Homework Week 5

Chris Messer

2022-09-25

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.

> I work in a consulting, and in the last two years, the job market has been very hot. Multiple times a day, offers to interview would come through, but there was always a nagging question- how much should I be asking for salary?
>
> A linear regression model would be highly effective at taking salary data from people in the same profession, and information regarding the individual, and create a model to predict salary. One such dataset can be found at https://www.big4transparency.com.
>
> Some of the model inputs could be:
>
> 1. Age
> 2. Years of Experience
> 3. Title
> 4. Ethnicity
> 5. Gender
> 6. Industry
> 7. Tier of FirmE
> 8. Experienced Hire/Homegrown
> 9. Hours per week worked

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 overfitting. We'll see ways of dealing with this sort of problem later in the course.

> The model I determined was best is:
>
> y= m105.02+Ed196.47+Po1*115.02+U2*89.37+Ineq67.65+Prob*-3801.84-5040.5
>
> The output of the model is below under subheading *Linear Model 2.* It has an R^2 valueof .69, meaning 69% of the variation in the observed values is due to the inputted dimensions.

First, lets import the data and look at a summary of it.

```
crime_data <- read.csv('uscrime.txt', sep="\t")
summary(crime_data)
```

```
##        M               So               Ed             Po1
##  Min.   :11.90   Min.   :0.0000   Min.   : 8.70   Min.   : 4.50
##  1st Qu.:13.00   1st Qu.:0.0000   1st Qu.: 9.75   1st Qu.: 6.25
##  Median :13.60   Median :0.0000   Median :10.80   Median : 7.80
##  Mean   :13.86   Mean   :0.3404   Mean   :10.56   Mean   : 8.50
##  3rd Qu.:14.60   3rd Qu.:1.0000   3rd Qu.:11.45   3rd Qu.:10.45
##  Max.   :17.70   Max.   :1.0000   Max.   :12.20   Max.   :16.60
##       Po2               LF              M.F              Pop
##  Min.   : 4.100   Min.   :0.4800   Min.   : 93.40   Min.   :  3.00
##  1st Qu.: 5.850   1st Qu.:0.5305   1st Qu.: 96.45   1st Qu.: 10.00
##  Median : 7.300   Median :0.5600   Median : 97.70   Median : 25.00
##  Mean   : 8.023   Mean   :0.5612   Mean   : 98.30   Mean   : 36.62
##  3rd Qu.: 9.700   3rd Qu.:0.5930   3rd Qu.: 99.20   3rd Qu.: 41.50
##  Max.   :15.700   Max.   :0.6410   Max.   :107.10   Max.   :168.00
##        NW              U1                U2            Wealth
##  Min.   : 0.20   Min.   :0.07000   Min.   :2.000   Min.   :2880
##  1st Qu.: 2.40   1st Qu.:0.08050   1st Qu.:2.750   1st Qu.:4595
##  Median : 7.60   Median :0.09200   Median :3.400   Median :5370
##  Mean   :10.11   Mean   :0.09547   Mean   :3.398   Mean   :5254
##  3rd Qu.:13.25   3rd Qu.:0.10400   3rd Qu.:3.850   3rd Qu.:5915
##  Max.   :42.30   Max.   :0.14200   Max.   :5.800   Max.   :6890
##       Ineq            Prob              Time           Crime
##  Min.   :12.60   Min.   :0.00690   Min.   :12.20   Min.   : 342.0
##  1st Qu.:16.55   1st Qu.:0.03270   1st Qu.:21.60   1st Qu.: 658.5
##  Median :17.60   Median :0.04210   Median :25.80   Median : 831.0
##  Mean   :19.40   Mean   :0.04709   Mean   :26.60   Mean   : 905.1
##  3rd Qu.:22.75   3rd Qu.:0.05445   3rd Qu.:30.45   3rd Qu.:1057.5
##  Max.   :27.60   Max.   :0.11980   Max.   :44.00   Max.   :1993.0
```

Linear Model - Method 1

Now, lets start building our linear regression model. For this first model we will use all columns for our predictors.

```
lm_model <- lm(Crime~., crime_data)
```

Now, lets go ahead and store our predictor point that the question asked us to predict a crime rate for into a new variable, so that we can call predict() on it later.

A note on scaling

Should we scale our data? Scaling is not particularly needed in a linear regression. In fact, I don't believe it will change the model performance at all! However, that doesn't mean it is *useless.* When we get our coefficients from our model, if we do not scale our data, we cannot compare our coefficients to one another to determine which coefficients are more impactfull to our model. However, if we scale them, we can directly compare our coefficients against one another to determine which ones are more important. For example, if a scaled coefficient was close to 0, it is not causing much variation in y.

However, if we did not scale our data, a coefficient of .0000001 may still be significant, if the input we are multiplying it by is recorded ina high degree. For example, if the input value was centimeters to the sun, multiplying it by .0000001 would still yield a large number and thus be relevant to our y value.

```
new_data <- 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)
```

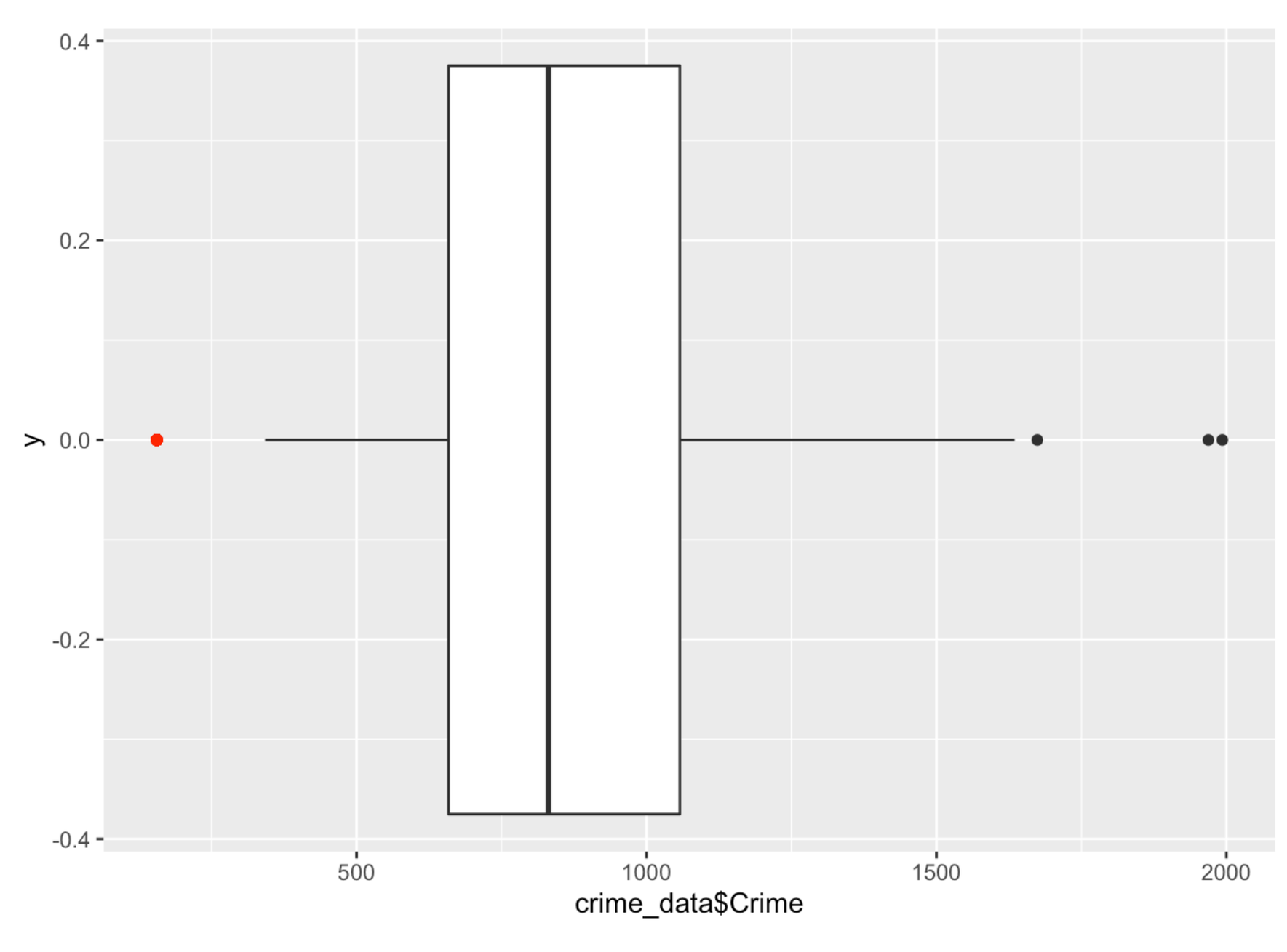
Now that our model is built and we have our point stored we want to predict, lets see what we get.

```
predict1<-predict(lm_model, new_data)
predict1
```

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

Interesting… our predicted crime rate seems pretty low compared to what was in our data set. Lets lay it over the other crime rates for a sanity check.

```
library('ggplot2')
p <- ggplot(as.data.frame(crime_data$Crime), aes(x=crime_data$Crime)) +
  geom_boxplot()
p + geom_point(aes(x=predict1[[1]], y = 0), colour="red")
```



As expected, it is almost half of what the next highest crime rate is.

Exploratory Data Analysis

Lets dive into this a little more. There are two potential explanations for this. Either the predictors we gave the linear model do in fact indicate our predicted crime rate should be lower, or we are overfitting our model using predictors that don't impact crime rate very well, and thus, randomness in our training data for these coefficients is causing an artificially low predicted crime.

Crime Rate *Should* be lower

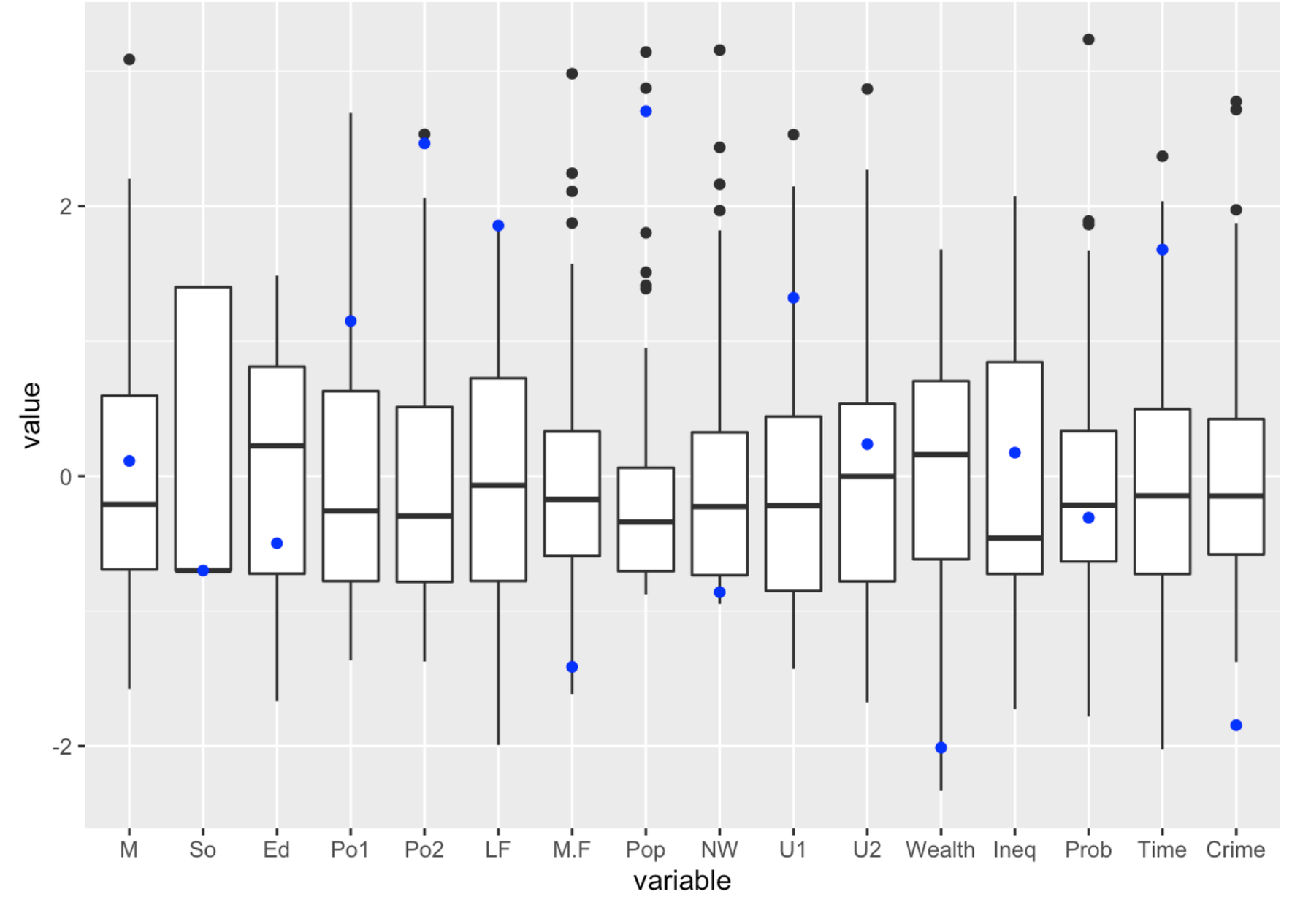
In order to see how our testing point compares to the training points, I've plotted them against the training data for each predictor.

```
#combine the datasets
new_data_pred <- new_data
new_data_pred$Crime <- predict1[[1]]
combined <- rbind(crime_data,new_data_pred)

#scale the data so we can plot on the same chart
scaled_crime <- data.frame(scale(combined))

#create the boxplot
library('reshape2')
ggplot(melt(scaled_crime[1:47,]), aes(x = variable, y = value)) +
  geom_boxplot() + geom_point(data = melt(scaled_crime[48,]), aes(x=variable, y = value), colour="blue")
```

```
## No id variables; using all as measure variables
## No id variables; using all as measure variables
```



None of the dimensions in out test data point are less than *or* greater than the lowest/highest datapoint in the training data. As such, it is easy to conclude that this was a result of an overfit model. If none of the input values of our test data are lower (or higher in the case of a negatively correlated factor) than the test point, it would not be reasonable to observe a predicted crime rate less than all other observed data points.

Model is over-fit, too many predictors

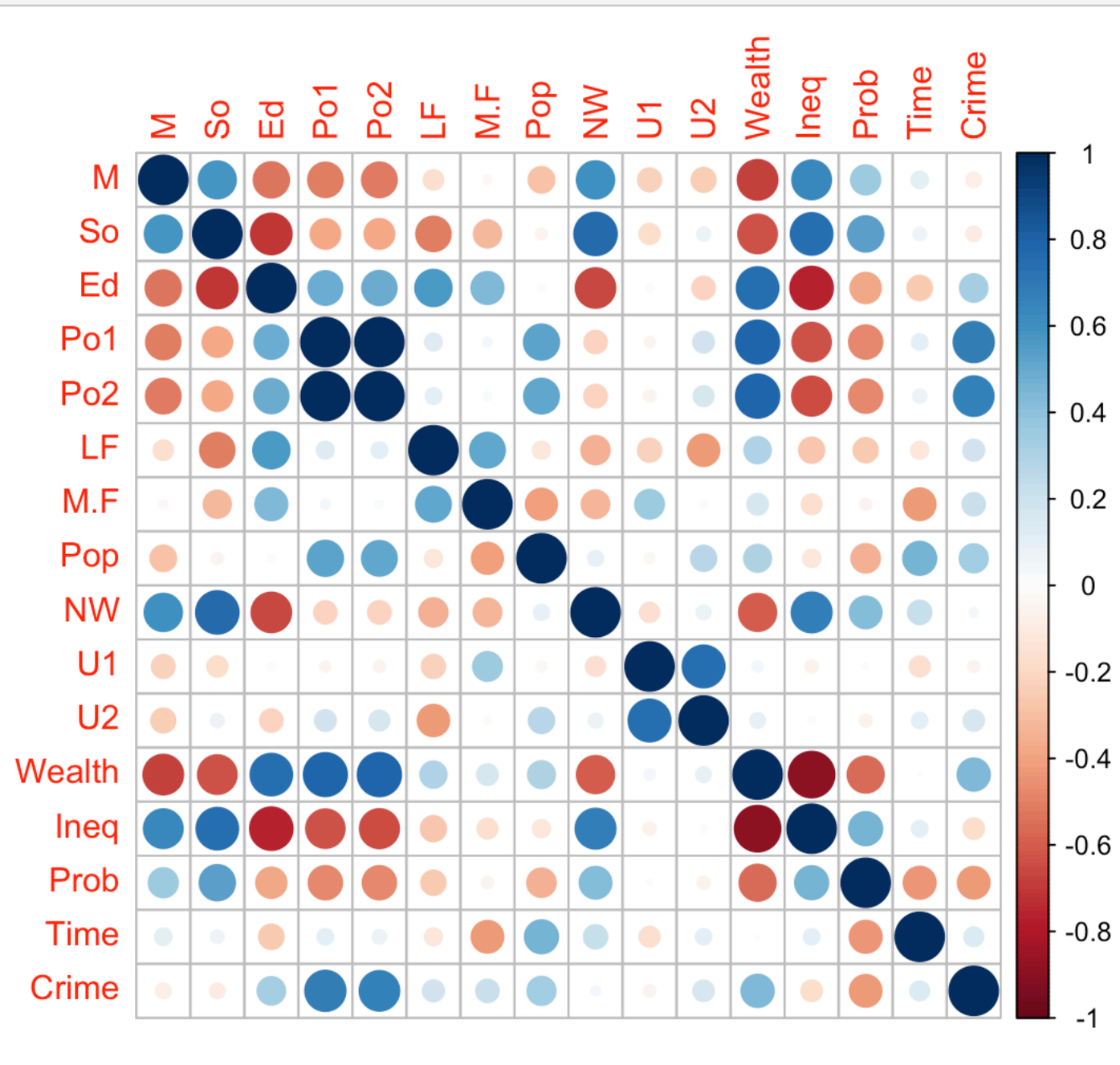
Since it is difficult to tell whether our model predicted a low crime rate appropriatly just based on comparing our test values to our training values, let's take a more statistical approach.

First I'd like to take a closer look at how are variables correlate with crime individually, to see if some of them are stronger than others.

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
crime_data.cor <- cor(crime_data)
corrplot(crime_data.cor)
```



If we look at the far right column/bottom row, we can see how each variable correlates with crime individually. Right off the bat, we can see some winners and some losers. It appears NW (percentage of non-white population), U1 (unemployment rate of youth aged 14- 24), both have a very weak correlation to the crime rate. Conversely, Po1 and Po2 (police spending in 1960 and 1959, respectively), and Wealth (median value of household assets) all have a strong positive correlation with crime, and Prob (probability of imprisonment) has a strong negative correlation with crime.

Additional Notes on correlation plot

It is important to note that correlation <> causation. For example, even though there is a strong correlation in police spending and crime, is it accurate to say that more police spending results in more crime? Or that more crime results in more police spending? Or does lowering police spending decrease crime? None of these are accurate. More likely, more police spending results in a higher *reported* crime rate.

Some other interesting things to point out:

1. U1 and U2 are highly positively correlated. This makes sense, because this is the unemployment rate between 14-24 yearolds and 24+ year olds. It is reasonable to assume that if unemployment is high in one of those groups, it is also high in the other group, since unemployment is caused by macroeconomic factors that impact both age groups.
2. PO1 and PO2 are highly correlated, since this is police spending in Y1 and Y2. This would be reasonably correlated, since high spending in one year liekly leads to high spending in the next in regards to police, as a drastic change would likely only come from sweeping political changes.
3. Wealth and Education have a strong positive correlation. This is reasonable, given higher education opens opportunities for more income.
4. Southern states have a higher probability of being committed to prison for an offense. This likely comes from "tough on crime" attitude of conservative lawmakers.
5. Higher wealth correlates strongly in a positive direction with police spending. Since police spending is a piblic funded expense, it is reasonable to assume that higher income areas would lead to higher budgets for police.

Reducing Dimensionality

To get a better model, lets remove some of the factors that are not heavily correlated with crime. While we could use the above correlation plot, lets take a more statistical approach - observing the p-values of the linear regression model, and removing the dimensions that are not relevant.

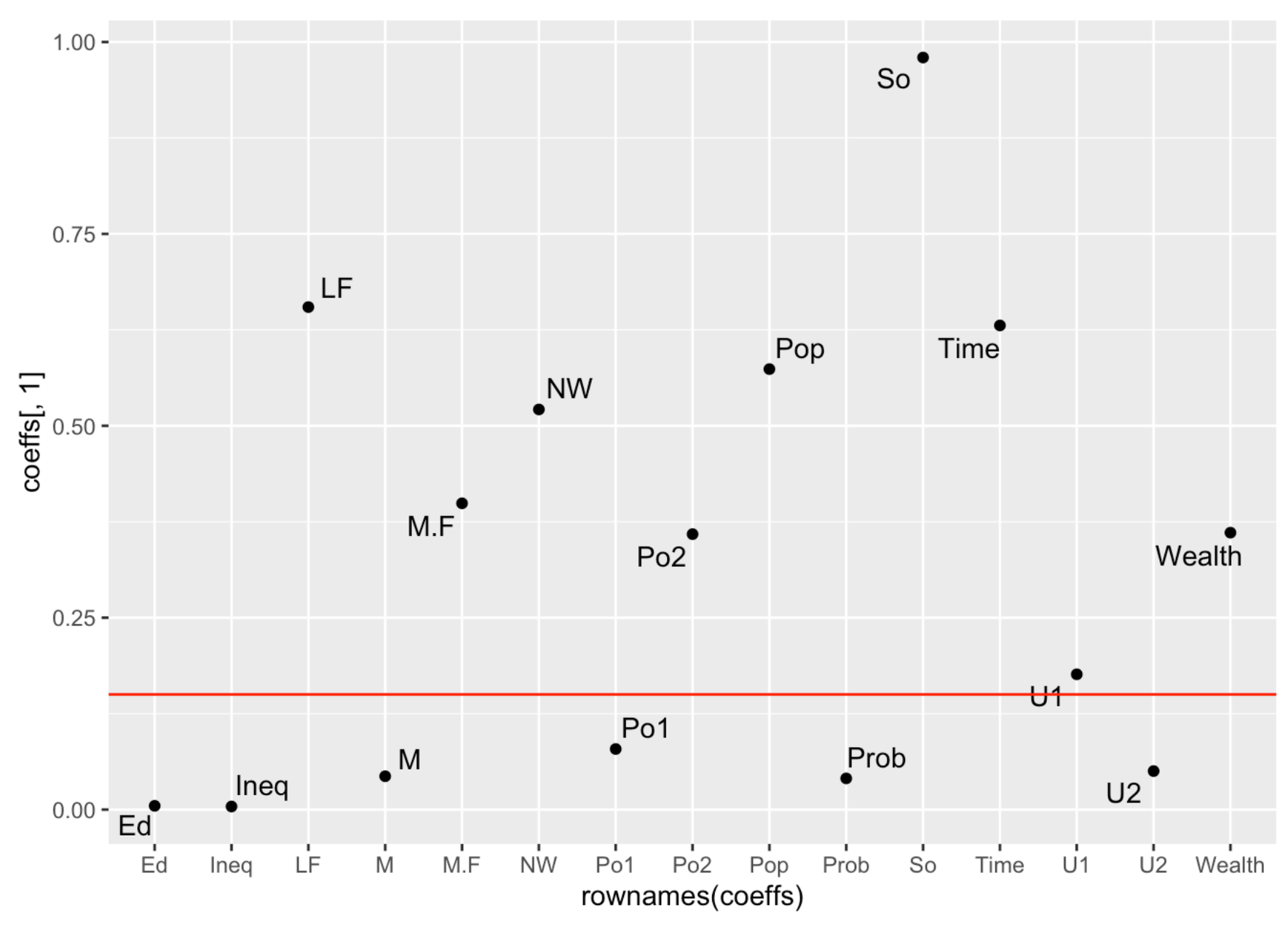
```
sum1 <- summary(lm_model)
sum1
```

```
##
## Call:
## lm(formula = Crime ~ ., data = crime_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
```

As seen in the above output, some of the p values are significantly higher than others. Lets plot them out to see if there is a clear pattern.

```
library(ggrepel)
coeffs<-as.data.frame(sum1$coefficients[2:16,4])

ggplot((coeffs), aes(rownames(coeffs),coeffs[,1])) +
  geom_point() + geom_text_repel(aes(label = rownames(coeffs)))  + geom_hline(yintercept=.15, linetype="solid", c
olor = "red")
```



Visually, it is pretty easy to separate these coefficients. Lets draw an arbitrary line, and toss out any predictors that have a p-value of .15 or greater. These are not particularly relevant to predicting crime, and thus including skews our prediction. This is because of the degree of randomness in our training data, if we included them we are baking in that randomness to our predicted point.

Linear Model 2

So now, lets make a second linear model and only use the predictors that were below p value of .15.

```
lm_model2 <- lm(Crime~M+Ed+Po1+U2+Ineq+Prob, crime_data)
summary(lm_model2)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = crime_data)
##
## 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 ***
## Po1           115.02      13.75   8.363 2.56e-10 ***
## U2             89.37      40.91   2.185  0.03483 *
## Ineq           67.65      13.94   4.855 1.88e-05 ***
## Prob        -3801.84    1528.10  -2.488  0.01711 *
## ---
## 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
```

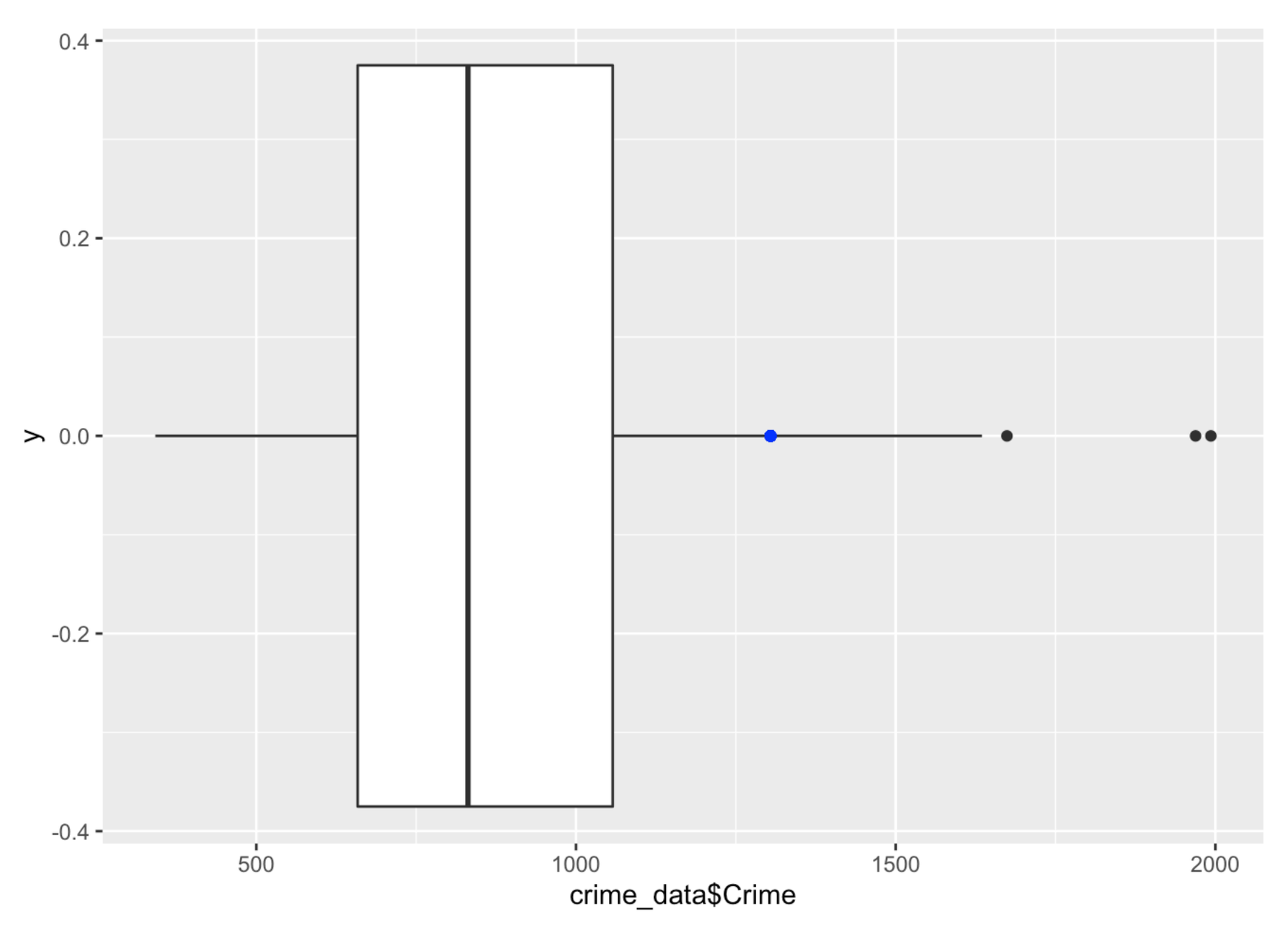
Looking at our summary, the Rsquared value is similar to our original model (this is expected) but our p value of the coefficients are all significantly closer to 0, with no obvious outliers. Now, lets again predict our test point.

y= m105.02+Ed196.47+Po1*115.02+U2*89.37+Ineq67.65+Prob*-3801.84-5040.5

```
predict2 <- predict(lm_model2, new_data)
predict2[[1]]
```

```
## [1] 1304.245
```

```
p <- ggplot(as.data.frame(crime_data$Crime), aes(x=crime_data$Crime)) +
  geom_boxplot()
p + geom_point(aes(x=predict2[[1]], y = 0), colour="blue")
```



At last, we have a crime rate that makes a little more sense!

Now that we have two models, we will perform cross validation on them in order to

```
#this library will perform CV for us
library(DAAG)
model1 <- cv.lm(crime_data, lm_model, m=5, seed=10, printit=F, plotit =F)
model2 <- cv.lm(crime_data, lm_model2, m=5, seed=10, printit=F, plotit =F)

#the model does not outpuut r^2, so we will have to calculate ourselves. The formula is 1-(Sum of Squared Errors/
sumofsquared differences)
SSres1 <- attr(model1,"ms")*nrow(crime_data)
SSres2 <- attr(model2, 'ms')* nrow(crime_data)
SStotal <- sum((crime_data$Crime - mean(crime_data$Crime))^2)

rs1 <- 1- SSres1/SStotal
rs2 <- 1-SSres2/SStotal

print(c(rs1,rs2))
```

```
## [1] 0.3984500 0.6949336
```

As we can see in the above output, model 1 (using all 15 factors), the given dimensions accounts for only 40% of the variance. The second model, with only dimensions with a p-value of .15 or less, gives us a model where 69% of the variance in the crime is predicted by the inputted dimensions. As such, I have concluded the second model is stronger and we will report using that model.