ISYE 6501 HW 5

Question 8.2

MODEL 1

> setwd("~/Desktop/ISYE 6501/ISYE HW 5/uscrime")

> library(tidyverse)

crime=read_tsv("uscrime.txt")

head(crime)

```
# A tibble: 6 × 16
      M    So    Ed   Po1   Po2    LF   M.F   Pop    NW    U1    U2 Wealth  Ineq
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>
1  15.1     1   9.1   5.8   5.6 0.51   95      33  30.1 0.108   4.1   3940  26.1
2  14.3     0  11.3  10.3   9.5 0.583 101.     13  10.2 0.096   3.6   5570  19.4
3  14.2     1   8.9   4.5   4.4 0.533  96.9    18  21.9 0.094   3.3   3180  25
4  13.6     0  12.1  14.9  14.1 0.577  99.4   157   8   0.102   3.9   6730  16.7
5  14.1     0  12.1  10.9  10.1 0.591  98.5    18   3   0.091   2     5780  17.4
6  12.1     0  11    11.8  11.5 0.547  96.4    25   4.4 0.084   2.9   6890  12.6
```

set.seed(0)

model_1_crime <- lm(Crime~.,data=crime)

summary(model_1_crime)

```
Call:
lm(formula = Crime ~ ., data = crime)

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
```

predict(model_1_crime)

155.4349

range(crime$Crime)

342 1993

data.frame(summary(model_1_crime)$coef[summary(model_1_crime)$coef[,4] <= .05, 4])

(Intercept)	0.0008929887
M	0.0434433942
Ed	0.0048614327
Ineq	0.0039831365
Prob	0.0406269260

A regression model is used estimate the relationship between independent and dependent variables, it is helpful in determining if the correlation can be used to predict data.

I set my session up with the crime data and used the package tidyverse, and read_tsv to visualize the data. I set the seed to 0, then used lm which is a function in R that fits linear models and carries out regression. Then I plugged in the values and used the predict function to figure out the predicted crime rate for my model 1.

Using predict, our regression model shows that 155 is the predicted crime rate from our data, and when using the range function, the lowest value from the data comes out to be 342. This predicted crime rate is below the minimum value of crime rate from our data set, most likely due to overfitting.

The lecture stated that if a p value of an attribute > 0.05, then it isn't significant to the data and therefore it doesn't need to be used in the model.

Below, I created a separate model which removes values > 0.05 which will yield a more accurate result. The values that have asterisk are the ones below 0.05, while the rest are above and will be removed in model 2.

MODEL 2

```
set.seed(0)

model_2_crime<-lm(Crime~M+Ed+Ineq+Prob,data=crime,x=TRUE,y=TRUE)

summary(model_2_crime)
```

```
Call:
lm(formula = Crime ~ M + Ed + Ineq + Prob, data = crime, x = TRUE,
    y = TRUE)

Residuals:
    Min      1Q  Median      3Q     Max
-532.97 -254.03  -55.72  137.80  960.21

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1339.35    1247.01  -1.074  0.28893
M              35.97      53.39   0.674  0.50417
Ed            148.61      71.92   2.066  0.04499 *
Ineq           26.87      22.77   1.180  0.24458
Prob        -7331.92    2560.27  -2.864  0.00651 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 347.5 on 42 degrees of freedom
Multiple R-squared:  0.2629,    Adjusted R-squared:  0.1927
F-statistic: 3.745 on 4 and 42 DF,  p-value: 0.01077
```

predict(model_2_crime)

898.1004

The predicted value of 898 is more accurate given our data, since it isn’t below the minimum value of crime rate from our data set, this is due to the removal of p values > 0.05.