DATA 22700 Spring 2023

Exercise 8: Uncertainty visualization and statistical modeling

Due May 19, 2023

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 model expansion workflow to attempt to describe that 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 responses as an iPython notebook on Gradescope.

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 may find it helpful to examine the joint distribution of variables in a variety of ways, including but not limited to these examples on the altair gallery.

Last, students should **use a model expansion workflow** to try to describe the relationship between two identified predictor variables and the outcome variable (i.e., charges). A model expansion workflow entails adding one predictor or interaction at a time to subsequent models such that you are gradually building up an aspirational or intended model. This workflow helps you understand more complex models in terms of simpler models, which can help you course correct when your intended model doesn't fit the data. We suggest using PyMC for modeling since it does a good job of estimating uncertainty.

Diagnostic visualizations play a key role in statistical modeling. **For each model they fit, students should create at least one visualization of inferential uncertainty and one visualization of predictive uncertainty**. Inferential uncertainty concerns the estimated relationship between variables in the model. Predictive uncertainty visualizations show how well a model's predictions match the empirical distribution of the data. We suggest using either altair or arviz for these visualizations. Hint: Some helpful arviz methods are plot_trace, pair_plot, plot_lm, and plot_ppc.

Update: Since we did not have time to work on this in class, students should simply try to do what is described above. Feel free to reference the lecture notes from May 11, which are on the course website. If you run into difficulties, no worries. Just turn in whatever you managed to do in the time you could afford to spend on the exercise. Anything you turn in will receive a score of S.