Homework #2

Eric Pettengill

3.7.10 This problem uses the `Carseats` data set in the ISLR package.

(a) Fit a multiple regression model to predict `Sales` using `Price`, `Urban`, and `US`.

```r
library(ISLR)
sales.fit <- lm(Sales ~ Price + Urban + US, data = Carseats)
```

(b) Provide an interpretation of each coefficient in the model.

```r
sales.fit$coefficients
```

```
## (Intercept)       Price    UrbanYes       USYes
## 13.04346894 -0.05445885 -0.02191615  1.20057270
```

For every dollar increase in price of the carseat, sales with go down by 0.0545 dollars. Urban = YES corresponds to 0.0219 dollars of sales lower than Urban = NO. US = YES indicates US made carseats have a Sales value of 1.2 dollars higher than non-US made carseats.

(c) Write out the model in equation form.

$$Sales = 13.04 - 0.054 \cdot Price - 0.0219 \cdot Urban + 1.201 \cdot US$$

$$Urban = \begin{cases} 1 & yes \\ 0 & no \end{cases}$$

$$US = \begin{cases} 1 & yes \\ 0 & no \end{cases}$$

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

```r
summary(sales.fit)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

Since the F-stat for the model is significant with p-value: $< 2.2e\text{-}16$ we can see that Price(p-value $< 2e\text{-}16$) and US(p-value $= 4.86e\text{-}06$) both reject the null hypothesis $H_0 : \beta_j = 0$, that is, they are significant predictors of Sales for this model.

(e) Fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

```r
sales.fit.update <- lm(Sales ~ Price + US, data = Carseats)
summary(sales.fit.update)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

Price($p < 2e\text{-}16$) and US($p = 4.71e\text{-}06$) are again significant.

(f) How well do the models in (a) and (e) fit the data?

We can see below that testing $H_0 : \beta_{urban} = 0$ fails to reject with p = 0.9357. Also, the model with only PRICE and US has a higher adjusted R-squared than the model with PRICE, URBAN, and US, as well as a slightly smaller RSE. So the simpler model provides a better fit.

```r
anova(sales.fit, sales.fit.update)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price + Urban + US
## Model 2: Sales ~ Price + US
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    396 2420.8
## 2    397 2420.9 -1  -0.03979 0.0065 0.9357
```

(g) Using the model from (e), obtain 95% confidence intervals for the coefficients.

```r
confint(sales.fit.update)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

3.7.13 Make sure to `set.seed(1)`.

(a) Using the `rnorm` function, create a vector, `x`, containing 100 observations drawn from a $N(0, 1)$.

(b) Using the `rnorm` function, create a vector, `eps`, containing 100 observations drawn from a $N(0, 0.25)$

```
set.seed(1)
x <- rnorm(100, mean = 0, sd = 1)
eps <- rnorm(100, mean = 0, sd = .5)
```

(c) Using `y` and `eps`, generate a vector Y according to the model $Y = -1 + 0.5X + \epsilon$. What is the length of vector Y? What are the values of β_0 and β_1?

```
y <- -1 + 0.5*x + eps

length(y)
```

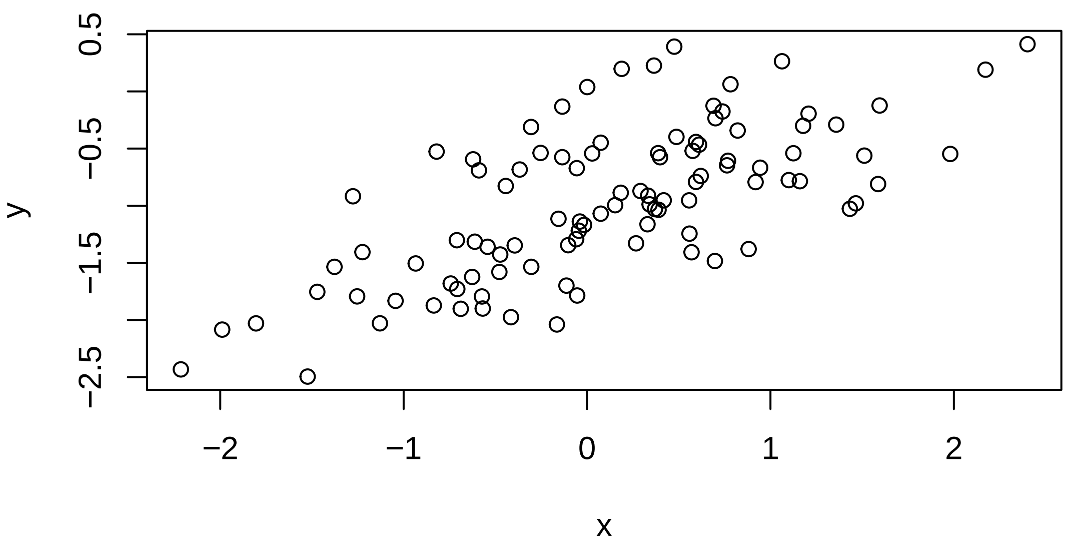
```
## [1] 100
```

$$\beta_0 = -1$$

$$\beta_1 = 0.5$$

(d) Create a scatterplot displaying the relationship between `x` and `y`. Comment.

```
plot(x,y)
```



We can see that `x` and `y` are linear in their relationship.

(e) Fit a least squares linear model to predict y using x. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1?

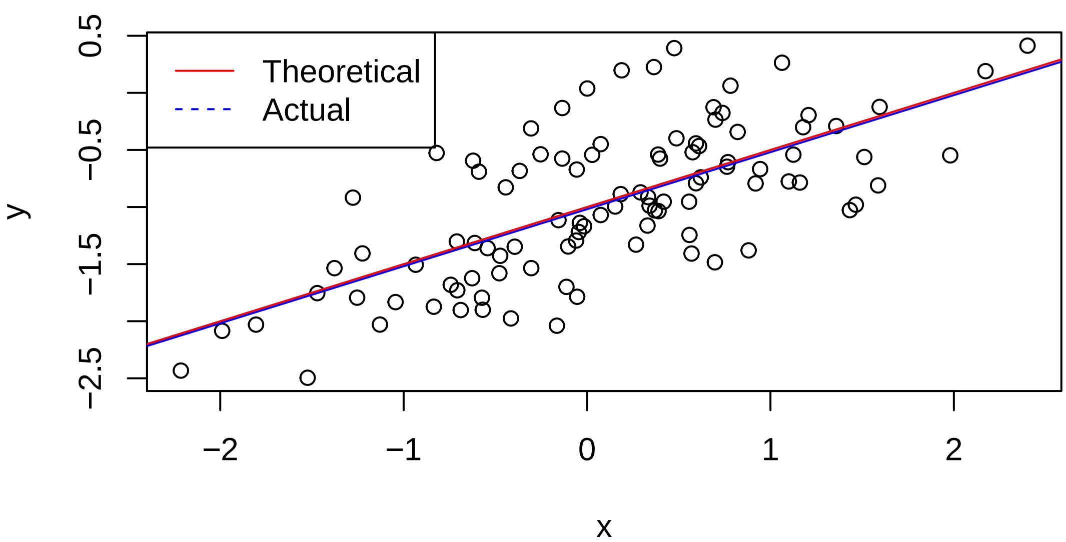
As we can see by the summary printed below $\hat{\beta}_0 \approx \beta_0 = -1$ and $\hat{\beta}_1 \approx \beta_1 = 0.5$.

```
mod <- lm(y ~ x)
summary(mod)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93842 -0.30688 -0.06975  0.26970  1.17309
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
## x            0.49947    0.05386   9.273 4.58e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4814 on 98 degrees of freedom
## Multiple R-squared:  0.4674, Adjusted R-squared:  0.4619
## F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

(f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Create a legend.

```
plot(x,y)
abline(a = -1, b = 0.5, col = "red")
abline(mod, col = "blue")
legend("topleft", legend = c("Theoretical", "Actual"), col = c("red", "blue"), lty = 1:2)
```



(g) Fit a polynomial regression model that predicts y using x and x^2. Is there evidence that the quadratic term improves the model fit? Explain.

```r
mod.quad <- lm(y ~ x + I(x^2))
summary(mod.quad)
```

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98252 -0.31270 -0.06441  0.29014  1.13500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.97164    0.05883 -16.517  < 2e-16 ***
## x            0.50858    0.05399   9.420  2.4e-15 ***
## I(x^2)      -0.05946    0.04238  -1.403    0.164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.479 on 97 degrees of freedom
## Multiple R-squared:  0.4779, Adjusted R-squared:  0.4672
## F-statistic:  44.4 on 2 and 97 DF,  p-value: 2.038e-14
```

```r
anova(mod, mod.quad)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ x + I(x^2)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     98 22.709
## 2     97 22.257  1   0.45163 1.9682 0.1638
```

No, there is not evidence that the quadratic term improves the model. Since p = 0.1638 for $H_0 : \beta_{x^2} = 0$, we fail to reject H_0.