Crimedataregression

Using Regression analysis to predict observed crime rate. The data set includes the effect of punishment regimes on crime rates in the United states on 47 states in the year of 1960.

```
crime <- read.delim("http://www.statsci.org/data/general/uscrime.txt")
head(crime)
```

```
##      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
```

The data set contains the following variables:

Variable Description M percentage of males aged 14-24 in total state population So indicator variable for a southern state Ed mean years of schooling of the population aged 25 years or over Po1 per capita expenditure on police protection in 1960 Po2 per capita expenditure on police protection in 1959 LF labour force participation rate of civilian urban males in the age-group 14-24 M.F number of males per 100 females Pop state population in 1960 in hundred thousands NW percentage of nonwhites in the population U1 unemployment rate of urban males 14-24 U2 unemployment rate of urban males 35-39 Wealth wealth: median value of transferable assets or family income Ineq income inequality: percentage of families earning below half the median income Prob probability of imprisonment: ratio of number of commitments to number of offenses Time average time in months served by offenders in state prisons before their first release Crime crime rate: number of offenses per 100,000 population in 1960

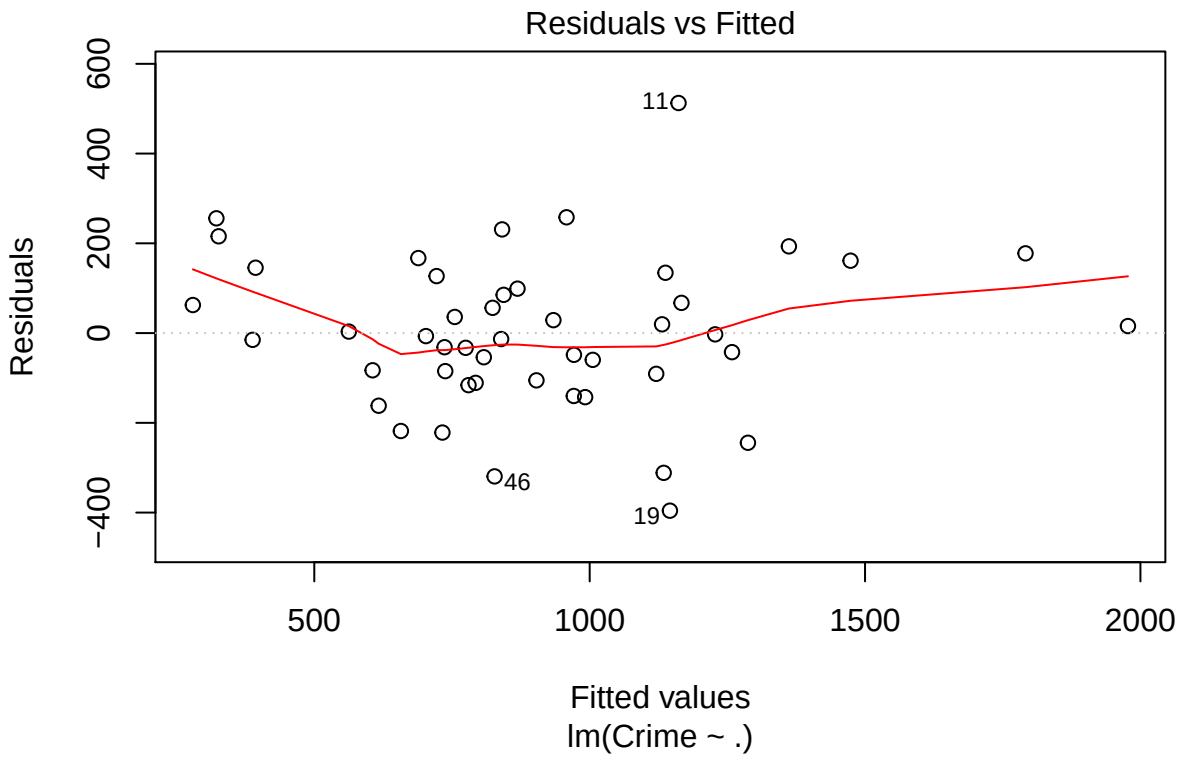
```
a <- data.frame(M = 14.0,So = 0,Ed = 10.0, Po1 = 12.0,Po2 = 15.5,
                LF = 0.640, M.F = 94.0,Pop = 150,NW = 1.1,U1 = 0.120,
                U2 = 3.6, Wealth = 3200,Ineq = 20.1,Prob = 0.04, Time = 39.0)

#This is the test data for regression model
```

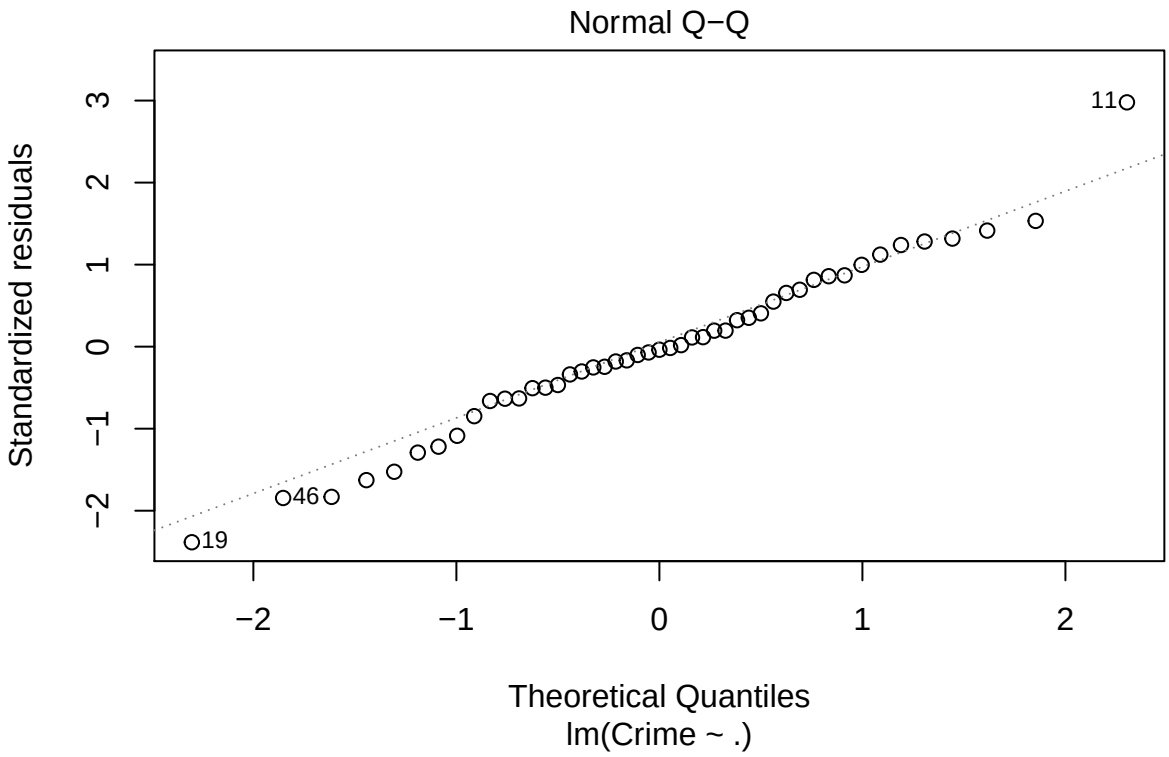
A linear regression is a statistical model that will analyze relationship between response variables compared to other variables. The goal is to find relevant variables to build a model that can be as accurate as possible in order to predict crime rate.

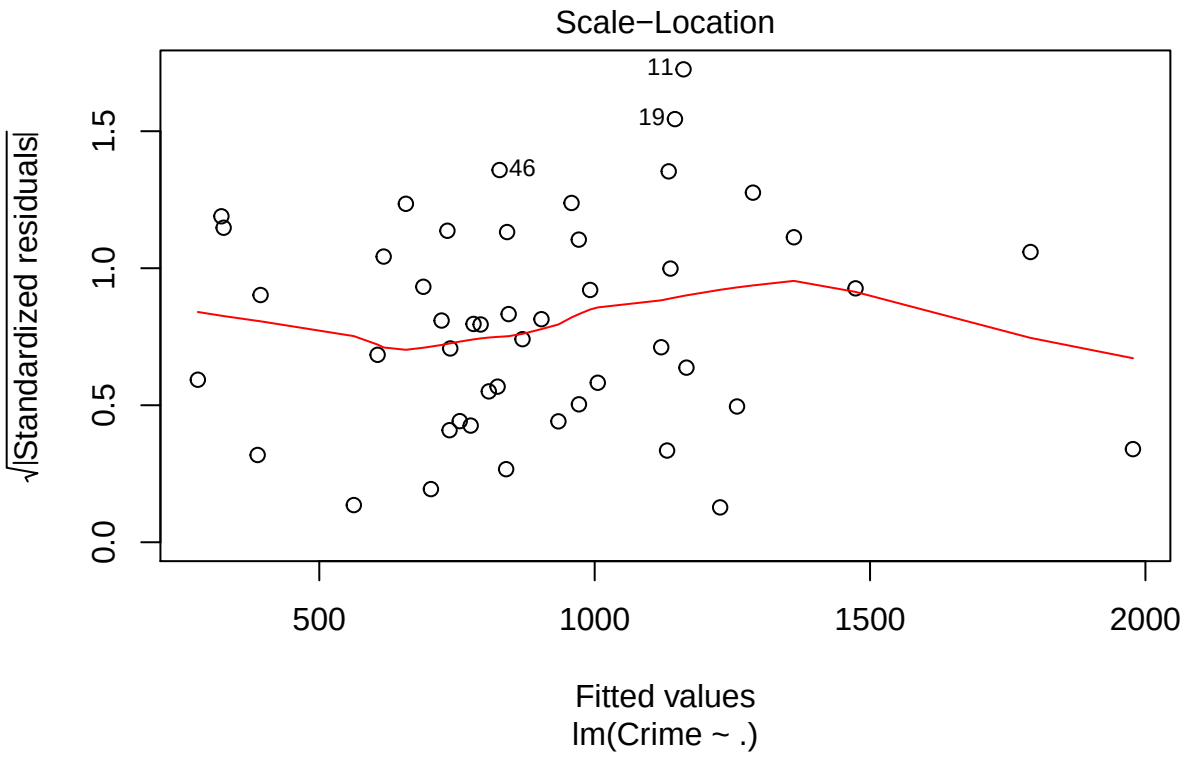
Below is the line creating the linear regression model.

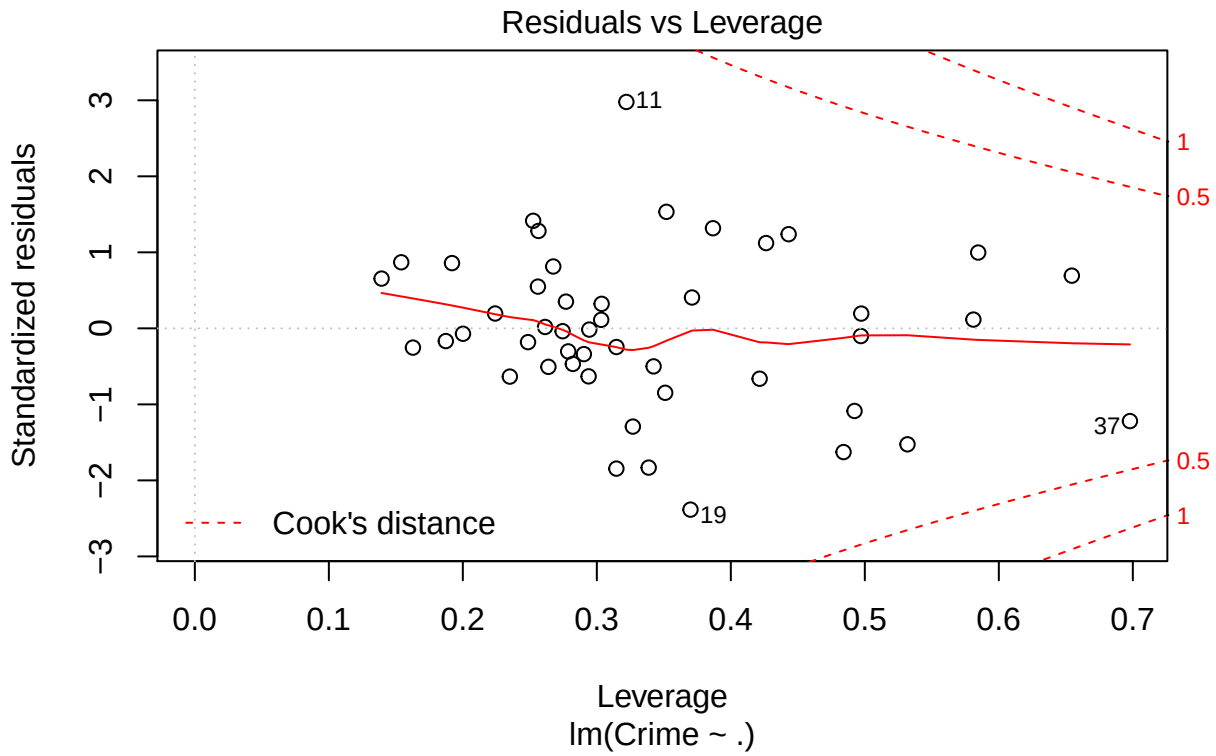
```
model1 <- lm(Crime~., data = crime)
```

```
plot(model1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Crime ~ .)
11
46
19
−2
−1
0
1
2
3


Scale−Location
11
19
46
√|Standardized residuals|
0.0
0.5
1.0
1.5
500
1000
1500
2000
Fitted values
lm(Crime ~ .)



You can see from the plot there a couple outliers but not hugely influential in data set. Looking at the model we can get an idea for accuracy from the R squared value and influential variables to have a p value less than 0.05. Lets predict a value from using this model we created.

```
predict(model1, a)
```

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

```
min(crime$Crime) # minimum value in response variable crime
```

```
## [1] 342
```

Predicting a value using this model we see that through linear regression that this value is way out of the range of data. It is far below the minimum. This model most likely is not going to predict accurately.

```
summary(model1)
```

```
##
## 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
```

Looking at the summary we can see what features have a higher significant impact on the response variable. This is determined by p value. We can just use variable with a p value less than 0.05.

The null hypothesis is basically that the predictor is not meaningful in the model. So if the p value is less than 0.05, we reject null and conclude that the variable is significant.

Below in the model we have only included variables that are significant.

```
model2 <- lm(Crime~M+Ed+Ineq+Prob, data = crime)
summary(model2)
```

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

Looking at the summary of model 2 we now see that Ed and Prob are the only variables with p value less than 0.05. Lets see what model 2 predicts.

```
predict(model2, a)
```

```
##        1
## 897.2307
```

This is the prediction for model 2. Lets look at the max value and see if the value is in range.

```
print(max(crime$Crime)) #max is 1993
```

```
## [1] 1993
```

```
print(min(crime$Crime)) #min is 342
```

```
## [1] 342
```

Model 2 value is within the range so the predicted value makes sense. Looking at R squared values between model1 and model2 you can see model1 has a better R squared value at 0.7 while model 2 is at 0.19. The higher the R squared value, it shows the model fits the observed data better. Since we are taking all variables into account in model1, this is most likely overfitting.

```
model3 <- lm(Crime~M+Ed+Ineq+Prob+Po1+U2, data = crime)
summary(model3)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Ineq + Prob + Po1 + U2, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -470.68  -78.41  -19.68  133.12  556.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5040.50     899.84  -5.602 1.72e-06 ***
## M             105.02      33.30   3.154  0.00305 **
## Ed            196.47      44.75   4.390 8.07e-05 ***
## Ineq           67.65      13.94   4.855 1.88e-05 ***
## Prob        -3801.84    1528.10  -2.488  0.01711 *
## Po1           115.02      13.75   8.363 2.56e-10 ***
## U2             89.37      40.91   2.185  0.03483 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 200.7 on 40 degrees of freedom
## Multiple R-squared:  0.7659, Adjusted R-squared:  0.7307
## F-statistic: 21.81 on 6 and 40 DF,  p-value: 3.418e-11
```

For model 3 I added in Po1 and U2 and I picked these two variables to be added since in the initial model 1 they were the two variables that still had a p value above 0.05 but only very slightly compared to other variables. Looking at the R squared value for model 3, we can see it is 0.7 which is pretty good so this model fits the data pretty well.

```
predict(model3, a)
```

```
##        1
## 1304.245
```

Model 3 gives a predicted value of 1304 which is within the range of values so it this data point does make sense.

```
AIC(model1)
```

```
## [1] 650.0291
```

```
AIC(model2)
```

```
## [1] 690.0666
```

```
AIC(model3)
```

```
## [1] 640.1661
```

We can also see that the AIC is smallest for model 3 which shows its the best model compared to the others.

```
crossvalidate <- trainControl(method = "repeatedcv", number = 10, repeats = 3)

model4 <- train(Crime~M+Ed+Ineq+Prob+Po1+U2, data = crime, method = "lm", trControl = crossvalidate)

print(model4)
```

```
## Linear Regression
##
## 47 samples
##  6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 41, 42, 43, 41, 44, 43, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   207.9865  0.7366771  165.9235
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

We can also use K-fold Cross validation to further validate data and reduce bias. We get a R squared with 0.75 which is also an improved R squared value.This model predicts the best according to R2 value considering overfitting is minimized and using feature selection.