Homework 3: Due Friday September 24 at 3:00

Searching for solutions on the web or elsewhere is forbidden.

1. Under Files/Data on Canvas you will find the dataset `auto-mpg.cvs`. Download it and load it into R:

df = read.csv("auto-mpg.cvs",head=TRUE)
names(df)
str(df)

Information on the data can be found at:

`https://archive.ics.uci.edu/ml/datasets/auto+mpg`

The response variable of interest is `mpg` (fuel consumption), measured in miles per gallon.

(a) Regress mpg on weight. (For now, ignore the rest of the variables.) Plot the residuals versus weight and versus the fitted values. What conclusions do you draw from these plots?

(b) Apply several transformations on weight including: square root, log and reciprocal. Refit the linear regression model for each transformation and produce new residual plots. What does the plot suggest about the linearity assumption now? What else does it suggest? What assumptions are violated?

(c) Do a qqplot of the residuals. Do they seem Normal?

(d) Use whatever transformation you think worked best. Get a 90 percent confidence interval for the slope. Make sure to use the sandwich standard error.

2. We will use the auto-mpg.csv data again. We will use all the variables except car.name. You should get rid of that variable:

```
df = read.csv("auto-mpg.cvs",head=TRUE)
df = df[,-9]
```

Using the command `str(df)` you can see that horsepower has been recorded as a factor for unknown reasons. Let's make it numeric.

```
df$horsepower = as.numeric(df$horsepower)
```

(a) Do a pairs plot. Comment on what you see in the plots.

(b) Now fit a regression model of mpg on all the other variables. Check the residuals, influence etc. Make any changes you think are reasonable (omitting observations, transformations etc). Then get the estimates and confidence intervals for the parameters. Briefly summarize what steps you did and why.

(c) Now do a regression of `mpg` on `horsepower`. Plot the data and add the regression line to the plot. Do you see anything in the plot that suggests that this is not a good model? Find a way to fit a model for these variables that avoids this problem.

3. Get the file gpa.txt from Canvas:

```
df = read.table("gpa.txt")
```

You might want to give the variables names. For example:

```
names(df) = c("GPA","ACT")
```

(a) Estimate the regression function using a polynomial. You will need to decide what order of polynomial to use. Give the estimates and confidence intervals for the parameters.

(b) Check the fit and summarize what you find. Adjust the model if you feel it is needed.

(c) Give a 60% prediction interval for the GPA of students with the mean ACT score.

(4) In this question you will examine the effects of measurement error, meaning that the X_i's are not measured precisely. Suppose we have data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X_i \in \mathbb{R}$ and $Y_i \in \mathbb{R}$. Suppose that

$$Y_i = \beta X_i + \epsilon_i$$

where $\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2)$. Note that this a regression through the origin (no intercept). The least squares estimator is

$$\widehat{\beta} = \frac{\sum_i Y_i X_i}{\sum_i X_i^2}.$$

Now suppose that we don't observe X_i; instead we observe

$$U_i = X_i + \delta_i$$

where $\delta_1, \ldots, \delta_n \sim N(0, \tau^2)$. The δ_i's are independent of the X_i's. We will estimate β by replacing X_i with U_i in the formula for $\widehat{\beta}$. In other words

$$\widehat{\beta} = \frac{\sum_i Y_i U_i}{\sum_i U_i^2}.$$

Find the mean of $\widehat{\beta}$ (conditional on $X_1, \ldots, X_n$). You may use the following fact: if $A_1, \ldots, A_n$ and $B_1, \ldots, B_n$ are random variables then

$$\mathbb{E}\left(\frac{\sum_i A_i}{\sum_i B_i}\right) \approx \frac{\mathbb{E}[\sum_i A_i]}{\mathbb{E}[\sum_i B_i]}.$$

How does the measurement error affect $\widehat{\beta}$? Does this bias disappear as n gets larger? What happens as $\tau \to \infty$?