Assignment #5: Chapter 11 Questions 1, 20, 22, 38, 44

Question 11.1: If $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimates for the intercept and slopes in a simple linear regression model, show that the least-squares equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ always goes through the point $(\bar{x}, \bar{y})$.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \left(\bar{y} - \hat{\beta}_1 \bar{x}\right) + \hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y}$$

Clearly $\langle \bar{x}, \bar{y} \rangle$ is on the line of the least squares equation.

Question 11.20: Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent normal random variables with $\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_i$ and $\mathrm{Var}[Y_i] = \sigma^2$, for $i = 1, 2, \ldots, n$. Show that the maximum likelihood estimators (MLEs) of β_0 and β_1 are the same as the least squares estimators.
If $f(\beta_0, \beta_1 \mid x_i . y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2}$, then the likelihood function is

$$L(\beta_0, \beta_1 \mid x_i, y_i) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_i^n (y_i - \beta_0 - \beta_1 x_i)^2}$$

By taking the logarithm of both sides, this simplifies to

$$\ln L(\beta_0, \beta_1 \mid x_i, y_n) = n \ln \sigma\sqrt{2\pi} - \frac{1}{2\sigma^2} \sum_i^n (y_i - \beta_0 - \beta_1 x_i)^2$$

To maximize the likelihood, the log likelihood must be minimized. In other words, minimize

$$\sum_i^n (y_i - \beta_0 - \beta_1 x_i)^2$$

This is the same optimization criteria for them least squares estimation problem. Therefore the MLEs of β_0 and β_1 will be the same as the least squares estimators.

Question 11.22: Under the assumptions of Exercise 11.20, find the MLE of σ^2.
The log likelihood function, where $\theta = \sigma^2$, is

$$\ln L(\theta) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \theta - \frac{1}{2\theta} \sum_i^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Differentiate this with respect to θ.

$$\frac{d}{d\theta} \ln L(\theta) = -\frac{n}{2\theta} + \frac{1}{2\theta^2} \sum_i^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Equate this to 0 and solve for θ.

$$\frac{d}{d\theta} \ln L(\theta) = -\frac{n}{2\theta} + \frac{1}{2\theta^2} \sum_i^n (y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

$$\frac{n}{2\theta} = \frac{1}{2\theta^2} \sum_i^n (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\hat{\theta} = \frac{1}{n} \sum_i^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Thus the MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Question 11.38: Fit a straight line to the five data points in the accompanying table.

y	3.0	2.0	1.0	1.0	0.5
x	-2.0	-1.0	0.0	1.0	2.0

Give the estimates of β_0 and β_1.

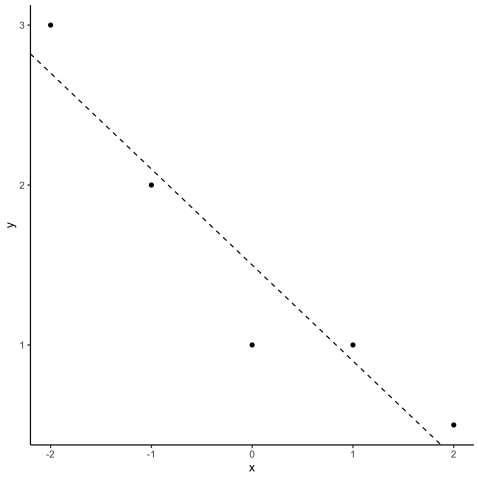
$$b_0 = 1.5 \quad b_1 = -0.6$$

R Code:

```
library(tidyverse)
df = data.frame(x = c(-2,-1,0,1,2),
                y = c(3, 2, 1, 1, 0.5))
b1 = sum((df$x - mean(df$x))*(df$y - mean(df$y))) /
  sum((df$x - mean(df$x))^2)
b0 = mean(df$y) - (b1 * mean(df$x))

ggplot(df, aes(x,y)) + geom_point() +
  geom_abline(intercept = b0, slope = b1,
              linetype = "dashed") +
  theme_classic()
```

Plot the points and sketch the fitted line as a check on the calculations.



Find a 90% confidence interval for E[Y] when $x^* = 0$. Then find 90% confidence intervals for E[Y] when $x^* = -2$ and $x^* = +2$. Compare the lengths of these intervals. Plot these confidence limits on the graph.

To calculate the confidence intervals, first calculate $\bar{x}$, S and S_{xx}.

$$\begin{aligned}
\bar{x} &= \frac{\sum_i^n x_i}{n} = 0 \\
S^2 &= \frac{\text{SSE}}{n-2} = \frac{\sum_i^n (y_i - \bar{y})^2 - \hat{\beta}_1 S_{xy}}{n-2} \\
&= \frac{\sum_i^n (y_i - \bar{y})^2 - \hat{\beta}_1 \sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n-2} \\
&= \frac{4 - (-0.6)(-6)}{3} = 0.13 \\
S_{xx} &= \sum_i^n (x_i - \bar{x})^2 = 10
\end{aligned}$$

The formula for finding the 95% confidence interval for $\text{E}[Y] = \beta_0 + \beta_1 x^*$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

where

$$t_{\alpha/2} = \texttt{qt(0.95, df = 5-2)} = 2.353363$$

at the 90% confidence level. Then the confidence intervals are:

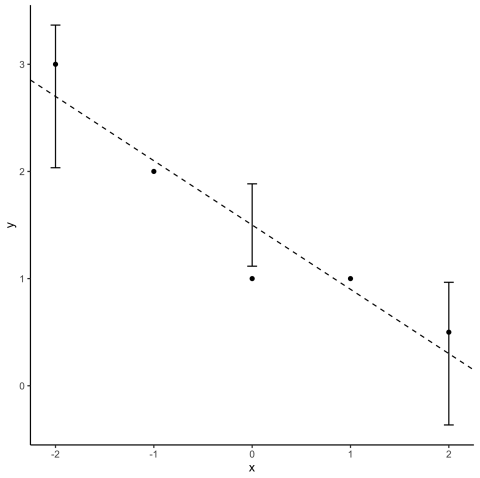
$$\text{when } x^* = 0,\ 1.5 + (-0.6)(0) \pm (2.353363)\sqrt{0.13}\sqrt{\frac{1}{5} + \frac{(0-0)^2}{10}}$$
$$1.5 \pm 0.3842546$$
$$(1.115745, 1.884255)$$

$$\text{when } x^* = -2,\ 1.5 + (-0.6)(-2) \pm (2.353363)\sqrt{0.13}\sqrt{\frac{1}{5} + \frac{(-2-0)^2}{10}}$$
$$2.7 \pm 0.6655485$$
$$(2.034452, 3.365548)$$

$$\text{when } x^* = 2,\ 1.5 + (-0.6)(2) \pm (2.353363)\sqrt{0.13}\sqrt{\frac{1}{5} + \frac{(2-0)^2}{10}}$$
$$0.3 \pm 0.6655485$$
$$(-0.3655485, 0.9655485)$$

The length of the confidence interval at $x^* = 0$ is 0.7685092 while for both $x^* = -2$ and $x^* = 2$, it is 1.3310970.

The confidence intervals are plotted on the graph as shown:



R Code:

```
Xnew = c(0, -2, 2)
CI_df = data.frame(x = Xnew,
        CI_left = (1.5 + (-0.6*Xnew)) -
        (qt(0.95, 5-2)*sqrt(0.1333)*sqrt(0.2 + (Xnew - 0)^2/10)),
        CI_right = (1.5 + (-0.6*Xnew)) +
        (qt(0.95, 5-2)*sqrt(0.1333)*sqrt(0.2 + (Xnew - 0)^2/10)))

left_join(df, CI_df) %>% ggplot(aes(x,y)) + geom_point() +
  geom_abline(intercept = b0, slope = b1,
              linetype = "dashed") +
  geom_errorbar(aes(ymin = CI_left,
                    ymax = CI_right,
                    x = x), width = 0.1) +
  theme_classic()
```

Question 11.44: What did housing prices look like in the "good old days"? The median sale prices for new single-family houses are given in the accompanying table for the years 1972 through 1979.

Year	Median Sales Price (x 1000)
1972 (1)	\$27.6
1973 (2)	\$32.5
1974 (3)	\$35.9
1975 (4)	\$39.3
1976 (5)	\$44.2
1977 (6)	\$48.8
1978 (7)	\$55.7
1979 (8)	\$62.9

Letting Y denote the median sales price and x the year (using integers $1, 2, \ldots, 8$), fit the model $Y = \beta_0 + \beta_1 x + \varepsilon$. Find a 95% prediction interval for the median sale price for the year 1981. Repeat for 1982. Would you feel comfortable in using this model to predict the median sale price for the year 1988?

The coefficient estimates are

$$\beta_0 = 21.575 \quad \beta_1 = 4.841667$$

The formula for finding the prediction interval is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Find $t_{\alpha/2}$, $\bar{x}$, S and S_{xx}.

$$\begin{aligned}
t_{\alpha/2} &= \texttt{qt(0.975, 6)} = 2.446912 \\
\bar{x} &= \frac{\sum_i^n x_i}{n} = 4.5 \\
S^2 &= \frac{\sum_i^n (y_i - \bar{y})^2 - \hat{\beta}_1 \sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n - 2} \\
&= \frac{1002.839 - (4.841667)(203.35)}{10 - 2} = 3.047639 \\
S_{xx} &= \sum_i^n (x_i - \bar{x})^2 = 42
\end{aligned}$$

The prediction interval for the median sale price for the year 1981 (or 10) is

$$\begin{aligned}
21.575 + 4.841667(10) &\pm (2.446912)\sqrt{3.047639}\sqrt{1 + \frac{1}{8} + \frac{(10 - 4.5)^2}{42}} \\
69.99167 &\pm 5.802649 \\
(64.18902, &\, 75.79432)
\end{aligned}$$

The prediction interval for the median sale price for the year 1982 (or 11) is

$$\begin{aligned}
21.575 + 4.841667(11) &\pm (2.446912)\sqrt{3.047639}\sqrt{1 + \frac{1}{8} + \frac{(11 - 4.5)^2}{42}} \\
74.83334 &\pm 6.235725 \\
(68.59761, &\, 81.06906)
\end{aligned}$$

As seen by these two prediction intervals, the range of predictions for the sales prices grows larger as the year goes further out from the dataset. This is known as extrapolation. Now, the year 1988 is much farther away from the realm of this dataset, (a whole 9 years away)! It is not practical to use this model to predict the mean sale price for the year 1988.