**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.**

**Answer:**

I once worked as an analytics point of contact for line of business in a firm that provides home loans. The name line of business is called Government Insuring. In order to explain how I would use linear regression here I need to first explain what Government Insuring is.

Government Insuring: When a borrower applies for a loan, a company provides the borrower different types of loans and the borrower would eligible to get certain types of loans based on certain criteria’s. Some of the loans are provided by government schemes. They are called as FHA loans (for a common man), VA loans (for Veterans) and USDA loans (for people associated with agriculture). Once a loan is funded to a borrower, the company intends to sell these loans to other companies called Investors at a small profit margin so that my company gets the money to fund more loans. The investors are paid the money monthly after they receive the money from borrower’s monthly payment with interest. Just in case if borrower fails to pay the loan, government provides an option of insuring a loan at a very small premium which works as a protection for a company. But in order to get a loan insured, the government needs all the documents associated with the loans without errors. If errors are encountered then Government calls it a defect. If the defect is not resolved then the loan cannot be insured and company is at a risk for that loan.

I would use linear regression here to predict how many government loans are insured.

**Response**: Total loans Insured per year

**Predictors**: Given below are the predictors

1. **Channel**: A loan can come through 4 channels. They are

* Wholesale – Where a bank gets bulk of loans from a wholesale realtor.
* Retail – Where loan originates from a retail medium.
* Correspondent – Where loan originates from another local bank later purchased.
* Consumer Direct – Where loan originates from the borrower directly applying loan online.

For all the above channels, we get different form of documentation leading to different defects

1. **Loan Type**: A borrower can be eligible for different types of loans based on different parameters like

income, job status etc. Different types of loans are USDA loans (for farmers or people in agriculture),

VA loans (for veterans), FHA loans etc. The loan type again leads to difference in the documents

Required

1. **Number of Defects :** As number of defects increases, the chances of being not fixed also increases which makes the loan fail to insure.
2. **State:** Different states in USA have different laws for granting of loans. Most of the laws per state

may be same but there may be only few laws which change causing a loan to fail being insured.

1. **Property Type:** A property can be a Condominium, or an apartment or a house etc. A property can

have number of units ranging from 1 to 5. For every different property type there is different

documentation in different states.

**Question 8.2**

**Using crime data from http://www.statsci.org/data/general/uscrime.txt (file uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html ), use regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the following data: 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 Show your model (factors used and their coefficients), the software output, and the quality of fit. Note that because there are only 47 data points and 15 predictors, you’ll probably notice some over fitting. We’ll see ways of dealing with this sort of problem later in the course.**

**Answer:**

Given below are the steps to Predict a crime rate in a city

**Step 1: Load the dataset**

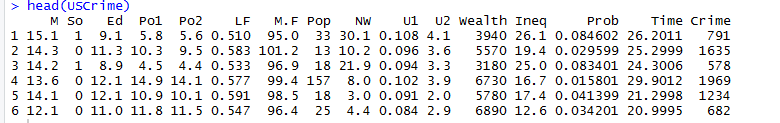
**Code:**

*library(data.table)*

*uscrime = read.table("C://Users/D100793/Desktop/Junk/Georgia Tech/uscrime.txt", header = TRUE,sep = '\t')*

*head(USCrime)*

**Output:**



**Step 2: Refining the Model**

We start this process with considering all the predictors and keeping the field Crime as a response.

**Code:**

##Running Linear Regression Model on all the data

*lm\_uscrime1 <- lm(Crime~.,data = uscrime)*

## Summary of the Model

*summary(lm\_uscrime1)*

##Setting up test points

*test\_point1 <- 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 = 0.040, Time = 39.0)*

##Predicting the Model

*Pred\_model1 <- predict(lm\_uscrime1,test\_point1)*

*Pred\_model1*

##Creating a qq plot

*qqnorm(uscrime$Crime)*

*qqline(uscrime$Crime)*

**Output:**

> lm\_uscrime1 <- lm(Crime~.,data = uscrime)

> summary(lm\_uscrime1)

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) -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

> test\_point1 <- 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,

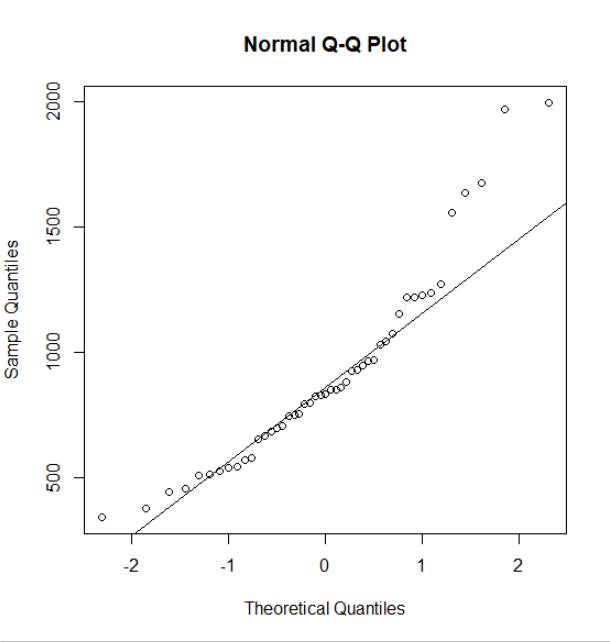
+ Prob = 0.040, Time = 39.0)

> Pred\_model1 <- predict(lm\_uscrime1,test\_point1)

> Pred\_model1

1

155.4349



**Analysis:**

A predictor whose P-Values is below 0.05 is said to be a significant predictor in prediction of the Response variable. Significance is denoted by \* (asterisk sign). More \* represent higher significance.

So looking at the summary of lm\_uscrime1 model there appears 5 significant variables. They are M, Ed , Ineq and Prob along with the intercept value. The R-squared is 0.8031 which means the model has lesser variability.

But when we predict the model based on the Test Points, it appears the predict output for prediction Pred\_model1 shows that the prediction 155 which does not appear to be true as per the qqplot shown above. The qqplot says that the Crime exists between 342 and 1960 where as the model predicts 155 which shows that this Model is not a good model. Also the P-Value of the of the model is 3.539e-07 which is very high to reject the null hypothesis which assumes that the model has all significant predictors.

So we need to refine the model even more. Lets consider this a new model with only significant predictors from lm\_uscrime1.

**Code:**

##Running Linear Regression Model on aSignificant values of lm\_uscrime1

lm\_uscrime2 <- lm(Crime~M+Ed+Ineq+Prob+Po1+U2,data = uscrime)

## Summary of the Model

summary(lm\_uscrime2)

##Setting up test points

test\_point2 <- data.frame(M = 14.0, Ed = 10.0, Ineq = 20.1,Po1 = 12.0,U2 = 3.6 ,Prob = 0.040)

##Predicting the Model

Pred\_model2 <- predict(lm\_uscrime2,test\_point2)

Pred\_model2

mean(uscrime$Crime)

median(uscrime$Crime))

**Output:**

|  |
| --- |
| > ##Running Linear Regression Model on aSignificant values of lm\_uscrime1  > lm\_uscrime2 <- lm(Crime~M+Ed+Ineq+Prob,data = uscrime)  > ## Summary of the Model  > summary(lm\_uscrime2)  Call:  lm(formula = Crime ~ M + Ed + Ineq + Prob + Po1 + U2, data = uscrime)  Residuals:  Min 1Q Median 3Q Max  -470.7 -78.4 -19.7 133.1 556.2  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -5040.5 899.8 -5.60 1.7e-06 \*\*\*  M 105.0 33.3 3.15 0.0031 \*\*  Ed 196.5 44.8 4.39 8.1e-05 \*\*\*  Ineq 67.7 13.9 4.85 1.9e-05 \*\*\*  Prob -3801.8 1528.1 -2.49 0.0171 \*  Po1 115.0 13.8 8.36 2.6e-10 \*\*\*  U2 89.4 40.9 2.18 0.0348 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 201 on 40 degrees of freedom  Multiple R-squared: 0.766, Adjusted R-squared: 0.731  F-statistic: 21.8 on 6 and 40 DF, p-value: 3.42e-11  > ##Setting up test points  > test\_point2 <- data.frame(M = 14.0, Ed = 10.0, Ineq = 20.1,Prob = 0.040)  > ##Predicting the Model  > Pred\_model2 <- predict(lm\_uscrime2,test\_point2)  > > Pred\_model2  1  1304  > mean(uscrime$Crime)  [1] 905.0851  > median(uscrime$Crime)  [1] 831 |
|  |
| |  | | --- | |  | |

**Analysis:**

So looking at the summary of lm\_uscrime2 model there appears to have all significant variables. They are Ed and Prob. The R-squared is 0.766 which means the model has low variability and a better accuracy.

When we look the predictions made in Pred\_model2 step, we get the value of 1304 which ranges between min and max value of 342 and 1960 respectively based on qqplot. This model predicts the data well. This appears to be a good Model.

**Step 3: Validating the refined model** lm\_uscrime2

Running cv.lm on the

**Code:**

*lm\_uscrime\_Cv = cv.lm(uscrime,lm\_uscrime2,m = 4)*

**Output:**

lm\_uscrime\_Cv = cv.lm(uscrime,lm\_uscrime2,m = 4)

Analysis of Variance Table

Response: Crime

Df Sum Sq Mean Sq F value Pr(>F)

M 1 55084 55084 0.46 0.5031

Ed 1 725967 725967 6.01 0.0185 \*

Ineq 1 37674 37674 0.31 0.5794

Prob 1 990334 990334 8.20 0.0065 \*\*

Residuals 42 5071868 120759

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

fold 1

Observations in test set: 11

2 6 12 18 24 25 26 27 28 32 39

Predicted 1159 818.4 981 257 1023.72 872 1033 875 1169.5 855.1 833.6

cvpred 1170 759.9 984 -102 962.11 717 928 819 1166.6 801.8 914.5

Crime 1635 682.0 849 929 968.00 523 1993 342 1216.0 754.0 826.0

CV residual 465 -77.9 -135 1031 5.89 -194 1065 -477 49.4 -47.8 -88.5

Sum of squares = 2713939 Mean square = 246722 n = 11

fold 2

Observations in test set: 12

1 9 10 11 17 22 23 29 35 40 42 45

Predicted 637 680 1059 1005 696 634 947 921 891.7 1056 593 726

cvpred 687 726 1117 853 830 706 798 787 742.4 1022 748 633

Crime 791 856 705 1674 539 439 1216 1043 653.0 1151 542 455

CV residual 104 130 -412 821 -291 -267 418 256 -89.4 129 -206 -178

Sum of squares = 1366338 Mean square = 113861 n = 12

fold 3

Observations in test set: 12

5 7 14 15 20 21 33 37 38 44 46 47

Predicted 1130.1 909.85 1006 703.2 906 1076 998.2 941.5 893 1177 777 969

cvpred 1191.5 953.29 1080 723.5 941 1144 1055.3 908.5 966 1236 800 1033

Crime 1234.0 963.00 664 798.0 1225 742 1072.0 831.0 566 1030 508 849

CV residual 42.5 9.71 -416 74.5 284 -402 16.7 -77.5 -400 -206 -292 -184

Sum of squares = 750639 Mean square = 62553 n = 12

fold 4

Observations in test set: 12

3 4 8 13 16 19 30 31 34 36 41 43

Predicted 554.4 1281 1011 1022 761 1075 665.24 776 1014.4 1047 1209 914.1

cvpred 563.8 1253 985 992 746 1084 693.03 787 1009.5 1064 1216 917.2

Crime 578.0 1969 1555 511 946 750 696.00 373 923.0 1272 880 823.0

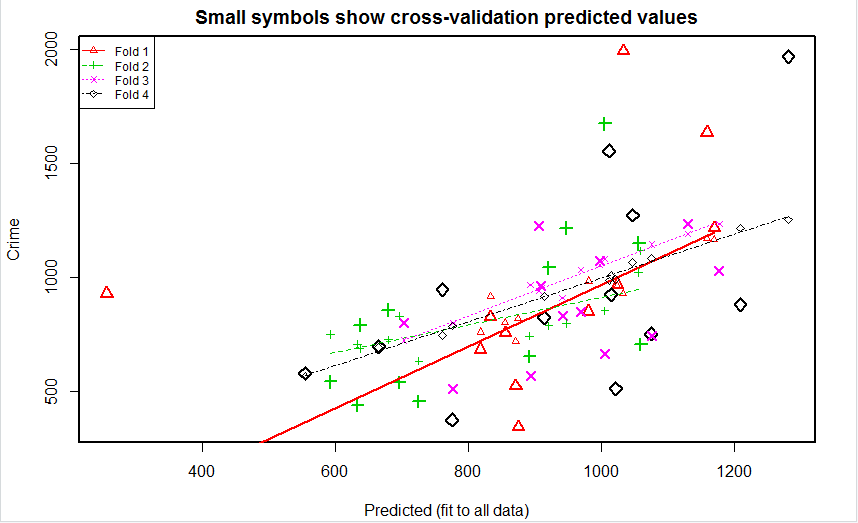
CV residual 14.2 716 570 -481 200 -334 2.97 -414 -86.5 208 -336 -94.2

Sum of squares = 1566064 Mean square = 130505 n = 12

Overall (Sum over all 12 folds)

ms

136106



**Step 3: Calculation of R-Squared**

**Code:**

*n = length(uscrime$Crime)*

*avg = mean(uscrime$Crime)*

*SSE<-0*

*SSR<-0*

*SST<-0*

*for(i in 1:n){*

*SST = SST + (uscrime$Crime[i] - avg)^2*

*SSE = SSE + (uscrime$Crime[i] - lm\_uscrime\_Cv$cvpred[i])^2*

*SSR = SSR + (lm\_uscrime\_Cv$cvpred[i] - avg)^2*

*}*

*SSE*

*SST*

*SSR*

*R\_Squared = 1- (SSE/SST)*

*R\_Squared*

**Output:**

> SSE

[1] 1932333

> SST

[1] 6880928

> SSR

[1] 5006846

>

> R\_Squared = 1- (SSE/SST)

> R\_Squared

[1] 0.719

**ANALYSIS: With an R\_Squared of 0.719 the model appears to be accurate enough and less variable**