Homework 1

STAT3530 Applied Linear Models

Due Thursday, September 3, in class.

You may collaborate, but write up your own solutions. To receive credit, you must submit a paper copy of this document with hand written solutions (by tablet if you like) below each prompt. You can append up to two additional pages if you need extra space; if so, please number clearly.

Throughout, $x_1, \ldots, x_n$ are fixed numbers and $Y_1, \ldots, Y_n$ are random variables.

A few identities

These are short, and will reappear later.

1. Show that $\sum_{i=1}^n (x_i - \bar{x}) = 0$.

2. Show that $\sum_{i=1}^n (x_i - \bar{x})^2 = (n-1)s_x^2$.

3. Show that $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$

4. Using the result of problem 1, show that

$$\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^n (x_i - \bar{x})Y_i$$

   That is, centering the Y_i makes no difference here.

Least squares estimates for the simple linear model

Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and the sum of squared errors

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

We minimize in two steps by profiling: first minimize over β_0 holding β_1 fixed, then over β_1.

5. Treating β_1 as a constant, differentiate $SSE(\beta_0, \beta_1)$ with respect to β_0, set the result to zero, and solve to obtain $\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{x}$.

6. Substitute your answer from problem 5 back into SSE and show that it simplifies to

$$\sum_{i=1}^{n} \left([Y_i - \bar{Y}] - \beta_1 [x_i - \bar{x}] \right)^2$$

7. Differentiate the expression in problem 6 with respect to β_1, set the result to zero, and solve to obtain

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Then write down the resulting estimate $\widehat{\beta}_0$.

8. The previous problem shows that $\widehat{\beta}_1$ is a *linear combination* of the responses:

$$\widehat{\beta}_1 = \sum_{i=1}^n w_i Y_i = w^T Y$$

Identify w_i, then show that $\sum_i w_i = 0$.

9. Briefly: using the result in the previous problem, what happens to $\widehat{\beta}_1$ if every response Y_i is increased by the same constant c? What happens to $\widehat{\beta}_0$?

The vector of ones

Let $\mathbf{1}$ be the n-dimensional vector with every entry equal to 1, let I be the $n \times n$ identity matrix, and let $y \in \mathbb{R}^n$.

10. Noting $\mathbf{1}^T\mathbf{1}$ and $\mathbf{1}^T Y = \sum_{i=1}^n Y_i$, write $\bar{Y}$ in terms of $\mathbf{1}$ and Y.

11. Define $J = \frac{1}{n}\mathbf{1}\mathbf{1}^T$. Show that J is symmetric and idempotent.

12. Show that $JY = \bar{Y}\mathbf{1}$. Describe in words what J does to a vector.

13. Define $M = I - J$. Show that M is symmetric and idempotent, and that $MY = Y - \bar{Y}\mathbf{1}$. Describe in words what M does to a vector.

14. Show that $Y^T M Y = \sum_{i=1}^n (Y_i - \bar{Y})^2$. What familiar quantity is $\frac{1}{n-1} Y^T M Y$?

15. Each matrix below has $\mathbf{1}$ as its first column. Determine the rank of each, and wherever the columns are linearly dependent give the dependency explicitly.

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Random variables and random vectors

Suppose $Y_1, \ldots, Y_n$ are independent, each with mean $\mathbb{E}[Y_i] = \mu$ and variance $\mathrm{Var}[Y_i] = \sigma^2$, and collect them into the random vector $Y = \begin{bmatrix} Y_1 & \cdots & Y_n \end{bmatrix}^T$.

16. Explain why $\mathbb{E}[Y] = \mu\mathbf{1}$ and $\mathrm{Var}[Y] = \sigma^2 I$. Say in particular why every off-diagonal entry of $\mathrm{Var}[Y]$ is zero.

17. The sample mean can be written $\bar{Y} = \frac{1}{n}\mathbf{1}^T Y$. Use the rules for $\mathbb{E}[AY]$ and $\mathrm{Var}[AY]$ to show that

$$\mathbb{E}[\bar{Y}] = \mu \qquad \text{and} \qquad \mathrm{Var}[\bar{Y}] = \frac{\sigma^2}{n}$$

18. Let a be any fixed $n \times 1$ vector. Show that

$$\mathrm{Var}[a^T Y] = \sigma^2 a^T a = \sigma^2 \sum_{i=1}^{n} a_i^2$$

19. Problem 8 establishes that $\hat{\beta}_1 = w^T Y$ with $w_i = (x_i - \bar{x}) \big/ \sum_j (x_j - \bar{x})^2$. Assume $\text{Var}[Y] = \sigma^2 I$ and use the result in the previous problem to show that

$$\text{Var}[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Fitting a model in R

The `galapagos` data record the area of 30 Galapagos islands and the number of plant species found on each, both on the log scale. Load it into R with

```
load(url("https://tdruiz-stat3530.share.connect.posit.cloud/_data/galapagos.RData"))
```

20. Make a scatterplot of `log.species` against `log.area`. Do the data show a linear trend? If so, describe it.

21. fit the simple linear model with `lm()`. Report the estimates from your fitted model.

    $\hat{\beta}_0 =$ ____________ , $\hat{\beta}_1 =$ ____________ , $\hat{\sigma} =$ ____________

22. Write out the fitted model equation.

23. Interpret each of the three estimates in the context of the data. Remember that both variables are on the log scale.