

HW1

STAT-340 Applied Regression Method

Yerim Oh

Details

Due Date

This assignment is due at 11:59 PM on the 27th of September.

Grading

The problems will be graded for correctness. An emphasis will be placed on full explanations of your thought process. You should write complete sentences and make sure the steps can be followed in the proofs.

Collaboration

You are allowed to work with others on this assignment, but you must complete and submit your own write up. You should not copy large blocks of code or written text from another student.

Submit in Gradescope

The problems in this assignment require writing mathematical formulas. It can be done directly in RMarkdown using LaTeX coding. For instance, the code:

```
\begin{flalign}
\frac{d}{dx} f(x) &= \frac{d}{dx} (10 x^2 + e^{3x} - \log(x)) \notag \\
&= (20 x + 3e^{3x} - 1/x). \notag
\end{flalign}
```

results in the following output:

$$\begin{aligned}\frac{d}{dx} f(x) &= \frac{d}{dx} (10x^2 + e^{3x} - \log(x)) \\ &= (20x + 3e^{3x} - 1/x).\end{aligned}$$

You may write your answers up using LaTeX. However, you may write your answers to all problems by hand:

- 1) on a tablet or
- 2) on a piece of paper, then scan your solutions and include the scan/picture in your Markdown document. The code to include a figure is as follows (from `\begin{figure}` to `\end{figure}`):

```
\begin{figure}[!ht]
\centering
\includegraphics[width=8cm]{NamePicture.png}
\end{figure}
```

I prefer you spend time understanding the problems rather than formatting them. Pick the easiest way for you, but make sure your answers are neatly presented and legible.

Problem 1: One-way ANOVA

Suppose you conduct an experiment where a total of n subjects are randomly assigned to one of two groups (control and treatment). For each subject i , you record a response variable y_i . The model may be written as follows:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where $x_i = 1$ if the subject was assigned to the treatment group, and $x_i = 0$, in the control group. Suppose that there are n_1 subjects in the control group and n_2 subjects in the treatment group, such that $n_1 + n_2 = n$.

(a) Write down the model in matrix form. As part of your answer, please define each matrix and vector carefully. For \mathbf{X} , you may assume the first n_1 subjects were assigned to the control group and the remaining n_2 subjects to the treatment group. Since you don't know the exact values for n_1 and n_2 , you'll likely have to use some dots \vdots in your specification of \mathbf{X} to indicate some rows that you haven't explicitly written down.

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

when $n = n_1 + n_2$

(b) By working through matrix calculations, demonstrate that $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)' = (\bar{y}_1, \bar{y}_2 - \bar{y}_1)'$, where $\bar{y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i$ and $\bar{y}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^n y_i$.

We know that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

$$\begin{aligned} \mathbf{X}'\mathbf{X} &= \begin{bmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 + \cdots + 1 + 1 + \cdots + 1 & 0 + \cdots + 0 + 1 + \cdots + 1 \\ 0 + \cdots + 0 + 1 + \cdots + 1 & 0 + \cdots + 0 + 1 + \cdots + 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 \cdot (n_1 + n_2) & 1 \cdot n_2 \\ 1 \cdot n_2 & 1 \cdot n_2 \end{bmatrix} = \begin{bmatrix} n_1 + n_2 & n_2 \\ n_2 & n_2 \end{bmatrix} \\ (\mathbf{X}'\mathbf{X})^{-1} &= \frac{1}{(n_1 + n_2)n_2 - (n_2)^2} \begin{bmatrix} n_2 & -n_2 \\ -n_2 & n_1 + n_2 \end{bmatrix} \\ &= \frac{1}{n_1 n_2} \begin{bmatrix} n_2 & -n_2 \\ -n_2 & n_1 + n_2 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{n_1} & -\frac{1}{n_1} \\ -\frac{1}{n_1} & \frac{n_1 + n_2}{n_1 n_2} \end{bmatrix} = \begin{bmatrix} \frac{1}{n_1} & -\frac{1}{n_1} \\ -\frac{1}{n_1} & \frac{1}{n_1} + \frac{1}{n_2} \end{bmatrix} \end{aligned}$$

$$\begin{aligned}
(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' &= \begin{bmatrix} \frac{1}{n_1} & -\frac{1}{n_1} \\ -\frac{1}{n_1} & \frac{1}{n_1} + \frac{1}{n_2} \end{bmatrix} \begin{bmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{n_1} & \cdots & \frac{1}{n_1} & 0 & \cdots & 0 \\ -\frac{1}{n_1} & \cdots & -\frac{1}{n_1} & \frac{1}{n_2} & \cdots & \frac{1}{n_2} \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \begin{bmatrix} \frac{1}{n_1} & \cdots & \frac{1}{n_1} & 0 & \cdots & 0 \\ -\frac{1}{n_1} & \cdots & -\frac{1}{n_1} & \frac{1}{n_2} & \cdots & \frac{1}{n_2} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{n_1}y_1 + \cdots + \frac{1}{n_1}y_{n_1} + 0 + \cdots + 0 \\ -\frac{1}{n_1}y_1 + \cdots + -\frac{1}{n_1}y_{n_1} + \frac{1}{n_2}(y_{n_1+1}) + \cdots + \frac{1}{n_2}(y_{n_1+n_2}) \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{n_1}(y_1 + \cdots + y_{n_1}) \\ -\frac{1}{n_1}(y_1 + \cdots + y_{n_1}) + \frac{1}{n_2}(y_{n_1+1} + \cdots + y_{n_1+n_2}) \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{n_1} \sum_{i=1}^{n_1} y_i \\ -\frac{1}{n_1} \sum_{i=1}^{n_1} y_i + \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} y_i \end{bmatrix} \\
&= \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}
\end{aligned}$$

(c) Interpret the coefficient β_0 .

The average response variable of the subject in the control group.

(d) Interpret the coefficient β_1 .

For the subject in the treatment group, the average response variable increases by β_1 on average.

Problem 2: Unique solution for $\hat{\beta}$

Assume you observe four response variables, y_i , for $i = 1, \dots, 4$. Further assume the model for the four observations can be written as follows:

$$\begin{aligned}
y_1 &= \beta_0 + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 + \varepsilon_2, \\
y_3 &= \beta_0 + \beta_2 + \varepsilon_3, \\
y_4 &= \beta_0 + \beta_1 + \beta_2 + \varepsilon_4,
\end{aligned}$$

and $E(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$.

a) Assuming $\beta = (\beta_0, \beta_1, \beta_2)'$, determine the design matrix X for that model.

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

b) Discuss whether or not there exists a unique solution for, $\hat{\beta}$, that minimizes $(y - X\beta)'(y - X\beta)$. Justify your answer.

Model in matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

To have a unique solution $\hat{\beta}$ that minimizes $(y - X\beta)'(y - X\beta)$, the matrix $X'X$ should be invertible.

$$X' = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad X'X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{bmatrix}$$

Then, check whether $\det(X'X) \neq 0$ to determine if $X'X$ is invertible.

$$\begin{aligned} \det(X'X) &= 4(2 \cdot 2 - 1 \cdot 1) - 2(2 \cdot 2 - 1 \cdot 2) + 2(2 \cdot 1 - 2 \cdot 2) \\ &= 12 - 4 + (-4) \\ &= 4 \end{aligned}$$

Since $\det(X'X) = 4 \neq 0$, $X'X$ is invertible.

Therefore, there exists a unique solution $\hat{\beta}$ that minimizes $(y - X\beta)'(y - X\beta)$.