36-617: Applied Linear Models
Fall 2021
HW03 – Due Mon Sept 20, 11:59pm

- Please turn the homework in online in our course webspace at canvas.cmu.edu.
  - There is a link to Gradescope in the description of this assignment on Canvas.
  - You should submit a single pdf to Gradescope. If you need help with this, please see https://www.cmu.edu/teaching/gradescope/index.html. Also, allow yourself some extra time to create the pdf & upload it in Gradescope.
  - Gradescope allows the TA to grade all the problem 1's together, then all the problem 2's, and so forth. This leads to more consistent grading and better comments for you.
- Reading:
  - You should have read Chapter 3 in Sheather for this week. We are skipping Sheather Ch 4 for now, and proceeding to Ch 5, for next week.
- There are five exercises; the first four are all related.

Exercises

1. [Based on Sheather, Ch 3, # 5, pp. 109ff.] An analyst for the auto industry has asked for your help in modeling data on the prices of new cars. Interest centers on modeling suggested retail price as a function of the cost to the dealer for 234 new cars. The data set, which is available in the file `cars04.dat` in the same folder as this hw, is a subset of the data described at `http://www.amstat.org/publications/jse/datasets/04cars.txt`. You can read the data into R with a command like `data <- read.table(cars04.dat,header=T)`.

   There are many interesting variables in this data set, but we are only interested in two, for now: `SuggestedRetailPrice`, and `DealerCost`.

   (a) Provide a good initial description of the data *(This would be, essentially, the* **Data** *section of an IDMRAD paper, minus a description of where the data came from)*. This should include
   - A table with the names and definitions of all variables in the data set.
   - The number of observations (rows) and variables (columns).
   - A brief discussion of whether there is any missing data, and if so, how much and where.
   - An appropriate summary of each individual variable. This might consist of means and standard deviations, and/or boxplots, and/or histograms, for example.
   - Scatterplots (perhaps in a scatterplot matrix) of all the pairs of variables.

   Since our only interest for now is in `SuggestedRetailPrice` and `DealerCost`, it would be useful to include an additional, separate description of those two variables.

   (b) Fit the model
   $$\text{SuggestedRetailPrice} = \beta_0 + \beta_1 \text{DealerCost} + \varepsilon$$
   Provide a `summary()` of the model, and also the four casewise residual diagnostic plots. Write five sentences:
   - One sentence for each of the four diagnostic plots, indicating the main feature of each, and whether this feature indicates good or bad fit of the model to the data; and

- One more sentence comparing your conclusions here with the description of the two variables `SuggestedRetailPrice` and `DealerCost` from part (a).

2. [Continuing problem #1: Box-Cox transformations]

   (a) Now explore power transformations of our two variables using `boxCox()` and `powerTransform()` from `library(car)`.

      i. Find the Box-Cox transformation for `DealerCost`.

      ii. Construct a new variable `TransCost` that is the transformation of `DealerCost` you found. Fit the regression `SuggestedRetailPrice` $= \beta_0 + \beta_1$`TransCost` $+ \varepsilon$, and use this to find the Box-Cox transformation for `SuggestedRetailPrice`.

   (b) Fit the model

$$\text{TransPrice} = \beta_0 + \beta_1 \text{TransCost} + \varepsilon$$

   Provide a `summary()` of the model, and also the four casewise residual diagnostic plots. Write five sentences:

   - One sentence for each of the four diagnostic plots, indicating the main feature of each, and whether this feature indicates good or bad fit of the model to the data.
   - One more sentence indicating whether you are happy with the fit of the model to the power-transformed data.

3. [Continuing problem #1: log transformations]

   Now fit the model suggested by Sheather,

$$\log(\text{SuggestedRetailPrice}) = \beta_0 + \beta_1 \log(\text{DealerCost}) + \varepsilon$$

   rovide a `summary()` of the model, and also the four casewise residual diagnostic plots. Write five sentences:

   - One sentence for each of the four diagnostic plots, indicating the main feature of each, and whether this feature indicates good or bad fit of the model to the data.
   - One more sentence indicating whether you are happy with the fit of the model to the log-transformed data.

4. [Continuing problem #1: comparisons and conclusions]

   (a) All of these models—(1b), (2b), and (3)—have incredibly high R^2. For which model are the assumptions of linear regression best satisfied, based on your analyses above?

   (b) Complete each sentence below, or explain why it is difficult to complete:

   - For the model in (1b), a _____ increase in `DealerCost` results in a _____ increase in the expected value of `SuggestedRetailPrice`.
   - For the model in (2b), a _____ increase in `DealerCost` results in a _____ increase in the expected value of `SuggestedRetailPrice`.
   - For the model in (3), a _____ increase in `DealerCost` results in a _____ increase in the expected value of `SuggestedRetailPrice`.

   (You may want to consult the file "`log xform and percent interpretation.pdf`" in the week03 folder in our files area on canvas.)

   (c) Which model—(1b), (2b), or (3)—would you use if you were the statistical consultant on this project? Why?

5. [Based on Gelman & Hill. Ch 3, #1, p. 49] The file `pyth.dat`, in the same folder as this hw, contains outcome `y` and inputs `x1`, `x2` for 40 data points, with a further 20 points with the inputs but no observed outcome (for this problem we will ignore these last 20 points). Save the file to your working directory and read it into R using the `read.table()` function.

   (a) Fit the two models

   $$\begin{aligned} \mathbf{M1}: \texttt{y} &= \beta_0 + \beta_1 \texttt{x1} + \varepsilon \\ \mathbf{M2}: \texttt{y} &= \beta_0 + \beta_1 \texttt{x2} + \varepsilon \end{aligned}$$

   Which model provides a better fit for `y`? Why?

   (b) Construct new variables `y2` $=$ `y`^{2}, `x12` $=$ `x1`^{2}, and `x22` $=$ `x2`^{2} and fit the models

   $$\begin{aligned} \mathbf{M3}: \texttt{y2} &= \beta_0 + \beta_1 \texttt{x12} + \varepsilon \\ \mathbf{M4}: \texttt{y2} &= \beta_0 + \beta_1 \texttt{x22} + \varepsilon \end{aligned}$$

   Compare the fits of these two models to the models in part (a). Which fits best? Why?

   (c) To fit the model

   $$\texttt{y2} = \beta_0 + \beta_1 \texttt{x1} + \beta_2 \texttt{x2} + \varepsilon\,,$$

   we just expand the R modeling language a little bit: `y ~ x1 + x2`. Fit both of the models

   $$\begin{aligned} \mathbf{M5}: \texttt{y} &= \beta_0 + \beta_1 \texttt{x1} + \beta_2 \texttt{x2} + \varepsilon \\ \mathbf{M6}: \texttt{y2} &= \beta_0 + \beta_1 \texttt{x12} + \beta_2 \texttt{x22} + \varepsilon \end{aligned}$$

   Compare these to the earlier models. Which fits best? Why?

   (d) Can you find a simple, recognizable function `x3 = (something involving both x1 and x2)`, so that

   $$\mathbf{M7}: \texttt{y} = \beta_0 + \beta_1 \texttt{x3} + \varepsilon$$

   provides a fit comparable to the best fitting models above? What is going on?