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 to Sheather Ch 3, p. 105, #3: Do the first three exercises below as "just hw problems" and then use the results to write a brief report as outlined in problem 4.

Exercises

1. Please do Sheather, Ch 3, p. 105, #3, **Part A**. Remember that the data is in the "0-textbook" folder in the files area on Canvas for this class.

   Feel free to explore common log or power transformations, Box-Cox, etc. Remember to "round" any transformations in your final model to something that will be meaningful to the publisher.

   For part (c), *at least* provide a `summary()` of your final model, and also the four casewise residual diagnostic plots, and write five sentences:

   - One sentence for each of the four diagnostic plots, indivating 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 your final model.

   If there are any interesting data points (e.g., outliers, high leverage points, or anything else that seems interesting), please use the `Magazine` and/or `PARENT.COMPANY` variables to identify them.

2. Please do Sheather, Ch 3, p. 105, #3, **Part B**.

   The language in part (a) is a little confusing. You are allowed to fit a multiple regression model of the form $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \varepsilon$, for as many positive integer powers of x as you think are needed. The hint suggests you may be able to stop at x^3, but if including more powers helps, then go ahead. (Note: y = `AdRevenue`, and x = `Circulation`.)

   For part (c), provide an answer similar to part (c) in problem 1.

   Because of the semantics of the modeling language in R, the command `lm(y ~ x + x^2)` produces the same model as `lm(y ~ x)` (try it!). To get powers, it is safest to construct new variables, e.g.

```
x2 <- x^2
lm(y ~ x + x2)
```

You can accomplish the same thing with the `I()` function: `lm(y ~ x + I(x^2))`, though constructing a separate variable and putting it in the data frame with the other variable(s) is usually better for using the `predict()` function, etc.

3. Please do Sheather, Ch 3, p. 105, #3, **Part C**.

   Note that for part (a) we are not doing any formal comparisons using likelihood ratio tests, AIC, BIC or anything like that yet. Your comparison should be based on looking at the data, the plausibility and explainability of the model, the casewise residual diagnostic plots, etc.

4. Write a brief IDMRAD paper based on your answers to problems 1–3. Remember to label the **Introduction**, **Data**, **Methods**, **Results** and **Discussion** and **Technical Appendix** sections. Here are some additional instructions:

   Title Choose a brief title that gives the reader some idea of what the data are and what the result is.

   Author Your name, your institution (Department of Statistics and Data Science, Carnegie Mellon University), and your email address.

   Abstract *We will skip the abstract for this exercise.*

   Introduction This can largely parrot information in the problem statement before **Part A** on p. 105 of Sheather. The main research problem is to develop a good regresson model to predict gross advertising revenue per advertising page in 2006 (in thousands of dollars) from circulation (in millions).

   Data First, say where the data came from (again, you can parrot part of the problem statement before **Part A**). Then, provide (a) a table with the variable names and their definitions (including units of measurement), (b) a set of summary statistics for the quantitative variables (e.g. mean, SD, etc.), and (c) any exploratory plots that seem useful (histogram, boxplots, scatter plot, etc.), but don't be repetitive (for example, for some distributions, histograms and boxplots tell the same story, so you wouldn't want to include both). *Include at least one sentence about each of these three elements in the Data section.*

   Methods[1] The methods section will read something like this (not much more than a paragraph or two, since this is a simple analysis): "To build a good regression model for predicting gross advertising revenue per advertising page from circulation, we considered two approaches. First we considered simple linear regression, with various transformations for the circulation variable (see Appendix, Part A). The methods we used to find suitable transformations were *blah-blah-blah*. The final transformation was adjusted to improve interpretation and clarity of communication. Then we considered polynomial regression models (see Appendix, Part B), i.e., regressing gross advertising revenue on circulation, circulation^2, and so forth, up to the power *blah-blah*. We selected the best model from both the transformed simple linear regression and the polynomial regression models by considering casewise residual diagnostic plots, *and blah-blah-blah* (see Appendix, Part C)."

   Results[2] This will be three short sections, perhaps just three paragraphs. First, remind the reader what methods you used in Part A, give the best model from Part A, and provide a table of coefficient estimates and standard errors. Second, same thing for Part B. Finally, summarize your findings for Part C. Normally I would *not* include regression diagnostic plots without a very good reason, but in this problem if there are a couple of plots that really explain why you chose the model you did, you can include it/them here.

[1] Note that I am citing parts of the Appendix for any reader who wants to check the technical details.

[2] Again, it would be appropriate to cite parts of the Appendix for any reader who wishes to examine the technical details.

Discussion Re-state the research question, and state the main result. Interpret the result: what does it say about the relationship between gross advertising revenue per advertising page, and circulation? Is there anything else the reader should know, or keep in mind when considering or using this result? Finally, discuss any weakneses of your analyses, and suggest possible future work (could be work that fixes the weakness, or work that extends these results to a larger data set or a different setting, etc.)

References *We will skip the references for this exercise.*

Technical Appendix For the technical appendix, include the questions and the answers to problems 1, 2 and 3 above.

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

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

   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

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

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

   (c) To fit the model

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

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

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

   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

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

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