Homework 1 solutions

STAT3530 Applied Linear Models

Solutions are shown in blue. Sums run from 1 to n throughout.

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

Split the sum and use $\sum_i x_i = n\bar{x}$:

$$\sum_i (x_i - \bar{x}) \;=\; \sum_i x_i - \sum_i \bar{x} \;=\; n\bar{x} - n\bar{x} \;=\; 0$$

The second sum has n identical terms because $\bar{x}$ does not depend on i.

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

This is the definition of the sample variance, rearranged. Since $s_x^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$, multiplying both sides by $n-1$ gives

$$(n-1)s_x^2 \;=\; \sum_i (x_i - \bar{x})^2$$

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

Expand the square:

$$\sum_i (x_i - \bar{x})^2 \;=\; \sum_i \left(x_i^2 - 2\bar{x}x_i + \bar{x}^2\right) \;=\; \sum_i x_i^2 \;-\; 2\bar{x}\sum_i x_i \;+\; n\bar{x}^2$$

The middle term is $2\bar{x}(n\bar{x}) = 2n\bar{x}^2$, so the last two combine to $-2n\bar{x}^2 + n\bar{x}^2 = -n\bar{x}^2$, leaving $\sum_i x_i^2 - n\bar{x}^2$.

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.

Distribute, then pull the constant $\bar{Y}$ out of the second sum:

$$\sum_i (x_i - \bar{x})(Y_i - \bar{Y}) \;=\; \sum_i (x_i - \bar{x})Y_i \;-\; \bar{Y}\underbrace{\sum_i (x_i - \bar{x})}_{=\,0\text{ by problem 1}} \;=\; \sum_i (x_i - \bar{x})Y_i$$

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 $\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{x}$.

With β_1 held fixed, SSE is a function of the single variable β_0. By the chain rule the derivative of $(Y_i - \beta_0 - \beta_1 x_i)^2$ is $2(Y_i - \beta_0 - \beta_1 x_i) \cdot (-1)$, so

$$\frac{d}{d\beta_0} SSE \;=\; -2 \sum_i (Y_i - \beta_0 - \beta_1 x_i)$$

Setting this to zero and dividing by -2,

$$\sum_i Y_i - n\beta_0 - \beta_1 \sum_i x_i \;=\; 0 \qquad \Longrightarrow \qquad n\bar{Y} - n\beta_0 - \beta_1 n\bar{x} \;=\; 0$$

Dividing by n and solving gives $\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{x}$.

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

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

Substitute $\beta_0 = \bar{Y} - \beta_1 \bar{x}$ into a single term and regroup:

$$Y_i - \beta_0 - \beta_1 x_i \;=\; Y_i - (\bar{Y} - \beta_1 \bar{x}) - \beta_1 x_i \;=\; (Y_i - \bar{Y}) - \beta_1 (x_i - \bar{x})$$

Squaring and summing gives the stated expression, which is now a function of β_1 alone.

7. Differentiate the expression in problem 6 with respect to β_1, set the result to zero, and solve to obtain
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
Then write down the resulting estimate $\hat{\beta}_0$.

The derivative of the inside with respect to β_1 is $-(x_i - \bar{x})$, so

$$\frac{d}{d\beta_1} \sum_i \left([Y_i - \bar{Y}] - \beta_1[x_i - \bar{x}]\right)^2 \;=\; -2\sum_i (x_i - \bar{x})\Big([Y_i - \bar{Y}] - \beta_1[x_i - \bar{x}]\Big)$$

Set this to zero and separate the two pieces:

$$\sum_i (x_i - \bar{x})(Y_i - \bar{Y}) \;=\; \beta_1 \sum_i (x_i - \bar{x})^2$$

Solving, then applying problem 4 to the numerator,

$$\hat{\beta}_1 \;=\; \frac{\sum_i (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_i (x_i - \bar{x})^2} \;=\; \frac{\sum_i (x_i - \bar{x})Y_i}{\sum_i (x_i - \bar{x})^2}$$

Substituting back into problem 5, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$.

8. The previous problem shows that $\hat{\beta}_1$ is a *linear combination* of the responses:
$$\hat{\beta}_1 = \sum_{i=1}^n w_i Y_i = w^T Y$$
Identify w_i, then show that $\sum_i w_i = 0$.

Reading the weights off problem 7,
$$w_i \;=\; \frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2}$$
The denominator does not depend on i, so it factors out of the sum:

$$\sum_i w_i \;=\; \frac{\sum_i (x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} \;=\; 0$$

using problem 1 in the numerator.

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

Replacing each Y_i by $Y_i + c$,

$$\sum_i w_i(Y_i + c) \;=\; \sum_i w_i Y_i \;+\; c\underbrace{\sum_i w_i}_{=0} \;=\; \hat{\beta}_1$$

so the **slope does not change**. The intercept becomes $(\bar{Y} + c) - \hat{\beta}_1 \bar{x} = \hat{\beta}_0 + c$, so the **intercept increases by** c. This is what it should be: shifting every response up by the same amount slides the line up without tilting it.

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.

Since $\mathbf{1}^T\mathbf{1} = n$ and $\mathbf{1}^T Y = \sum_i Y_i$,

$$\bar{Y} \;=\; \left(\mathbf{1}^T\mathbf{1}\right)^{-1}\mathbf{1}^T Y$$

Written this way it is already in the shape of the general formula $\hat{\beta} = (X^T X)^{-1} X^T Y$, with $X = \mathbf{1}$.

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

Symmetric, using $(AB)^T = B^T A^T$:

$$J^T \;=\; \tfrac{1}{n}\left(\mathbf{1}\mathbf{1}^T\right)^T \;=\; \tfrac{1}{n}\mathbf{1}\mathbf{1}^T \;=\; J$$

Idempotent, regrouping so the scalar $\mathbf{1}^T\mathbf{1} = n$ appears in the middle:

$$JJ \;=\; \tfrac{1}{n^2}\mathbf{1}(\mathbf{1}^T\mathbf{1})\mathbf{1}^T \;=\; \tfrac{1}{n^2}\cdot n \cdot \mathbf{1}\mathbf{1}^T \;=\; J$$

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

Group the product as $\mathbf{1}(\mathbf{1}^T Y)$, noting $\mathbf{1}^T Y$ is a scalar:

$$JY \;=\; \tfrac{1}{n}\mathbf{1}(\mathbf{1}^T Y) \;=\; \mathbf{1}\cdot\tfrac{1}{n}\sum_i Y_i \;=\; \bar{Y}\mathbf{1}$$

In words: J replaces every entry of a vector by the average of its entries. It projects onto the set of vectors with all entries equal.

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.

$M^T = I^T - J^T = I - J = M$ by problem 11. For idempotence, expand and use $JJ = J$:

$$MM \;=\; (I-J)(I-J) \;=\; I - J - J + JJ \;=\; I - J \;=\; M$$

And $MY = (I-J)Y = Y - JY = Y - \bar{Y}\mathbf{1}$ by problem 12. In words: M centers a vector, subtracting the mean from every entry.

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

Because M is symmetric and idempotent, $M = M^T M$, so

$$Y^T M Y \;=\; Y^T M^T M Y \;=\; (MY)^T(MY) \;=\; \|MY\|^2 \;=\; \sum_i (Y_i - \bar{Y})^2$$

using problem 13 in the last step. Therefore $\frac{1}{n-1} Y^T M Y = s_Y^2$, the **sample variance**.

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.

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

$\mathrm{rank}(A) = 2$. One column is constant and the other is not, so neither is a multiple of the other; a 4×2 matrix can have rank at most 2.

$\mathrm{rank}(B) = 2$. The columns are dependent — the second and third add to the first:

$$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad \text{that is,} \qquad B_{\cdot 1} - B_{\cdot 2} - B_{\cdot 3} = 0$$

Any two of the three are independent, so the rank is exactly 2. This is what a design matrix looks like with an intercept *and* an indicator for every group: the indicators sum to the intercept column, so $B^T B$ has no inverse.

Random variables and random vectors

Suppose $Y_1, \dots, Y_n$ 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.

Expectation is taken entrywise, and the ith entry of $\mathbb{E}[Y]$ is $\mathbb{E}[Y_i] = \mu$. A vector with every entry μ is $\mu\mathbf{1}$.

For the variance, the (i, i) entry is $\mathrm{Var}[Y_i] = \sigma^2$ and the (i, j) entry for $i \neq j$ is $\mathrm{Cov}[Y_i, Y_j]$. The Y_i are independent, and independent random variables have zero covariance, so every off-diagonal entry is 0. That matrix is $\sigma^2 I$.

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

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

Apply the rules with the fixed $1 \times n$ matrix $A = \frac{1}{n}\mathbf{1}^T$:

$$\mathbb{E}[\bar{Y}] \;=\; A\,\mathbb{E}[Y] \;=\; \tfrac{1}{n}\mathbf{1}^T(\mu\mathbf{1}) \;=\; \tfrac{\mu}{n}\,\mathbf{1}^T\mathbf{1} \;=\; \mu$$

$$\mathrm{Var}[\bar{Y}] \;=\; A\,\mathrm{Var}[Y]\,A^T \;=\; \tfrac{1}{n}\mathbf{1}^T(\sigma^2 I)\tfrac{1}{n}\mathbf{1} \;=\; \tfrac{\sigma^2}{n^2}\,\mathbf{1}^T\mathbf{1} \;=\; \frac{\sigma^2}{n}$$

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

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

Here a^T plays the role of A:

$$\mathrm{Var}[a^T Y] \;=\; a^T\,\mathrm{Var}[Y]\,a \;=\; a^T(\sigma^2 I)a \;=\; \sigma^2 a^T a \;=\; \sigma^2 \sum_i a_i^2$$

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 $\mathrm{Var}[Y] = \sigma^2 I$ and use the result in the previous problem to show that

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

Take $a = w$ in problem 18. Every w_i has the same denominator, so

$$\sum_i w_i^2 \;=\; \frac{\sum_i (x_i - \bar{x})^2}{\left(\sum_j (x_j - \bar{x})^2\right)^2} \;=\; \frac{1}{\sum_i (x_i - \bar{x})^2}$$

with one factor cancelling. Therefore

$$\mathrm{Var}[\hat{\beta}_1] \;=\; \mathrm{Var}[w^T Y] \;=\; \sigma^2 \sum_i w_i^2 \;=\; \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

The estimate is more precise when the predictors are more spread out.

Fitting a model in R

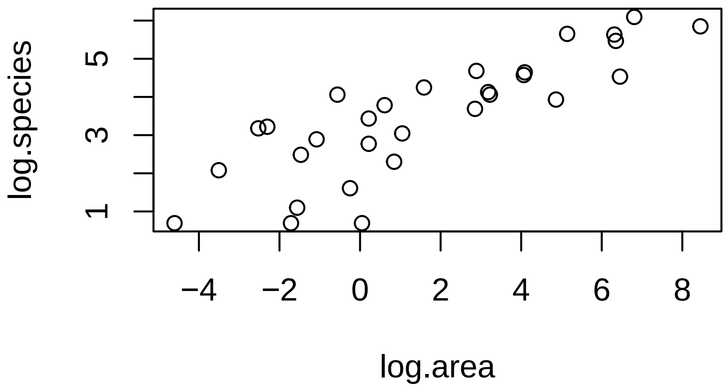
The `galapagos` data record the area of 30 Galapagos islands and the number of plant species found on each, both on the log scale. Load it into R with

```
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.

Yes. The trend is positive and reasonably linear: islands with larger area have more plant species. The scatter is moderate, with no obvious curvature or outliers.

```
plot(log.species ~ log.area, data = galapagos)
```



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

```
fit <- lm(log.species ~ log.area, data = galapagos)
coef(fit)
```

```
(Intercept)    log.area
  2.8705854   0.3847802
```

```
sigma(fit)
```

```
[1] 0.8668994
```

$\hat{\beta}_0 = 2.871, \quad \hat{\beta}_1 = 0.385, \quad \hat{\sigma} = 0.867$

22. Write out the fitted model equation.

$$\widehat{\text{log.species}} = 2.871 + 0.385 \times \text{log.area}$$

The fitted value is an estimate of the **mean** log species count among islands at that log area.

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

All three describe the *mean* log species count — the model says how the average of `log.species` behaves, not how any individual island behaves.

Slope. For each one-unit increase in log area, the mean log species count increases by an estimated 0.385.

Intercept. Among islands with log area 0, the estimated mean log species count is 2.871. Log area ranges from about -4.6 to 8.4 in these data, so 0 is well inside the observed range and this is a genuine interpretation rather than an extrapolation.

Residual standard deviation. After accounting for log area, log species varies by an estimated 0.867.