Homework 6

Miles Tweed

3/21/2021

Problem 1

```
crime <- read.csv("../../Data/fl_crime.csv")
names(crime) <- c('country','crime','education','urban','income')
head(crime)
```

```
##    country crime education urban income
## 1  Alachua   104      82.7  73.2   22.1
## 2    Baker    20      64.1  21.5   25.8
## 3      Bay    64      74.7  85.0   24.7
## 4 Bradford    50      65.0  23.2   24.6
## 5  Brevard    64      82.3  91.9   30.5
## 6  Broward    94      76.8  98.9   30.6
```

Part 1

$$\text{crime}_i = \beta_{education} \cdot \text{education}_i + \beta_{urban} \cdot \text{urban}_i + \beta_{income} \cdot \text{income}_i + \epsilon_i,\ \epsilon_i \sim_{i.i.d.} \mathbb{N}(0, \sigma^2)$$

Part 2

i

$$\begin{aligned} E[Y_i] &= E[\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \epsilon_i] && (1)\\ &= E[\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i}] + E[\epsilon_i] && (2)\\ &= E[\text{Constant}] + E[\epsilon_i \sim \mathbb{N}(0, \sigma^2)] && (3)\\ &= \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + 0 && (4) \end{aligned}$$

ii

$$\begin{aligned} V[Y_i] &= V[\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \epsilon_i] && (5)\\ &= V[\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i}] + V[\epsilon_i] && (6)\\ &= V[\text{Constant}] + V[\epsilon_i \sim \mathbb{N}(0, \sigma^2)] && (7)\\ &= 0 + \sigma^2 && (8) \end{aligned}$$

iii The expected value of Y_i is constant and given by $\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i}$ so because the variance of Y_i is defined solely by the irreducible error term ϵ_i and we make the assumption that $\epsilon_i \sim \mathbb{N}(0, \sigma^2)$ it stands to reason that $Y_i \sim \mathbb{N}(\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i},\ \sigma^2)$

Part 3

```r
lm.obj <- lm(crime~education + urban + income, data = crime)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = crime ~ education + urban + income, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.407 -15.080  -6.588  16.178  50.125
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.7147    28.5895   2.089   0.0408 *
## education    -0.4673     0.5544  -0.843   0.4025
## urban         0.6972     0.1291   5.399 1.08e-06 ***
## income       -0.3831     0.9405  -0.407   0.6852
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.95 on 63 degrees of freedom
## Multiple R-squared:  0.4728, Adjusted R-squared:  0.4477
## F-statistic: 18.83 on 3 and 63 DF,  p-value: 7.823e-09
```

Fitted Equation

$$\hat{\text{crime}} = 59.7 - 0.467 \cdot \text{education} + 0.697 \cdot \text{urban} - 0.383 \cdot \text{income}$$

Part 4

For education

$H_0 : \beta_{education} = 0$ $H_a : \beta_{education} \neq 0$. Since the p-value 0.403 is greater than the significance level of 0.05 we would fail to reject the null hypothesis.

For urban

$H_0 : \beta_{urban} = 0$ $H_a : \beta_{urban} \neq 0$. Since the p-value 1.08e-6 is much less than the significance level of 0.05 we would reject the null hypothesis in favor of the alternative which suggests that there may be a linear relationship between urbanization and crime holding all other variables constant.

For income

$H_0 : \beta_{income} = 0$ $H_a : \beta_{income} \neq 0$. Since the p-value 0.685 is greater than the significance level of 0.05 we would fail to reject the null hypothesis.

Part 5

For counties that have the same percentage of residents aged at least 25 in the county with at least a high school diploma as well as the same median income, an increase of one percent of the population who lives in urban areas results in an increase of 0.697 crimes in the past year per 1000 residents on average across all such counties were the percentage of residents living in urban areas differs by one percent.

Part 6

$$H_0 : \beta_{education} = \beta_{urban} = \beta_{income} = 0 \quad H_a : \{\text{at least one } \beta_j \neq 0 \;\; j = education, urban, income\}$$

The F-statistic of 18.83 with 63 degrees of freedom led to a p-value of 7.823e-09. This small p-value indicates statistical significance so we would reject the null hypothesis in favor of the alternative, that at least one of the explanatory variables has a linear relationship to the response variable.

Problem 2

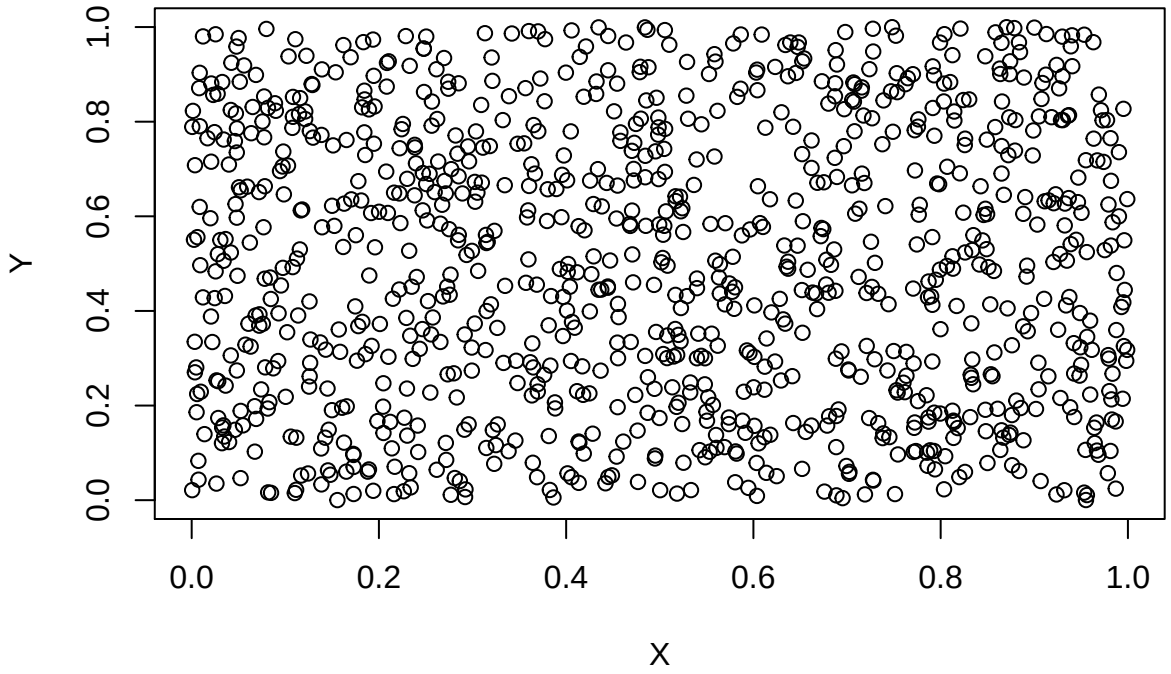
Part 1

I generated the X and Y data from a uniform distribution since this would lead to the most evenly spread out data. This led to a very low correlation and no visible trend in the scatter plot.

```
Y <- runif(1000)
X <- runif(1000)
cor(X,Y)
```

```
## [1] -0.00311659
```

```
plot(X,Y)
```



Part 2

```
Y <- runif(200)
X <- matrix(nrow=200,ncol=50)

for (i in 1:50){
  temp <- runif(200)
  X[,i] <- temp
}

df <- data.frame(X,Y)
```

Part 3

```
lm.obj <- lm(Y~., df)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = Y ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5366 -0.1744  0.0043  0.1588  0.4857
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.165026   0.299370   0.551  0.58229
## X1           0.101279   0.076493   1.324  0.18752
## X2          -0.012807   0.078922  -0.162  0.87131
## X3          -0.135769   0.077574  -1.750  0.08215 .
## X4           0.082875   0.074773   1.108  0.26950
## X5          -0.073506   0.080712  -0.911  0.36392
## X6           0.165508   0.084373   1.962  0.05167 .
## X7           0.026532   0.082802   0.320  0.74909
## X8           0.038947   0.073174   0.532  0.59535
## X9           0.105924   0.079978   1.324  0.18740
## X10         -0.015659   0.074941  -0.209  0.83477
## X11          0.042536   0.075263   0.565  0.57281
## X12          0.118507   0.078723   1.505  0.13435
## X13          0.080990   0.075637   1.071  0.28600
## X14         -0.063652   0.085361  -0.746  0.45703
## X15          0.030273   0.074516   0.406  0.68514
## X16          0.043088   0.081864   0.526  0.59943
## X17          0.047504   0.078930   0.602  0.54818
## X18          0.006966   0.084294   0.083  0.93425
## X19         -0.040058   0.077734  -0.515  0.60709
## X20         -0.113559   0.079763  -1.424  0.15662
## X21          0.016631   0.082181   0.202  0.83990
## X22         -0.020660   0.072689  -0.284  0.77663
## X23          0.073995   0.077602   0.954  0.34187
## X24         -0.045032   0.081683  -0.551  0.58225
## X25         -0.096981   0.074877  -1.295  0.19725
## X26          0.113153   0.079737   1.419  0.15797
## X27          0.085851   0.072598   1.183  0.23887
## X28          0.016035   0.073203   0.219  0.82691
## X29          0.002670   0.077084   0.035  0.97242
## X30         -0.024897   0.071435  -0.349  0.72794
## X31          0.026383   0.075835   0.348  0.72840
## X32         -0.103990   0.076400  -1.361  0.17553
## X33          0.082077   0.074173   1.107  0.27027
## X34         -0.029785   0.085051  -0.350  0.72669
## X35         -0.023849   0.078408  -0.304  0.76143
## X36         -0.065836   0.078777  -0.836  0.40465
## X37          0.069814   0.078044   0.895  0.37248
## X38         -0.013065   0.078464  -0.167  0.86798
## X39          0.015579   0.076618   0.203  0.83915
## X40          0.159143   0.075319   2.113  0.03627 *
## X41          0.011684   0.076705   0.152  0.87913
## X42         -0.178962   0.074165  -2.413  0.01704 *
## X43         -0.041464   0.076245  -0.544  0.58737
```

```
## X44          0.085577   0.074067   1.155  0.24978
## X45         -0.120031   0.080120  -1.498  0.13621
## X46          0.011993   0.080083   0.150  0.88116
## X47          0.218469   0.079858   2.736  0.00698 **
## X48          0.026158   0.077959   0.336  0.73769
## X49          0.092254   0.075875   1.216  0.22596
## X50         -0.036606   0.077202  -0.474  0.63608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2747 on 149 degrees of freedom
## Multiple R-squared:  0.2591, Adjusted R-squared:  0.01047
## F-statistic: 1.042 on 50 and 149 DF,  p-value: 0.4141
```

(a) There were 5 predictors that were statistically significant at the $\alpha = 0.05$ level. The choice to reject the null hypothesis for these predictors would be a Type I error because these would be false positive where an effect is seen and no linear relationship truly exists. We know that to be the case because the data is all generated using a uniform distribution where the data points will be evenly distributed over the range. Although it is not the case here, if we failed to reject H_0 when a linear relationship truly exists we would be committing a Type II error.

(b) The appropriate procedure to test if at least one of the explanatory variables has a linear relationship with the response variable is analysis of variance for regression (ANOVA). The F statistic is a measure that informs us of the strength of the relationship and is used to generate a p-value based on its distribution. For this model the F statistic is 1.141 at 149 degrees of freedom which corresponds to a p-value of 0.2703 which is not statistically significant indicating that it is unlikely that one of the explanatory variables has a strong relationship with the response variable. In other words, it is unlikely that the null hypothesis that $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{50} = 0$ is false. A higher F-statistic or smaller p-value would indicate more evidence against this null hypothesis.

Problem 3

```
library(ISLR)
fix(Auto)
lm.obj <- lm(mpg~.-name, data = Auto)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
```

```
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

Part 1

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

or...

$$H_0 : \beta_{cylinders} = \beta_{displacement} = \beta_{horsepower} = \beta_{weight} = \beta_{acceleration} = \beta_{year} = \beta_{origin} = 0$$
$$H_a : \{\text{at least one } \beta_j \neq 0\},\ j = 1, 2, 3, 4, 5, 6, 7$$

The F-Statistic corresponds to this test. Yes it is significant because the p-value is $<$ 2.2e-16. Therefore we reject the null hypothesis in favor of at least one of the coefficients being not equal to zero.

Part 2

The significant predictors are displacement, weight, year, and origin.

Part 3

The two most significant predictors are year and weight with the coefficients 0.751 and -0.006 respectively.

For year

For vehicles with the same number of cylinders, displacement, horsepower, weight, acceleration, and country of origin, a vehicle that is one year older will have a 0.751 miles per gallon increase on average over all such vehicles that difer by one year in age.

For weight

For vehicles of the same model year andd with the same number of cylinders, displacement, horsepower, acceleration, and country of origin, a vehicle that is one pound heavier will have a 0.006 mile per gallon decrease on average over all such vehicles that difer in weight by one pound.

Part 4

```
confint(lm.obj)
```

```
##                     2.5 %       97.5 %
## (Intercept)  -26.349864469 -8.087004775
## cylinders     -1.129001385  0.142248747
## displacement   0.005119788  0.034671499
## horsepower    -0.044058392  0.010156103
## weight        -0.007756074 -0.005192013
## acceleration  -0.113769257  0.274920933
## year           0.650551315  0.850994041
## origin         0.879280169  1.973000822
```

For weight

We can be 95% confident that an increase in weight of one pound will result in a decrease of between 0.005 and 0.008 miles per gallon on average across all vehicles of the same model year that differ in weight by one pound and have the same displacement, horsepower, acceleration, number of cylinders, and country of origin.

For year

We can be 95% confident that an increase of one year in age will result in an increase of between 0.65 and 0.85 miles per gallon on average across all vehicles that have the same weight, displacement, horsepower, acceleration, number of cylinders, and country of origin but differ in age by one year.

Part 5

The residual standard error (RSE) for this model is 3.328 which indicates that the model misses the true value of the response by 3.328 miles per gallon on average across all datapoints in the sample.

The R^2 value is 0.82 which indicates that this linear model can explain about 82% of the variability of the response data around the mean.