Crime Rate Dataset analysis

Patrick Cullinane

05/16/2022

Crime Rate Dataset

```
library(dplyr)
library(ggplot2)
library(car)
library(ggpubr)
library(tidyverse)
library(rstatix)
library(ggstatsplot)

setwd("C:/Users/Laura/Documents/code/MA-541")
df <- read.csv("Crime_R.csv")


#str(df)
```

Introduction

The Crime Rate dataset is comprised of data crime rate and associated variables collected in the United States at two different periods of time. The dataset is broken down into 27 columns and 47 rows. The dataset was developed by the University of Sheffield

```
dim(df)
```

```
## [1] 47 27
```

```
names(df)
```

```
##  [1] "CrimeRate"          "Youth"              "Southern"
##  [4] "Education"          "ExpenditureYear0"   "LabourForce"
##  [7] "Males"              "MoreMales"          "StateSize"
## [10] "YouthUnemployment"  "MatureUnemployment" "HighYouthUnemploy"
## [13] "Wage"               "BelowWage"          "CrimeRate10"
## [16] "Youth10"            "Education10"        "ExpenditureYear10"
## [19] "LabourForce10"      "Males10"            "MoreMales10"
## [22] "StateSize10"        "YouthUnemploy10"    "MatureUnemploy10"
## [25] "HighYouthUnemploy10" "Wage10"            "BelowWage10"
```

The first 13 columns consist of crime rate data and other measurements taken at a point in time, and the next set consists of the same data taken a decade later. Additionally there is a column called Southern which consists of a binary variable, 1 if a Southern state 0 otherwise. The Southern column applys to both sets of columns within the dataset. We are not told the exact date at which the data has been collected. Overall the data is a mixture of discrete, binary, continuous variables.

```
# split data into year 0 and year + 10

df0 <- df %>%
  select(-ends_with('10'))
df10 <- df %>%
  select(ends_with('10'))

head(df0,2)
```

```
##   CrimeRate Youth Southern Education ExpenditureYear0 LabourForce Males
## 1      45.5   135        0      12.4               69         540   965
## 2      52.3   140        0      10.9               55         535  1045
##   MoreMales StateSize YouthUnemployment MatureUnemployment HighYouthUnemploy
## 1         0         6                80                 22                 1
## 2         1         6               135                 40                 1
##   Wage BelowWage
## 1  564       139
## 2  453       200
```

```
tail(df0,2)
```

```
##    CrimeRate Youth Southern Education ExpenditureYear0 LabourForce Males
## 46     157.7   136        0      15.1              149         577   994
## 47     161.8   131        0      13.2              160         631  1071
##    MoreMales StateSize YouthUnemployment MatureUnemployment HighYouthUnemploy
## 46         0       157               102                 39                 0
## 47         1         3               102                 41                 0
##    Wage BelowWage
## 46  673       167
## 47  674       152
```

Experiment 1

The first question we will examine will be whether there is a relationship between Males per 1000 females and states classified as Southern. To accomplish this we will use 3 columns; Southern, Males, and Males10. As mentioned previously Southern is a binary variable while Males and Males10 are discrete variables. Males and Males10 both refer to the number of males per 1000 females in counted in a US State.

To get an idea of the make-up of the data we look at the summary statistics.

```
str(df[,c("Southern","Males","Males10")])
```

```
## 'data.frame':    47 obs. of  3 variables:
##  $ Southern: int  0 0 1 1 0 0 0 0 0 0 ...
##  $ Males   : int  965 1045 962 968 989 972 984 977 968 1024 ...
##  $ Males10 : int  974 1039 959 983 989 983 993 973 968 1024 ...
```

```
summary(df[,c("Southern","Males","Males10")])
```

```
##     Southern          Males           Males10
##  Min.   :0.0000   Min.   : 934.0   Min.   : 935.0
##  1st Qu.:0.0000   1st Qu.: 964.5   1st Qu.: 969.5
##  Median :0.0000   Median : 977.0   Median : 983.0
##  Mean   :0.3404   Mean   : 983.0   Mean   : 986.9
##  3rd Qu.:1.0000   3rd Qu.: 992.0   3rd Qu.: 994.0
##  Max.   :1.0000   Max.   :1071.0   Max.   :1079.0
```

Next let's look at how the data's shape using some plots.

```
df1 <- df[,c("Southern","Males")]
df2 <- df[,c("Southern","Males10")]
stem(df$Southern)
```

```
##
##   The decimal point is 1 digit(s) to the left of the |
##
##    0 | 0000000000000000000000000000000
##    2 |
##    4 |
##    6 |
##    8 |
##   10 | 0000000000000000
```

```
stem(df$Males)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    92 | 48
##    94 | 803356
##    96 | 24445688992223478
##    98 | 1244556690468
##   100 | 228
##   102 | 498
##   104 | 59
##   106 | 1
```

```
stem(df$Males10)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    92 | 5
##    94 | 5899269
##    96 | 28890134688
##    98 | 022233377992333359
##   100 | 131
##   102 | 4499
##   104 | 08
##   106 | 9
```

We can see from the stem plots that we have more Southern states. We should note that although it is not explicity stated in the data what constitutes a Southern state the data seems to align with the US census bureau’s definition of a Southern state. The US census counts 16 states in total as Southern States, which appears to align with our data. Alternatively we can see there are 31 “0” states which in total make 47 states counted in this dataset. We are not told what comprise the 31 non-Southern states.

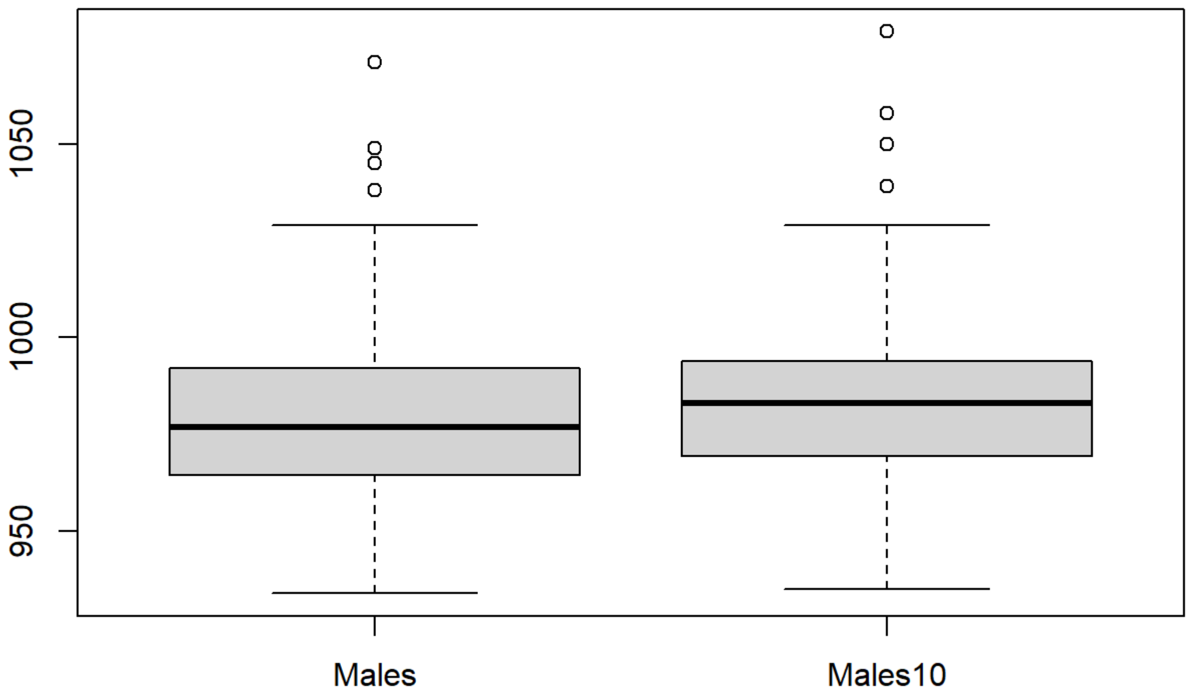
```
sum(df$Southern == 1)
```

```
## [1] 16
```

```
sum(df$Southern == 0)
```

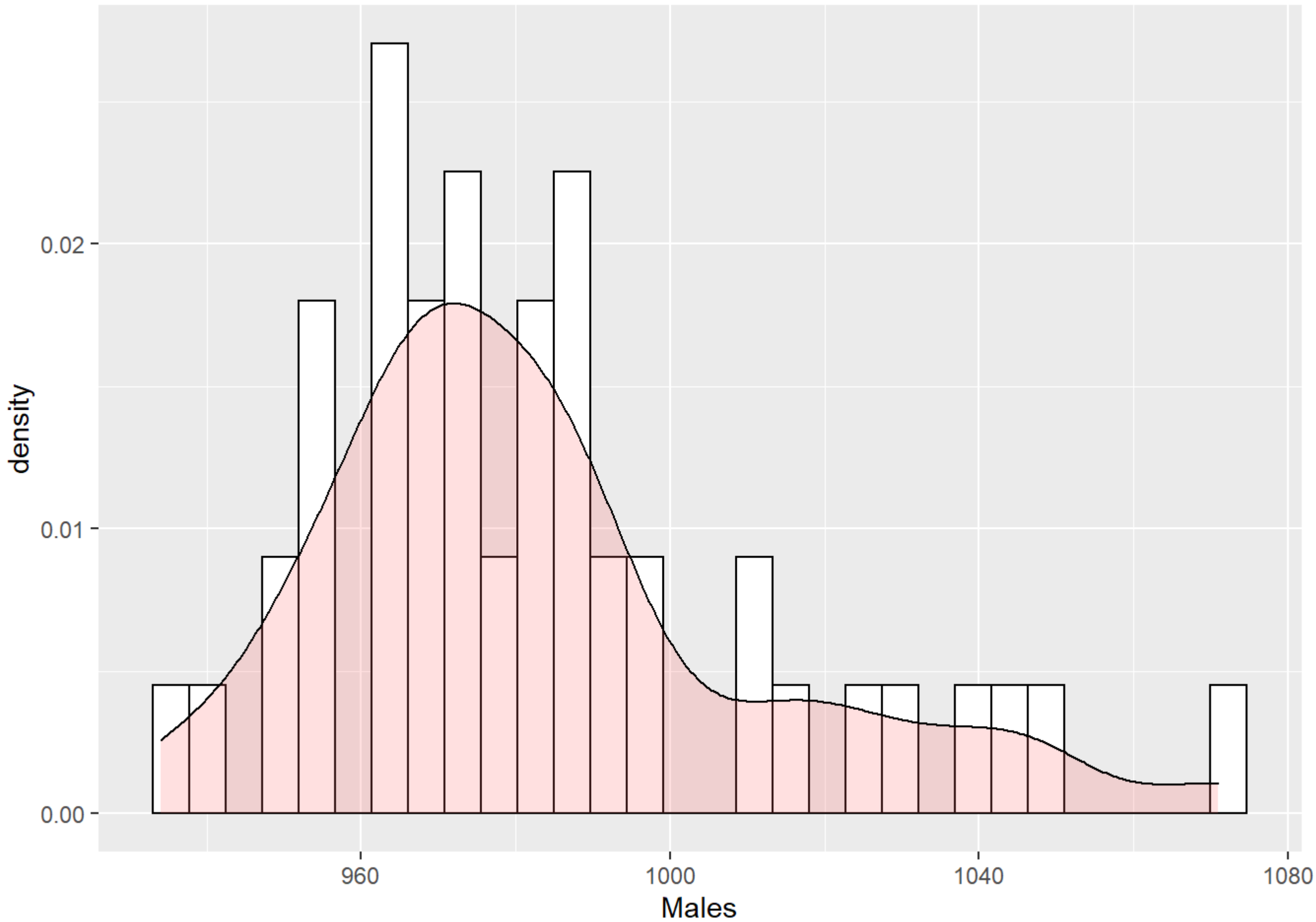
```
## [1] 31
```

```
boxplot(df[,c("Males","Males10")])
```

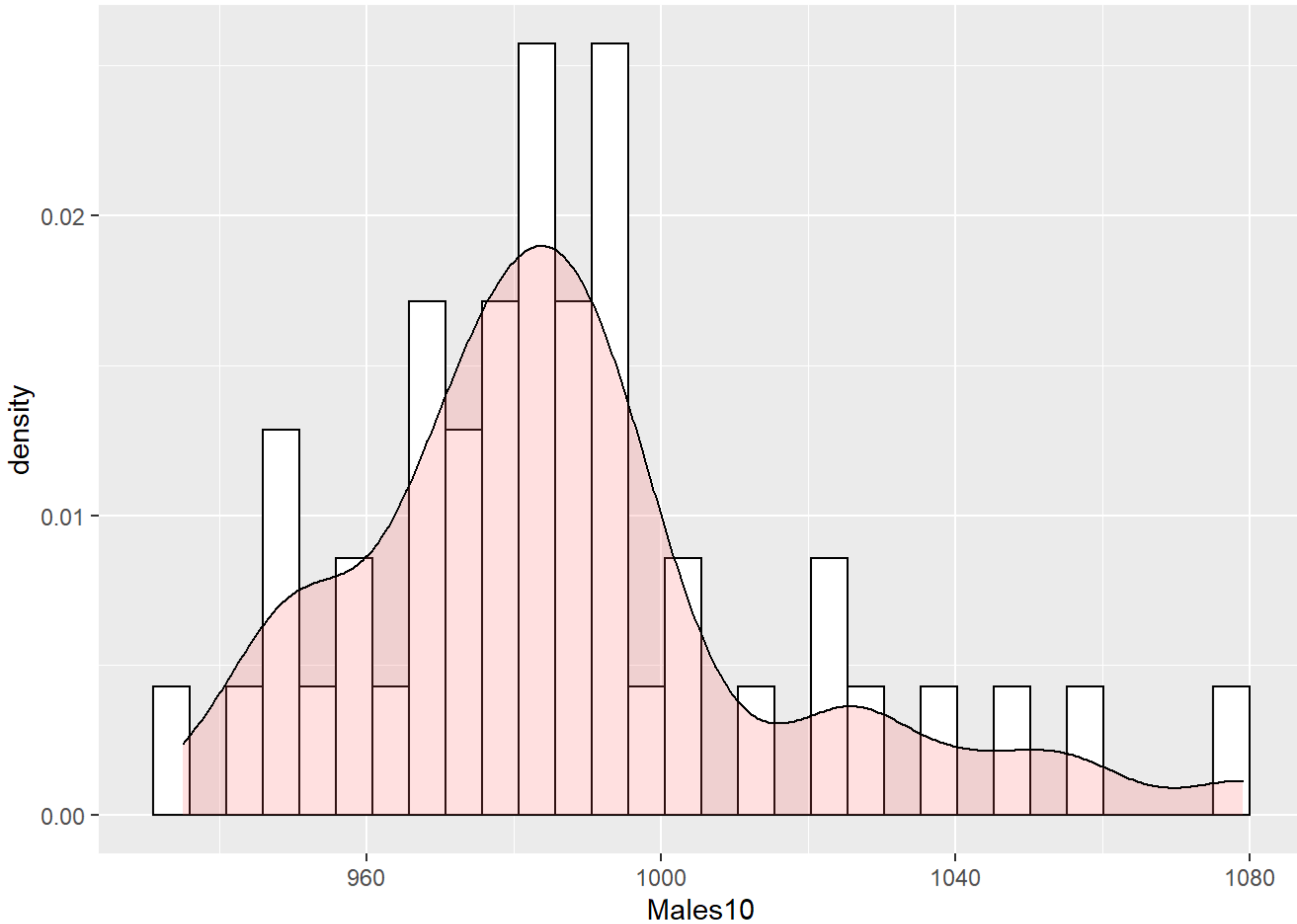


From the stem plots we can see that Males and Males10 both appear to be rightly skewed. We can examaine this in more detail using a histogram.

```
ggplot(df, aes(x=Males)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")
```




```r
ggplot(df, aes(x=Males10)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")
```



More detailed examination with our histograms and density plots again show us that data may be skewed to the right. This will be important to note as we further examine whether there is a difference in means between groups of data.

With this intuition we will examine using anova whether there is a significant difference in group means. To accomplish this we must first encode our the Southern column to match Males and Males10, and then stack the data frames on top of eachother. Next we will run an anova test on the data.

Next we formulate the following hypothesis:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$$

$$H_1 : \alpha_l \neq \alpha_k$$

The null hypothesis states that all groups measured will be the same while the alternative hypothesis states there is a difference in atleast two of the groups.

To proceed further we need to make four groups of data to compare. To do this we will map a new variable from Southern to Males10. We will stick with the current mapping of the Males and Southern column which is “1” if Southern otherwise “0”.

In the second group we will use “3” to denote a Southern state and “2” otherwise. In this way we will have 4 groups of Males. At this point we will have four groups: 0:Non-Southern Males, 1:Southern Males, 2:Non-Southern Males10, 3:Southern Males. To run the test we then stack the data on top of eachother to create one long column of Males and Southern variables. Additionally we establish that we will reject the null hypothesis at an alpha of 0.05.

```
df2$Southern <- ifelse(df2$Southern == 0, 2,3)
names(df2)[2] <- "Males"

df_stack <- rbind(df1,df2)
df_stack$Southern <- as.factor(df_stack$Southern)

aov_test <- aov(Males ~ Southern, data=df_stack)
summary(aov_test)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Southern     3   8765  2921.8   3.627  0.016 *
## Residuals   90  72501   805.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

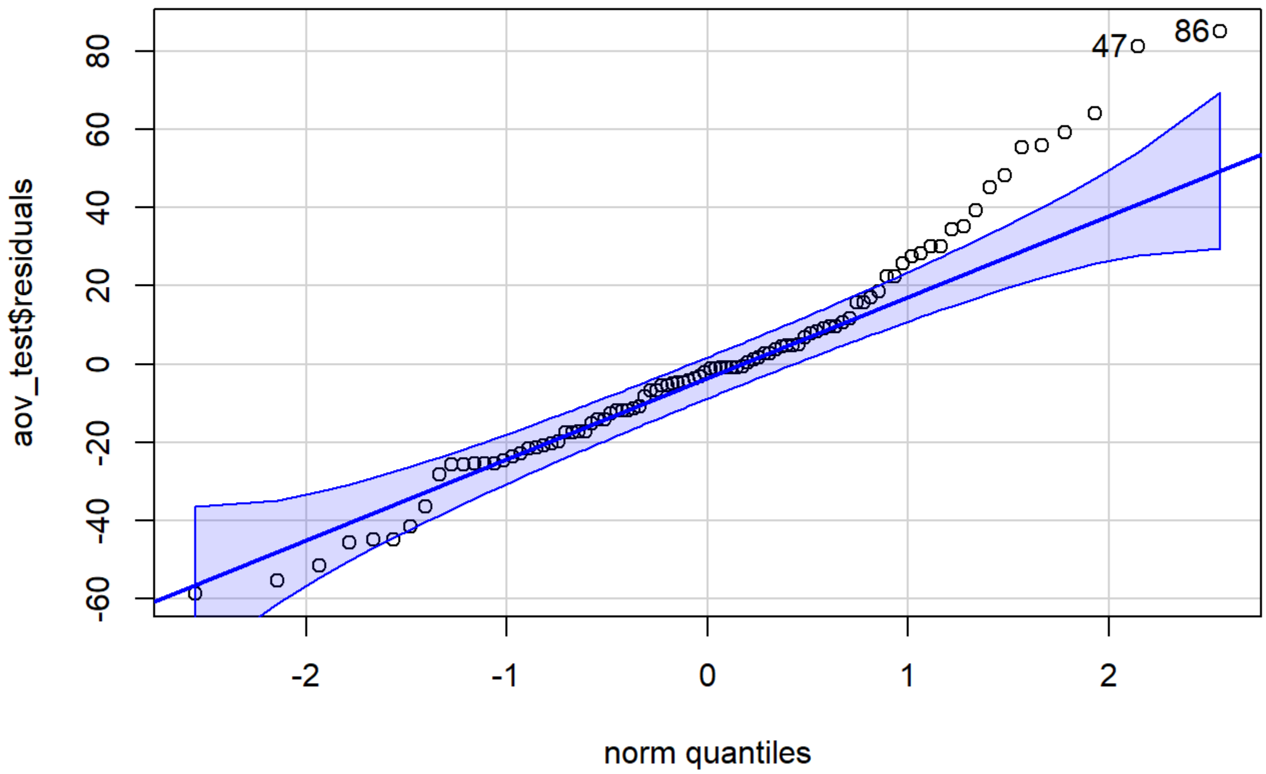
As we can see from our test there appears to be a signficant difference between atleast two of the groups as the p-value is below our threshold of 0.05. Our test does not tell us which groups are different so we need to perform further tests to gain more insight.

Our next step is to perform post-hoc teting to see which groups are different. First though we would like to examine our results in more detail to see what tests we need to perform.

We examine the residuals of the anova results to see if our data meets the normality assumption needed for tukey or bonferroni. We will do this graphically through a histogram of the residuals and qqplot.

```
hist(aov_test$residuals)
```


```
qqPlot(aov_test$residuals)
```



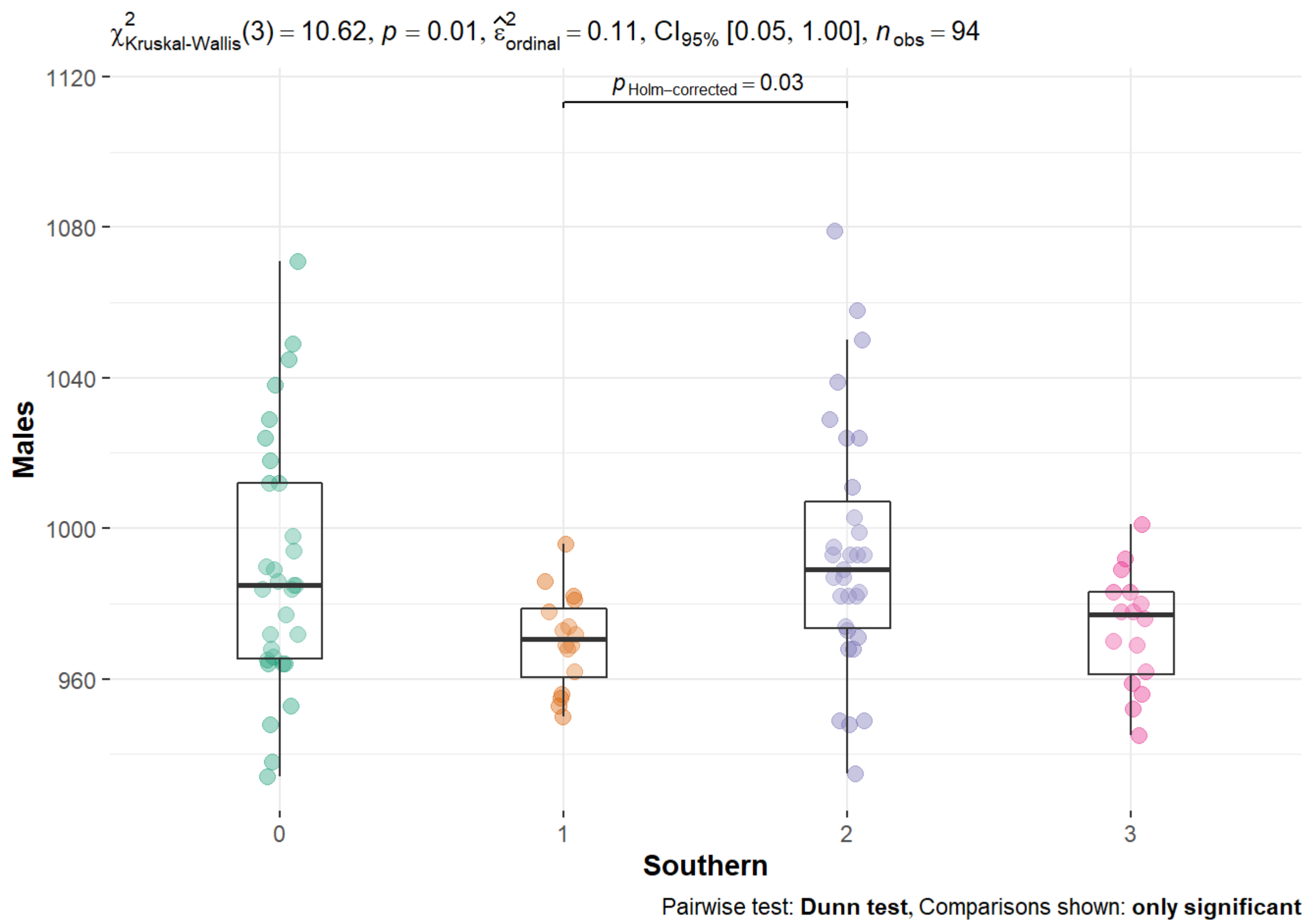

```
## [1] 86 47
```

Graphically it appears that the residuals are not nornally distributed but we need to examine this more formally with a kruskal test

```
kruskal.test(Males ~ Southern, data = df_stack)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Males by Southern
## Kruskal-Wallis chi-squared = 10.624, df = 3, p-value = 0.01394
```

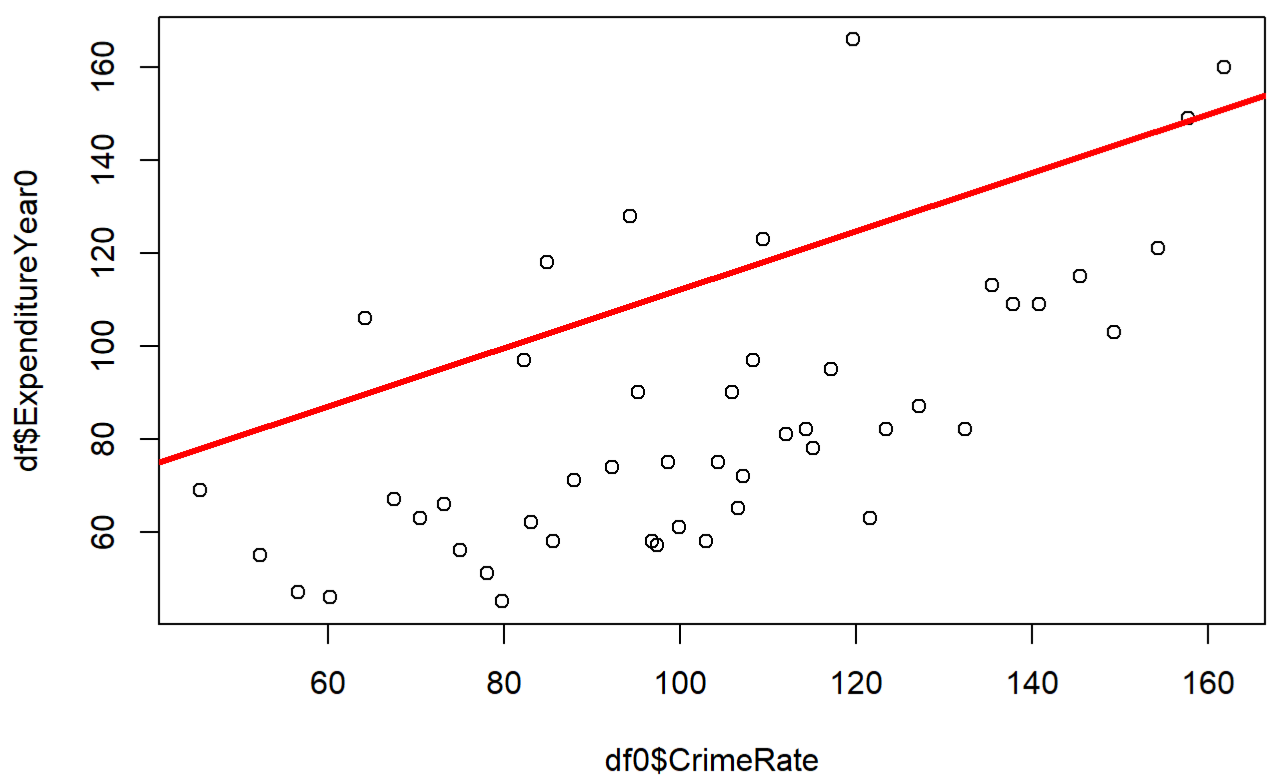
```
ggbetweenstats(
  data = df_stack,
  x = "Southern",
  y = "Males",
  type = "nonparametric", # ANOVA or Kruskal-Wallis
  plot.type = "box",
  pairwise.comparisons = TRUE,
  pairwise.display = "significant",
  centrality.plotting = FALSE,
  bf.message = FALSE
)
```



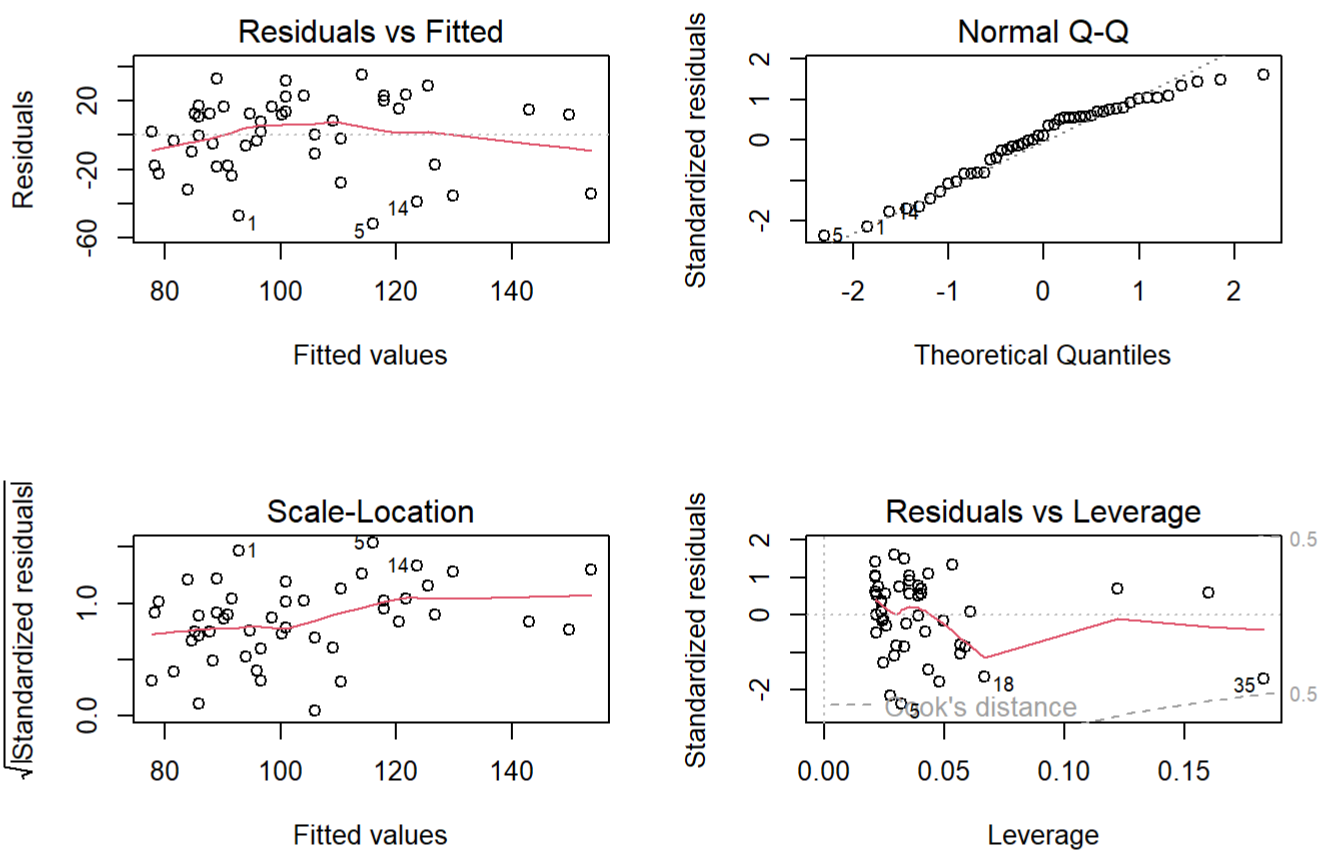

```
lm.fit <- lm(CrimeRate ~ ExpenditureYear0, data=df0)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = CrimeRate ~ ExpenditureYear0, data = df0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.802 -17.477   2.174  15.728  35.183
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       49.4067     9.9479   4.967 1.03e-05 ***
## ExpenditureYear0   0.6283     0.1106   5.680 9.29e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.29 on 45 degrees of freedom
## Multiple R-squared:  0.4176, Adjusted R-squared:  0.4046
## F-statistic: 32.26 on 1 and 45 DF,  p-value: 9.293e-07
```

```
plot(df0$CrimeRate, df$ExpenditureYear0)
abline(lm.fit, lwd=3, col="red")
```



```
par(mfrow = c(2,2))
plot(lm.fit)
```




```
model <- lm(CrimeRate~.,data=df0)
summary(model)
```

```
##
## Call:
## lm(formula = CrimeRate ~ ., data = df0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.204 -10.557   2.919  10.391  32.707
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -258.30363  192.43539  -1.342  0.18866
## Youth                 0.86498    0.35319   2.449  0.01980 *
## Southern              0.56966   12.04365   0.047  0.96256
## Education             6.43119    3.75033   1.715  0.09575 .
## ExpenditureYear0      0.71271    0.20199   3.528  0.00125 **
## LabourForce           0.10771    0.12281   0.877  0.38680
## Males                -0.18383    0.23656  -0.777  0.44265
## MoreMales            17.33920   15.83577   1.095  0.28147
## StateSize            -0.09895    0.11444  -0.865  0.39349
## YouthUnemployment    -0.09173    0.46132  -0.199  0.84361
## MatureUnemployment    0.68776    0.99491   0.691  0.49423
## HighYouthUnemploy    -4.49806   10.82134  -0.416  0.68035
## Wage                  0.19189    0.08950   2.144  0.03950 *
## BelowWage             0.55336    0.20693   2.674  0.01156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.17 on 33 degrees of freedom
## Multiple R-squared:  0.6842, Adjusted R-squared:  0.5598
## F-statistic:   5.5 on 13 and 33 DF,  p-value: 3.616e-05
```

```
lm.fit <- lm(
  CrimeRate ~
    Education + Youth + Wage + BelowWage
    + ExpenditureYear0, data=df0)

lm.fit2 <- lm(
  CrimeRate ~
    Youth + Wage + BelowWage
    + ExpenditureYear0, data=df0
)

summary(lm.fit)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Education + Youth + Wage + BelowWage +
##     ExpenditureYear0, data = df0)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -43.32 -12.69   3.12  10.78  32.52
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -338.74486   90.91882  -3.726 0.000588 ***
## Education           4.72597    3.05412   1.547 0.129450
## Youth               0.78508    0.29627   2.650 0.011387 *
## Wage                0.20208    0.08097   2.496 0.016679 *
## BelowWage           0.55952    0.15831   3.534 0.001029 **
## ExpenditureYear0    0.69979    0.15487   4.519  5.2e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.55 on 41 degrees of freedom
## Multiple R-squared:  0.6326, Adjusted R-squared:  0.5878
## F-statistic: 14.12 on 5 and 41 DF,  p-value: 4.872e-08
```

```
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Youth + Wage + BelowWage + ExpenditureYear0,
##     data = df0)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46.02 -12.06   3.09  12.70  33.83
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -265.89320   79.06065  -3.363 0.001653 **
## Youth               0.76376    0.30082   2.539 0.014913 *
## Wage                0.21169    0.08206   2.580 0.013475 *
## BelowWage           0.49014    0.15432   3.176 0.002797 **
## ExpenditureYear0    0.66540    0.15579   4.271 0.000109 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.86 on 42 degrees of freedom
## Multiple R-squared:  0.6111, Adjusted R-squared:  0.5741
## F-statistic:  16.5 on 4 and 42 DF,  p-value: 3.367e-08
```

```
anova(lm.fit, lm.fit2)
```

```
## Analysis of Variance Table
##
## Model 1: CrimeRate ~ Education + Youth + Wage + BelowWage + ExpenditureYear0
## Model 2: CrimeRate ~ Youth + Wage + BelowWage + ExpenditureYear0
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     41 14110
## 2     42 14934 -1   -824.05 2.3945 0.1294
```

```
lm.fit3 <- lm(
  CrimeRate ~
    + log10(Youth)
    + log10(Wage)
    + log10(BelowWage)
    + log10(ExpenditureYear0), data=df0
)

lm.fit4 <- lm(
  CrimeRate ~
  + log10(Youth)
  + log10(BelowWage)
  + log10(ExpenditureYear0), data=df0
)

summary(lm.fit3)
```

```
##
## Call:
## lm(formula = CrimeRate ~ +log10(Youth) + log10(Wage) + log10(BelowWage) +
##     log10(ExpenditureYear0), data = df0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.094 -12.065   0.593  12.248  27.813
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -1719.42     364.93  -4.712 2.70e-05 ***
## log10(Youth)              248.00      89.06   2.785 0.008003 **
## log10(Wage)               168.54      75.88   2.221 0.031799 *
## log10(BelowWage)          218.64      56.90   3.843 0.000405 ***
## log10(ExpenditureYear0)   176.39      29.61   5.958 4.57e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.07 on 42 degrees of freedom
## Multiple R-squared:  0.6815, Adjusted R-squared:  0.6512
## F-statistic: 22.47 on 4 and 42 DF,  p-value: 5.63e-10
```

```r
summary(lm.fit4)
```

```
##
## Call:
## lm(formula = CrimeRate ~ +log10(Youth) + log10(BelowWage) + log10(ExpenditureYear0),
##     data = df0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.539 -12.120   3.539  11.659  28.879
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1032.55     202.39  -5.102 7.25e-06 ***
## log10(Youth)               195.90      89.76   2.182  0.03458 *
## log10(BelowWage)           134.11      44.18   3.035  0.00407 **
## log10(ExpenditureYear0)   215.46      24.88   8.659 5.64e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.83 on 43 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6193
## F-statistic: 25.94 on 3 and 43 DF,  p-value: 9.768e-10
```

```r
anova(lm.fit3, lm.fit4)
```

```
## Analysis of Variance Table
##
## Model 1: CrimeRate ~ +log10(Youth) + log10(Wage) + log10(BelowWage) +
##     log10(ExpenditureYear0)
## Model 2: CrimeRate ~ +log10(Youth) + log10(BelowWage) + log10(ExpenditureYear0)
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     42 12231
## 2     43 13668 -1   -1436.5 4.9328 0.0318 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

References:

1. www.statstutor.ac.uk
2. US Census