36-617: Applied Linear Models
Fall 2020
HW04 – Due Mon Sept 28, 11:59pm Pgh Time

- Please turn the homework in, as a single pdf, online in GradeScope using the link provided on the HW04 assignment page on canvas.cmu.edu, under Assignments. Upload *one* file per person.
- This week we are discussing Ch 5 of Sheather. Next week we will move on to Ch 6.
- There are three major exercises below; each one has "parts".

Exercises

1. Let $y = X\beta + \epsilon$, where $y = (y_1, \ldots, y_n)^T$ is an $n \times 1$ column vector, X is an $n \times (p+1)$ matrix whose first column is all 1's, $\beta = (\beta_0, \ldots, \beta_p)^T$ is a $(p+1) \times 1$ column vector, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \sim N(0, \sigma^2 I)$ is an $n \times 1$ random column vector, following a multivariate Normal distribution with mean vector 0 and variance-covariance matrix $\sigma^2 I$, where I is the $n \times n$ identity matrix.

   (a) Use properties of the hat matrix $H = X(X^T X)^{-1} X^T$ and the multivariate Normal distribution as discussed in class, to show
   $$\hat{e} \sim N(0, (I - H)\sigma^2)$$

   (b) Let H be the hat matrix for the multivariate regression model $y = X\beta + \epsilon$ as in part (a), and let H_1 be the hat matrix for the intercept-only model $y = \beta_0 + \epsilon$.

      i. Show that the fitted values $\hat{y}$ for the intercept-only model is an $n \times 1$ column vector, all of whose entries are $\overline{y}$, that is,
      $$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \overline{y} \\ \vdots \\ \overline{y} \end{bmatrix} \qquad (*)$$
      (where the first "=" is the definition of $\hat{y}$ and the second "=" is what I want you to show).

      ii. Find a simple expression, in terms of (some or all of) y, I, H and H_1, for the sample covariance
      $$\text{Cov}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})(\hat{y}_i - \overline{y}).$$
      *(***Hint:** *We can rewrite* $Cov(y, \hat{y}) = \frac{1}{n}(y - \overline{y})^T(\hat{y} - \overline{y})$*, where* $\hat{y}$ *is the column vector of fitted values from* $y = X\beta + \epsilon$ *and, abusing notation slightly,* $\overline{y}$ *is the column vector in* $(*)$ *above, i.e. the fitted values from the intercept-only model* $y = \beta_0 + \epsilon$*.)*

      iii. Continue along the lines of the calculations in part (ii) to show that the sample correlation between y and $\hat{y}$ can be written as
      $$\text{Corr}(y, \hat{y}) = \sqrt{\frac{SS_{reg}}{SST}}$$
      and hence R^2 for the regression model $y = X\beta + \epsilon$ really is the squared correlation between y and $\hat{y}$:
      $$R^2 = \text{Corr}(y, \hat{y})^2 \, .$$

   (c) Show that $\hat{e}$ and $\hat{y}$ have sample correlation 0, and hence a scatter plot of $\hat{e}$ vs $\hat{y}$ should show no increasing or decreasing trend, when the model $y = X\beta + \epsilon$ is true.

2. [Based on Gelman & Hill (2009), p. 51, #5] The subfolder `beauty` in the hw04 folder in the "Files" area for our course on canvas contains data from Hamermesh and Parker (2005) on student evaluations of instructors beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students who had not attended the classes and were not aware of the course evaluations. Various documents in the folder give background and some variable definitions (some variables are defined in the ".log" file there, others' definitions you will have to deduce from pdf's in the subfolder).

   (a) Fit a regression model predicting `courseevaluation` (average student evaluations) from `btystdave` (the average of 6 standardized beauty ratings for each instructor) and `female`. Then fit the same model with the interaction between `btystdave` and `female` added in.

      i. Graph each fitted model on a scatter plot of `courseevaluation` vs `btystdave`. Indicate clearly in the graph what the various parameters in the model represent geometrically.

      ii. Display the four standard diagnostic plots in R and comment on their features, for each model. *N.b. the interpretation of these plots is exactly the same as it was for simple regression.* Comment on whether the fit seems adequate from the evidence in these plots, for either model. In case there are problems with the fit, indicate what they are and how you might improve things.

      iii. Produce summaries of the two fitted models; comment on the coefficient estimates and their standard errors, and on R^2, for each model Use a partial F test to determine whether the interaction should be kept. *Your comments should include not only technical points ("B" in the "ABA^{-1}" metaphor for applied statistics from the course syllabus), but also what it means for understanding how factors may influence course evaluations ("A^{-1}").*

   (b) Now what happens when you try to control for other variables by adding them to the better of these two models (no more interactions, just use additional main effects, for now)? Find the best such model, and comment on its fit, and the interpretation of the estimated coefficients, using the same tools you used in part (a). *Don't forget A^{-1} from the "ABA^{-1}" metaphor.*

3. An IDMRAD paper has the following sections:

   Abstract Summarize I, D, M R and D sections of the paper (typically one sentence each).

   Introduction Why would anyone want to read this paper? What questions will be addressed?

   Data What data set was used in this study? Typically, include variable definitions, sample size, quick numerical summaries of the variables and initial EDA, but *no model fitting or analysis*.

   Methods List the methods and/or analyses that will be used to answer the questions stated in the **Introduction**. *No data analysis, graphing, model fitting, etc. appears here*; you just say what methods and analyses you will use with which variables, to answer each question.

   Results Here you *finally* get to show the data analyses (model fitting, graphics, etc.) that you did, and what the results were. Don't overload the reader: put the highlights here so the reader understands what you did and why, and refer the reader to specific pages or sections of the technical appendix for more details. It should be clear which data analyses and results go with which question from the **Introduction**. *Every analysis that is presented here should have been mentioned in the* **Methods** *section.*

   Discussion What does it all mean? Typically you will say, for each question from the **Introduction**, how the analyses that you did the **Results** section answers that question. You might also mention EDA and so forth from the **Data** section if that makes clearer to the reader what answers you found for one (or more) of the questions. Then you will talk about the big picture, what future work or generalizations of your work might look like, and any limitations of your study. But *there should be no additional analyses or results in this section; just use the analyses you did for the* **Results** *section (and possibly the* **Data** *section).*

References All the works you relied on to write your paper, in ASA citation format (see the section entitled "The Reference List" in https://amstat.tfjournals.com/asa-style-guide/).

Technical Appendix This contains complete versions of the analyses listed in the **Methods** section and presented in the **Results** section. There may be additional analyses here (e.g. to support the **Data** section of the paper, or to show why the methods and analyses that you chose for the paper were the right ones).

Review the paper "menu pricing IDMRAD - version 2 with appx.pdf" in the week01 folder in the files area for our class on Canvas, to see an example of what goes in each section, and how the technical appendix is used to contain details that would be "too much" in the main sections the paper.

For this exercise, please write

(a) The **Data** section

(b) The **Methods** section

(c) The **Results** section

(d) The **Technical Appendix**

of an IDMRAD paper, based on your work with the `cars04.dat` data from HW03. For the **Technical Appendix**, use what you turned in for problems 1–4 on HW03. To get you started on the other parts, here is an **Introduction** for the paper. Since I cited some sources in the introduction, I also include a **Reference** section.

Introduction

The automobile industry is a major segment of the US economy. A part of understanding this segment of the economy is understanding the factors that influence the suggested retail price of autos in the US. I have been asked to examine the variables in an extract of automobile data from The Kiplinger (2003); the particular extract was provided by Sheather (2009). In particular the questions I will answer in this report are (a) What can we learn from descriptive statistics and exploratory data analysis of all of the variables in the Kiplinger data extract? (b) What is the best/most interpretable model that relates Dealer Cost to Suggested Retail Price in the data, ingoring all other variables? (c) What does this model tell us (if anything) about the relationship between Dealer Cost and Suggested Retail Price?

References

Sheather, S.J. (2009), *A Modern Approach to Regression with R.* New York: Springer Science + Business Media LLC.

The Kiplinger (2003), *Kiplinger's Personal Finance, December 2003, vol. 57, no. 12,* pp. 104–123. New York: PARS International Corporation. Accessed at http://www.kiplinger.com.