HomeWork 5

# HomeWork 3

# ISYE 6501

### Question 8.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a linear regression model would be appropriate. List some (up to 5) predictors that you might use.

### Answers

I have recently been getting a lot of robo calls and I could use regression to predict how many robo calls I will get tomorrow. Some of the predictors I could use are 1. Number of robo calls recieved the previous day 2. Number of robo call numbers blocked 3. Wether or not personal information has been hacked 4. use of scam call detecting app 5. if phone number entered in an untrust worthy website

### Question8.2

Using crime data, use regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the following data:

library(kernlab)  
library(kknn)  
library(lattice)  
library(ggplot2)

##   
## Attaching package: 'ggplot2'

## The following object is masked from 'package:kernlab':  
##   
## alpha

library(caret)

##   
## Attaching package: 'caret'

## The following object is masked from 'package:kknn':  
##   
## contr.dummy

library(e1071)  
library(outliers)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(olsrr)

##   
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':  
##   
## rivers

library(leaps)

Next we will load the data and look at the data structure.

rm(list=ls())  
uscrime <- read.table("uscrime.txt",stringsAsFactors = FALSE, header = TRUE)

#Fit a linear regression model  
lm\_uscrime <- lm(Crime~.,data = uscrime)  
  
# the code below makes the summary results easier to read  
options(scipen=4)  
summary(lm\_uscrime)

##   
## Call:  
## lm(formula = Crime ~ ., data = uscrime)  
##   
## 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) -5984.28760 1628.31837 -3.675 0.000893 \*\*\*  
## M 87.83017 41.71387 2.106 0.043443 \*   
## So -3.80345 148.75514 -0.026 0.979765   
## Ed 188.32431 62.08838 3.033 0.004861 \*\*   
## Po1 192.80434 106.10968 1.817 0.078892 .   
## Po2 -109.42193 117.47754 -0.931 0.358830   
## LF -663.82615 1469.72882 -0.452 0.654654   
## M.F 17.40686 20.35384 0.855 0.398995   
## Pop -0.73301 1.28956 -0.568 0.573845   
## NW 4.20446 6.48089 0.649 0.521279   
## U1 -5827.10272 4210.28904 -1.384 0.176238   
## U2 167.79967 82.33596 2.038 0.050161 .   
## Wealth 0.09617 0.10367 0.928 0.360754   
## Ineq 70.67210 22.71652 3.111 0.003983 \*\*   
## Prob -4855.26582 2272.37462 -2.137 0.040627 \*   
## Time -3.47902 7.16528 -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: 0.0000003539

# coefficients  
lm\_uscrime$coef

## (Intercept) M So Ed Po1   
## -5984.28760450 87.83017324 -3.80345030 188.32431475 192.80433828   
## Po2 LF M.F Pop NW   
## -109.42192538 -663.82614508 17.40685553 -0.73300815 4.20446100   
## U1 U2 Wealth Ineq Prob   
## -5827.10272440 167.79967222 0.09616624 70.67209945 -4855.26581548   
## Time   
## -3.47901784

#round(summary(lm\_uscrime)$coef, 3)

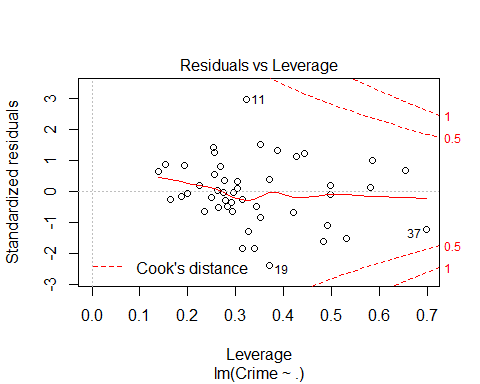
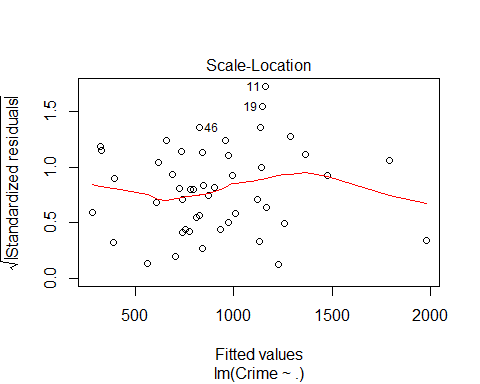
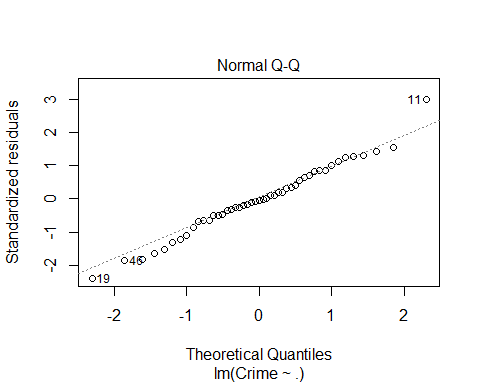
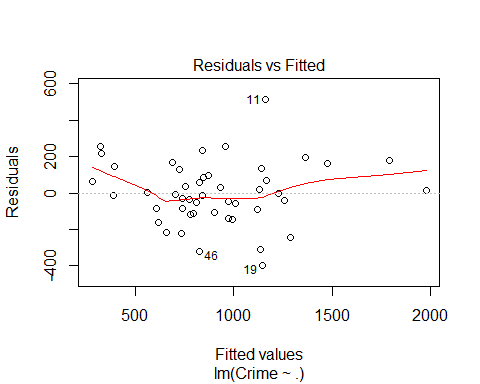
Looking at the p-values, it looks like number of families per 1000 earning below 1/2 the median income, Unemployment rate of urban males per 1000 of age 35-39, Mean # of years of schooling 25 or older, and number of males of age 14-24 per 1000 population are all statistically significant predictors of crime rate. Let us try and predict using the model above.

# Create the test datapoint mannually ising the data.frame() function  
test\_point <- 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 the crime rate for test data point  
pred\_model <-predict(lm\_uscrime, test\_point)  
pred\_model

## 1   
## 155.4349

Below we make a few plots to see how well our model is doing.

#Check how good this basic predictor is  
plot(lm\_uscrime)



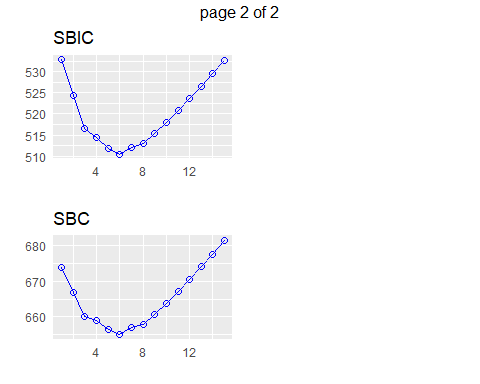
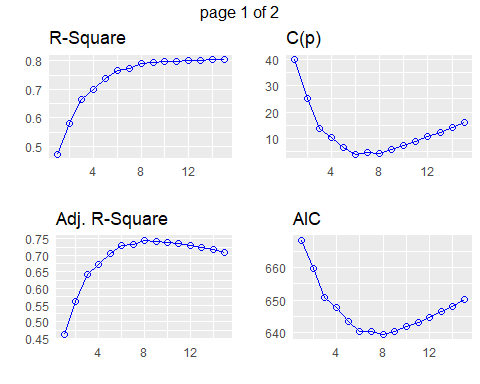
The model created above is very simple and the prediction doesnt seem very good. As our data set is small it is likely we are over fitting. Also the qqplot is not a straight line. This can indicate the data is not normal.

It looks like sme of our coefficients are not statistically significant. We want to use the attributes that are significant so our predictions are more meaningful.

olsmodel<-lm(Crime~.,data = uscrime)  
ols\_step\_best\_subset(olsmodel)

## Best Subsets Regression   
## ------------------------------------------------------------------------  
## Model Index Predictors  
## ------------------------------------------------------------------------  
## 1 Po1   
## 2 Po1 Ineq   
## 3 Ed Po1 Ineq   
## 4 M Ed Po1 Ineq   
## 5 M Ed Po1 Ineq Prob   
## 6 M Ed Po1 U2 Ineq Prob   
## 7 M Ed Po1 U2 Wealth Ineq Prob   
## 8 M Ed Po1 M.F U1 U2 Ineq Prob   
## 9 M Ed Po1 M.F U1 U2 Wealth Ineq Prob   
## 10 M Ed Po1 M.F Pop U1 U2 Wealth Ineq Prob   
## 11 M Ed Po1 Po2 M.F Pop U1 U2 Wealth Ineq Prob   
## 12 M Ed Po1 Po2 M.F Pop NW U1 U2 Wealth Ineq Prob   
## 13 M Ed Po1 Po2 LF M.F Pop NW U1 U2 Wealth Ineq Prob   
## 14 M Ed Po1 Po2 LF M.F Pop NW U1 U2 Wealth Ineq Prob Time   
## 15 M So Ed Po1 Po2 LF M.F Pop NW U1 U2 Wealth Ineq Prob Time   
## ------------------------------------------------------------------------  
##   
## Subsets Regression Summary   
## ---------------------------------------------------------------------------------------------------------------------------------------------  
## Adj. Pred   
## Model R-Square R-Square R-Square C(p) AIC SBIC SBC MSEP FPE HSP APC   
## ---------------------------------------------------------------------------------------------------------------------------------------------  
## 1 0.4728 0.4611 0.3926 39.9970 668.3155 532.8242 673.8659 3789009.5774 84044.2865 1832.1343 0.5741   
## 2 0.5803 0.5612 0.4856 25.0706 659.5957 524.3170 666.9963 3086424.3129 69821.2571 1526.3252 0.4769   
## 3 0.6656 0.6423 0.5748 13.6394 650.9145 516.6543 660.1652 2517546.1300 58059.6879 1273.9521 0.3966   
## 4 0.7004 0.6719 0.6089 10.1620 647.7503 514.3357 658.8512 2310597.8418 54301.0868 1197.0690 0.3709   
## 5 0.7379 0.7060 0.6412 6.2577 643.4641 511.7153 656.4151 2071865.4454 49597.5016 1099.5673 0.3388   
## 6 0.7659 0.7307 0.6662 3.8596 640.1661 510.4861 654.9673 1898463.0711 46275.0374 1032.7288 0.3161   
## 7 0.7746 0.7341 0.6491 4.4889 640.3850 511.9841 657.0364 1875967.4762 46542.8778 1046.6580 0.3179   
## 8 0.7888 0.7444 0.6676 4.2449 639.3151 513.0400 657.8166 1804845.5315 45560.9155 1033.4764 0.3112   
## 9 0.7927 0.7422 0.6504 5.6388 640.4503 515.4829 660.8019 1821158.9635 46759.4869 1071.0019 0.3194   
## 10 0.7959 0.7392 0.6367 7.1276 641.7082 518.0836 663.9100 1843851.0485 48135.5148 1114.4676 0.3288   
## 11 0.7984 0.7350 0.6212 8.7453 643.1457 520.8365 667.1976 1875500.1742 49765.2559 1165.9856 0.3399   
## 12 0.8000 0.7295 0.6001 10.4782 644.7486 523.7026 670.6507 1916074.8266 51658.8801 1226.2461 0.3529   
## 13 0.8016 0.7234 0.5721 12.2372 646.3874 526.6271 674.1396 1960824.7558 53697.1089 1292.9119 0.3668   
## 14 0.8031 0.7169 0.5239 14.0007 648.0301 529.6000 677.6324 2008747.1419 55856.6552 1365.9016 0.3815   
## 15 0.8031 0.7078 0.4856 16.0000 650.0291 532.6322 681.4816 2075661.6070 58587.2228 1456.9309 0.4002   
## ---------------------------------------------------------------------------------------------------------------------------------------------  
## AIC: Akaike Information Criteria   
## SBIC: Sawa's Bayesian Information Criteria   
## SBC: Schwarz Bayesian Criteria   
## MSEP: Estimated error of prediction, assuming multivariate normality   
## FPE: Final Prediction Error   
## HSP: Hocking's Sp   
## APC: Amemiya Prediction Criteria

k <- ols\_step\_best\_subset(olsmodel)  
plot(k)



Looks like the best model uses the top 8 significant coefficients.After 8 the R squared value does not increase and the AIC and BIC values are lowest of 8. lets use this model to make our prediction and see how it compares with the original simple linear regression model.

ollm\_uscrime <- lm(Crime~ M+Ed+Po1+M.F+U1+U2+Ineq+Prob ,data = uscrime)  
olpred\_model <-predict(ollm\_uscrime, test\_point)  
  
# coefficients  
  
ollm\_uscrime$coef

## (Intercept) M Ed Po1 M.F U1   
## -6426.10102 93.32155 180.12011 102.65316 22.33975 -6086.63315   
## U2 Ineq Prob   
## 187.34512 61.33494 -3796.03183

# Prediction using the model selected   
cat("The predicted crime rate for the test values is ", olpred\_model)

## The predicted crime rate for the test values is 1038.413

The final prediction of 1038 seems much more in line with the data.

(Intercept)

-5726.224

(Intercept)

-5489.906