HW #1
BIOSTAT 705 Spring 2024
Due on Feb 1st

1. Given the Motor Trend Car Road Tests (mtcars) dataset (`mtcars.csv`):

   Let

$$X_i = \begin{cases} 1 & \text{if wt} > \text{median} \\ 0 & \text{if wt} \leq \text{median} \end{cases}$$

   wt = weight (1000 lbs)

   where X is called an *indicator* or *dummy* variable
   In R type help(mtcars) will provide more details on the dataset and variables name.

   a) Fit a simple linear regression, using $\text{mpg}_i = \beta_0 + \beta_1 X_i + \epsilon_i$; how you interpret $\hat{\beta}_1$ in the context of the probelm?

   b) Write the above model in a matrix form (ie, $Y = X\beta + \epsilon$). Identify the design matrix X.

   c) What is the estimated mean square errors (MSE)?

   d) Re-do part a using $\text{mpg}_i = \beta_0 + \beta_1 wt_i + \epsilon_i$. How this results differ from part a? Which model would you recommend? why?

   e) Suppose we fitted a model in part a by adding *hp* (horsepower) and inetraction, ie $\text{mpg}_i = \beta_0 + \beta_1 X_i + \beta_2 hp_i + \beta_3 X_i * hp_i + \epsilon_i$. How would you interpret term? is this model additive?

   f) Similarly, we added *hp* and interaction to the model in part e, ie $\text{mpg}_i = \beta_0 + \beta_1 wt_i + \beta_2 hp_i + \beta_3 wt_i * hp_i + \epsilon_i$. How should the interaction term be interpreted in this model? Can we interpret the main effects? why or why not?

2. Stigler, History of Statistics pg.285 gives Galton's famous data on heights of sons (Y in inches) and average parents height (X in inches) scaled to represent a male height (essentially sons' heights versus fathers' heights). Data are given in `parents_offsprings.csv`. Consider a statistical model for these data, randomly sampled from some population of interest. In particular, choose a model which accounts for the apparent linear dependence of the mean height of sons on midparent height X.

   a) Construct a scatterplot of these data.

   b) What is β_1? Is this statistic or parameter?

   c) What is the estimated slope ($\hat{\beta}_1$)? Is this statistic or parameter?

   d) How much does $\hat{\beta}_1$ vary about β_1 from sample to sample? (Provide an estimate of the standard error, as well as an expression indicating how it was computed)

   e) What is a region of plausible values for β_1 suggested by the data?

   f) What is the line that best fits these data, using criterion that smallest sum of squared residuals is "best?"

   g) How much of the observed variation is the heights of sons (the y-axis) is explained by this "best" line?

   h) What is the estimated average height of sons whose midparent height is $x = 68$?

   i) Is this the true average height in the whole population of sons whose midparent height is $x = 68$?

   j) Under the model, what is the true average height of sons with midparent height is $x = 68$?

   k) What is the estimated standard deviation among the population of sons whose midparent height is $x = 68$? Would you call this standard deviation a "standard error"?

   l) What is the estimated standard deviation among the population of sons whose midparent height is $x = 72$? Bigger, smaller, or the same that for $x = 68$? Is your answer obviously supported or refuted by inspection of the scatterplot?

   m) What is the estimated standard error of the estimated average for sons with midparent height $x = 68, \hat{\mu}(68) = \hat{\beta}_0 + 68\hat{\beta}_1$? Provide an expression for this standard error.

   n) Is the estimated standard error of $\hat{\mu}(72)$ bigger, smaller, or the same as that for $\hat{\mu}(68)$?

   o) Is the observed linear association between son's height and midparent height strong? Report a test statistic.

   p) What quantity can you use to describe or characterize the linear association between son's height and midparent height in the whole population? Is this a parameter or a statistic?

3. Consider data on corn yield Y (bushels/acre) and rainfall X (inches/yr) in six Midwestern states recorded from 1890 to 1927. Data are given in `corn_yield_and_rainfall.csv`.

   a) Construct a scatterplot of corn yield and rainfall measurements.

   b) How can we describe the association between yield and rainfall? Does it appear *linear*?

   c) How can we measure the strength of linear association?

   d) To what degree the variability in yield described or explained by its association with rainfall? In what units this variability measured?

   e) How can we use this association to estimate average yield, given a certain level of rainfall, say 14 inches/yr?

   f) How can we use this association to predict future yield, if we have an idea about what the rainfall will be? use $x_0 = 14$ (inches/yr).

   g) What is the difference between e and f?

   h) What proportion of variance in yield is explained by the linear regression model?.

4. An investigator wants to examine the relationship between body temperature y and heart rate x. Further, he would like to use heart rate to predict the body temperature. (Data BodyTemperature.csv).

   a) Construct a simple linear regression model for body temperature using heart rate as the predictor.

   b) Interpret the estimated slope ($\hat{\beta}_1$) and examine its statistical significance.

   c) Construct the 95% confidence interval for the population slope β_1. How would you interpret this CI and does it agree with your conclusion in part (b)?

   d) Calculate adjusted R^2, what is tell us about the contribution of heart rate in the model?

   e) If someones heart rate is 75, what would be your estimate of this persons body temperature?

   f) Suppose the investigator believe that adding sex to the model in part (a) will improve model prediction, is he correct? how would you quantify the contribution of sex to the model?

   g) Suppose theinvestigator added an interaction between heart rate and sex to the model in part (f), how would you Interpret this interaction term?