Crime Data Analysis

Andrew Wang

07/09/2020

Question of Interest

I was interested in how the level of crime in NZ is effected by unemployment rate, immigration, and the season.

Therefore, I collected victimisation data from NZ's Police government database for the period July 2014 - June 2020. I also collected data on unemployment rates, and immigration from NZ stats.

The variables of interest are:

- Number_of_Victims: The number of victimisations for a given month of a year.
- Unemployment_Rate: A number for the unemployment rate for a given month of a year.
- Net_Migration_Arrival: A number for the net amount of people that have come into NZ for a given month of a year.
- Season: A four-level factor which describes the Season for a given date.
    - It has levels "Summer", "Autumn", "Winter", and "Spring"

Read in and Inspect the Data

```r
library(MASS)
library(s20x)
library(zoo)
```

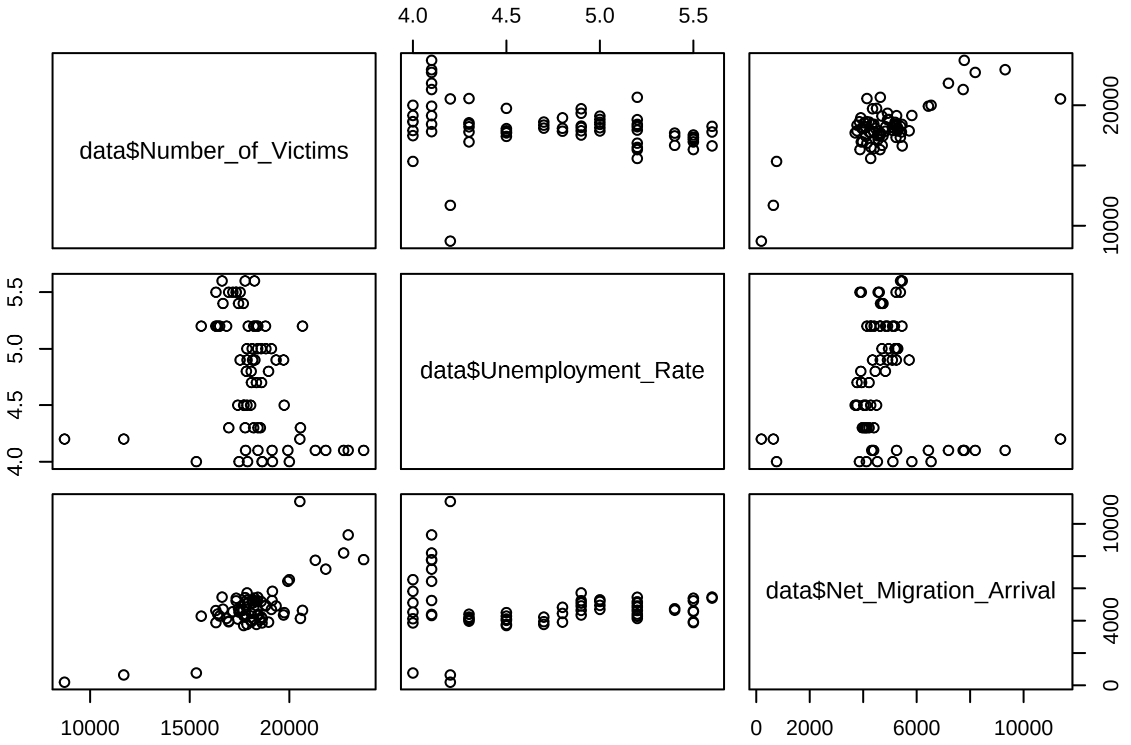
```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
data = read.csv("crimedata.csv")
head(data)
```

```
##   X     Date Number_of_Victims Month Season Unemployment_Rate
## 1 1 Jul 2014             15579     7 Winter               5.2
## 2 2 Aug 2014             16312     8 Winter               5.2
## 3 3 Sep 2014             16503     9 Spring               5.2
## 4 4 Oct 2014             16402    10 Spring                5.2
## 5 5 Nov 2014             16851    11 Spring                5.2
## 6 6 Dec 2014             16955    12 Summer                5.5
##   Net_Migration_Arrival
## 1                  4280
## 2                  4630
## 3                  4290
## 4                  4410
## 5                  4140
## 6                  3930
```

```
# Pairs Plot
pairs(~ data$Number_of_Victims+data$Unemployment_Rate+data$Net_Migration_Arrival, data = data)
```



```
# Trying a Poisson glm
poisson.mod = glm(Number_of_Victims ~ Unemployment_Rate + Net_Migration_Arrival + as.factor(Season), da
summary(poisson.mod)
```

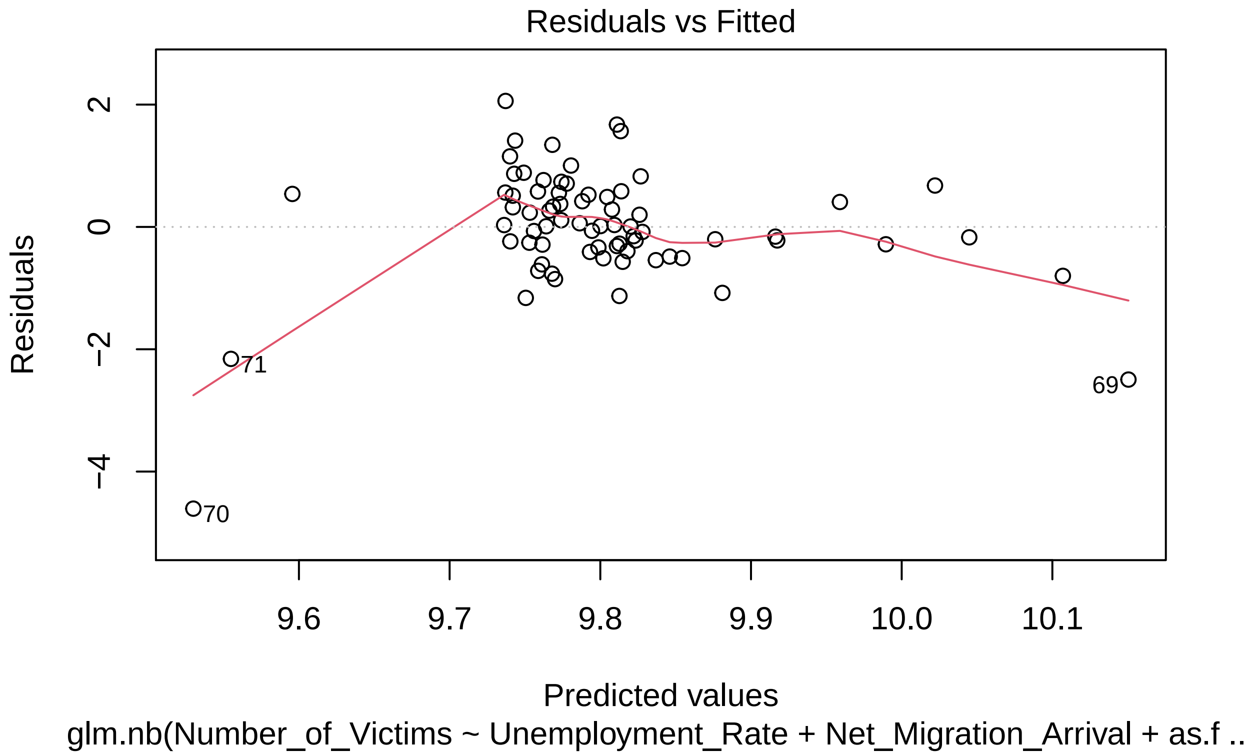
```
##
## Call:
## glm(formula = Number_of_Victims ~ Unemployment_Rate + Net_Migration_Arrival +
##     as.factor(Season), family = "poisson", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -49.437   -3.747   -0.340    5.391   20.421
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               9.703e+00  8.860e-03 1095.13   <2e-16 ***
## Unemployment_Rate        -3.637e-02  1.684e-03  -21.59   <2e-16 ***
## Net_Migration_Arrival     4.947e-05  5.441e-07   90.92   <2e-16 ***
## as.factor(Season)Spring   3.848e-02  2.503e-03   15.38   <2e-16 ***
## as.factor(Season)Summer   6.969e-02  2.487e-03   28.02   <2e-16 ***
## as.factor(Season)Winter   2.525e-02  2.519e-03   10.02   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 18088.6  on 71  degrees of freedom
## Residual deviance:  7142.9  on 66  degrees of freedom
## AIC: 7992.8
##
## Number of Fisher Scoring iterations: 4
```

The summary of the Poisson model shows the residual deviance to be vastly different from the degrees of freedom, so I know the Poisson model is not an adequate fit for the data.

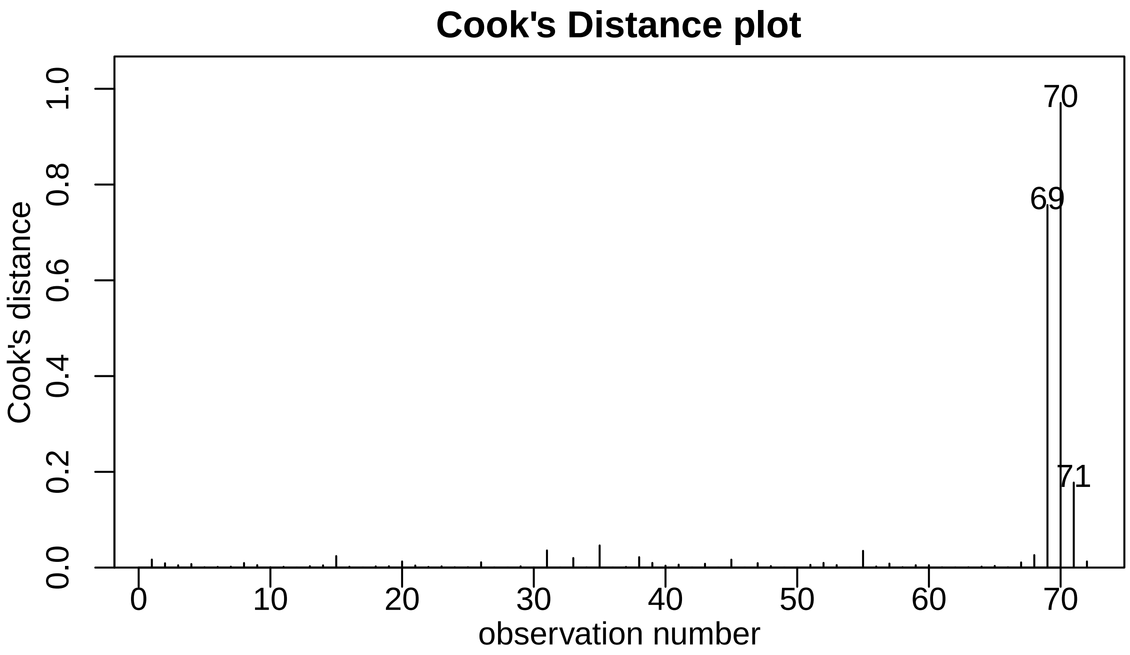
Model Building and Assumption checks

```
# Model Building and Assumption Checks
# Trying a negative binomial glm (because we need to allow for more variance)
nb.mod = glm.nb(Number_of_Victims ~ Unemployment_Rate + Net_Migration_Arrival + as.factor(Season), data
```

```
plot(nb.mod, which = 1)
```



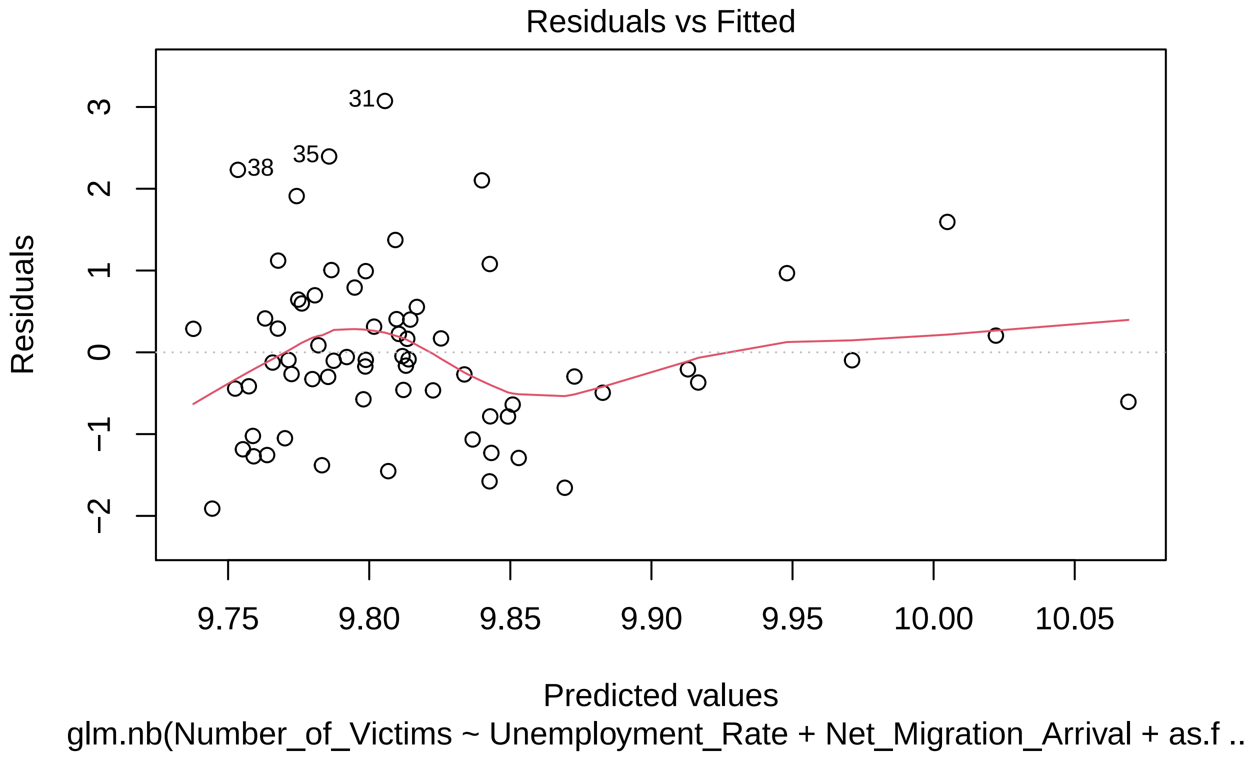
```
cooks20x(nb.mod)
```



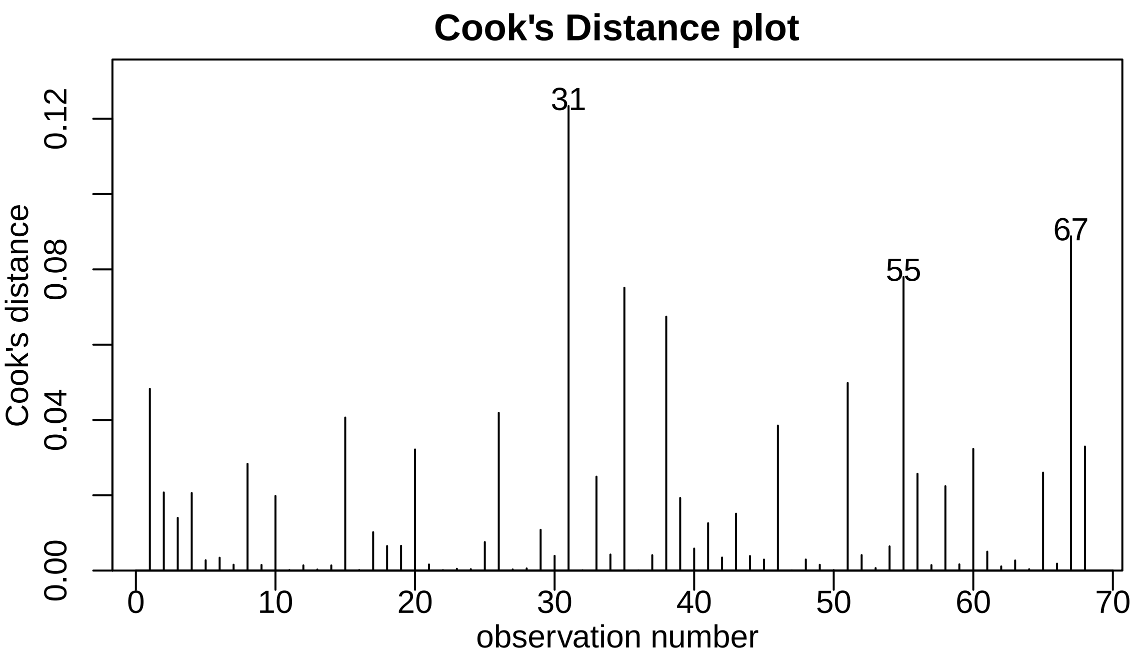
From the Cook's distance plot I can see there are a few influential data points. These data points are more recent months and are due to the COVID-19 outbreak in NZ. I have decided to remove data points later than February 2020, as the outbreak started around March 2020, and it is quite possible the normal trend in crime has been effected by COVID-19.

```
# Removing Covid months
data.1 = data[1:68,]
nb.mod = glm.nb(Number_of_Victims ~ Unemployment_Rate + Net_Migration_Arrival + as.factor(Season), data
```

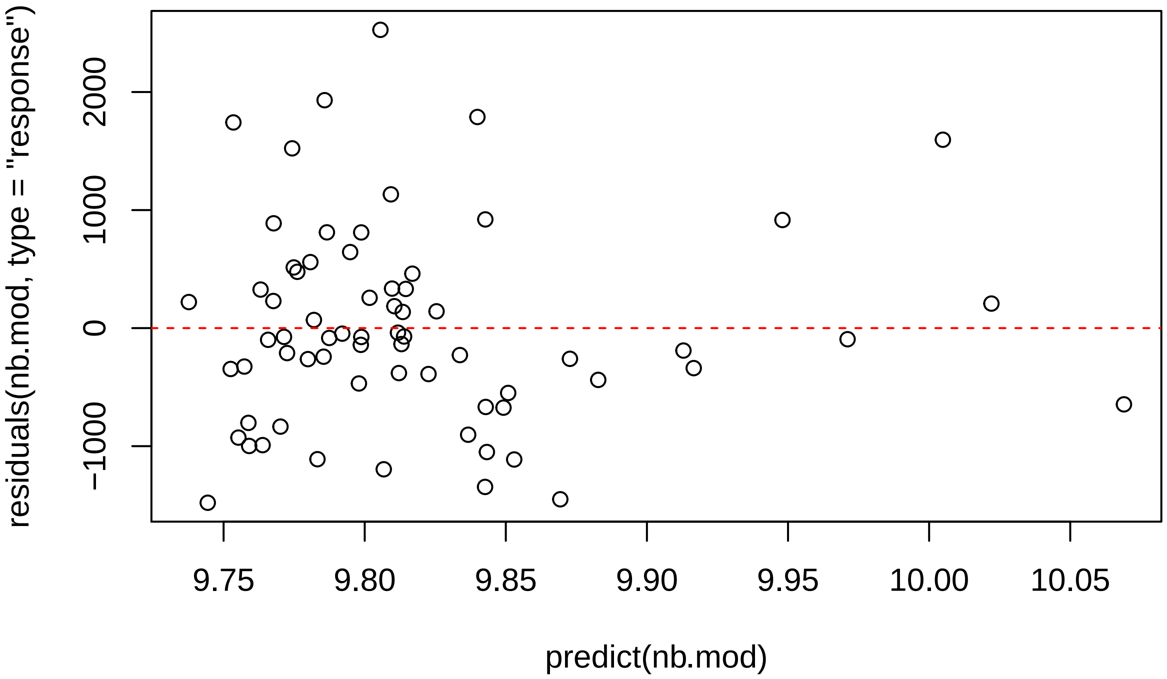
```
plot(nb.mod, which = 1)
```



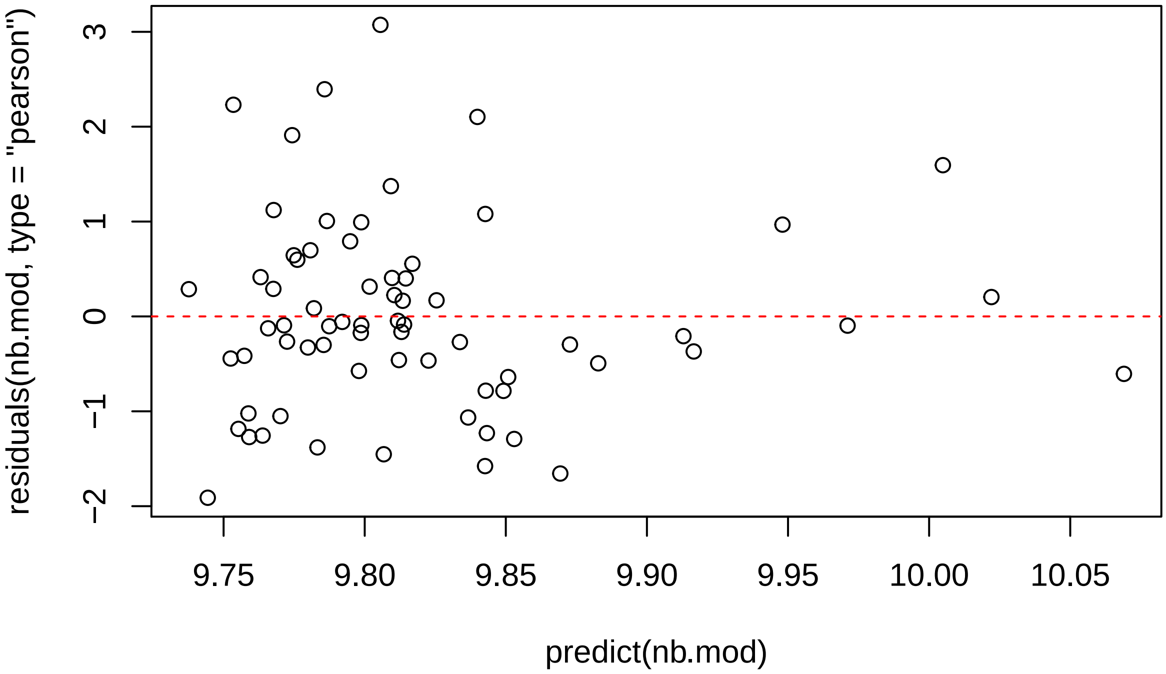
```
cooks20x(nb.mod)
```



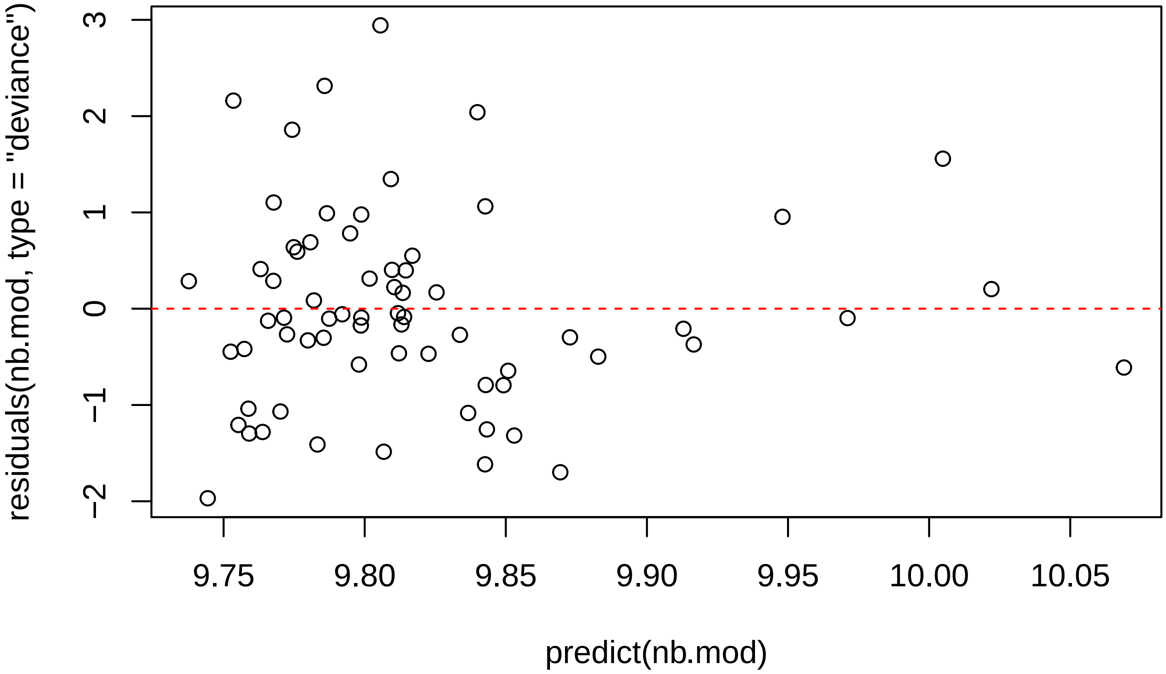
```
plot(predict(nb.mod), residuals(nb.mod, type = "response"))
abline(0,0, lty = 2, col = "red")
```



```
plot(predict(nb.mod), residuals(nb.mod, type = "pearson"))
abline(0,0, lty = 2, col = "red")
```



```r
plot(predict(nb.mod), residuals(nb.mod, type = "deviance"))
abline(0,0, lty = 2, col = "red")
```



We can see that deviance residuals and the pearson residual plots are very similar which indicates are assumptions are reasonable.

```
summary(nb.mod)
```

```
##
## Call:
## glm.nb(formula = Number_of_Victims ~ Unemployment_Rate + Net_Migration_Arrival +
##     as.factor(Season), data = data.1, init.theta = 500.0221811,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9686  -0.5875  -0.1004   0.4462   2.9429
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               9.905e+00  6.361e-02 155.707  < 2e-16 ***
## Unemployment_Rate        -6.149e-02  1.090e-02  -5.641 1.69e-08 ***
## Net_Migration_Arrival     4.193e-05  5.294e-06   7.921 2.37e-15 ***
## as.factor(Season)Spring -6.037e-03  1.597e-02  -0.378    0.705
## as.factor(Season)Summer  2.606e-02  1.608e-02   1.621    0.105
## as.factor(Season)Winter -2.002e-02  1.609e-02  -1.244    0.213
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## (Dispersion parameter for Negative Binomial(500.0222) family taken to be 1)
##
##     Null deviance: 215.123  on 67  degrees of freedom
## Residual deviance:  68.021  on 62  degrees of freedom
## AIC: 1121.4
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  500.0
##           Std. Err.:  88.1
##
##  2 x log-likelihood:  -1107.376
```

```
# Performing Chisq goodness of fit test
1 - pchisq(nb.mod$deviance,nb.mod$df.residual)
```

```
## [1] 0.2797595
```

```
# Rotating Factors
data.1 = within(data.1, {SeasonRotate=factor(Season,levels=c("Summer" ,"Winter" , "Autumn", "Spring"))}
nb.mod = glm.nb(Number_of_Victims ~ Unemployment_Rate + Net_Migration_Arrival + SeasonRotate, data = dat
summary(nb.mod)
```

```
##
## Call:
## glm.nb(formula = Number_of_Victims ~ Unemployment_Rate + Net_Migration_Arrival +
##     SeasonRotate, data = data.1, init.theta = 500.0221811, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9686  -0.5875  -0.1004   0.4462   2.9429
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            9.931e+00  6.478e-02 153.292  < 2e-16 ***
## Unemployment_Rate     -6.149e-02  1.090e-02  -5.641 1.69e-08 ***
## Net_Migration_Arrival  4.193e-05  5.294e-06   7.921 2.37e-15 ***
## SeasonRotateWinter    -4.608e-02  1.544e-02  -2.985  0.00284 **
## SeasonRotateAutumn    -2.606e-02  1.608e-02  -1.621  0.10507
## SeasonRotateSpring    -3.210e-02  1.513e-02  -2.122  0.03385 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(500.0222) family taken to be 1)
##
##     Null deviance: 215.123  on 67  degrees of freedom
## Residual deviance:  68.021  on 62  degrees of freedom
## AIC: 1121.4
##
## Number of Fisher Scoring iterations: 1
##
##
```

```
##                Theta:  500.0
##            Std. Err.:  88.1
##
##  2 x log-likelihood:  -1107.376
```

```
data.1 = within(data.1, {SeasonRotate=factor(Season,levels=c("Spring" ,"Winter" , "Autumn", "Summer"))}
nb.mod = glm.nb(Number_of_Victims ~ Unemployment_Rate + Net_Migration_Arrival + SeasonRotate, data = da
summary(nb.mod)
```

```
##
## Call:
## glm.nb(formula = Number_of_Victims ~ Unemployment_Rate + Net_Migration_Arrival +
##     SeasonRotate, data = data.1, init.theta = 500.0221811, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9686  -0.5875  -0.1004   0.4462   2.9429
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            9.899e+00  6.445e-02 153.578  < 2e-16 ***
## Unemployment_Rate     -6.149e-02  1.090e-02  -5.641 1.69e-08 ***
## Net_Migration_Arrival  4.193e-05  5.294e-06   7.921 2.37e-15 ***
## SeasonRotateWinter    -1.399e-02  1.537e-02  -0.910   0.3628
## SeasonRotateAutumn     6.037e-03  1.597e-02   0.378   0.7054
## SeasonRotateSummer     3.210e-02  1.513e-02   2.122   0.0339 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(500.0222) family taken to be 1)
##
##     Null deviance: 215.123  on 67  degrees of freedom
## Residual deviance:  68.021  on 62  degrees of freedom
## AIC: 1121.4
##
## Number of Fisher Scoring iterations: 1
##
##
##                Theta:  500.0
##            Std. Err.:  88.1
##
##  2 x log-likelihood:  -1107.376
```

```
data.1 = within(data.1, {SeasonRotate=factor(Season,levels=c("Winter", "Summer", "Autumn", "Spring"))})
nb.mod = glm.nb(Number_of_Victims ~ Unemployment_Rate + Net_Migration_Arrival + SeasonRotate, data = da
summary(nb.mod)
```

```
##
## Call:
## glm.nb(formula = Number_of_Victims ~ Unemployment_Rate + Net_Migration_Arrival +
##     SeasonRotate, data = data.1, init.theta = 500.0221811, link = log)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.9686  -0.5875  -0.1004   0.4462   2.9429
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            9.885e+00  6.433e-02 153.665  < 2e-16 ***
## Unemployment_Rate     -6.149e-02  1.090e-02  -5.641 1.69e-08 ***
## Net_Migration_Arrival  4.193e-05  5.294e-06   7.921 2.37e-15 ***
## SeasonRotateSummer     4.608e-02  1.544e-02   2.985  0.00284 **
## SeasonRotateAutumn     2.002e-02  1.609e-02   1.244  0.21346
## SeasonRotateSpring     1.399e-02  1.537e-02   0.910  0.36278
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for Negative Binomial(500.0222) family taken to be 1)
##
##     Null deviance: 215.123  on 67  degrees of freedom
## Residual deviance:  68.021  on 62  degrees of freedom
## AIC: 1121.4
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  500.0
##           Std. Err.:  88.1
##
##  2 x log-likelihood:  -1107.376
```

```
(exp(confint(nb.mod)) - 1)*100
```

```
## Waiting for profiling to be done...
```

```
##                               2.5 %        97.5 %
## (Intercept)            1.729411e+06  2.227537e+06
## Unemployment_Rate     -7.961472e+00 -3.922603e+00
## Net_Migration_Arrival  3.159086e-03  5.232081e-03
## SeasonRotateSummer     1.592976e+00  7.934758e+00
## SeasonRotateAutumn    -1.143854e+00  5.292374e+00
## SeasonRotateSpring    -1.601698e+00  4.509988e+00
```

Methods and Assumptions Check

The response variable Number_of_Victims is a count, therefore I first fit a generalised linear model with a Poisson reponse distribution. However, after comparing the residual deviance and the degrees of freedom for this model, I knew it was not an adequate fit as the two numbers were very different. The model needed to allow for more variance so I fitted a GLM with a Negative Binomial reponse distribution.

When checking the residuals and cooks plot, I could see there were a couple of influential data points at 69 and 70. These data points were due to the COVID-19 outbreak as the date of these data points was March and April of 2020. I decided to remove all data points after February 2020 as I believe the COVID-19 outbreak has had an effect on the data used in this study. All other assumptions were satisfied, and we can trust the results from this Negative Binomial model (P-value = 0.28).

Our final model is:

$$log(\mu_i) = \beta_0 + \beta_1 \times URate_i + \beta_2 \times Net\_Mig_i + \beta_3 \times Summer_i + \beta_4 \times Autumn_i + \beta_5 \times Spring_i$$

Where μ_i is the mean number of Victimisations with a Negative Binomial distribution, at a given unemployment rate, net migration number, and in Winter. $URate_i$ is the unemployment rate for observation i. Net_Mig_i is the net migration into NZ for observation i. $Summer_i$, $Autumn_i$ and $Spring_i$ are dummy variables which take the value 1 if the observation is in that particular season, otherwise it is 0.

Executive Summary

I was interested in how the level of crime in NZ is effected by unemployment rate, immigration, and the season.

I can conclude that the higher the unemployment rate and net migration into NZ, the higher the number of victimisations. The mean number of Victimisations will