Homework 2

STAT3530 Applied Linear Models

Due Thursday, September 10, in class.

You may collaborate, but write up your own solutions. Submit a paper copy of this document with handwritten solutions below each prompt.

Throughout, $Y = X\beta + \epsilon$ with $\mathbb{E}[\epsilon] = 0$ and $\text{Var}[\epsilon] = \sigma^2 I$, and X has full column rank unless stated otherwise.

Recovering the simple linear regression estimates

On Homework 1 you derived the least squares estimates for the simple linear model one parameter at a time. Here you will verify that the matrix formula $\hat{\beta} = (X^T X)^{-1} X^T Y$ recovers the same estimates.

Take the design matrix with one predictor,

$$X = \begin{bmatrix} \mathbf{1} & x \end{bmatrix}$$

You may use the identities $\sum_{i=1}^n (x_i - \bar{x})^2 = (n-1)s_x^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$ from Homework 1 and the formula for the inverse of a 2×2 matrix,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

where $ad - bc$ is the matrix determinant.

1. Compute $X^T X$ and $X^T Y$, writing each entry as a sum. Then rewrite the two entries with $\sum_i x_i = n\bar{x}$ and $\sum_i Y_i = n\bar{Y}$.

2. Show that $\det(X^T X) = n(n-1)s_x^2$, then write down $(X^T X)^{-1}$.

3. Multiply out $\hat{\beta} = (X^T X)^{-1} X^T Y$ and show that its second entry is

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

which is what you found on Homework 1. *Hint: expand* $\sum_i (x_i - \bar{x}) Y_i$.

4. Show that the first entry simplifies to $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$. *Hint: put* $\bar{Y} - \hat{\beta}_1 \bar{x}$ *over the common denominator* $(n-1)s_x^2$*, then work backwards.*

Centering the predictors

In class we centered the predictor, giving the design matrix

$$X = \begin{bmatrix} \mathbf{1} & \tilde{x} \end{bmatrix}, \qquad \tilde{x}_i = x_i - \bar{x}$$

This is the model you just worked through, with x replaced by $\tilde{x}$. Notice how much the algebra simplifies.

5. Show that $\mathbf{1}^T\tilde{x} = 0$.

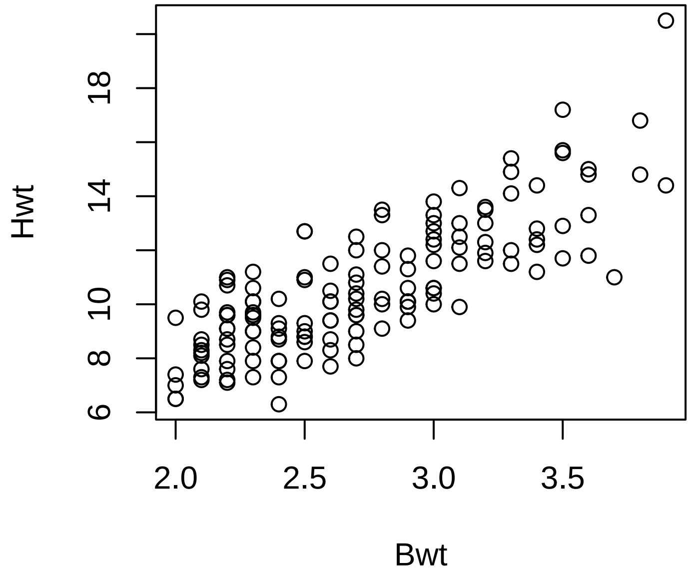
6. Using problem 5, show that

$$X^TX = \begin{bmatrix} n & 0 \\ 0 & (n-1)s_x^2 \end{bmatrix} \qquad \text{and} \qquad (X^TX)^{-1} = \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{(n-1)s_x^2} \end{bmatrix}$$

7. What is $\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1]$, and what does that say about the two estimators?

8. The `cats` data has heart weights in grams and body weights in kilograms for 144 adult cats. A plot of the data is shown below.

```
data(cats, package = 'MASS')
plot(Hwt ~ Bwt, data = cats)
```



a. Fit the model `Hwt ~ Bwt` with and without centering the predictor. Report and interpret the intercept estimate in each case.

b. Estimate $\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1)$ for each model.

c. Now refit the model after centering the *response*. What happens to the estimates? Explain.

The hat matrix

Let $H = X(X^TX)^{-1}X^T$ and $M = I - H$.

9. Show that H is symmetric and idempotent, *i.e.*, that it is a projection.

10. Show that M is symmetric and idempotent, *i.e.*, that it is also a projection.

11. Show that $HX = X$, $MX = 0$, and $MH = 0$. Interpret each result briefly in words.

Least squares for a population mean

Suppose we have a random sample $Y_1, \ldots, Y_n$, each with mean μ and variance σ^2, and we want to estimate μ. Express this via the linear model

$$Y = X\beta + \epsilon \qquad \text{with} \qquad X = \mathbf{1}\,, \quad \beta = \mu$$

12. Compute $X^T X$, $(X^T X)^{-1}$, and $X^T Y$ for this model and show that the least squares estimator is the sample mean $\bar{Y}$.

13. Use $\mathbb{E}[\hat{\beta}] = \beta$ and $\text{Var}[\hat{\beta}] = \sigma^2 (X^T X)^{-1}$ to show that the sample mean is unbiased with variance $\text{Var}[\bar{Y}] = \frac{\sigma^2}{n}$.

14. Compute the hat matrix for this model.

15. The `temps` dataset contains 130 measurements of body temperatures in Fahrenheit for a sample of adults. In R, you can fit an intercept-only model via the formula `lm(y ~ 1)`. Verify that the least squares estimate recovers the sample mean, and its estimated variance recovers the usual standard error for a sample mean. Then construct a 95% confidence interval for the mean body temperature. Is 98.6 a plausible value?

```r
load(url("https://tdruiz-stat3530.share.connect.posit.cloud/_data/temps.RData"))
hist(temps$body.temp, xlab = 'Body temperature', main = NULL)
```