**PROG8430**

**MLR - Example**

**Author:** David Marsh, 90131080

**Course:** PROG8430

**Background**

The causes of differing neighbourhood crime rates in urban areas is a concern for law enforcement, policy makers and the public at large. This analysis will seek to discover some variables that help explain the differing crime rates in different Denver neighbourhoods.

**Data Source**

Data was obtained from each of several neighbourhoods in Denver which includes information on population, demographics and crime rates.

**Data Transformation and Cleaning (Description)**

**Dates**

Dates were transformed from text to dates and then to Julien calendar dates so ratios could be constructed.

**Income**

Transformed Income from factor to numeric (needed to use some substrings).

**Counsellor**

The data identifying the city counsellor responsible for the area was transformed to four dummy variables.

I have included all the code used for transformations in the Appendix.

**Descriptive Data Analysis**

Pop PopChg Child Lunch IncChg

Min. : 2.100 Min. :-3.300 Min. :10.40 Min. : 5.70 Min. :11.70

1st Qu.: 4.200 1st Qu.: 1.950 1st Qu.:21.00 1st Qu.:29.40 1st Qu.:23.60

Median : 7.300 Median : 4.600 Median :29.10 Median :57.20 Median :26.30

Mean : 7.165 Mean : 7.265 Mean :27.73 Mean :50.99 Mean :26.41

3rd Qu.: 9.350 3rd Qu.: 7.950 3rd Qu.:34.00 3rd Qu.:74.80 3rd Qu.:29.35

Max. :18.700 Max. :68.600 Max. :41.50 Max. :88.50 Max. :39.80

Income Crime CrmChg Year Counsel

Min. :22904 Min. : 27.10 Min. :-45.60 Min. : 1158 Johnson :10

1st Qu.:35113 1st Qu.: 62.05 1st Qu.:-25.05 1st Qu.: 7420 Jones : 8

Median :44819 Median : 77.50 Median :-13.40 Median :14333 Langlois:13

Mean :43368 Mean : 94.78 Mean :-13.08 Mean :12952 Williams:12

3rd Qu.:50534 3rd Qu.:117.10 3rd Qu.: -2.65 3rd Qu.:18206

Max. :67432 Max. :258.00 Max. : 45.40 Max. :23318

John Jone Lang Will

Min. :0.0000 Min. :0.000 Min. :0.0000 Min. :0.0000

1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000

Median :0.0000 Median :0.000 Median :0.0000 Median :0.0000

Mean :0.2326 Mean :0.186 Mean :0.3023 Mean :0.2791

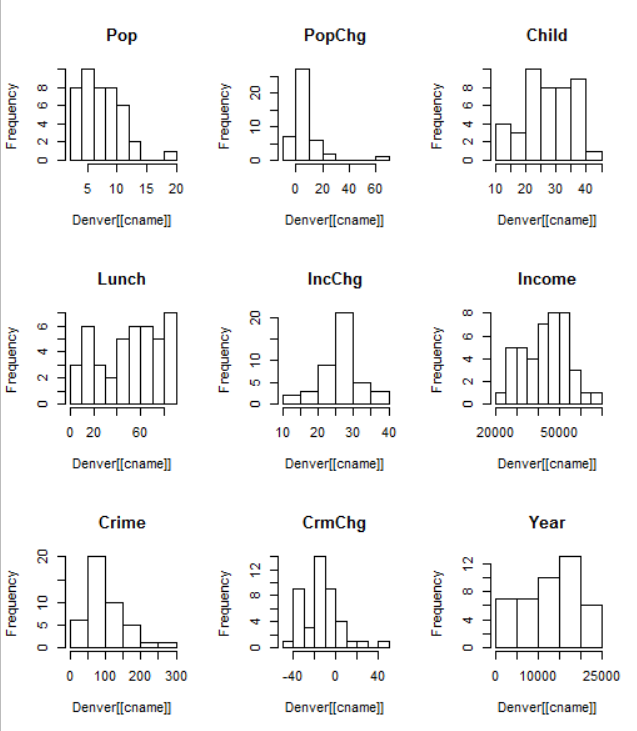
3rd Qu.:0.0000 3rd Qu.:0.000 3rd Qu.:1.0000 3rd Qu.:1.0000

Max. :1.0000 Max. :1.000 Max. :1.0000 Max. :1.0000

From the summary statistics we conclude that the transformation of counsellors, dates and income worked properly.

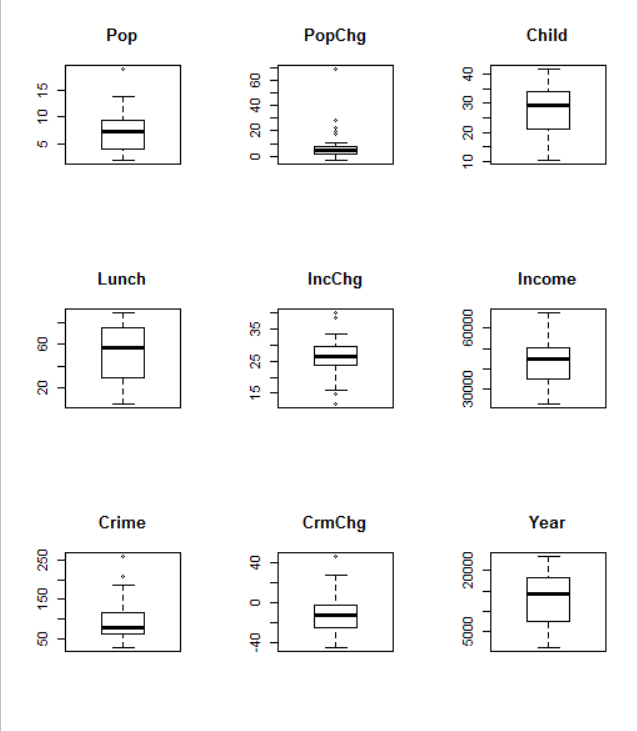
Also, all of the data looks reasonable (that is, there are no values that seem like they are necessarily wrong).

Population change seems very tightly clustered as does population but both seem to have some extreme values.



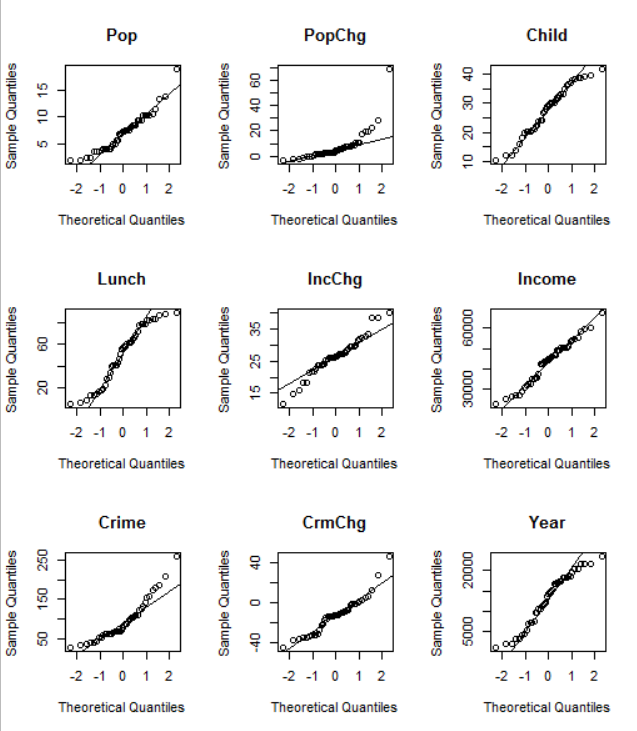
The histograms show reasonable distributions. CrmChg, Pop and PopChg may have some extreme values that will influence the outcomes.

**Outlier**



There seem to be outliers in Population and Population Change. I will leave them in until for now carefully to determine the effect on the outcome model.

**Exploratory Data Analysis**



**statistic p.value**

Pop 0.9375896 0.02134848

PopChg 0.6237796 0.00000000287914

Child 0.9619144 0.1628618

Lunch 0.9332436 0.01504336

IncChg 0.9674727 0.2580618

Income 0.9738707 0.4260645

Crime 0.9014753 0.001387222

CrmChg 0.9373063 0.02086331

Year 0.9475033 0.04835184

John 0.5237307 0.0000000001290699

Jone 0.4747685 0.0000000000331353

Lang 0.5776868 0.0000000006479525

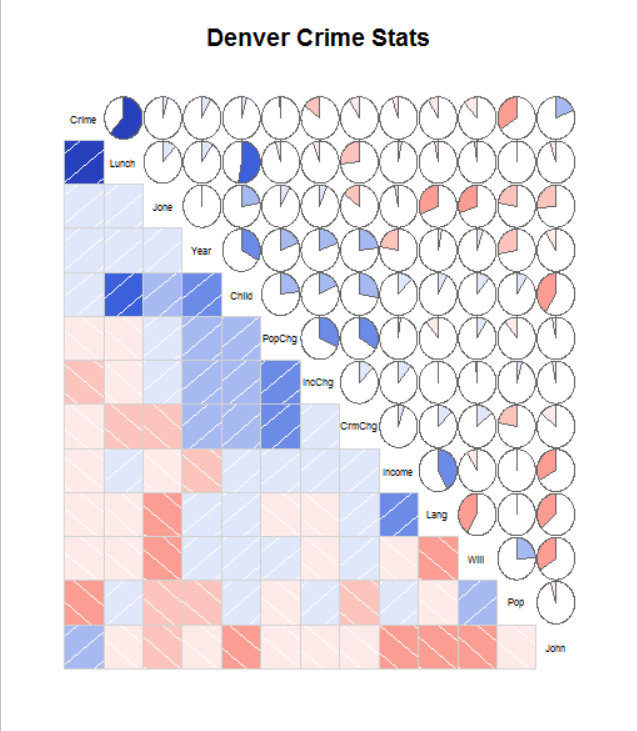
Will 0.5619874 0.0000000003996874

Child, Income Change and income all appear to be approximately normally distributed.

Also, PopChg really does seem to have a significant outlier.

Correlations

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Pop | PopChg | Child | Lunch | IncChg | Income | Crime | CrmChg | Year | John | Jone | Lang | Will |
| Pop | 1.00 | -0.06 | 0.12 | -0.02 | 0.09 | 0.00 | -0.36 | -0.31 | -0.18 | -0.14 | -0.24 | 0.01 | 0.33 |
| PopChg | -0.06 | 1.00 | 0.13 | 0.20 | -0.10 | 0.03 | 0.17 | -0.10 | 0.04 | 0.10 | 0.00 | 0.01 | -0.10 |
| Child | 0.12 | 0.13 | 1.00 | 0.60 | 0.12 | 0.11 | 0.09 | 0.20 | 0.35 | -0.38 | 0.22 | 0.04 | 0.13 |
| Lunch | -0.02 | 0.20 | 0.60 | 1.00 | -0.08 | 0.06 | 0.70 | -0.17 | 0.14 | -0.09 | 0.14 | -0.01 | -0.03 |
| IncChg | 0.09 | -0.10 | 0.12 | -0.08 | 1.00 | 0.09 | -0.20 | 0.02 | 0.12 | 0.07 | -0.03 | -0.01 | -0.03 |
| Income | 0.00 | 0.03 | 0.11 | 0.06 | 0.09 | 1.00 | -0.09 | 0.01 | -0.19 | -0.36 | -0.03 | 0.44 | -0.08 |
| Crime | -0.36 | 0.17 | 0.09 | 0.70 | -0.20 | -0.09 | 1.00 | -0.03 | 0.05 | 0.20 | 0.10 | -0.10 | -0.17 |
| CrmChg | -0.31 | -0.10 | 0.20 | -0.17 | 0.02 | 0.01 | -0.03 | 1.00 | 0.21 | -0.08 | -0.14 | 0.13 | 0.07 |
| Year | -0.18 | 0.04 | 0.35 | 0.14 | 0.12 | -0.19 | 0.05 | 0.21 | 1.00 | -0.10 | 0.00 | 0.02 | 0.07 |
| John | -0.14 | 0.10 | -0.38 | -0.09 | 0.07 | -0.36 | 0.20 | -0.08 | -0.10 | 1.00 | -0.26 | -0.36 | -0.34 |
| Jone | -0.24 | 0.00 | 0.22 | 0.14 | -0.03 | -0.03 | 0.10 | -0.14 | 0.00 | -0.26 | 1.00 | -0.31 | -0.30 |
| Lang | 0.01 | 0.01 | 0.04 | -0.01 | -0.01 | 0.44 | -0.10 | 0.13 | 0.02 | -0.36 | -0.31 | 1.00 | -0.41 |
| Will | 0.33 | -0.10 | 0.13 | -0.03 | -0.03 | -0.08 | -0.17 | 0.07 | 0.07 | -0.34 | -0.30 | -0.41 | 1.00 |



Crime rates seem to be more strongly positively correlated with Lunch and counsellor Johnson and negatively correlated with Population. I expect to see these in the final model.

Other correlations to notice are:

1. Lunch and Child
2. Counsellor Lang and Income
3. Population Change and Income Change
4. Population Change and Crime Change
5. Counsellor Johnson and Income and Child.

Some of these (like Lunch and Child) may have confounding effects in the model.

**Models**

**Model 1: All Variables included**

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 53.76% of variation is explained by the model.
3. The residuals look approximately symmetrical.
4. Three variables (and the intercept) look significant (p-values of t-test <0.10)
5. Variable child is negatively correlated with crime instead of positively.

Call:

lm(formula = Crime ~ Pop + PopChg + Child + Lunch + IncChg +

CrmChg + Year + John + Jone + Lang, data = Denver, na.action = na.omit)

Residuals:

Min 1Q Median 3Q Max

-61.071 -18.649 -8.337 14.490 101.448

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 117.8887633 34.7218068 3.395 0.00185 \*\*

Pop -2.7978287 1.9416137 -1.441 0.15930

PopChg 0.1536696 0.5229723 0.294 0.77078

Child -3.2217685 1.1448579 -2.814 0.00830 \*\*

Lunch 1.9161267 0.3105066 6.171 0.000000662 \*\*\*

IncChg -0.3957626 1.0173455 -0.389 0.69984

CrmChg 0.8110489 0.4340789 1.868 0.07088 .

Year 0.0004003 0.0009908 0.404 0.68886

John 8.8272678 16.4137034 0.538 0.59444

Jone 8.3699529 18.4199612 0.454 0.65261

Lang -0.9189266 14.3687422 -0.064 0.94941

---

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

Residual standard error: 34.95 on 32 degrees of freedom

Multiple R-squared: 0.6477, Adjusted R-squared: 0.5376

F-statistic: 5.884 on 10 and 32 DF, p-value: 0.00005471

**Model 2: Backward Selection**

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 59.95% of variation is explained by the model.
3. The residuals look approximately symmetrical.
4. Four variables (and the intercept) look significant (p-values of t-test <0.10). Pop, Child, Lunch and CrmChg. Pop did not appear as significant in the all variables model.
5. Variable child is still negatively correlated with crime instead of positively.

Call:

lm(formula = Crime ~ Pop + Child + Lunch + CrmChg, data = Denver,

na.action = na.omit)

Residuals:

Min 1Q Median 3Q Max

-62.77 -17.96 -10.45 14.63 105.76

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 119.6135 20.5912 5.809 0.0000010410 \*\*\*

Pop -3.4053 1.5305 -2.225 0.032097 \*

Child -3.1877 0.8496 -3.752 0.000585 \*\*\*

Lunch 1.9228 0.2710 7.096 0.0000000182 \*\*\*

CrmChg 0.7711 0.3642 2.117 0.040846 \*

---

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

Residual standard error: 32.53 on 38 degrees of freedom

Multiple R-squared: 0.6376, Adjusted R-squared: 0.5995

F-statistic: 16.72 on 4 and 38 DF, p-value: 0.00000005514

**Model 3: Forward Selection**

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 51.07% of variation is explained by the model.
3. The residuals look approximately symmetrical.
4. Three variables (and the intercept) look significant (p-values of t-test <0.10). Pop, Lunch and Counsellor John.
5. All coefficients match the correlation table.

Call:

lm(formula = Crime ~ Lunch + Pop + John, data = Denver, na.action = na.omit)

Residuals:

Min 1Q Median 3Q Max

-49.28 -22.34 -10.72 18.35 131.00

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 59.185 17.148 3.451 0.00136 \*\*

Lunch 1.261 0.214 5.893 0.00000073 \*\*\*

Pop -4.814 1.588 -3.032 0.00430 \*\*

John 24.787 13.014 1.905 0.06422 .

---

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

Residual standard error: 35.96 on 39 degrees of freedom

Multiple R-squared: 0.5456, Adjusted R-squared: 0.5107

F-statistic: 15.61 on 3 and 39 DF, p-value: 0.0000007979

**Model Evaluation**

**Verifying Assumptions**

1. **Independence of Predictors**

The Spearman rho value for Lunch, Pop and John are all very low (-0.02 and -0.14) suggesting that the predictors are independent.

1. **Distribution of Error Terms**

The error terms seem to be approximately normally distributed.

Shapiro-Wilk normality test

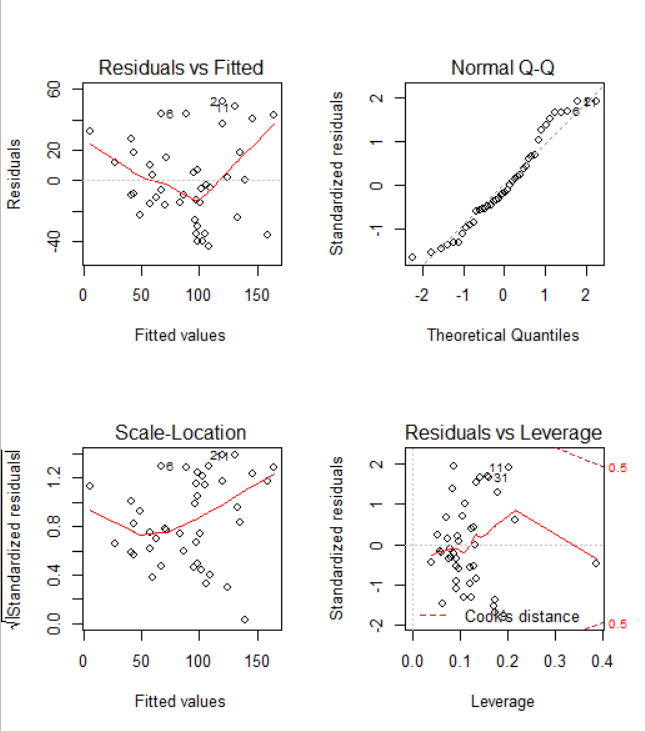
data: FwdDenRes

W = 0.95362, p-value = 0.08711

1. **Non-AutoCorrelation and Homoscedasticity**

Based on Residuals vs. Fitted and Scale-Location, there appears to be no explicit pattern to the residuals. Therefore, no there is no appearance of autocorrelation.

Based on Residuals vs. Leverage and Cook’s Distance, there is no data point exerting undue influence or leverage on the model.



**Final Model, Recommendation and Interpretation**

Based on the above, I recommend the following model (developed with forward selection):

Crime Rate =

1.261 \* Participation in the School Lunch program +

(-4.814) \* Total Population +

(24.787) \* Neighbourhoods with Johnson as the Counsellor

**APPENDIX 1: Data Transformation**

**Transforming Dates**

Denver$Year <- as.Date(Denver$Year,format="%d/%m/%Y")

#Converts to Days Since Jan 1, 1970 and adds 10000 to eliminate negative numbers

Denver$Year <- julian(Denver$Year)+10000

**Transforming Income**

Denver$Income <- as.numeric(gsub('[$,]', '', Denver$Income))

**Transforming Counsellors**

Cnc\_Dummies <- model.matrix(~Counsellor -1, data=Denver)

#Combine the Datasets again

Denver <- cbind(Denver, Cnc\_Dummies)