Homework 3

Problem 1- Insurance Claims! (IGLM 9.2)

The attached `insurance.csv` dataset contains information on the numbers of automobile insurance policies (`n`) and the number of claims (`y`), tabulated by the car's insurance category (`car`), the age category of the policy-holder (`age`), and the region where the policy-holder lives (`dist`, equal to 1 if the policy-holder lives in a major city, and 0 elsewhere). This data is derived from another dataset in Aitkin et. al. (1989).

a. Use *Poisson regression* to fit a model relating the number of claims `y`, against all tabulating variables (`car`, `age`, and `dist`), as well as all interaction terms between the categorical variables. Be aware that by default `R` will read the category labels as numbers! Use eg. `as.factor` to convert them to categorical.

b. Based on the modeling in (a) above, Aitkin et. al. determined that all the interaction terms were insignificant, and that both `car` and `age` could be treated as continuous variables rather than categories. Fit this new model, and compare to the model in (a) above. What conclusions do you reach?

Problem 2- Rats!

The attached `rats.csv` contains birthweights for 30 baby rats (in grams), measured every week for 5 weeks. For this problem we'll denote by Y_{jk} the weight of the j-th rat at age x_{jk} (in days) where $j = 1, ..., 30$ and $k = 1, .., 5$

a. Conduct an exploratory analysis of the data. Provide plots of rat-specific growth trajectories (piecewise-linear connections between observations for each rat, in a single figure). If you're using R this will probably be easiest in `ggplot2`, while if you're using Python you may want to checkout `plotnine` or `altair`.

b. Fit the linear model $E[Y_{jk}] = \alpha + \beta x_{jk}$, assuming the variables Y_{jk} are all independent (ie. ignoring the fact that measurements corresponding to the same rat will have some correlation). Provide a summary of this model, check all diagnostics, and discuss the adequacy of the linear regression assumptions.

c. Fit the linear model $E[Y_{jk}] = \alpha_j + \beta_j x_{jk}$ (note that here the intercept and slope vary across individuals j). Provide a summary of this model, check all diagnostics, and discuss the adequacy of the linear regression assumptions.

d. Fit the hierarchical model $E[Y_{jk}] = \alpha_j + \beta_j x_{jk}$ where $\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$ and $\beta_j \sim N(\mu_\beta, \sigma_\beta^2)$. Provide a summary table.

e. For at least two rats, plot the observed the data, as well as the predictions obtained from the hierarchical model in (d) and those from the two linear models fit in (b) and (c). Using these plots discuss which model you would prefer the most.

Problem 3- Students!

In this problem we will work through some fundamental concepts in Bayesian analysis. This problem largely follows Richard McElreath's "Statistical Rethinking", section 4.7.

a. A class of n students is measured for height (in cm) each year for k years. Write down both a likelihood and a prior for a Bayesian simple linear regression of student height (as dependent variable) against

year of measurement (as independent variable). Give a short interpretation of all components of the model (including likelihood and prior), and justify your choice of prior.

b. Now, pretend you are told (before collecting any data) that each student eats a diet plentiful in spinach, and thus it is guaranteed that each will get taller every year. Does this information change any of your priors, and if so how?

c. In addition to the information in (b) above (and before collecting any data) you learn that at outset of the study the average height of students is 120cm. Does this change any of your priors, and if so how?

d. In addition to (b) and (c) above (and still before collecting any data) you learn that the variance of heights among students in the same year cannot be more than 64cm. Does this information change any of your priors, and if so how?

e. Finally, using the priors you selected at the end of (d) above, describe the prior distribution of the students' heights at the end of the first year of the study. With the aid of a software tool such as R or Python (or a physical tool such as a coin, a bag of dice, or a FERMIAC), draw 100 samples from this prior distribution and plot them in a histogram. Does this prior distribution accurately reflect your beliefs about the students' heights at the end of the first year? Does it accurately reflect what is physiologically possible? If not, what would you change in your priors to resolve the discrepancy?