DATA 23700 Winter 2025

Exercise 8: Uncertainty visualization and statistical modeling

Due March 7, 2025

In this exercise, students will use uncertainty visualizations to explore and model a given dataset. Students will find a trivariate relationship (i.e., involving two predictor variables and one outcome variable) in the dataset through visual inspection, and they will then use a Bayesian regression model to describe that trivariate relationship, including whether there seems to be an interaction between the two predictors. The purpose of this exercise is to put some of what we've been discussing this week about uncertainty visualization into practice.

Students should submit their work on Gradescope as a knitted RMarkdown notebook (a .Rmd file compiled to a PDF). To knit a .Rmd file into a PDF, select the "Knit" dropdown menu in RStudio, and select the "Knit to PDF" option.

Instructions

Students should first **download this dataset on medical costs** from Kaggle. The data contain records about what people pay for healthcare and factors that might help to predict that.

Students should next **visually explore the dataset** looking for at least two other variables that seem predictive of medical costs. Using uncertainty visualizations is key here, and student should examine the joint distribution of variables in a variety of ways.

Students should **use a Bayesian regression model expansion workflow** to describe the relationship between two identified predictor variables and the outcome variable (i.e., charges). If students struggle to fit their intended model, they might find it helpful to build the model up in stages, adding one predictor or interaction at a time. The idea is that a sequence of models of increasing complexity can help you understand more complex models in terms of simpler models. We ask that you use brms for modeling since it does a good job of estimating uncertainty.

Diagnostic visualizations play a key role in statistical modeling. **For each model you fit, students should output a fit summary, trace plots, pairs plots, and a posterior predictive check**. See the demonstration from in class for examples.

Finally, students should **plot predictive and inferential uncertainty against predictors of interest.** This exercise requires students should prepare three of each of the following kinds of visualizations (i.e., one for each predictor in the model and one for their interaction):

- *Predictive uncertainty visualizations* (3 required) show how well a model's predictions match the empirical distribution of the data. These can help identify opportunities to improve the model.
- *Inferential uncertainty visualizations* (3 required) show the estimated relationship between predictors in the model and the outcome variable. These can help identify

what we can reliably conclude from a fitted model (e.g., smoking is associated with a "significant" increase in cancer).

We suggest using ggplot2 and ggdist for these visualizations. To extract samples from Bayesian regression models, we can use a few helpful tidybayes functions:

- `add_predicted_draws` for predictive uncertainty.
- `add_epred_draws` or `add_linpred_draws` for inferential uncertainty.

The difference between these functions is explained here. See the demonstration from in class for examples of their use.

Students should work in a literate programming style, interleaving blocks of code with expository text explaining their reasoning so that someone else could repruduce not only their work but their thought process. *Skipping on the text blocks will result in a score less than S.*