**Project Details**

1. The dataset “healthcare.csv” contains simulated data on annual health care   
   utilization in dollars “Y” and annual household income “income”.   
   healthcaredata <- read.csv(…)
2. Fit a GLM that predicts the health utilization from log(income); use a log link and   
   select what seems to you to be the best variance function (and give a reason for your   
   choice). Try the Poisson, Gamma, and Inverse Gaussian variance assumptions, and don’t   
   forget to consider over or under dispersion. Make appropriate residual plots. Test if coef for log(income) is equivalent to .5
3. Compute leverages and Cook’s distances using model from a. Choose reasonable thresholds for the leverages and for Cook’s distance. Use these thresholds to find data points of interest. Plot the points (log(income), Y) with leverages greater than your chosen threshold. What do you find?
4. The dataset “phweight.csv” is a subset of data from a public health study on children from a country in South Asia and has 7 variables: id (child identifier), sex (1=M, 2=F), wt (child’s weight in kg), mage (mother’s age in years), lit (indicator of mother’s literacy 0=no 1=yes), died (number of children the mother has had that died), and age (of the child, in months). Data was collected at multiple points through the child’s lifetime and each child is identified by their id value.

Study1007data <- read.csv(……)

1. We need to figure out if there is a link between weight gain as the child ages and the sex of the child. Create a plot that we can use to informally determine whether this relationship exists. Explain your conclusion.
2. The primary goal in this study is to determine what factors influence the weight of children in this country. Fit a fixed effects model with weight as the response and age, sex, mother’s age, literacy, and other child deaths as predictors. Which terms are significant in this model?
3. Also, fit a mixed effects model with weight as the response. Include an interaction term between age and sex, and main effects for *mage, lit, and died*. Use a random intercept term for the child.
4. What is the predicted difference in child weight between a 15- and 25-year-old mother?
5. What is the expected difference in weights for identical twins (i.e., same fixed effects) according to the model? You may need to use a function/simulation/for-loop
6. For this data, does it make more sense to use a fixed effects model (part b) or a mixed effects model (part c)? Why?
7. Make the following diagnostic plots and interpret:
8. Residuals vs. Fitted plot
9. ii. QQ plot of the residuals
10. iii. QQ plot of the random effects
11. Fit a model with age and mother’s age as the only fixed effects and compare it to the previous model. Is the smaller model acceptable to use instead?