B27
Hector G. Flores Rodriguez

Stats 110 - HW 3
October 24, 2017

1. Create new data set called *NewPrices* by removing the row with the outlier

```
textprices = read.csv("../data/TextPrices.csv")
NewPrices = textprices[-4,]
```

1-(a). Use summary command in R to provide summaries of the two variables for the original data set and the new data set.

```
# With outlier
summary(textprices)
```

```
##      Pages            Price
##  Min.   :  51.0   Min.   :  4.25
##  1st Qu.: 212.0   1st Qu.: 17.59
##  Median : 456.5   Median : 55.12
##  Mean   : 464.5   Mean   : 65.02
##  3rd Qu.: 672.0   3rd Qu.: 95.75
##  Max.   :1060.0   Max.   :169.75
```

```
# Without outlier
summary(NewPrices)
```

```
##      Pages            Price
##  Min.   :  51.0   Min.   :  4.25
##  1st Qu.: 200.0   1st Qu.: 16.95
##  Median : 488.0   Median : 51.50
##  Mean   : 466.8   Mean   : 62.83
##  3rd Qu.: 696.0   3rd Qu.: 95.00
##  Max.   :1060.0   Max.   :169.75
```

1-(b).

The mean number of pages for the data without outlier is: **466.8**

The predicted price for the mean number of pages would be: **$62.83**

2. Find the R^2 using the TextPrices data and again using the NewPrices data.

```
# R-squared value for textprices
tp_fit = lm("Price ~ Pages", data=textprices)
summary(tp_fit)
```

```
##
## Call:
## lm(formula = "Price ~ Pages", data = textprices)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -65.475 -12.324  -0.584  15.304  72.991
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.42231   10.46374  -0.327    0.746
## Pages        0.14733    0.01925   7.653 2.45e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.76 on 28 degrees of freedom
## Multiple R-squared:  0.6766, Adjusted R-squared:  0.665
## F-statistic: 58.57 on 1 and 28 DF,  p-value: 2.452e-08
```

```
# R-squared value for NewPrices
np_fit = lm("Price ~ Pages", data=NewPrices)
summary(np_fit)
```

```
##
## Call:
## lm(formula = "Price ~ Pages", data = NewPrices)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63.23 -12.29   2.16  13.25  53.46
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.8926     9.4775  -0.727    0.473
## Pages         0.1494     0.0173   8.634    3e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.72 on 27 degrees of freedom
## Multiple R-squared:  0.7341, Adjusted R-squared:  0.7243
## F-statistic: 74.55 on 1 and 27 DF,  p-value: 3.001e-09
```

2-(a).

$R^2_{TextPrices} = 0.6766$

$R^2_{NewPrices} = 0.7341$

73% of the variation in book prices (in NewPrice data set) is explained by the number of pages.

2-(b).

The NewPrice data set does a better job in prediction since it captures more of the variance than the original data set.

3. Exercise 2.15 with NewPrices data

3-(a).

```
mu_price = data.frame(Pages=450)
predict(np_fit, mu_price, interval="confidence")
```

```
##        fit      lwr      upr
## 1 60.32434 50.12599 70.52269
```

A 95% C.I. is [50.126, 70.523]

3-(b).

```
predict(np_fit, mu_price, interval="predict")
```

```
##        fit      lwr      upr
## 1 60.32434 4.557604 116.0911
```

A 95% P.I. is [4.558, 116.091]

3-(c).

```
mean(c(4.558, 116.091))
```

```
## [1] 60.3245
```

```
mean(c(50.126, 70.523))
```

```
## [1] 60.3245
```

The midpoint of the two intervals are the same since they both should be the value of $\hat{y}$ given by OLS.

3-(d).

The P.I. width is much larger than the C.I. This makes sense since the prediction interval takes into account uncertainty for individual observations rather than the mean value

3-(e).

The mean number of pages: 466.8

3-(f).

```
pages1500 = data.frame(Pages=1500)
predict(np_fit, pages1500, interval="predict")
```

```
##        fit      lwr      upr
## 1 217.1638 150.4203 283.9074
```

A 95% P.I. for a 1500-page textbook is [150.420, 283.907]. We don't have a 95% C.I since we are predicting outside the range of the data we have available.

4. Exercise 2.10 pg 82

4-(a). Decrease the width of the prediction interval

4-(b). Increase the width of the P.I.

4-(c). Increase the width of the P.I.

4-(d). Increase the width of the P.I.

5. Exercise 2.17, (b) and (c)

5-(b).

```
sparrows = read.csv("../data/Sparrows.csv")
sparrow_fit = lm("Weight ~ WingLength", data=sparrows)
summary(sparrow_fit)
```

```
##
## Call:
## lm(formula = "Weight ~ WingLength", data = sparrows)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5440 -0.9935  0.0809  1.0559  3.4168
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.36549    0.95731   1.426    0.156
## WingLength   0.46740    0.03472  13.463   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.4 on 114 degrees of freedom
## Multiple R-squared:  0.6139, Adjusted R-squared:  0.6105
## F-statistic: 181.3 on 1 and 114 DF,  p-value: < 2.2e-16
```

61.39% of the variation in weight is explained by WingLength

5-(c).

Provide the ANOVA table that partitions the total variablility in weight and interpret the F-test.

```
anova(sparrow_fit)
```

```
## Analysis of Variance Table
##
## Response: Weight
##             Df Sum Sq Mean Sq F value    Pr(>F)
## WingLength   1 355.05  355.05  181.25 < 2.2e-16 ***
## Residuals  114 223.31    1.96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F statistic is the same as the t^2 for $\beta_1 = 0$ vs $\beta_1 \neq 0$. Hence, there is significant evidence that there is a linear trend

6. Exercise 2.44

6-(a).

$\hat{\beta}_1 = r\frac{s_x}{s_y} = (0.701)(\frac{104807}{657}) = 111.826$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 247235 - (111.826)(2009) = 22576.57$

$\therefore\ \hat{y} = 22576.57 + 111.826x$

6-(b).

$r^2 = 0.491401$

49.1% of the variation in gate counts is explained by enrollements.

6-(c).

$\hat{y} = 22576.57 + 111.826(1445) = 184165.1$

6-(d).

$e = 130000 - (22576.57 + 111.826(2200)) = -138593.8$

7. Skin Cancer example shown in class on Oc 16

7-(a).

$\hat{\beta}_1 = -5.9776$

$SE\{\hat{\beta}_1\} = 0.5984$

$MSE = RSE^2 = 365.5744$

$SSX = \frac{MSE}{SE\{\hat{\beta}_1\}^2} = 610.9198$

$\hat{y} = 389.894 - (5.9776)(33.7) = 188.44888$

7-(b).

$(t_{0.025,47})\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(33.7-39.52)^2}{SSX}} = (2.011741)(5.272538) = 10.60698$

$\implies \hat{\beta}_1 \pm 10.60698$

$\therefore [-16.58458, 4.62938]$