MTH 427 - Spring 2023
Assignment #1
Due: Monday, February 20th 2023 (11:59PM)

1 Text Book Problems

- **11.1** If $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the least-squares estimates for the intercept and slope in a simple linear regression model, show that the least-squares equation $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ always goes through the point $(\bar{x}, \bar{y})$. [*Hint:* substitute $\bar{x}$ for x in the least squares equation and use the fact that $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$.]

  Proof. Following the hint, $\widehat{y}(\bar{x}) = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}$. But also by the hint, $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$, so substituting in for $\widehat{\beta}_0$ gives
  $$\widehat{y}(\bar{x}) = \bar{y} - \widehat{\beta}_1 \bar{x} + \widehat{\beta}_1 \bar{x} = \bar{y}$$
  Hence the point $(\bar{x}, \bar{y})$ is a solution to the least-squares equation for any $\widehat{\beta}_0, \widehat{\beta}_1$. □

- **11.5 (use R)** 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.[1] Letting Y denote the median sales price and x the year (using integers $1, 2, \ldots, 8$), fit the model $Y = \beta_0 + \beta_1 x + \varepsilon$. What can you conclude from the results?

Year	Median Sales Price (×1000)
1972 (1)	$27.60
1973 (2)	$32.50
1974 (3)	$35.90
1975 (4)	$39.30
1976 (5)	$44.20
1977 (6)	$48.80
1978 (7)	$55.70
1979 (8)	$62.90

  Solution: The summary shows that our equation is $\widehat{y} = 4841.7x + 21575$. This means that in 1971 (year 0) the expected value of the median sales price was $21575 and increased annually by an average of $4841.70.

```
> #11.5
> x <- c(1, 2, 3, 4, 5, 6, 7, 8)
> y <- c(27.6, 32.5, 35.9, 39.3, 44.2, 48.8, 55.7, 62.9)
> y = y * 1000 # fix cost scaling
> linear_regression_model = lm(y~x)
> summary(linear_regression_model)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-1825.00 -1597.92    16.67  1197.92  2591.67 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  21575.0     1360.3   15.86 3.99e-06 ***
x             4841.7      269.4   17.97 1.91e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1746 on 6 degrees of freedom
Multiple R-squared:  0.9818,	Adjusted R-squared:  0.9787 
F-statistic: 323.1 on 1 and 6 DF,  p-value: 1.908e-06
```

- **11.17 (use R)**

  a Calculate SSE and S^2 for Exercise 1.5.

  Solution: According to the R summary, the residual error is 1.746 so $S^2 = 1746^2 = 3048516$. Then SSE $= (n-2)S^2 = 6 \times 3048516 = 18291096$.

  b It is sometimes convenient, for computational purposes, to have x-values spaces symmetrically and equally about zero. The x-values can be rescaled (or coded) in any convenient manner, with no loss of information in the statistical analysis. Refer to Exercise 1.5. Code the x-values (originally given on a scale of 1 to 8) by using the formula

$$x^* = \frac{x - 4.5}{0.5}.$$

  Then fit the model $Y = \widehat{\beta}_0^* + \widehat{\beta}_1^* x^* + \varepsilon$. Calculate SSE. (Notice that the x^*-values are integers symmetrically spaced about zero.) Compare the SSE with the value obtained in part (a).

  Solution: Using R to compute a new summary,

```
> #11.17b
> x <- c(1, 2, 3, 4, 5, 6, 7, 8)
> y <- c(27.6, 32.5, 35.9, 39.3, 44.2, 48.8, 55.7, 62.9)
> y = y * 1000 # fix cost scaling
> x = (x - 4.5) / 0.5;
> linear_regression_model = lm(y~x)
> summary(linear_regression_model)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-1825.00 -1597.92    16.67  1197.92  2591.67 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  43362.5      617.2   70.25 5.60e-10 ***
x             2420.8      134.7   17.97 1.91e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1746 on 6 degrees of freedom
Multiple R-squared:  0.9818,	Adjusted R-squared:  0.9787 
F-statistic: 323.1 on 1 and 6 DF,  p-value: 1.908e-06
```

Our equation is now $Y = 43362.5 + 2420.8x$. Then the residual error is 1.746 so $S^2 = 1746^2 = 3048516$. Then SSE $= (n-2)S^2 = 6 \times 3048516 = 18291096$. This is the same as (a), which makes sense since translating and dialating the plot along the x-values doesn't change the residual size for each data point, which determines SSE.

- **11.20** Suppose that $Y_1, Y_2, \dots, Y_n$ are independent normal random variables with $\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i$ and $\mathrm{Var}(Y_i) = \sigma^2$, for $i = 1, 2, \dots, n$. Show that the maximum-likelihood estimators (MLEs) of β_0 and β_1 are the same as the least-squares estimators of section 11.3

  Solution:

$$f(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

$$\begin{aligned} L(Y_1, \dots, Y_n \mid \beta_0, \beta_1) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^{n} \cdot \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\} \end{aligned}$$

$$\begin{aligned}
l(Y_1, \ldots, Y_n \mid \beta_0, \beta_1) &= \ln\left[\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}\right] \\
&= \ln\left(\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\right) + \ln\left(\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}\right) \\
&= n\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2
\end{aligned}$$

Then

$$\frac{\partial l}{\partial \beta_0} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \implies -n\beta_0 + \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i = 0$$

So $\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$. Solving for β_0,

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \\
n\beta_0 &= \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i \\
\beta_0 &= \bar{y} - \beta_1 \bar{x}
\end{aligned}$$

Next

$$\frac{\partial l}{\partial \beta_1} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i(y_i - \beta_0 - \beta_1 x_i)\right) = 0$$

So $\sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$

Now solving for β_1,

$$\begin{aligned}
\sum x_i y_i &= \beta_0 \sum x_i + \beta_1 \sum x_i^2 \\
\beta_1 &= \frac{\sum x_i y_i - \beta_0 \sum x_i}{\sum x_i^2} \\
&= \frac{\sum x_i y_i - (\bar{y} - \beta_1 \bar{x}) \sum x_i}{\sum x_i^2} \\
&= \frac{\sum x_i y_i - \bar{y} \sum x_i + \beta_1 \bar{x} \sum x_i}{\sum x_i^2} \\
\beta_1 - \frac{\beta_1 \bar{x} \sum x_i}{\sum x_i^2} &= \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2} \\
\beta_1 \left(\frac{\sum x_i^2 - \bar{x} \sum x_i}{\sum x_i^2} \right) &= \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2} \\
\beta_1 &= \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2 - \bar{x} \sum x_i} \\
&= \frac{\sum x_i y_i - \frac{1}{n} \sum y_i \sum x_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2} \\
&= \frac{S_{xy}}{S_{xx}}
\end{aligned}$$

Which is the same as the least-squares method achieved.

- **11.21** Under the assumptions of Exercise 11.20, find $\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1)$. Use this answer to show that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are independent if $\sum_{i=1}^{n} x_i = 0$. [*Hint:* $\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \text{Cov}(\bar{Y} - \widehat{\beta}_1 \bar{x},\ \widehat{\beta}_1)$. Use Theorem 5.12 and the results of this section.]

Solution:

$$\begin{aligned}
\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) &= \text{Cov}(\bar{Y} - \widehat{\beta}_1 \bar{x}, \widehat{\beta}_1) && \text{(using the hint)} \\
&= \text{Cov}(\bar{Y}, \widehat{\beta}_1) + \text{Cov}(-\widehat{\beta}_1 \bar{x}, \widehat{\beta}_1) && \text{(separate sum)} \\
&= \underbrace{\text{Cov}(\bar{Y}, \widehat{\beta}_1)}_{0} + \text{Cov}(-\widehat{\beta}_1 \bar{x}, \widehat{\beta}_1) && \text{(by Theorem 5.12 and page 579)} \\
&= \text{Cov}(-\widehat{\beta}_1 \bar{x},\ \widehat{\beta}_1) \\
&= \text{Cov}\left(-\widehat{\beta}_1 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i,\ \widehat{\beta}_1 \right) && \text{(definition of } \bar{x}\text{)} \\
&= \text{Cov}\left(-\widehat{\beta}_1 \cdot \frac{1}{n} \cdot 0,\ \widehat{\beta}_1 \right) && \text{(by hypothesis)} \\
&= \text{Cov}(0, \widehat{\beta}_1) \\
&= 0 && \text{(by Covariance of a constant)}
\end{aligned}$$

Therefore $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are independent since their Covariance equals zero.

- **11.30a** In both cases, $H_0 : \widehat{\beta}_1 = 0$ vs $H_a : \widehat{\beta}_1 \neq 0$ with $\alpha = 0.05$.

  Small

  $$S^2 = \frac{\text{SSE}}{n-2} = \frac{2.04}{29} = 0.070345 \implies S = 0.26523$$

  $$\text{Var}(\widehat{\beta}_1) = c_{11}S^2 = (0.0202)^2 = c_{11} \cdot 0.070345 \implies c_{11} = 0.0058$$

  Computing the test statistic,

  $$T = \frac{\widehat{\beta}_1 - 0}{S\sqrt{c_{11}}} = \frac{0.155}{0.26523\sqrt{0.0058}} = 7.674$$

  Using the stat tables, $t_{0.025,\,29} = 2.045$, hence we are in the rejection region and the slope is nonzero.

  Large

  $$S^2 = \frac{\text{SSE}}{n-2} = \frac{1.86}{9} = 0.20667 \implies S = 0.4546$$

  $$\text{Var}(\widehat{\beta}_1) = c_{11}S^2 = (0.0193)^2 = c_{11} \cdot 0.20667 \implies c_{11} = 0.0018$$

  Computing the test statistic,

  $$T = \frac{\widehat{\beta}_1 - 0}{S\sqrt{c_{11}}} = \frac{0.190}{0.4546\sqrt{0.0018}} = 9.851$$

  Using the stat tables, $t_{0.025,\,9} = 2.262$, hence we are in the rejection region and the slope is nonzero.

  Thus both slopes are significantly far different than 0.

2 Additional Exercises Using R

2.1 Exercise 1

This exercise relates to the "Hwk-data2" dataset One study enrolled a group of 10 nurses, ages 50-54 years, who had smoked at least 1 pack per day and quit for at least 6 years. The nurses reported their weight before and 6 years after quitting smoking. A commonly used measure of obesity is BMI = w/h^2 (weight/height^2). The BMI of the 10 women before and 6 years after quitting smoking are given in the last two columns of: "Hwk-data2.csv"

(a) What test can be used to asses whether the mean BMI changed among heavy-smoking women 6 years after quitting smoking? Specify the hypotheses.

Solution: A paired t-test can be used with $H_0 : \mu_d = 0$ vs $H_a : \mu_d \neq 0$.

(b) Implement the test in part(a). (Is there sufficient evidence that the mean BMI changed among heavy-smoking women 6 years after quitting smoking?)

Solution: Yes, there is sufficient evidence that the BMI changed after 6 years (since the p-value is very small).

```r
# 2.1 (b)
library(readr)

df = read.csv("O:/Arr Matey/Hwk-data2.csv", header=T, na.strings="?")
df = na.omit(df)

baseline_before = df$BMI_baseline_never_smoking_women
baseline_after = df$BMI_6year_follow_up_never_smoking_women
smokers_before = df$BMI_baseline_heavy_smoking_women
smokers_after = df$BMI_6years_after_quitting_heavy_smoking_women

t.test(smokers_before, smokers_after, paired=TRUE)
var.test(smokers_before, smokers_after)


```

```
> t.test(smokers_before, smokers_after, paired=TRUE)

        Paired t-test

data:  smokers_before and smokers_after
t = -4.3145, df = 9, p-value = 0.001949
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -5.121709 -1.598291
sample estimates:
mean difference
          -3.36
```

(c) Provide a 98% confidence interval for the true mean change in BMI among heavysmoking women.

Solution: The data is normal, so the 98% confidence interval is $(1.162738, 5.557262)$.

```r
### 2.1 (c) ###
smoker_difference = smokers_after - smokers_before
shapiro.test(smoker_difference)
# W = 0.9138, p-value = 0.3081; hence normal

t.test(smoker_difference, conf.level=0.98)$"conf.int"
# 1.162738 5.557262
```

```
> ### 2.1 (c) ###
> smoker_difference = smokers_after - smokers_before
> shapiro.test(smoker_difference)

        Shapiro-Wilk normality test

data:  smoker_difference
W = 0.9138, p-value = 0.3081

> t.test(smoker_difference, conf.level=0.98)$"conf.int"
[1] 1.162738 5.557262
attr(,"conf.level")
[1] 0.98
```

One issue is that there has been a secular change in weight in society. For this purpose, a control group of 50-to 54 year old never-smoking women were recruited and their BMI was reported at baseline (ages 50-54) and 6 years later at a follow-up visit. The results are given in the first two columns of: "Hwk-data2.csv"

(d) What test can be used to assess whether the mean change in BMI over 6 years is different between women who quit smoking and women who have never smoked? Specify the hypotheses.

Solution: A two sample t-test, pooled (as the following R code shows the variances are equal, since the p-value is large). $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$, where μ_1 is the mean difference of the baseline group, and μ_2 is the mean difference of the smokers.

```
> ### 2.1 (d) ###
> baseline_difference = baseline_after - baseline_before
> var.test(smoker_difference, baseline_difference)

        F test to compare two variances

data:  smoker_difference and baseline_difference
F = 1.1627, num df = 9, denom df = 9, p-value = 0.826
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2888038 4.6811133
sample estimates:
ratio of variances
          1.162722
```

(e) Implement the test in part (d) (Do the data provide sufficient evidence to indicate a difference in mean BMI between the heavy-smoking women 6 years after quitting smoking and the never-smoking women at 6-year follow-up.)

Solution: Since the p value is large (> 0.1 in this case), we fail to reject the null hypothesis, and hence there is insufficient evidence that the smokers and non smokers BMI is difference.

```
> ### 2.1 (e) ###
> t.test(smoker_difference, baseline_difference, var.equal=TRUE)

        Two Sample t-test

data:  smoker_difference and baseline_difference
t = 1.7041, df = 18, p-value = 0.1056
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4214324  4.0414324
sample estimates:
mean of x mean of y
     3.36      1.55
```

(f) Provide a 90% Confidence interval for the difference in mean BMI between the heavysmoking women 6 years after quitting smoking and the never-smoking women at 6-year follow-up.

Solution: A 90% confidence interval is $(-0.03178454, 3.65178454)$.

```
> ### 2.1 (f) ###
> t.test(smoker_difference, baseline_difference, var.equal=TRUE, conf.level=.9)$"conf.int"
[1] -0.03178454  3.65178454
attr(,"conf.level")
[1] 0.9
```

2.2 Exercise 2

This exercise relates to the "Auto" dataset

(a) Use the appropriate function in R to perform a simple linear regression with *mpg* as the response variable and *horsepower* as the predictor.

Solution:

```
> library(readr)
> df = read.csv("O:/Arr Matey/Auto.csv", header=T, na.strings="?")
> df = na.omit(df)
> mpg = df$mpg;
> hp = df$horsepower
> linear_regression_model = lm(mpg~hp)
> # 2.2 (a)
> linear_regression_model = lm(mpg~hp)
> summary(linear_regression_model)

Call:
lm(formula = mpg ~ hp)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
hp          -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

(b) Give an interpretation of the coefficients in term of *mpg* and *horsepower*

Solution: The intercept β_0 means that a car with zero horsepower is expected to get 39.9 mpg. The slope coefficient β_1 means that for every unit increase of horsepower, the car's mpg will drop an average of 0.1578 units.

(c) Test whether there is a linear relationship between the predictor and the response? (i.e test whether the regression coefficient (slope) is zero: $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$)

Solution: Since the R summary says $p < 2 \times 10^{-16}$, we are extremely confident the slope is non-zero.

(d) Use the appropriate function in R to obtain 98% confidence intervals of the coefficient(s).

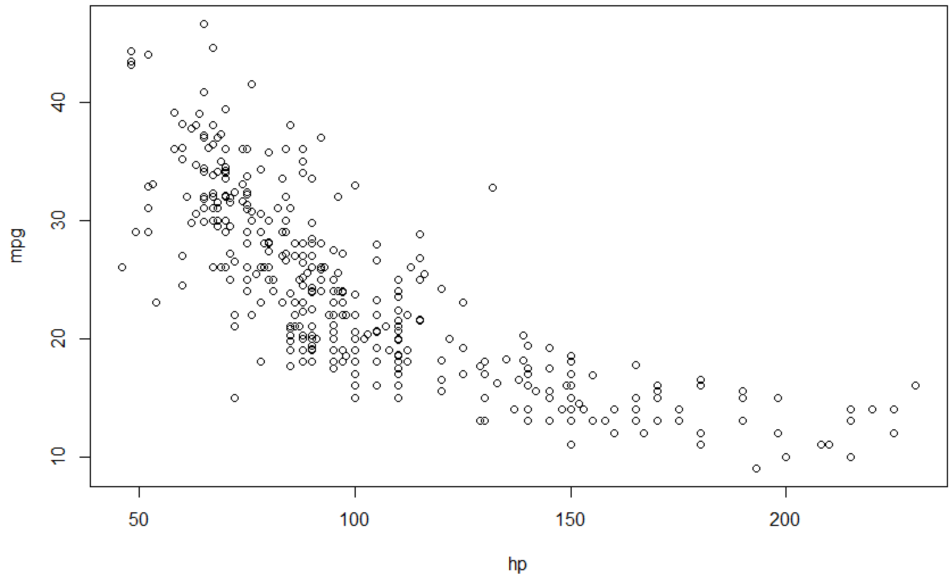
Solution: The intercept $\beta_0 \in (38.2598220, 41.6119001)$ and slope $\beta_1 \in (-0.1729011, -0.1427884)$

```
> # 2.2 (d)
> confint(linear_regression_model, level=0.98)
                   1 %       99 %
(Intercept) 38.2598220 41.6119001
hp          -0.1729011 -0.1427884
> |
```

(e) Display a scatter plot between **mpg** and **horsepower**. Does the scatter plot suggest a linear relationship between the two variables? Explain why?

Solution: The plot suggests some linearity since the data is clustered in a downwards trend, but it looks closer to a graph of $y(x) = 1/x$. The correlation coefficient from the summary is 0.6049, so it has a regular amount of correlation (not strong, not weak).

```
> # 2.2 (e)
> plot(hp, mpg)
```



(f) Display the least square regression line in the scatter plot in (a).

Solution:

```
> # 2.2 (f)
> abline(linear_regression_model)
```