HW4

Yu Fung David Wang

HW4: 5.1

Set Up Packages and Data

```
library(outliers)
crime_data <- read.table("uscrime.txt")
head(crime_data)
```

```
##     V1 V2   V3   V4   V5    V6    V7  V8   V9   V10 V11    V12  V13      V14
## 1    M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq     Prob
## 2 15.1  1  9.1  5.8  5.6  0.51    95  33 30.1 0.108 4.1   3940 26.1 0.084602
## 3 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4 0.029599
## 4 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180   25 0.083401
## 5 13.6  0 12.1 14.9 14.1 0.577  99.4 157    8 0.102 3.9   6730 16.7 0.015801
## 6 14.1  0 12.1 10.9 10.1 0.591  98.5  18    3 0.091   2   5780 17.4 0.041399
##       V15   V16
## 1    Time Crime
## 2 26.2011   791
## 3 25.2999  1635
## 4 24.3006   578
## 5 29.9012  1969
## 6 21.2998  1234
```

```
nrow(crime_data)
```

```
## [1] 48
```

Run the Grubbs Test on Data

```
# We only care about the last column (# of crimes per 100,000 people)
crimesperppl <- as.numeric(crime_data$V16[2:nrow(crime_data)])

# Check for (highest value outlier
hivalue_outlier_test <- grubbs.test(crimesperppl)
hivalue_outlier_test
```

```
##
##  Grubbs test for one outlier
##
## data:  crimesperppl
## G = 2.81287, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier
```

```
# Check for lowest value outlier
lowvalue_outlier_test <- grubbs.test(crimesperppl, opposite = TRUE)
lowvalue_outlier_test
```

```
##
```

```
##  Grubbs test for one outlier
##
## data:  crimesperppl
## G = 1.45589, U = 0.95292, p-value = 1
## alternative hypothesis: lowest value 342 is an outlier
```

Based on the Grubbs Outlier Test, we see that both p-values of the highest (p-val = 0.07887) and lowest (p-val = 1) values are both greater than 0.05 (at the 5% significance level), so we fail to reject the grubbs test for one outlier hypothesis that the highest and lowest values are not outliers. Thus, the null hypothesis holds and the alternative hypothesis do not hold. Hence, the highest and lowest values of 1993 and 342 are not outliers in the number of crimes per 100,000 people.

HW4 8.2

Set Up Data

```
crime2_data <- read.table("uscrime2.txt", header = TRUE)
head(crime2_data)
```

```
##      M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq     Prob
## 1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1 0.084602
## 2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4 0.029599
## 3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0 0.083401
## 4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7 0.015801
## 5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4 0.041399
## 6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6 0.034201
##      Time Crime
## 1 26.2011   791
## 2 25.2999  1635
## 3 24.3006   578
## 4 29.9012  1969
## 5 21.2998  1234
## 6 20.9995   682
```

```
nrow(crime2_data)
```

```
## [1] 47
```

Fit Linear Model - We want to find "Crime" in the data of unpredicted points hence, we want the formula to be the a linear combination of the other independent variables - Will check a summary of the fitted model to see the factors used, coefficients of the model, and quality of fit (adjusted R-squared value)

```
crime_lm <- lm(formula =  Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 + Wealth + In
                  data = crime2_data)

summary(crime_lm) # factors and coefficient of the model
```

```
##
## Call:
## lm(formula = Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop +
##     NW + U1 + U2 + Wealth + Ineq + Prob + Time, data = crime2_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -395.74  -98.09   -6.69  112.99  512.67
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -5.984e+03  1.628e+03  -3.675 0.000893 ***
## M            8.783e+01  4.171e+01   2.106 0.043443 *
## So          -3.803e+00  1.488e+02  -0.026 0.979765
## Ed           1.883e+02  6.209e+01   3.033 0.004861 **
## Po1          1.928e+02  1.061e+02   1.817 0.078892 .
## Po2         -1.094e+02  1.175e+02  -0.931 0.358830
## LF          -6.638e+02  1.470e+03  -0.452 0.654654
## M.F          1.741e+01  2.035e+01   0.855 0.398995
## Pop         -7.330e-01  1.290e+00  -0.568 0.573845
## NW           4.204e+00  6.481e+00   0.649 0.521279
## U1          -5.827e+03  4.210e+03  -1.384 0.176238
## U2           1.678e+02  8.234e+01   2.038 0.050161 .
## Wealth       9.617e-02  1.037e-01   0.928 0.360754
## Ineq         7.067e+01  2.272e+01   3.111 0.003983 **
## Prob        -4.855e+03  2.272e+03  -2.137 0.040627 *
## Time        -3.479e+00  7.165e+00  -0.486 0.630708
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209.1 on 31 degrees of freedom
## Multiple R-squared:  0.8031, Adjusted R-squared:  0.7078
## F-statistic: 8.429 on 15 and 31 DF,  p-value: 3.539e-07
```

We can see the given intercepts for each variable, now with the given model, we can use the given data to predict crime rate in the new city

```
# Predict crime rate of the new city
new_crimedata <- data.frame(M = 14.0,
So = 0,
Ed = 10.0,
Po1 = 12.0,
Po2 = 15.5,
LF = 0.640,
M.F = 94.0,
Pop = 150,
NW = 1.1,
U1 = 0.120,
U2 = 3.6,
Wealth = 3200,
Ineq = 20.1,
Prob = 0.04,
Time = 39.0)

predict(object = crime_lm, newdata = new_crimedata)
```

```
##        1
## 155.4349
```

Based on the linear regression model, the new data will have a crime rate of 155.4349 crimes per 100,000 people.

Finally, here is a software output of the linear regression graph based on fitted values by the model and the true values

```
plot(x = crime2_data$Crime,
     y = crime_lm$fitted.values,
     xlab = "True Vals",
```

```
     ylab = "Model Fitted Vals",
     main = "Crime Rate Based on Regression")

# also add the linear regression line
abline(b = 1, a = 0)

# add the R^2 value (goodness of fit too)
legend("topleft", legend = paste("R2:", format(summary(crime_lm)$adj.r.squared,digits=3)))
```