Exercises for Linear Models with Matrices

1. Suppose we have data $x_1, \ldots, x_n$ and $y_i, \ldots, y_n$, which we model as

$$Y_i = b_0 + b_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2).$$

   (a) Write down the formula for the sum of squares criterion.

   (b) Take the partial derivatives of the sum of squares criterion to derive the normal equations.

   (c) Solve the normal equations to show that the least squares estimates are

$$\begin{aligned}\widehat{b}_1 &= \frac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} \\ \widehat{b}_0 &= \overline{y} - \widehat{b}_1 \overline{x}\end{aligned}$$

2. In terms of the estimates $\widehat{b}_0$ and $\widehat{b}_1$, write down the formulas for the fitted values, the residuals, and $\widehat{\sigma}^2$ (the variance estimate).

3. Explain the difference between an estimate and an estimator.

   Answer:

   An estimate is a function of the data, for example $\widehat{b} = \overline{y}$, whereas an estimator is a random variable, a function of the random variables used to model the data, for example $\widehat{B} = \overline{Y}$.

4. Write down how you would explain all of the assumptions of the simple linear model for $y_1, \ldots, y_n$ given $x_1, \ldots, x_n$ to your grandmother (who was a chemical engineer).

5. Suppose we have data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. The quantity sxy is given to the left of the equals sign and is always equal to **exactly one** of the three expressions to the right of the equals sign. Circle the correct expression and show why sxy is equal to it.

$$\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \quad = \quad \sum_{i=1}^{n} (x_i - \overline{x}) y_i \qquad \sum_{i=1}^{n} (x_i - \overline{x}) \overline{y} \qquad \sum_{i=1}^{n} \overline{x}(y_i - x_i)$$

   Answer:

   The first one is the correct answer. To see why, we subtract zero in a clever way and regroup the terms:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \overline{x})y_i &= \sum_{i=1}^{n}(x_i - \overline{x})y_i - \overline{y}\sum_{i=1}^{n}(x_i - \overline{x}) \\
&= \sum_{i=1}^{n}(x_i - \overline{x})y_i - \sum_{i=1}^{n}(x_i - \overline{x})\overline{y} \\
&= \sum_{i=1}^{n}\left[(x_i - \overline{x})y_i - (x_i - \overline{x})\overline{y}\right] \\
&= \sum_{i=1}^{n}\left[(x_i - \overline{x})(y_i - \overline{y})\right]
\end{aligned}$$

6. Derive a formula for $\mathrm{Cov}(\widehat{B}_0, \widehat{B}_1)$

   Answer:

$$\begin{aligned}
\mathrm{Cov}(\widehat{B}_0, \widehat{B}_1) &= \mathrm{Cov}(\overline{Y} - \widehat{B}_1\overline{x}, \widehat{B}_1) \\
&= \mathrm{Cov}(\overline{Y}, \widehat{B}_1) - \overline{x}\mathrm{Cov}(\widehat{B}_1, \widehat{B}_1) \\
&= \mathrm{Cov}(\overline{Y}, \widehat{B}_1) - \overline{x}\mathrm{Var}(\widehat{B}_1) \\
&= \mathrm{Cov}(\overline{Y}, \widehat{B}_1) - \overline{x}\sigma^2/sxx
\end{aligned}$$

   The last step uses the formula for $\mathrm{Var}(\widehat{B}_1)$. Now we must consider the first term.

$$\begin{aligned}
\mathrm{Cov}(\overline{Y}, \widehat{B}_1) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \frac{1}{sxx}\sum_{i=1}^{n}(x_i - \overline{x})Y_i\right) \\
&= \frac{1}{n}\frac{1}{sxx}\sum_{i=1}^{n}(x_i - \overline{x})\mathrm{Var}(Y_i) \\
&= \frac{1}{n}\frac{1}{sxx}\sum_{i=1}^{n}(x_i - \overline{x})\sigma^2 \\
&= 0
\end{aligned}$$

   The second-to-last step uses the fact that the Y_i's are indepdent. So the answer is $-\overline{x}\sigma^2/sxx$.

7. Use the formulas for the variances and covariances of $\widehat{B}_0$ and $\widehat{B}_1$ to derive a simplified expression for the prediction variance $\mathrm{Var}(\widehat{B}_0 + \widehat{B}_1 x_i)$.

8. Suppose we have data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. Which of the following equations

qualify as statistical models for $y_1, \dots, y_n$?

$$
\begin{aligned}
&Y_i = b_0 + b_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2) \\
&Y_i = b_0 + b_1 x_i \\
&Y_i = e^{\cos(x_i) + \varepsilon_i}, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma^2) \\
&0 \le Y_i \le 10 \\
&Y_i \overset{ind}{\sim} \text{Uniform}(0, 10)
\end{aligned}
$$

Answer:

Yes to the first one. It is the simple linear model that fully expresses Y_i as a random variable and explains the assumptions underlying the random term ε_i.

No to the second because because the right side of the equation is not random.

Yes to the third one. Even though this is a nonlinear model, it still specifies Y_i as a random variable and delineates all of the assumptions. This happens to be a linear model for $\log(Y_i)$, a transformation of the response.

No to the fourth one because it merely says that Y_i is between 0 and 10, without specifying the distribution on that interval.

Yes to the fifth one, because it specifies the distribution–uniform–on the interval 0 to 10.

9. Under the simple linear model, there is a formula for the least squares estimators of the regression coefficients. Which property of the estimators allows us to conclude that the estimators are normally distributed?

10. Select all that are correct: A confidence interval for the slope b_1 . . .

    (a) contains the slope values that we rejected using a hypothesis test

    (b) contains the slope values that we failed to reject using a hypothesis test

    (c) is a set of plausible values of the slope given the data

    (d) has endpoints that are random variables

    Answer:

    (a) is wrong. It contains values that we fail to reject.

    (b) is right

    (c) is right, this is good interpretation of the definition of a confidence interval

    (d) is wrong. We *model* the endpoints as relizations of random variables. The endpoints themselves are numbers that we can write down and therefore are non-random.

11. Name two desirable properties of estimators, and explain why they are desirable.