36-617: Applied Linear Models
Fall 2022
HW02 – Due Mon Sept 12, 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.
  - Don't forget to identify which pages each (part of each) problem appears on in your solitions. 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.
  - Remember to list who you worked with, on this and every assignment.
- Reading:
  - For next week: Sheather Ch 6 (supplemental: ISLR 3.3.3; G&H Ch 4)
- There are three exercises below. No IMRAD this week; that will appear on HW03.

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) Let

   $$U = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

   be an $n \times 1$ column of all 1's.

   i. Using the definition of the hat matrix H, show $HX = X$

   ii. Using your result in (i), show $HU = U$

   (b) Use properties of the hat matrix $H = X(X^T X)^{-1} X^T$ to show that the column vector

   $$y - \hat{y} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}$$

   can be written as $y - \hat{y} = (I - H)y$.

(c) For any $n \times 1$ column vector

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

we know that $a^T U = \sum_{i=1}^n a_i$. Use this fact, together with your results in (a) and (b), to show that

$$\sum_{i=1}^n (y_i - \hat{y}_i) = 0.$$

Aside: *Hence y and $\hat{y}$ have the same sample mean $\overline{y}$, and the residuals vs fitted plot should always be "balanced", in some sense, around the horizontal line at $\hat{e} = 0$.*

2. 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[1]
   $$\hat{e} \sim N(0, (I - H)\sigma^2)$$
   where $\hat{e}$ is the column vector $\hat{e} = y - \hat{y}$.

   (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 $\hat{y}$:
      $$R^2 = \text{Corr}(y, \hat{y})^2 .$$

[1] Very similar to problem 1(b).

(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 overall trend.

Aside: *It is interesting to note that the results of problem #1 (c) and problem #2 (b) and (c) did not depend at all on the normality assumption, or even on whether the model fits the data well or not. In other words, even if the fit is terrible, it will still be true that the sum of $\hat{e}$ is zero, $R^2 = [Corr(y, \hat{y})]^2$, and the sample correlation between $\hat{e}$ and $\hat{y}$ is zero. The only thing that matters is that the model have an intercept, i.e. the X matrix should have a column of 1's.*

3. [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 $\mathtt{y2} = \mathtt{y}^2$, $\mathtt{x12} = \mathtt{x1}^2$, and $\mathtt{x22} = \mathtt{x2}^2$ 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) 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?