare the fits using leave-one-out cross validation, and in one sentence discuss what you found

Chapter 12 (15 points):

- Fit a linear regression including a log transformation of predictors, outcome, or both. The model should make sense in the applied context. Graph the data and fitted model, and in a few sentences explain the result including the transformation.
- Fit a linear regression with multiple predictors. Take one of the continuous predictors, bin it into discrete categories, and use these discrete indicators as predictors. Plot the data and both fitted models on the same graph.

Chapter 13 (10 points):

- Fit a logistic regression that makes sense in the applied context. Plot the data and fitted regression line. Interpret the coefficients and their uncertainties on the probability scale using the divide-by-4 rule.

Chapter 14 (20 points):

- Fit a negative binomial regression that makes sense intEvaluate logistic regression using fake-data simulation. Check that the coefficients parameter estimates are approximately unbiased and that 95% intervals have approximate 95% coverage. Your simulation should be realistic in that the simulated data should be similar to the real data you are working with.

Chapter 15 (15 points):

- Fit a logistic regression that makes sense in the applied context. Plot the data and fitted regression line. Interpret the coefficients and their uncertainties. Simulate replicated data from the fitted model, graph them, and compare and compare to your graph of data.

Chapter 16 (15 points):

- Design a new study, including decisions about measurements and sample size. Use existing data and knowledge to come up with reasonable assumptions about effect size and variation. Then simulate fake data from this design, analyze the fake data, and discuss the results.

Chapter 17 (20 points):

- Estimate a treatment effect with interactions and then poststratify, explaining where you got your poststratification numbers to estimate the average treatment effect for the population. Give an estimate and a standard error.

Chapter 18 (10 points):

- Frame a question of interest in terms of the effect of a binary treatment. For this example, explain what are the outcome variable y, the treatment variable z, the pre-treatment variables x, and the potential outcomes y0 and y1. Be precise.

Chapter 19 (10 points):

- Estimate a causal effect by linear or logistic regression of an outcome on a binary treatment variable, some pre-treatment predictors, and at least one treatment interaction. Interpret the coefficient estimates and standard errors from your fitted model. What are the assumptions required for you to interpret the coefficients on the treatment indicator as causal effects? Are these assumptions reasonable here?

Chapter 20 (20 points):

- Set up a causal inference problem with your data where you have an outcome variable, a binary treatment indicator, and some pre-treatment predictors, where there is noticeable lack of overlap, comparing treatment and control groups. Assess balance and overlap with graphs. Perform matching to get a subset of data with adequate overlap, and then fit a regression model on the subset. Discuss your results and their interpretation. In particular, how is your interpretation affected by the fact that you've only analyzed this subset?

Chapter 21 (20 points, either-or):

- Clearly define a causal effect of interest and perform an instrumental variables analysis using your data using the two regressions. Interpret each regression result along with the instrumental variables estimate. Compare to the result of a direct regression. In the context of this example, discuss the assumptions under which the direct and instrumental-variables regressions give good estimates of the causal effect.

- Perform a regression discontinuity analysis using your data. Discuss the choices involved in setting up this regression. Include other pre-treatment predictors in your model, not just the assignment variable. Interpret your result and discuss the assumptions required for it to represent a good causal inference.