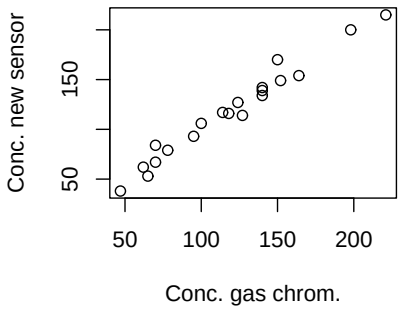
MATH356H Assignment #1
Solutions

1. Chapter 12, #3.

   A scatter plot for the data can be found below. The points seem to fall close to the line $y = x$, so that both methods do appear to be measuring roughly the same quantity.



2. Consider the simple regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

discussed in sections 12.1 and 12.2. Assume that $X = 0$ is within the scope of the model.

   (a) What is the implication for the regression function if $\beta_0 = 0$ so that the model is

$$Y = \beta_1 X + \varepsilon?$$

   That is, how would the regression function plot on a graph?
   The model forces the line to go through the point $(0, 0)$, so that the graph is a line that crosses the origin.

   (b) Find the least squares estimator of β_1 for the model $Y = \beta_1 X + \varepsilon$ given n points (x_i, y_i). (Follow the same procedure as in page 498 of Devore).

   Write $Q = \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2$. Then

$$\frac{dQ}{d\beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_1 X_i).$$

   Setting that to 0 and solving we get:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

   This is indeed a minimum because of the second derivative criterion:

$$\frac{d^2 Q}{d\beta_1^2} = \sum_{i=1}^{n} X_i^2 > 0.$$

   (c) Show that the estimator you obtained in (b) for β_1 is unbiased.

$$\begin{aligned} E\{\hat{\beta}_1\} &= E\left\{\frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}\right\} \\ &= E\left\{\frac{\sum_{i=1}^n X_i(\beta_1 X_i + \varepsilon_i)}{\sum_{i=1}^n X_i^2}\right\} \\ &= E\left\{\beta_1 \frac{\sum_{i=1}^n X_i^2}{\sum_{i=1}^n X_i^2}\right\} + \frac{1}{\sum_{i=1}^n X_i^2}\sum_{i=1}^n E(\varepsilon_i) = \beta_1, \end{aligned}$$

where we have used the fact that $E(\varepsilon_i) = 0$. Thus $\hat{\beta}_1$ is unbiased.

3. Show that the vector of residuals and the vector of fitted values are orthogonal to each other. That is, let e_i denote the i^{th} residual:

$$e_i \doteq y_i - \hat{y}_i,$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Then show that

$$\sum_{i=1}^n \hat{y}_i e_i = 0.$$

We have that $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ and $e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$. Substitution yields

$$\begin{aligned} \sum_{i=1}^n \hat{Y}_i e_i &= \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 X_i)(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\ &= \sum_{i=1}^n \left\{\hat{\beta}_0 Y_i - \hat{\beta}_0^2 - \hat{\beta}_0\hat{\beta}_1 X_i + \hat{\beta}_1 X_i Y_i - \hat{\beta}_0\hat{\beta}_1 X_i - \hat{\beta}_1^2 X_i^2\right\} \\ &= \hat{\beta}_0\left\{\sum_{i=1}^n Y_i - n\hat{\beta}_0 - \hat{\beta}_1\sum_{i=1}^n X_i\right\} + \hat{\beta}_1\left\{\sum_{i=1}^n X_i Y_i - \hat{\beta}_0\sum_{i=1}^n X_i - \hat{\beta}_1\sum_{i=1}^n X_i^2\right\} \\ &= 0 \end{aligned}$$

because of the normal equations.

4. Chapter 12, #12

All of the calculations below were done using `R`.

(a)

$$\begin{aligned} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{25,825 - (517)(346)/14}{39,095 - (517)^2/14} = .65299. \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1\bar{X} = \frac{346}{14} - (.65229)\left(\frac{517}{14}\right) = .6261. \end{aligned}$$

(b)

$$\hat{Y} = .6261 + .65229(35) = 23.45625 \qquad e = 21 - 23.45625 = -2.45625.$$

(c)

$$\begin{aligned} SSE &= \sum Y_i^2 - \hat{\beta}_0\sum Y_i - \hat{\beta}_1\sum X_i Y_i \\ &= 17454 - (.6261)(346) - (.65229)(25825) = 391.98015. \\ S^2 &= \frac{SSE}{n-2} = \frac{391.98015}{12} = 32.665 \\ S &= \sqrt{32.665} = 5.715. \end{aligned}$$

(d)

$$SST = S_{yy} = 17454 - (346)^2/14 = 8902.857.$$
$$r^2 = 1 - \frac{391.98015}{8902.857} = .95597.$$

Thus 95.6% of observed variation is explained by the model.

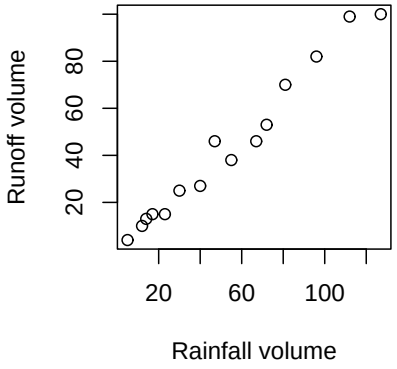
(e) If we delete 103 and 142 the new equation of the line is

$$E(Y) = 2.2891 + .56445X,$$

with $r^2 = .6879$, so only 68.79% of variation is explained by the fitted line.

5. Chapter 12, #16

(a) Yes, the scatter plot suggests a strong linear relationship between rainfall volume and runoff volume.



(b) $\hat{\beta}_0 = -1.128304771$ and $\hat{\beta}_1 = 0.826973147$.

(c) $\hat{Y} = -1.128304771 + 0.826973147(50) = 40.22$.

(d) $S = 5.240461548$

(e) $r^2 = 0.975268892$, so 97.5% of variation is explained by the regression line.

6. Chapter 12, #28

(a) Subtracting $\bar{x}$ from each x_i shifts the plot in a rigid fashion x units to the left without otherwise altering its character. The last squares line for the new plot will thus have the same slope as the one for the old plot. Since the new line is x units to the left of the old one, the new y intercept (height at $x = 0$) is the height of the old line at $x = \bar{x}$, which is $\hat{\beta}_0 + \hat{\beta}_1\bar{x} = \bar{y}$ (since $(\bar{x}, \bar{y})$ is on the old line - you can verify this on Exercise 26 (D)). Thus the new y intercept is $\bar{y}$.

(b) Let us call the estimators b_0 and b_1. We wish b_0 and b_1 to minimize

$$f(b_0, b_1) = \sum [y_i - (b_0 + b_1(X_i - \bar{x}))]^2.$$

Equating $\partial f/\partial b_0$ and $\partial f/\partial b_1$ to 0 yields

$$nb_0 + b_1 \sum (x_i - \bar{x}) = \sum y_i,$$

$$b_0 \sum (x_i - \bar{x}) + b_1 \sum (x_i - \bar{x})^2 = \sum (x_i - \bar{x})^2 = \sum (x_i - \bar{x}) y_i.$$

Since $\sum(x_i - \bar{x}) = 0$, $b_0 = \bar{y}$, and since $\sum(x_i - \bar{x})y_i = \sum(x_i - \bar{x})(y_i - \bar{y})$ [because $\sum(x_i - \bar{x})\bar{y} = \bar{y}\sum(x_i - \bar{x}) = 0$], $b_1 = \hat{\beta}_1$. Thus $\hat{\beta}_0^* = \bar{y}$ and $\hat{\beta}_1^* = \hat{\beta}_1^*$.