Christopher Giegel
Brandon Butch
Richard Miktus

Section 1:
Introduction:

Our group hopes to achieve a definite correlation between a group of data sets involving specific counties in the state of New Jersey. These data sets concern children (under the age of 18) and families in the state that have children. They are: Juvenile Arrests, Child Poverty, Population Density, Child Abuse/Neglect Referrals, Median Family Income, Total School Enrollment, and Special Education Enrollments. Using the population, we need to transform all of our data into percentages in order to better compare each county's statistics. We predict that child abuse, family income, and poverty numbers will contribute to the level of school enrollment in each area. We further predict that the school enrollment, combined with special education enrollment and population density, will explain the number of arrests in each county. Using several statistical functions in R, we will be able to test our predictions and evaluate the relevance of our data.

Findings and Conclusion:

After performing some analysis on our data, we found that some of our predictions held, while others did not. Our prediction that school enrollment could be described as a function of child abuse, family income and poverty was not entirely correct. In fact, it seems that school enrollment is relatively unrelated to many of the other variables. On the other hand, we found that there was a strong correlation between juvenile arrests and a number of the other variables, namely: the number of reported cases of child abuse, the number of children enrolled in special education and the number of children enrolled in school. Specifically, the number of arrests is strongly, positively correlated with reported cases of child abuse; weakly, positively correlated with children enrolled in special education; and weakly, negatively correlated with children enrolled in school. The specific model we found is:

$\log(\text{Arrests}) = 0.956301 + 0.086790(\text{sped}) +0.236819(\text{abuse}) -0.028434(\text{school})$

Section 2:
Data Acquisition:

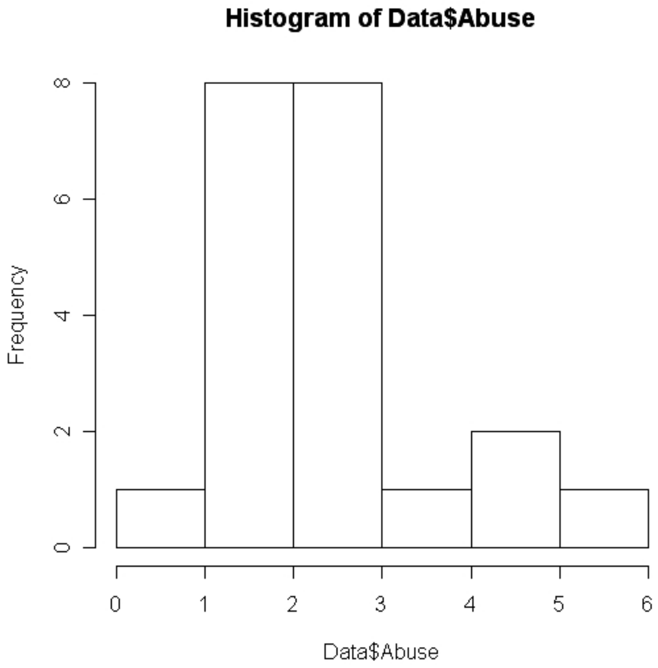
We gathered the raw data from the web site:
http://www.kidscount.org/cgi-bin/cliks.cgi?action=rawdata_results&subset=NJ
and put it into the data into a basic spreadsheet. (attached as rawData.csv)
We then adjusted the data by dividing most of the columns by their respective total populations to get percentages that can apply to one another. (attached as basicData.csv)
We then edited the data to make it more easily managed by R. (attached as RData.csv)

```
> data = read.csv('c:\\tmp\\RData.csv')
> attach(data)
```

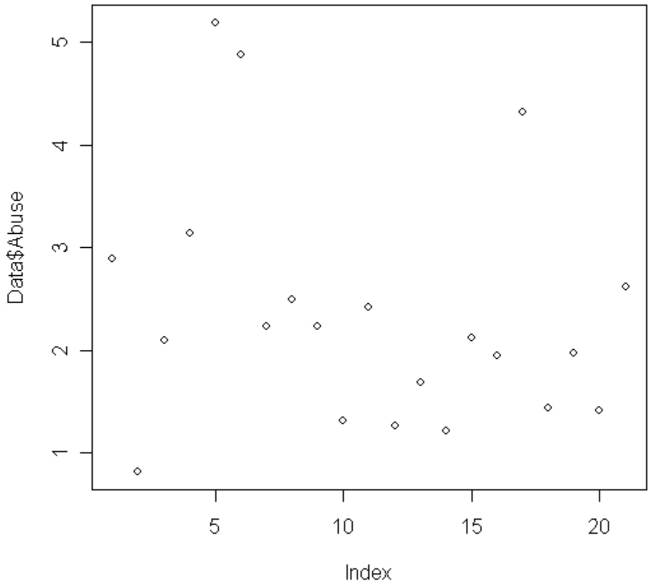
Section 3:
Single variable study:

First we do a preliminary study of each variable that we are going to analyze. This will give us a good understanding of each variable, its distribution, and what it may be used for. It will also let us know of possible problems we may run into.

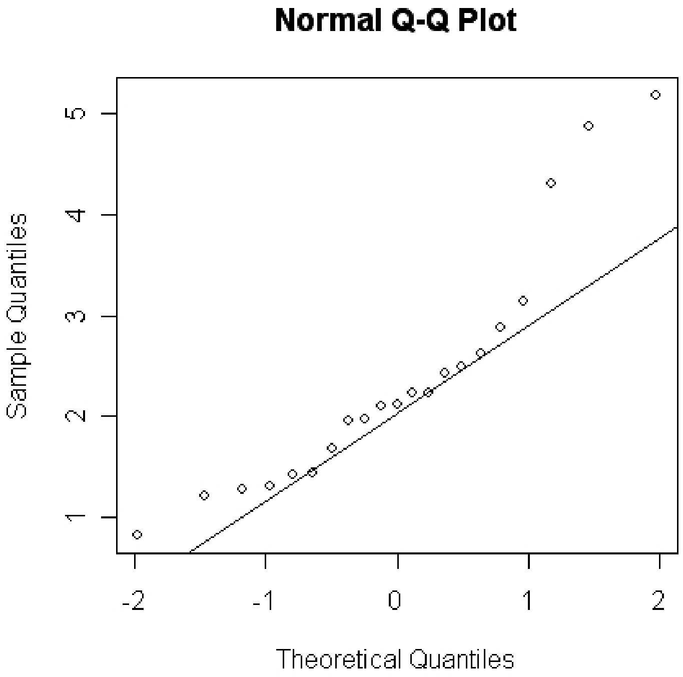
```
> hist(Data$Abuse)
```



```
plot(Data$Abuse)
```



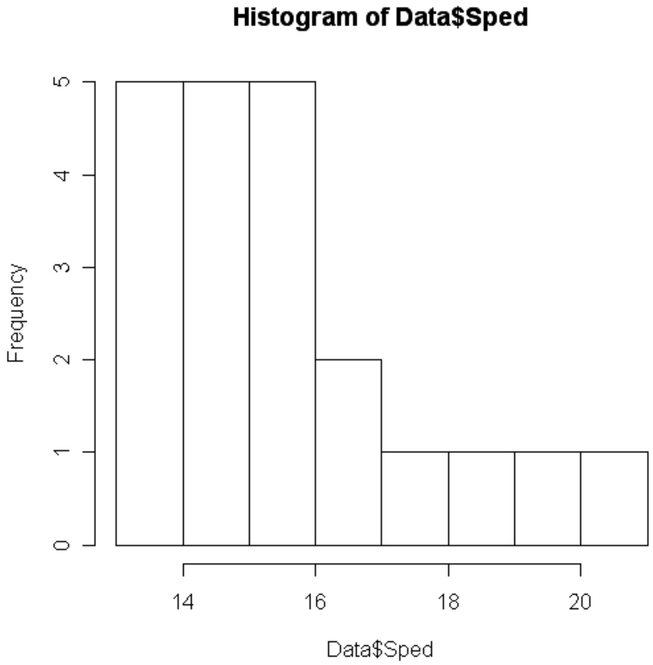
```
> qqnorm(Data$Abuse)
> qqline(Data$Abuse)
```



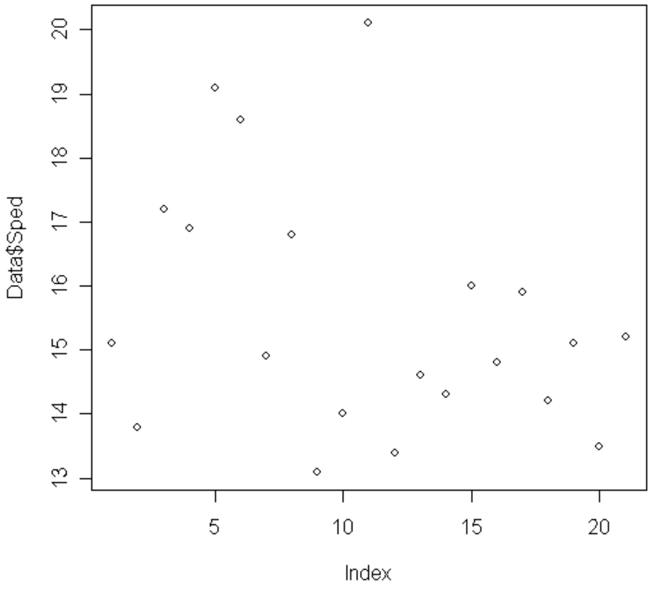
Our group believes that child abuse is a good indicator of school enrollment because that in a household bolds for a terrible environment in which to raise children. This environment leads to family problems and definitely does not help the enrollment level.

The histogram and scatter plot for child abuse/neglect referrals show a skew to the left. There are a much higher percentage of counties with between 1% and 3% referrals than any other amount. The Q-Q plot shows us that the distribution is close to normal in the middle, but very far from normal at the ends.

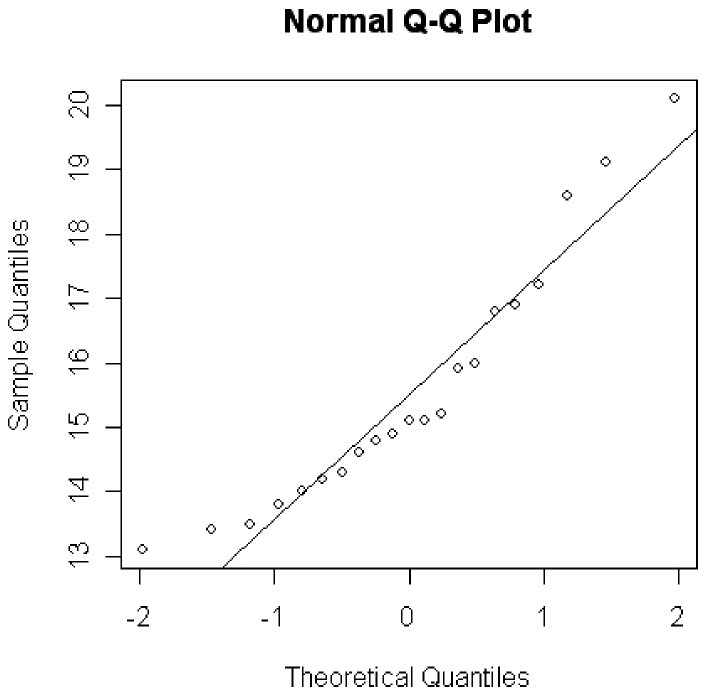
```
hist(Data$Sped)
```



```
plot(Data$Sped)
```



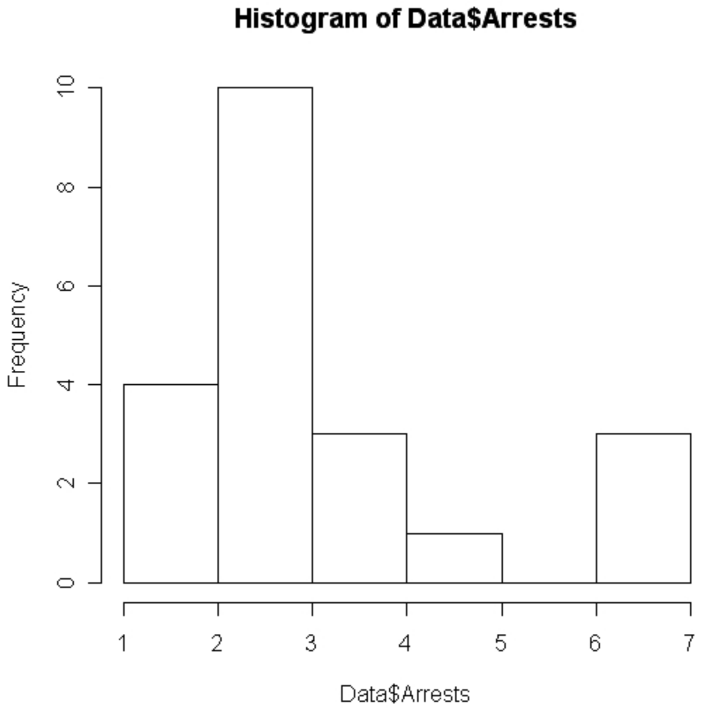
```
> qqnorm(Data$Sped)
> qqline(Data$Sped)
```



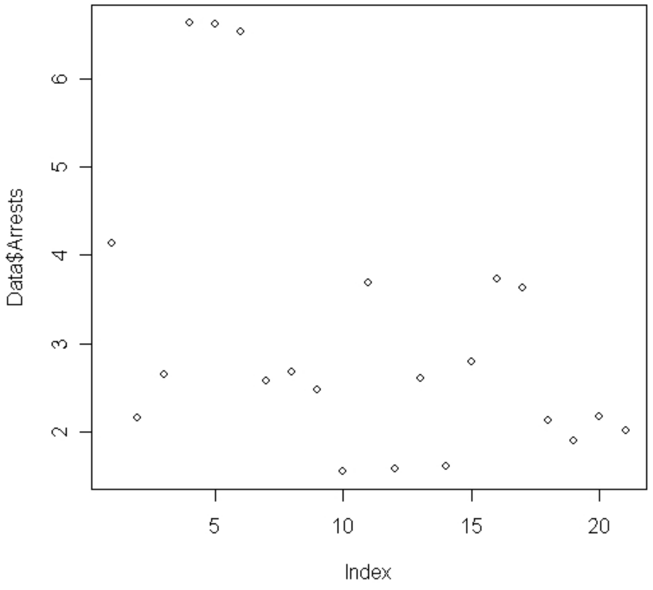
We believe the percentage of children enrolled in special education correlates to the number of juvenile arrests in any particular county. This is because we associate special education needs with poor family situations and shabby upbringings. These factors lead to bad influences like drugs, gangs, and violence.

The histogram and scatter plot for Special Education Enrollment show a very sharp skew to the left. There are many more instances of counties having smaller than 16% of children enrolled than that which have more than 16%. The Normal Q-Q plot shows close to a normal distribution with slight skews on both ends of the plot.

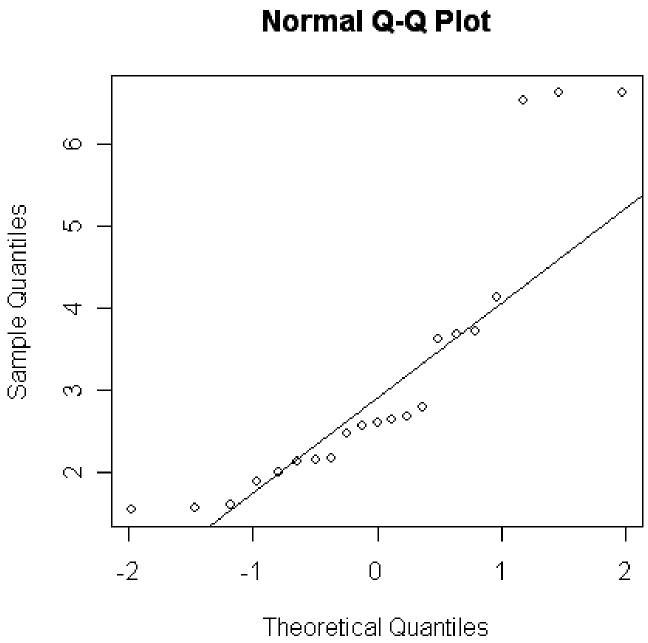
```
hist(Data$Arrests)
```



```
plot(Data$Arrests)
```



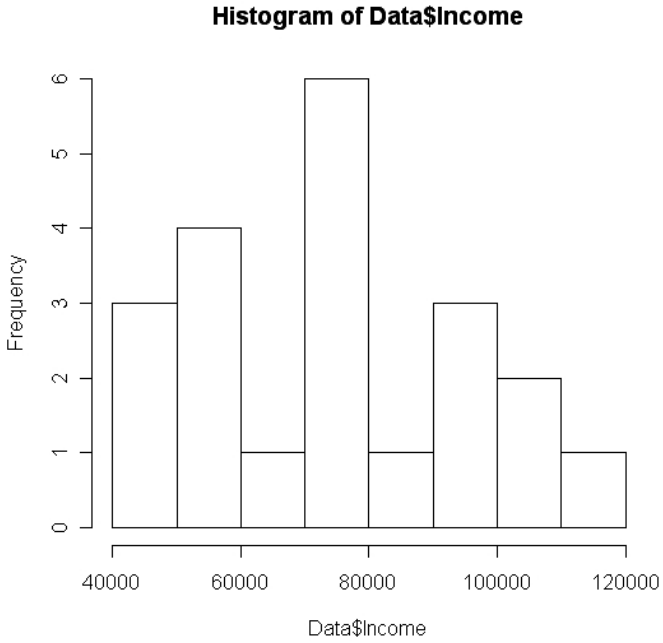
```
> qqnorm(Data$Arrests)
> qqline(Data$Arrests)
```



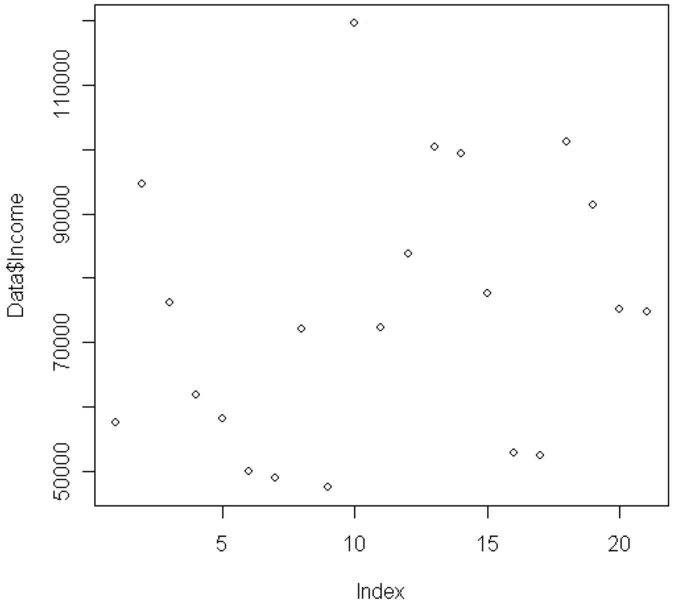
Our group aims to show that the level of juvenile arrests in any county can be explained as a function of a specific combination of our variable data sets. We believe the variables we have chosen show a special correlation to arrests and, when combined, will justify the differences in percentage of juvenile arrests in each county.

The histogram and scatter plot for juvenile arrests show a very uneven distribution. There is quite a random assortment of values for percent of juvenile arrests in each county, with between 2 and 3 being the highest by a lot. The Q-Q plot stays very true to normal until the end when it is not even close to a normal distribution.

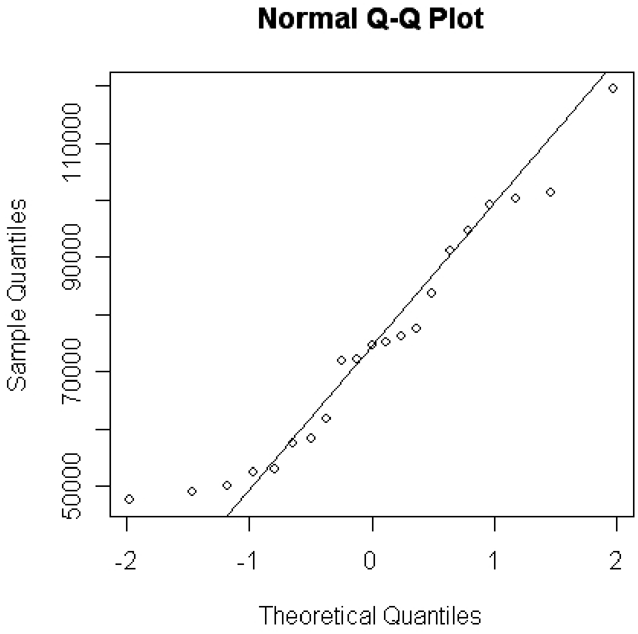
```
hist(Data$Income)
```



```
plot(Data$Income)
```



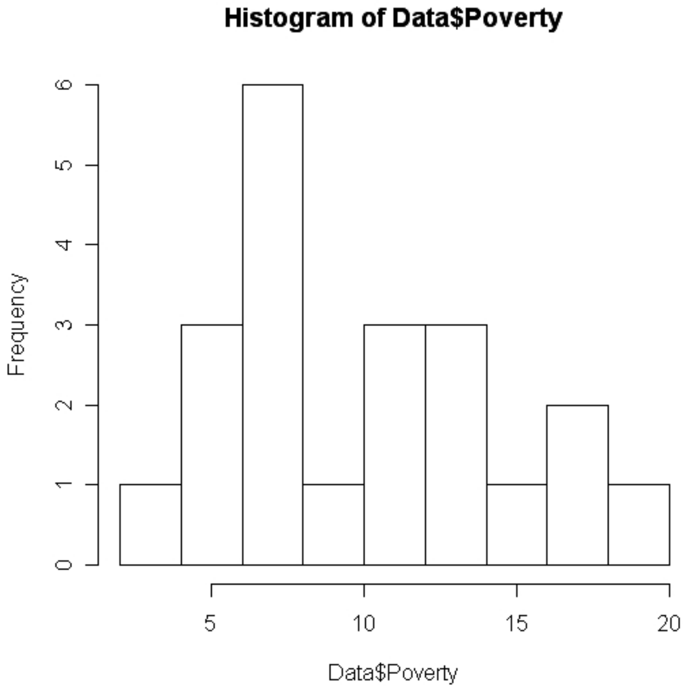
```
> qqnorm(Data$Income)
> qqline(Data$Income)
```



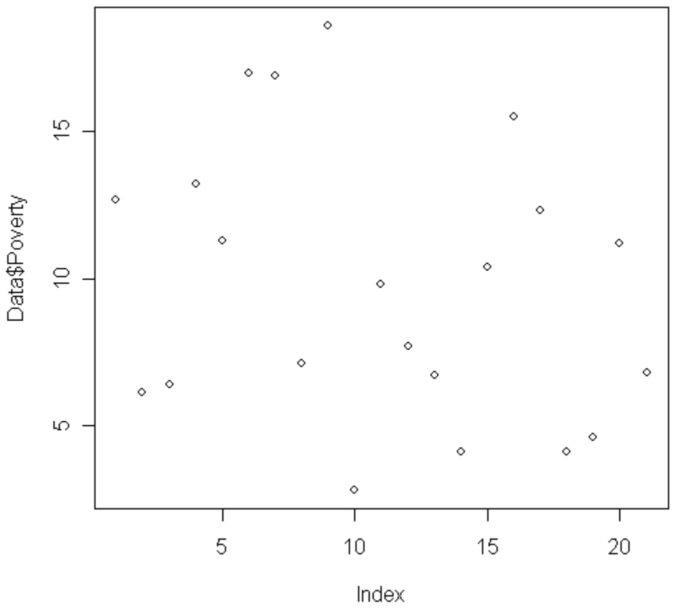
The average family income in each county will provide more reasoning for the differences in school enrollment. We believe poorer areas will have a lower percentage of enrollments.

The histogram and scatter plot for average family income show almost a bell curve with two huge valleys on either side of between 7000 and 8000. The Normal Q-Q plot shows a fairly normal distribution in the middle with a few outliers at the beginning.

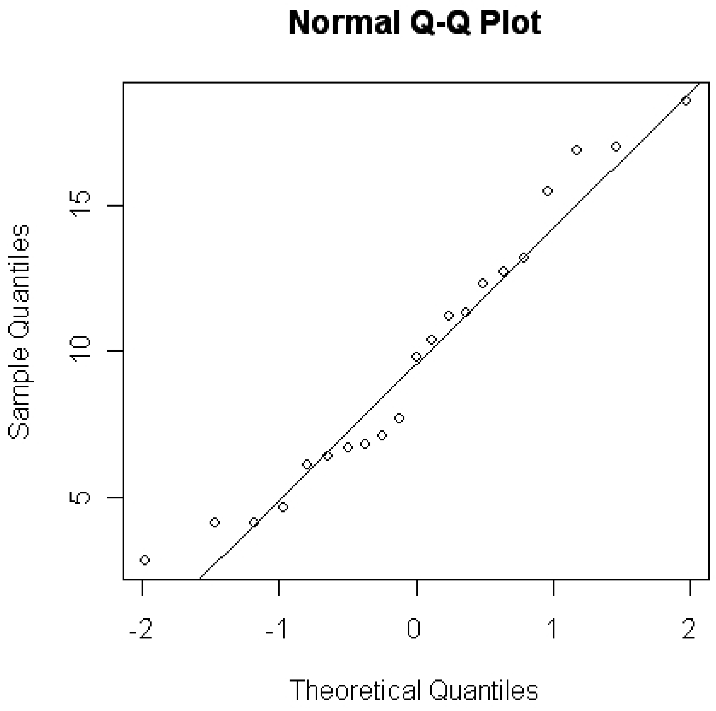
```
hist(Data$Poverty)
```



```
plot(Data$Poverty)
```



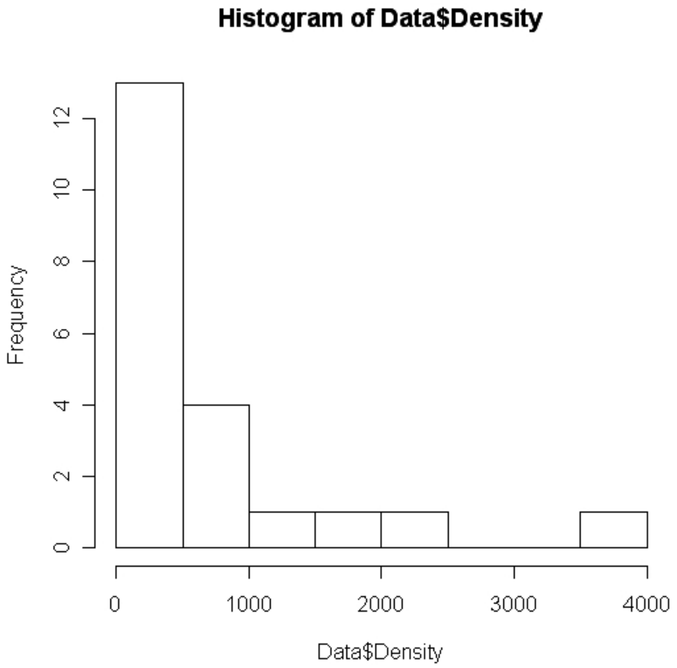
```
> qqnorm(Data$Poverty)
> qqline(Data$Poverty)
```



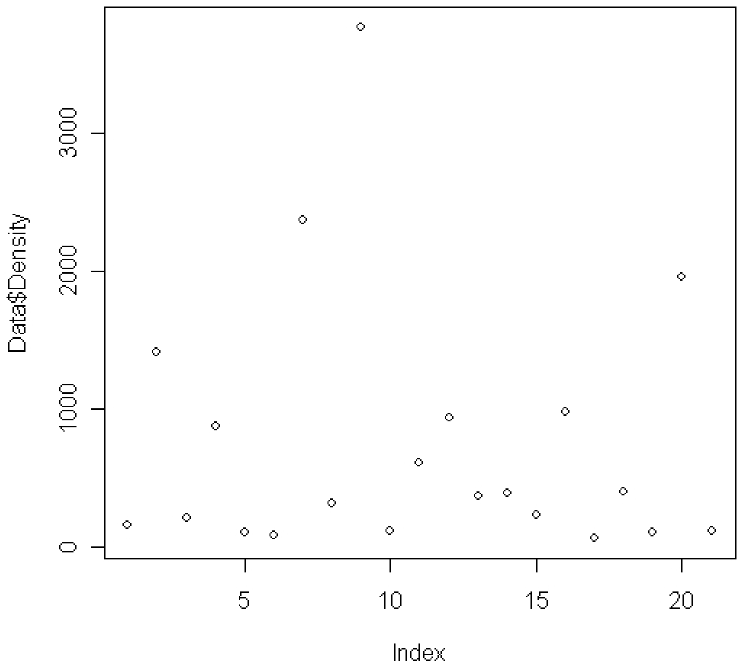
We believe that the level of child poverty will directly coincide with the percentage of juvenile arrests per county. Areas with more poverty and poor living conditions will almost certainly provide more instances of crime.

The histogram and scatter plot for child poverty show a small skew to the left, with most of the counties' poverty frequencies between 5% and 8%. The Q-Q plot shows close to a normal distribution except for an outlier in the very beginning.

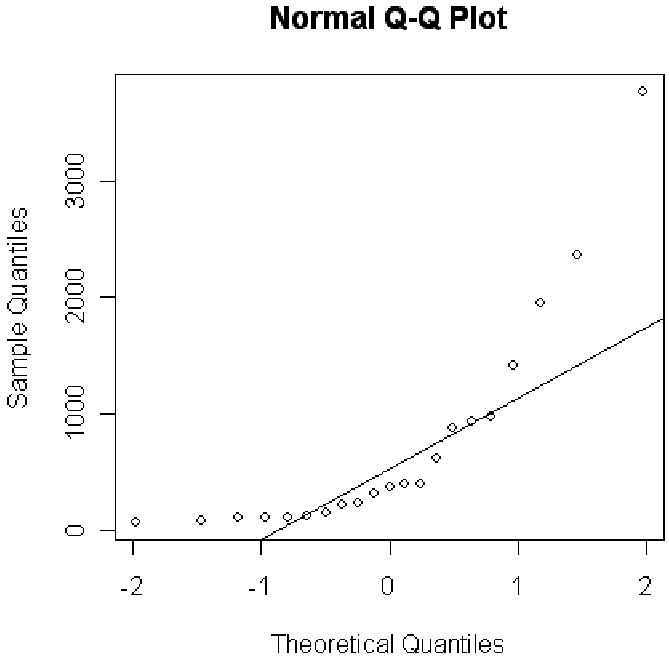
```
hist(Data$Density)
```



```
plot(Data$Density)
```



```
> qqnorm(Data$Density)
> qqline(Data$Density)
```

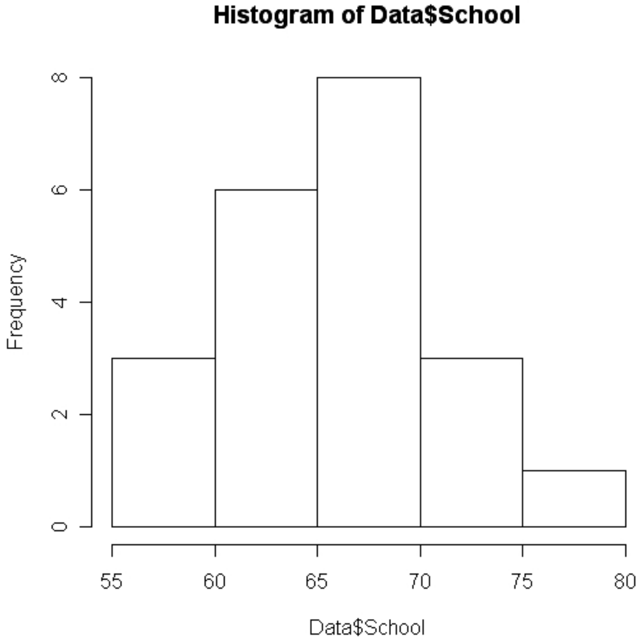


Population is an important statistic for our purposes. To compare the counties, every variable must be changed to a percentage. One of the reasons we believe population density contributes to the educational enrollment is because more densely populated areas are usually inner cities which have a lower percentage of enrollment.

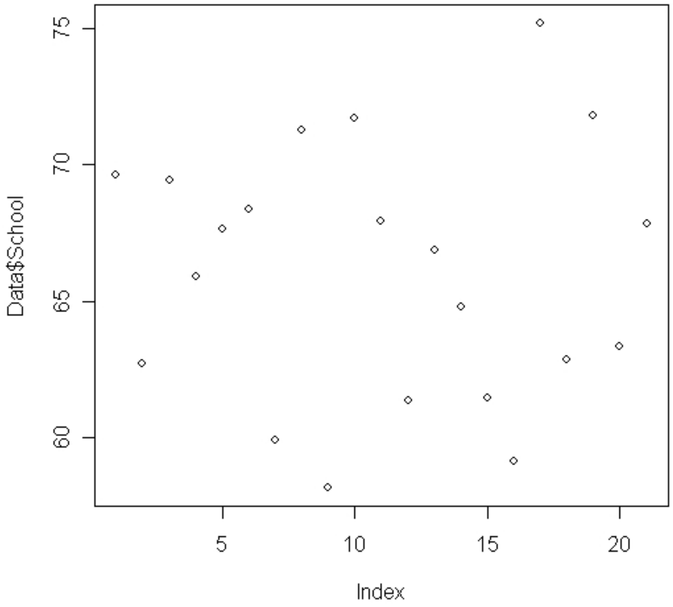
The histogram and scatter plot for population density show a huge skew to the left. Almost every county has a population density under 500. There is an extreme outlier above 3500. The Q-Q plot shows nothing near a normal distribution, with both ends being completely skewed.

Some of the population density numbers may be misleading due to the way we divided up the state of New Jersey. Some counties with cities exhibiting extreme population densities might also include a lot of rural, area skewing the statistics. A more accurate measure of population density would have to be done using individual towns and cities.

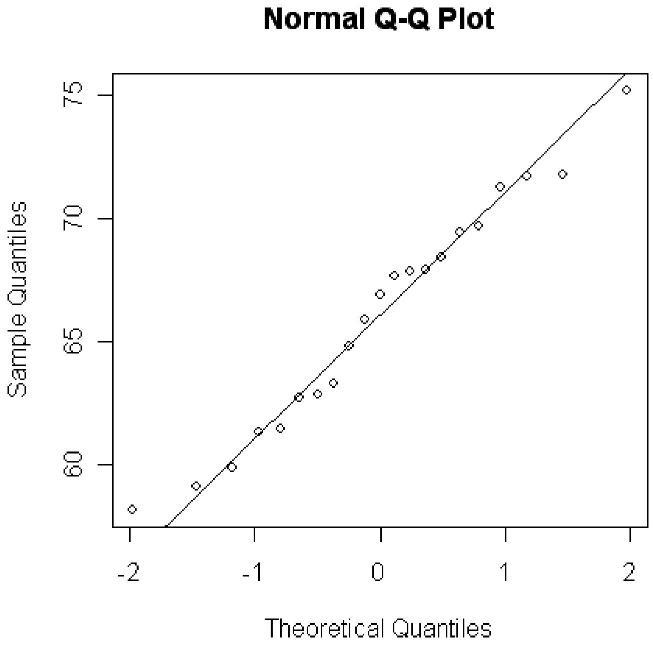
```
hist(Data$School)
```



```
> plot(Data$School)
```



```
> qqnorm(Data$School)
> qqline(Data$School)
```



Our group aims to show that the level of school enrollment can be explained as a function of the variables income, poverty, and child abuse. We believe that these variables exhibit a strong correlation to each other and enrollment. We hope to then use school enrollment as a variable to explain the number of juvenile arrests per county. Areas with a lower percentage of children enrolled in school will have higher crime rates and more arrests.

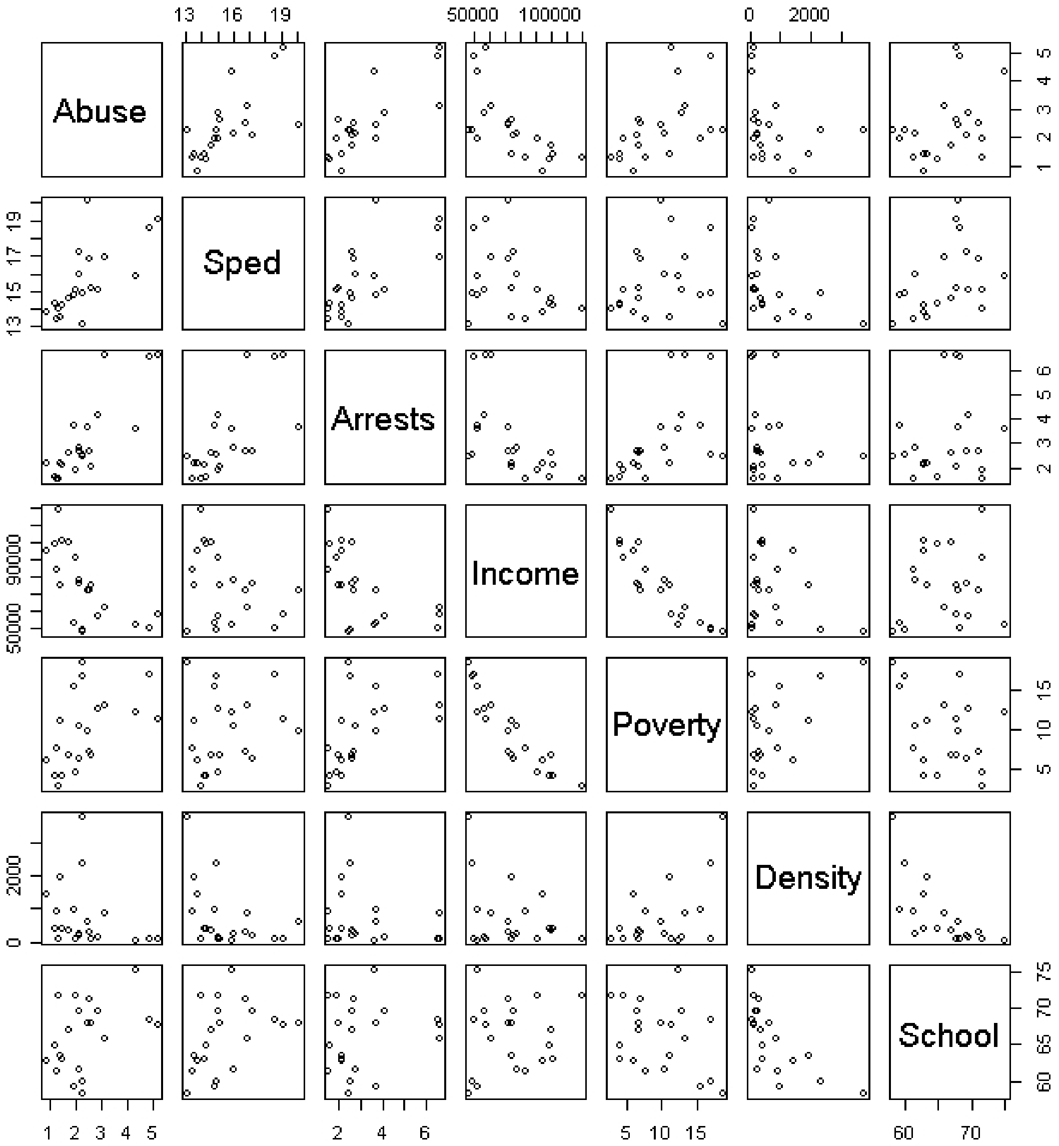
The histogram and scatter plot for school enrollment are fairly close to a bell curve, with a little skew to the left. The Q-Q plot shows the closest thing we have so far to a normal distribution, with a slight outlier in the beginning.

The percentage of school enrollment is skewed for every county because children aged 1-4 are obviously not enrolled in school yet. Although this changes the actual data, it remains relatively accurate for our project because it has similar effect on each county.

	Minimum	Maximum	Mean	Median	Standard Deviation	95% Confidence Interval
Abuse	0.8183	5.1840	2.3660	2.1150	1.178709	(1.861,2.869)
Sped	13.10	20.10	15.55	15.10	1.9382	(14.723,16.381)
Arrests	1.557	6.622	3.134	2.606	1.615998	(2.443,3.825)
Income	47570	119600	74640	74720	20587.14	(65833.31,83443.83)
Poverty	2.800	18.600	9.776	9.800	4.687953	(7.771,11.781)
Density	67.65	3765.00	741.90	369.50	941.8733	(339.059,1144.752)
School	58.20	75.18	66.07	66.90	4.668894	(64.073,68.067)

Section 4:
Two variable relationships:

```
> plot(data[2:8])
```



We explore any interesting 2 variable relationships before creating the linear models. Also we use these relationships to choose what variables should be used as explanatory variables for school and arrests.

Table of Correlation Values

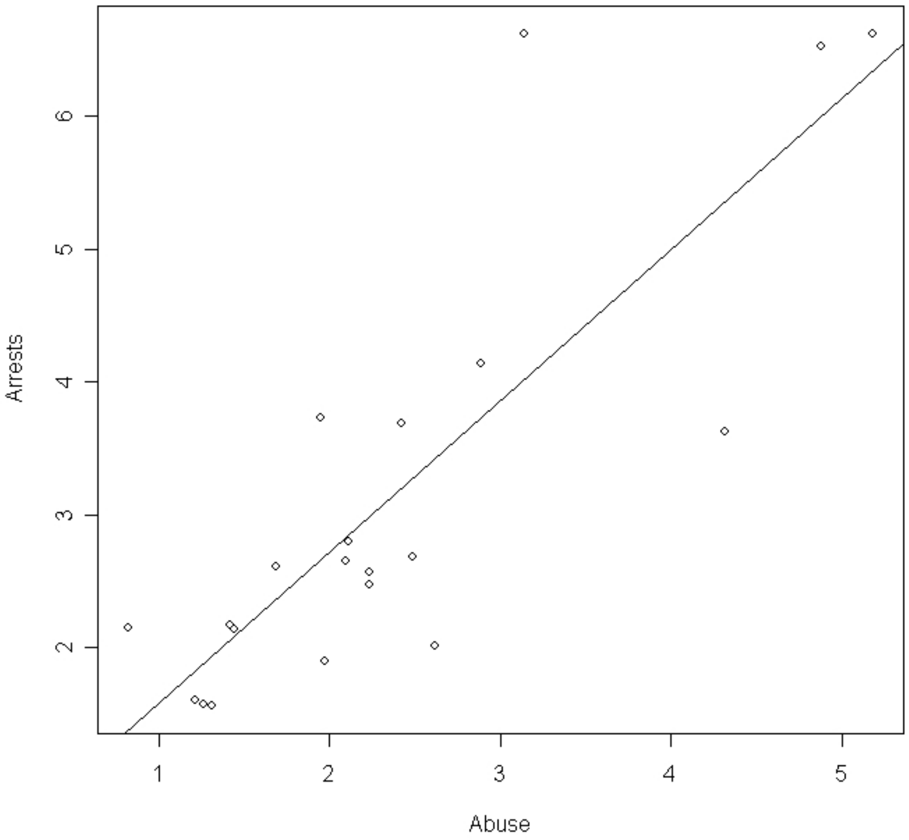
	Abuse	**Sped**	**Arrests**	**Income**	**Poverty**	**Density**	**School**
Abuse	1	0.6966361	0.8320943	-0.6636096	0.5130734	-0.2758241	0.4045335
Sped	0.6966361	1	0.7002274	-0.3572419	0.1783266	-0.4452980	0.4181188
Arrests	0.8320943	0.7002274	1	-0.6010875	0.5566217	-0.1872739	0.1591841
Income	-0.6636096	-0.3572419	-0.6010875	1	-0.9110014	-0.3356800	0.1533906
Poverty	0.5130734	0.1783266	0.5566217	-0.9110014	1	0.5321432	-0.3710788
Density	-0.2758241	-0.4452980	-0.1872739	-0.3356800	0.5321432	1	-0.7015329
School	0.4045335	0.4181188	0.1591841	0.1533906	-0.3710788	-0.7015329	1

By evaluating the correlation table, we can see which pairs of variables are related. There are several pairs which seem to be strongly correlated: abuse-sped, abuse-arrests, abuse-income, sped-arrests, density-school. Our initial objective was to explore school enrollment as a function of abuse, income and poverty, and arrests as a function of school enrollment, population density, and sped. However, looking at our correlation table, it seems that our original model must be revised. There is a very strong correlation between school enrollment and population density, so it makes sense to revise our model of the school enrollment function to include population density.

Now, our immediate goal is to discover whether or not arrests is a function of abuse, poverty, sped, and school enrollments; and whether or not school is a function of income, poverty, abuse and density. Before doing any further graphical or numerical analysis, we predict that arrests is strongly influenced by abuse and sped, moderately influenced by poverty and not influenced by school. Also, we predict that school is strongly influenced by density, and minimally influenced by abuse, income and poverty.

Also note that income and poverty are extremely correlated, with a correlation value of -0.911. This makes sense as the number of families with children living in poverty is directly and very strongly influenced by the general income of the county.

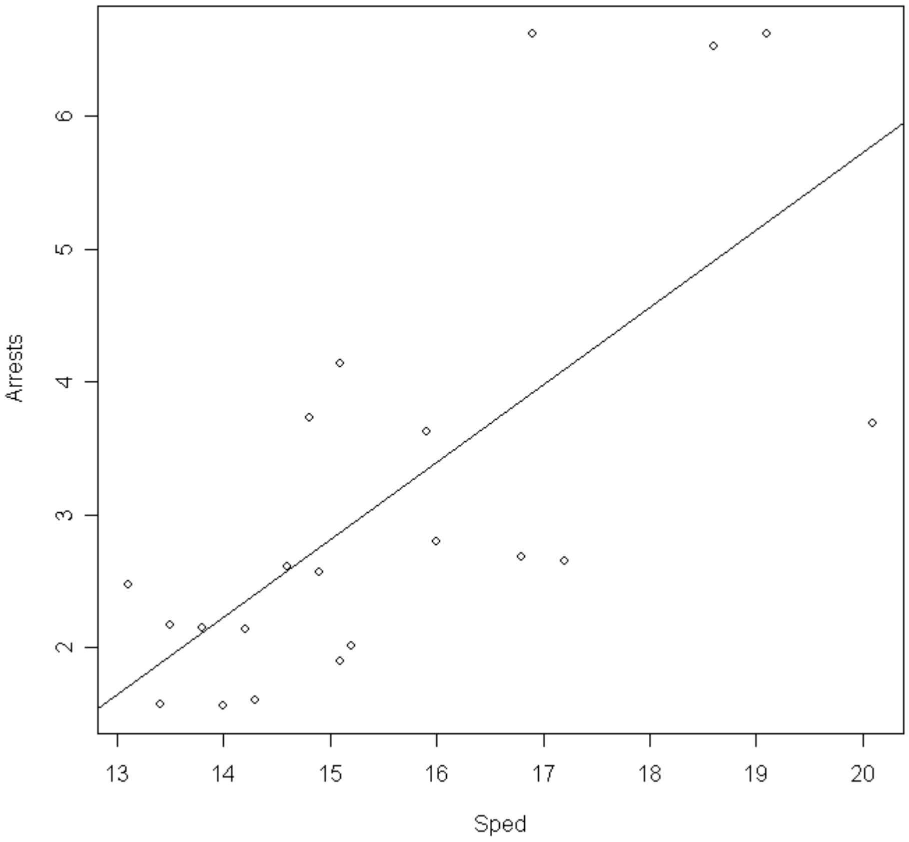
```
> var = lm(Arrests~Abuse)
> plot(Abuse,Arrests)
> abline(coef(var))
```



```
> round(var$residuals, 4)
0.4447, -0.2772, 0.1626, 2.8310, 3.1887, 2.0055, -1.9322,
0.0593, -2.3546, -0.2386, 0.5470, -1.1618, 0.0614, -0.4416,
-0.4605, -0.5023, 0.0063, 0.087, -0.2403, -1.2344, -0.5499
```

As indicated by our correlation table, abuse and arrests are very strongly related. The residuals for this data pair are relatively small, as many points lie very close to the regression line. Intuitively, these variable are very closely related because it seems that children brought up in an unstable environment are more likely to resort to crime. It is unlikely that children in an abusive household have been taught correct moral values by their parents. Because the environment in which children are raised has a significant impact on their social outlook, it is reasonable to assume that children coming from an abusive household are more likely to abuse others. Furthermore, a child being abused by his parents will be likely to spend less time at home. While this may be an effective way to minimize abuse, it is possible that the child will now spend this time getting into trouble.

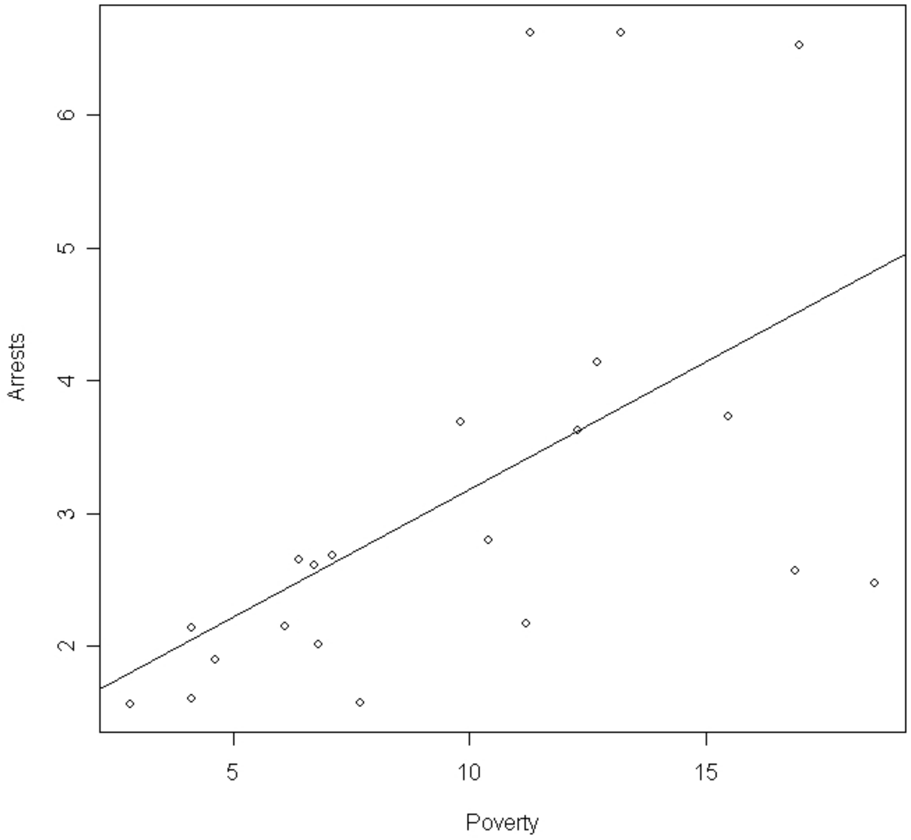
```
> var = lm(Arrests~Sped)
> plot(Sped,Arrests)
> abline(coef(var))
```



```
> round(var$residuals, 4)
1.2698, 0.0405, -1.4471, 2.7012, 1.4099, 1.6123, -0.1844,
-1.1826, 0.7702, -0.6708, -2.1035, -0.3036, 0.0272,
-0.7996, -0.6021, 1.0352, 0.2876, -0.2125, -0.9694, 0.2371,
-0.9152
```

The correlation between children in special education and number of arrests for each county is very strong. Although special education is often associated with learning disorders, a fair number of children in special education are placed there because of a simple refusal to learn. They may be disobedient and defiant in the normal classroom, and thus require special attention. Such behavior can be linked to illegal activity outside of school.

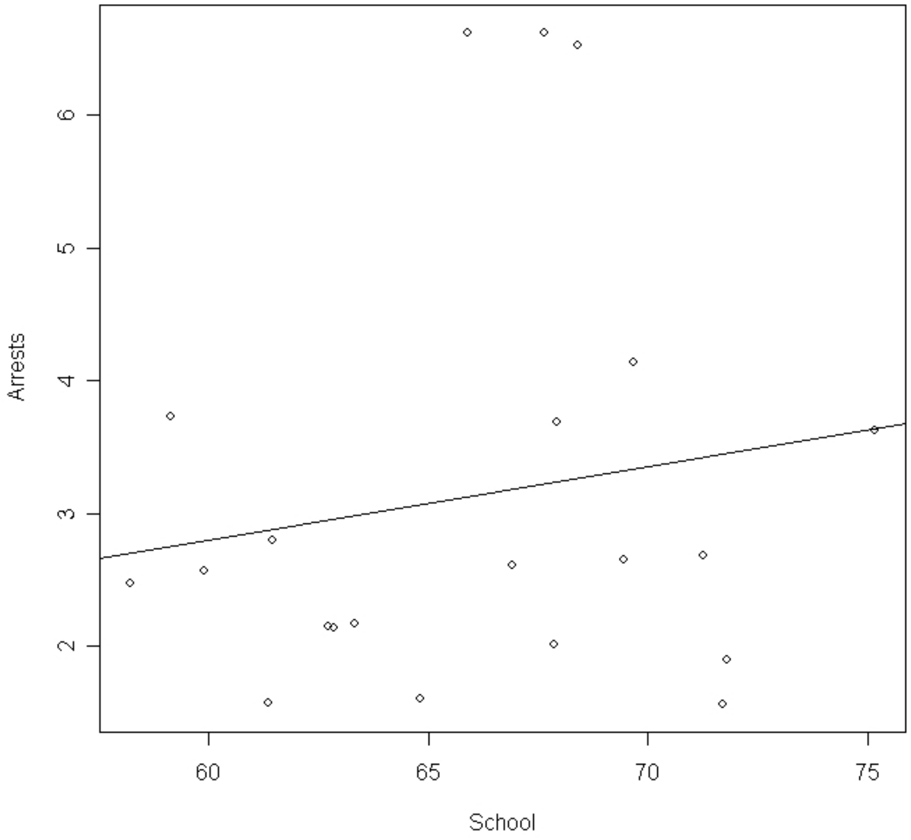
```
> var = lm(Arrests~Poverty)
> plot(Poverty,Arrests)
> abline(coef(var))
```



```
> round(var$residuals, 4)
0.4447, -0.2772, 0.1626, 2.8310, 3.1887, 2.0055, -1.9322,
0.0593, -2.3546, -0.2386, 0.5470, -1.1618, 0.0614, -0.4416,
-0.4605, -0.5023, 0.0063, 0.0871, -0.2403, -1.2344, -0.5499
```

Again, we see a clear correlation between these two variables. Although the relation is not as strong as the relations between arrests and abuse, and arrests and sped, it is still moderately strong. Once again, thinking about these two variables from a humanistic point of view, it makes sense that they should be related. Poverty can drive a person to commit illegal acts that that person would not normally commit. For example, a person living in poverty may be forced to steal food in order to provide for his family. So, it follows that the number of arrests in that area would be higher.

```
> var = lm(Arrests~School)
> plot(School,Arrests)
> abline(coef(var))
```



```
> round(var$residuals, 4)
0.8081, -0.7981, -0.6714, 3.4969, 3.3943, 3.2634, -0.2255,
-0.7402, -0.2279, -1.8883, 0.4496, -1.3009, -0.5744,
-1.4619, -0.0864, 0.9779, -0.0111, -0.8251, -1.5484,
-0.8104, -1.2200
```

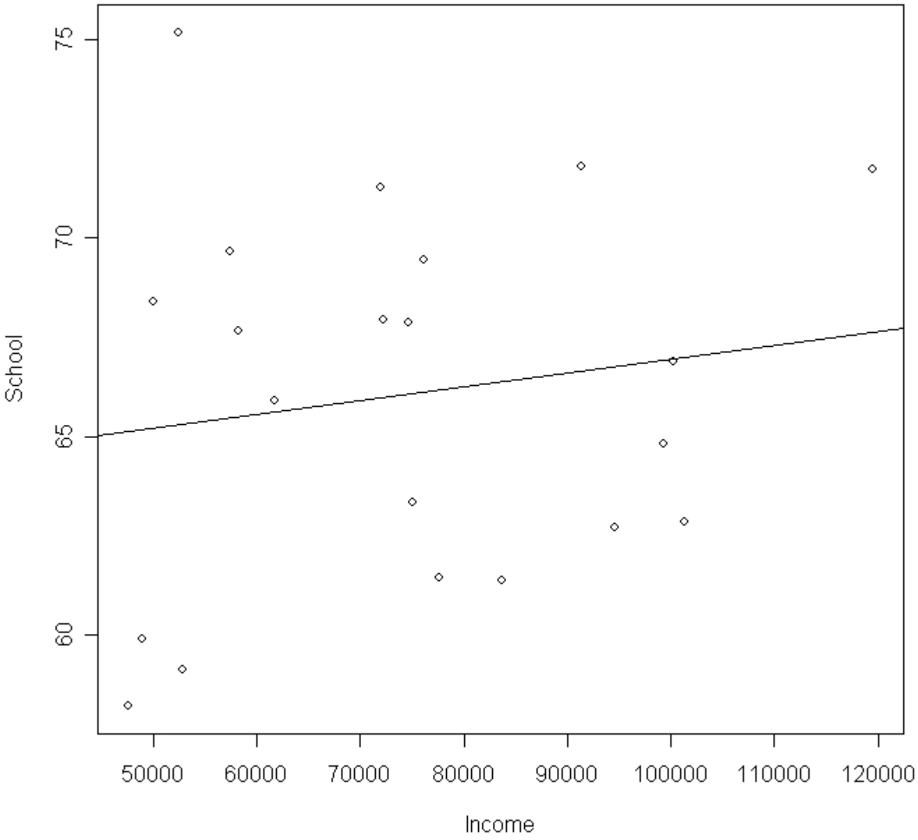
The graph of our arrests and school variables confirms our suspicions about their correlation from the table. The two variables seem to be very minimally correlated, if at all. This goes a bit against intuition because it seems that a low percentage of children enrolled in school would imply a higher number of juvenile arrests (school is often a strong deterrent against crime for children). However, in this case, the data tells us that the two are unrelated. It should be noted that this data is a bit skewed. The school variable is the percentage of all children – including those too young to attend school – who are enrolled. The percentage of children who are actually eligible to attend school is most likely higher.

	F Statistic	P Value	R^2
Arrests v. Abuse	42.76	2.911e-06	0.6924
Arrests v. Sped	18.28	0.0004087	0.4903
Arrests v. Poverty	8.529	0.008774	0.3098
Arrests v. School	0.494	0.4907	0.02534

These figures serve to confirm our analysis above. arrests v. abuse has an extremely small P Value, indicating that there is a definite relationship between them. The P Values for arrests v. sped and arrests v. poverty are moderate values, and the P Value for arrests v. school is very high. From this, we can conclude that there is a very good chance that a legitimate relationship exists for arrests v. sped and arrests v. poverty, but that the variables arrests and school are most likely unrelated. The high F statistics from the first three comparisons tell us that our data is related, while the low F statistic for arrests v. school tells us that no connection exists between the two. Our predictions from above appear to be justified: arrests is strongly correlated with abuse, moderately correlated with sped and poverty, and weakly (if at all) correlated with school.

Note that in all four plots above, there are three data points which do not conform with the rest of the data. All three points are from the same three counties and represent outliers in the study. For some reason, those three counties suffer from unusually high arrest rates.

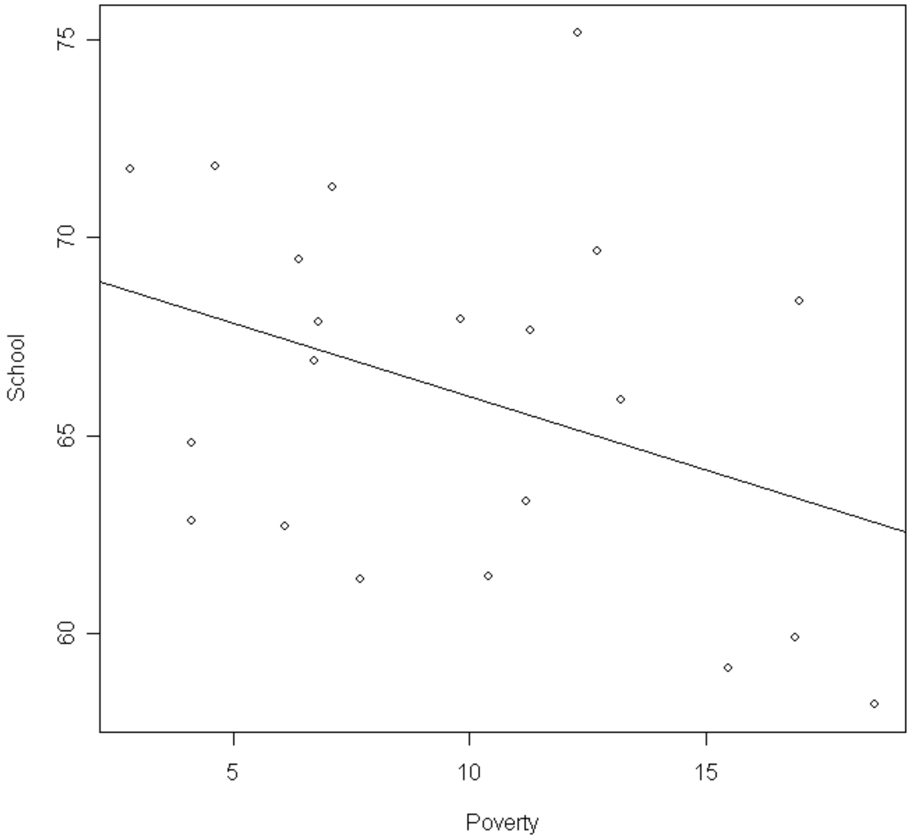
```
> var = lm(School~Income)
> plot(Income,School)
> abline(coef(var))
```



```
> round(var$residuals, 4)
4.1848, -4.0425, 3.3234, 0.2841, 2.1480, 3.1826, -5.2718,
5.2792, -6.9282, 4.0841, 1.9304, -5.0229, -0.0671, -2.1062,
-4.7212, -6.1739, 9.8763, -4.1391, 5.1355, -2.7516, 1.7963
```

The relationship here is extremely weak. The points vary from the regression line at great extent, which can be easily seen from the relatively high residuals. This makes sense because public school systems make it possible for every child to attend school, regardless of his family's income. One possible explanation for a weak positive correlation is the following: families with children who have a low income may require their children to drop out of school in order to find a job to help support the family. We suspect that this is not a commonality, thus yielding a very weak relationship between these data.

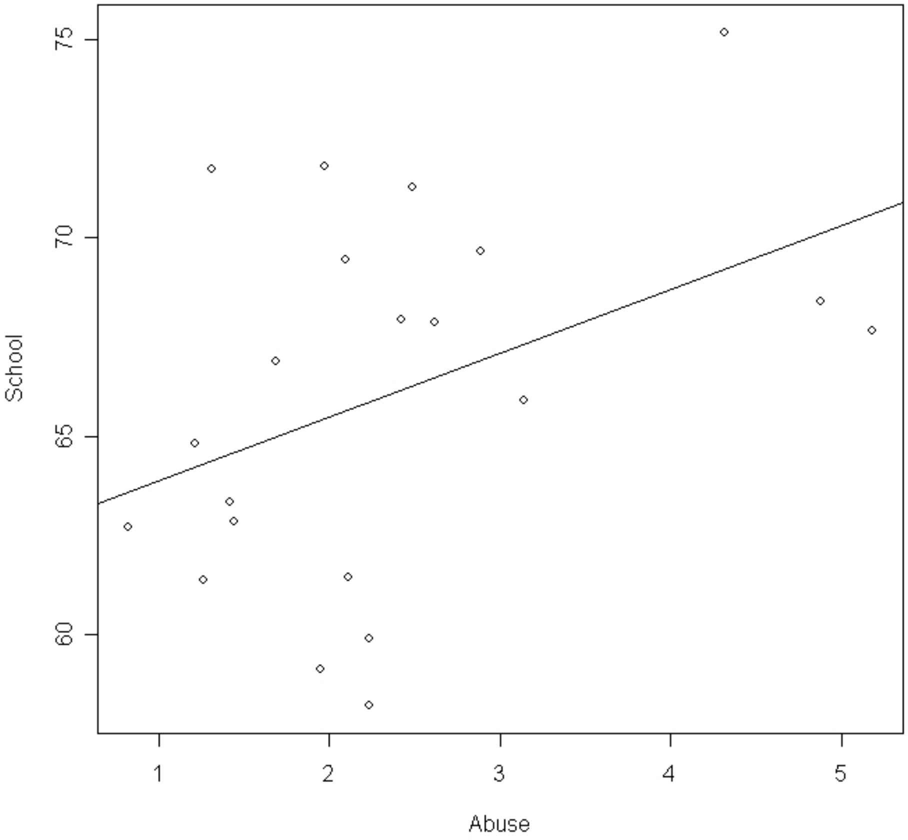
```
> var = lm(School~Poverty)
> plot(Poverty,School)
> abline(coef(var))
```



```
> round(var$residuals, 4)
4.6679, -4.7068, 2.1308, 1.1028, 2.1394, 4.9950, -3.5342,
4.2002, -4.6087, 3.0686, 1.8586, -5.4729, -0.3110, -3.3467,
-4.3862, -4.8167, 10.0376, -5.3097, 3.8027, -2.2100, 0.6992
```

From this data, we see that school enrollment is very weakly related to poverty. The analysis is extremely similar to above because of the intimate connection between poverty and income.

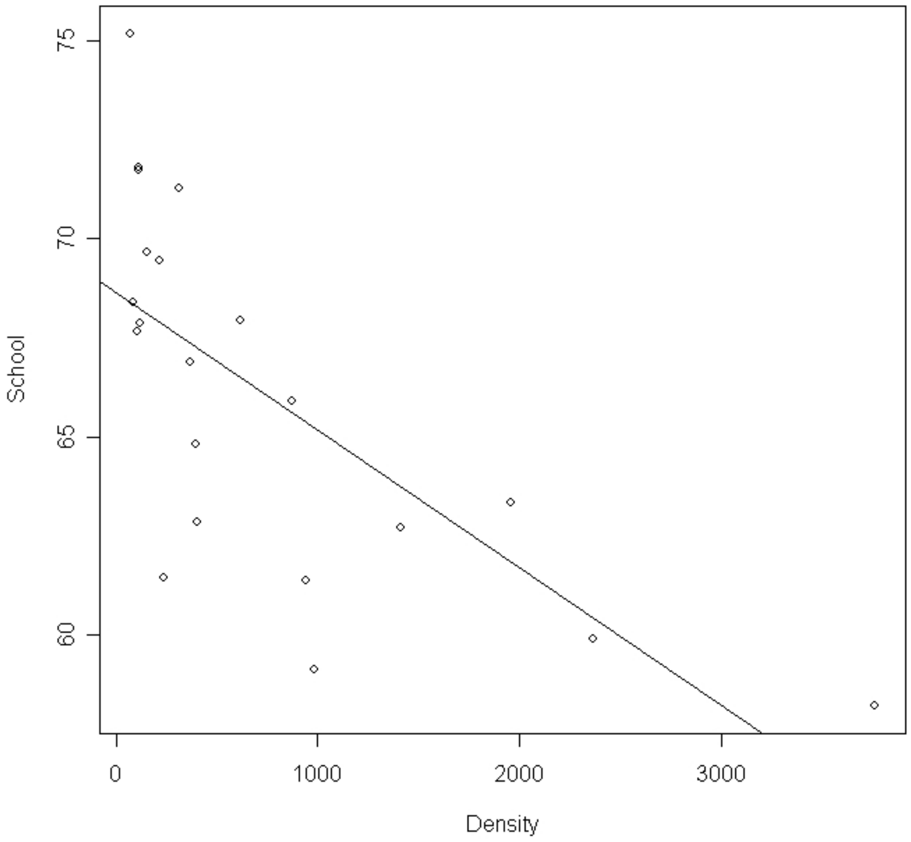
```
> var = lm(School~Abuse)
> plot(Abuse,School)
> abline(coef(var))
```



```
> round(var$residuals, 4)
2.7558, -0.8687, 3.8062, -1.4050, -2.9404, -1.7006,
-5.9588, 4.9886, -7.6625, 7.3396, 1.7602, -2.9413, 1.9156,
0.6015, -4.2143, -6.2677, 5.9821, -1.7264, 6.3471, -1.2107,
1.3995
```

Again, the residuals from these data are fairly high, and we can see from the plot that the points are scattered at great distances from the expected value. On the surface, there is no strong connection between the two variables.

```
> var = lm(School~Density)
> plot(Density,School)
> abline(coef(var))
```



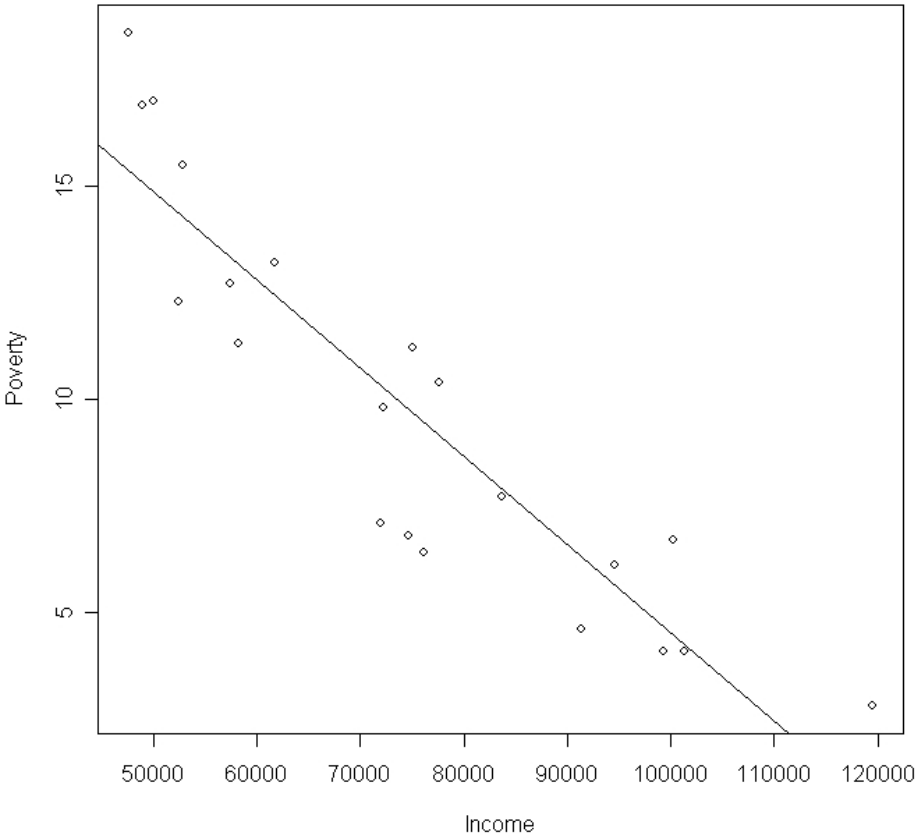
```
> round(var$residuals, 4)
1.5450, -1.0216, 1.5359, 0.2953, -0.6329, 0.0452, -0.5259,
3.6920, 2.6448, 3.4647, 1.4118, -4.0238, -0.4691, -2.4665,
-6.3856, -6.0997, 6.7601, -4.3976, 3.5171, 1.4849, -0.3739
```

The connection between population density and school enrollment is the only moderate-strong one. Notice that the relationship is negative, indicating that counties with a higher population density have lower school enrollment. This implies that children in urban areas are less likely to attend school.

	F Statistic	**P Value**	**R^2**
School v. Income	0.4578	0.5068	0.02353
School v. Poverty	3.034	0.0977	0.1377
School v. Abuse	3.718	0.06892	0.1636
School v. Density	18.41	0.0003944	0.4921

The largest F Statistic comes from school v. density. The other three F Statistics are relatively low. Likewise, the smallest P value is associated with school v. density. The other three P Values are large enough so as to question the validity of the relationship between the respective variables. This confirms our suspicions about density being a moderately-strong influence on school enrollment. The other relationships are probably negligible.

```
> var = lm(Poverty~Income)
> plot(Income,Poverty)
> abline(coef(var))
```



```
> round(var$residuals, 4)
-0.6390, 0.4645, -3.0473, 0.7603, -1.8855, 2.1114, 1.7859,
-3.2126, 3.2094, 2.3426, -0.4572, -0.1840, 2.2491, -0.5642,
1.2471, 1.2030, -2.0762, -0.1472, -1.7163, 1.5158, -2.9595
```

As noted above, there is a very strong correlation between income and poverty. This is obvious because the income of the family directly affects whether or not they are living below the poverty line.

	F Statistic	**P Value**	**R^2**
Poverty v. Income	92.71	9.645e-09	0.8299

The extremely small P Value and large F Statistic indicate an extremely strong relationship between poverty and income – as expected.

Section 5:
Possible correlation of explanatory variables and creation of the simple models.

We want to do two analyses, one with school as the response variable and one with arrests as the response variable. Regardless of what we found to be unlikely variables in section 4 for the two models, here we are going to try every plausible variable. We will search for explanatory variables that should be removed, and for collinearity as we use all possible regressions to find the best possible model.

School model:
Everything that could reasonably be used to explain school enrollment from our data is: abuse, income, poverty, density.
Check if any of these variables should not be used for the model of school.
Test the correlation and the colinearity. Also test the possibility of multilinearity.

```
> round(cor(data[c(2,5,6,7)]), 3)
         Abuse Income Poverty Density
Abuse    1.000 -0.664   0.513  -0.276
Income  -0.664  1.000  -0.911  -0.336
Poverty  0.513 -0.911   1.000   0.532
Density -0.276 -0.336   0.532   1.000
```

There is a very high correlation between poverty and income, very close to negative 1, so we may eliminate one of them.

```
> round(diag(solve(cor(data[c(2,5,6,7)]))),3)
  Abuse  Income Poverty Density
  3.730   8.489   9.278   3.248
```

Both Income and Poverty have high VIFs, close to 10, so one or both of them may be prime targets for removal. This makes sense because having a high average income will usually imply a low percentage of families living below the poverty line.

```
> eigs = eigen(cor(data[c(2,5,6,7)]))
> round(eigs$values, 3)
[1] 2.514 1.292 0.131 0.063
```

The fourth eigen value is very small, so there may be a multi colinearity problem.

```
> round(eigs$vectors, 2)
      [,1]  [,2]  [,3]  [,4]
[1,]  0.43  0.61  0.66 -0.01
[2,] -0.61 -0.06  0.43 -0.66
[3,]  0.61 -0.14 -0.27 -0.73
[4,]  0.27 -0.77  0.54  0.18
```

When we look at the fourth eigen vector however it tells us as did the other three tests that income and poverty are closely related. In order to solve this we think we should remove income. We have the opportunity with R however to test each possible linear model with these four explanatory variables. Using the function from lecture 11:

```
> all.poss.regs(School~Income+Poverty+Abuse+Density, data)
     rssp sigma2 adjRsq    Cp    AIC    BIC     CV Income Poverty Abuse Density
1 221.409 11.653  0.465 5.306 26.306 28.395 27.326      0       0     0       1
2 166.499  9.250  0.576 1.774 22.774 25.907 25.416      0       1     1       0
3 161.254  9.486  0.565 3.245 24.245 28.423 29.644      0       1     1       1
4 158.818  9.926  0.545 5.000 26.000 31.223 33.415      1       1     1       1
```

For AIC and rssp, the best fit is just the single variable density as the only explanatory variable for schooling. This would lead to an under fitted model. We seem to lack the data required to make a fair model of the school enrollment. Both the schooling and density variables are slightly misleading as they were discussed to be in section 3. The correlation between schooling and the other variables of this project were found to be generally low. The relationship between them was not promising as we discussed in section 4. All these factors lead us to feel that our analysis of school enrollment as a function of income, poverty, abuse, and population density may be flawed. Here we decide that we have found no strong relations to build a model of school enrollment with. Abandoning the school model, we shift our focus to the arrests model.

Arrests Model:
Arrests can potentially be a function of all of the other variables, so we must do a multilinearity test for all of them.

```
> round(cor(data[c(2,3,5,6,7,8)]), 3)
         Abuse   Sped Income Poverty Density School
Abuse    1.000  0.697 -0.664   0.513  -0.276  0.405
Sped     0.697  1.000 -0.357   0.178  -0.445  0.418
Income  -0.664 -0.357  1.000  -0.911  -0.336  0.153
Poverty  0.513  0.178 -0.911   1.000   0.532 -0.371
Density -0.276 -0.445 -0.336   0.532   1.000 -0.702
School   0.405  0.418  0.153  -0.371  -0.702  1.000
```

The only closely correlated variables are poverty and income, as from last time. The test from last time showed that the VIF was high for them as well, so before moving on, we deiced to remove income as an explanatory variable.

```
> cor = cor(data[c(2,3,6,7,8)])
> round(diag(solve(cor)), 2)
  Abuse    Sped Poverty Density  School
   5.27    2.25    5.06    3.56    2.71
```

None of the VIFs are very large.

```
> eigs = eigen(cor)
> round(eigs$values, 3)
[1] 2.486 1.792 0.401 0.231 0.090
> round(eigs$vectors, 2)
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,] -0.44 -0.50  0.26 -0.29  0.64
[2,] -0.50 -0.30 -0.67  0.46 -0.09
[3,]  0.11 -0.71  0.20 -0.17 -0.64
```

```
[4,]  0.52 -0.33  0.19  0.69  0.34
[5,] -0.53  0.22  0.64  0.45 -0.25
```

The 5th eigen value is small, and the corresponding eigen vector shows us we should look into collinearity. Once we decide on what model to use, we will look back into the possibility of multicollinearity. It is not a problem until we are trying to find how each individual variable affects the resulting arrests.

Now to create the model for arrests, we again use the function from lecture 11.

```
> all.poss.regs(Arrests~Abuse+Sped+Poverty+Density+School, data)
    rssp sigma2 adjRsq    Cp    AIC    BIC    CV Abuse Sped Poverty Density School
1 16.067  0.846  0.676 5.109 26.109 28.198 1.884     1    0       0       0      0
2 14.101  0.783  0.700 4.404 25.404 28.538 1.814     1    0       0       0      1
3 11.731  0.690  0.736 3.143 24.143 28.321 1.649     1    1       0       0      1
4 11.558  0.722  0.723 4.905 25.905 31.127 1.906     1    1       1       0      1
5 10.900  0.727  0.722 6.000 27.000 33.267 1.899     1    1       1       1      1
```

Using the AIC goodness test, the best model would be using all 5 variables. Using the adjusted R squared value though the best model would be Arrests as function of just abuse, sped and school. Doing the step process uses the adjusted R squared value as the signifier so we get.

```
> step(lm(Arrests~1), Arrests~Abuse+Sped+Poverty+Density
+School, trace=0)

Call:
lm(formula = Arrests ~ Abuse + School + Sped)

Coefficients:
(Intercept)        Abuse       School         Sped
    2.82510      0.99600     -0.09055      0.25306
```

With a very close AIC value we have pretty much the same explaining power. Given the choice between high and low number of variables, we pick the simpler model of only abuse school and sped. We may try adding in variations of the other variables later and see if we get a higher R squared value.

```
> model1 = lm(Arrests~Abuse+School+Sped)
> summary(model1)
Residuals:
    Min      1Q  Median      3Q     Max
-1.1172 -0.4875 -0.1767  0.3792  2.3599

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.82510    2.98349   0.947  0.35695
Abuse        0.99600    0.22305   4.465  0.00034 ***
School      -0.09055    0.04447  -2.036  0.05764 .
Sped         0.25306    0.13656   1.853  0.08132 .

Residual standard error: 0.8307 on 17 degrees of freedom
Multiple R-Squared: 0.7754,     Adjusted R-squared: 0.7358
F-statistic: 19.56 on 3 and 17 DF,  p-value: 9.42e-06
```

Section 6:

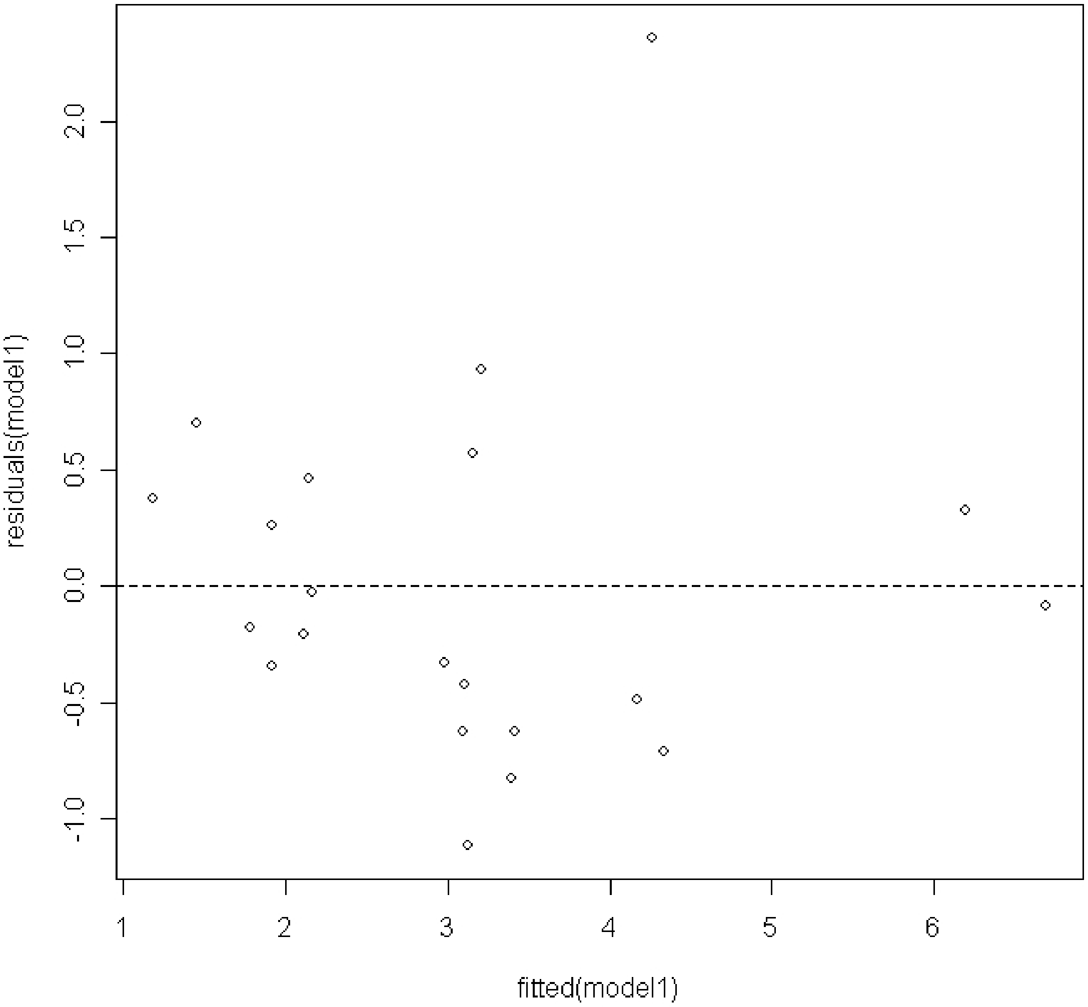
Residual Diagnostics and variation of the model:

Now that the model has less explanatory variables we can check for collinearity again

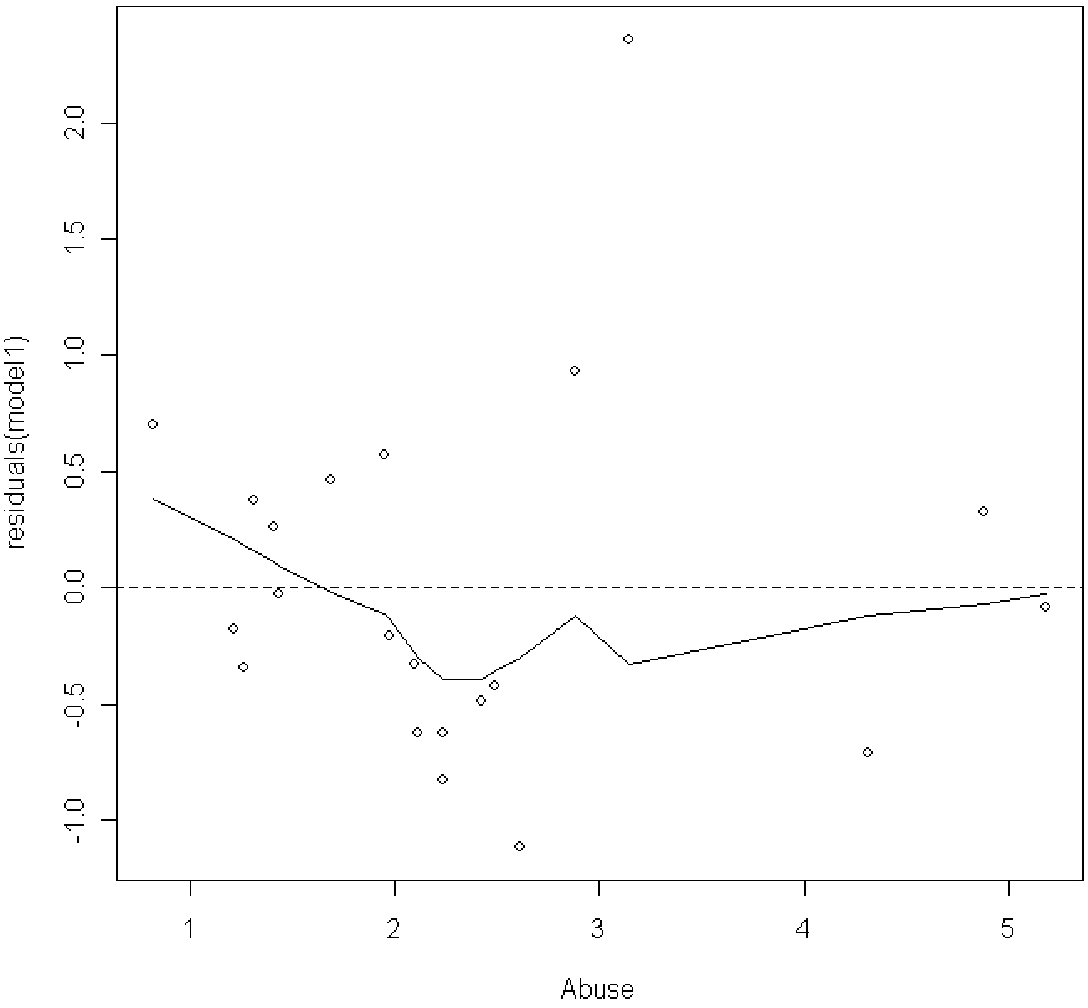
```
> cor = cor(data[c(2,3,8)])
> round(diag(solve(cor)),2)
 Abuse   Sped School
  2.00   2.03   1.25
> eigs = eigen(cor)
> round(eigs$values, 3)
[1] 2.026 0.670 0.303
```

and we find there is none, so we can move on to the residuals.

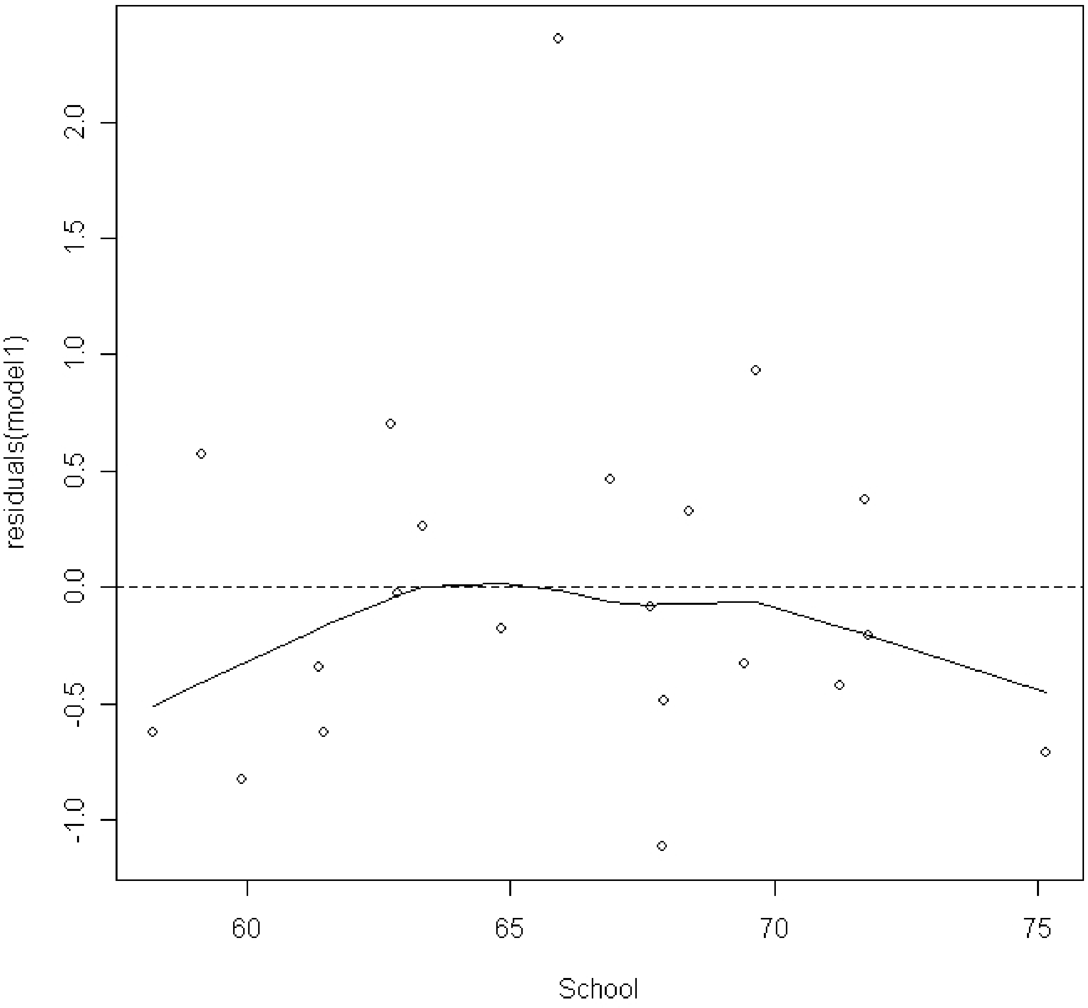
```
> plot(fitted(model1), residuals(model1))
> abline(0,0,lty=2)
```



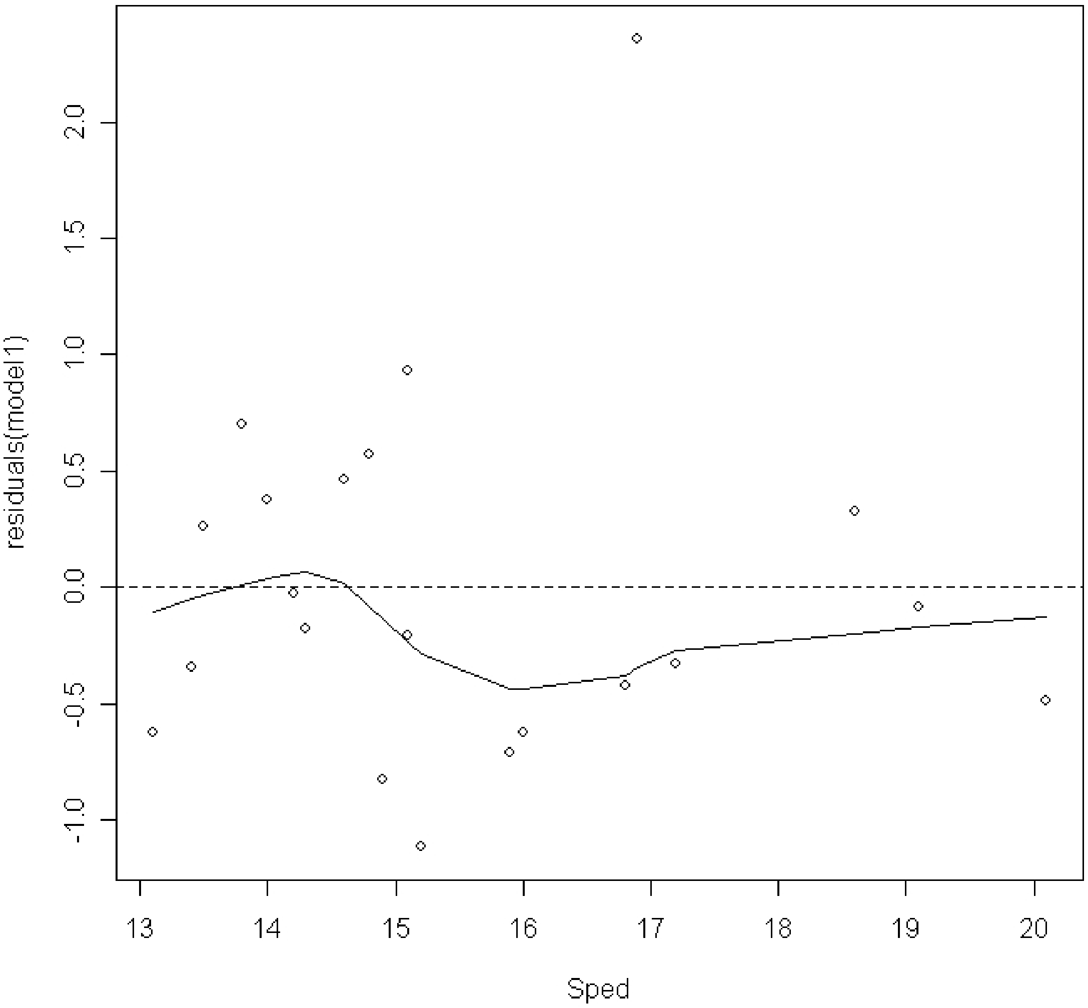
```
> plot(Abuse, residuals(model1))
> lines(lowess(Abuse, residuals(model1)))
> abline(0,0,lty=2)
```



```
> plot(School, residuals(model1))
> lines(lowess(School, residuals(model1)))
> abline(0,0,lty=2)
```

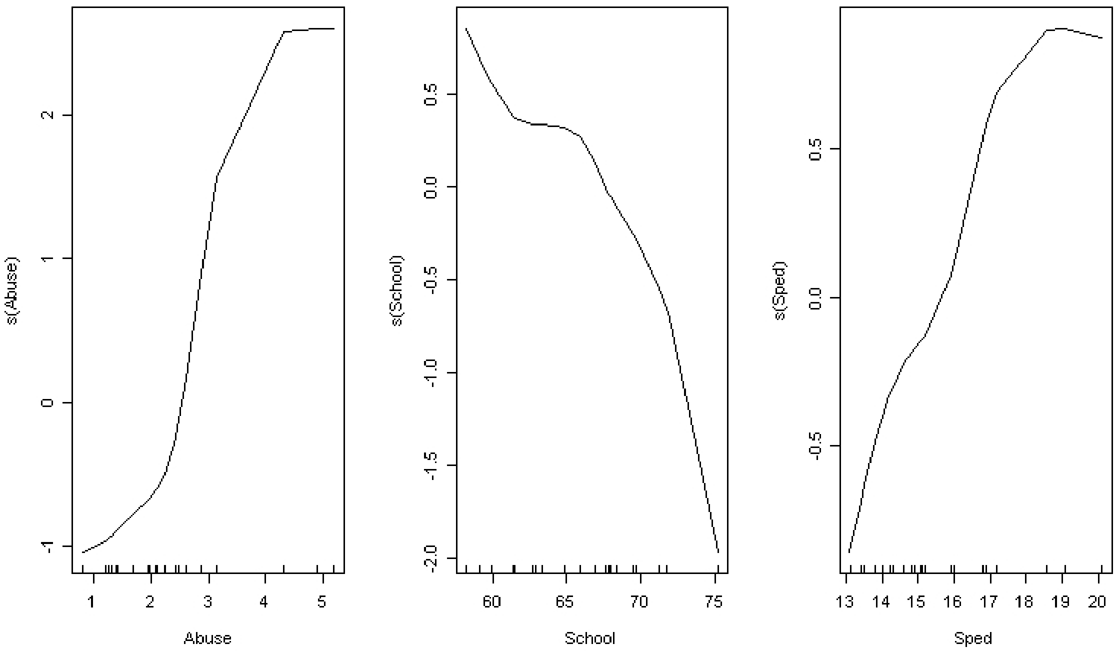


```
> plot(Sped, residuals(model1))
> lines(lowess(Sped, residuals(model1)))
> abline(0,0,lty=2)
```



There is no obvious relationship between the residuals and any of the explanatory variables or the fitted values. There is no obvious way to change the model to reduce the residuals based on this. The only variable that looks like it could use transformation because it has a relatively smooth curve is school. We use GAM to see what transformation might be applicable.

```
> par(mfrow = c(1,3))
> plot(gam(Arrests~s(Abuse)+s(School) + s(Sped)))
```



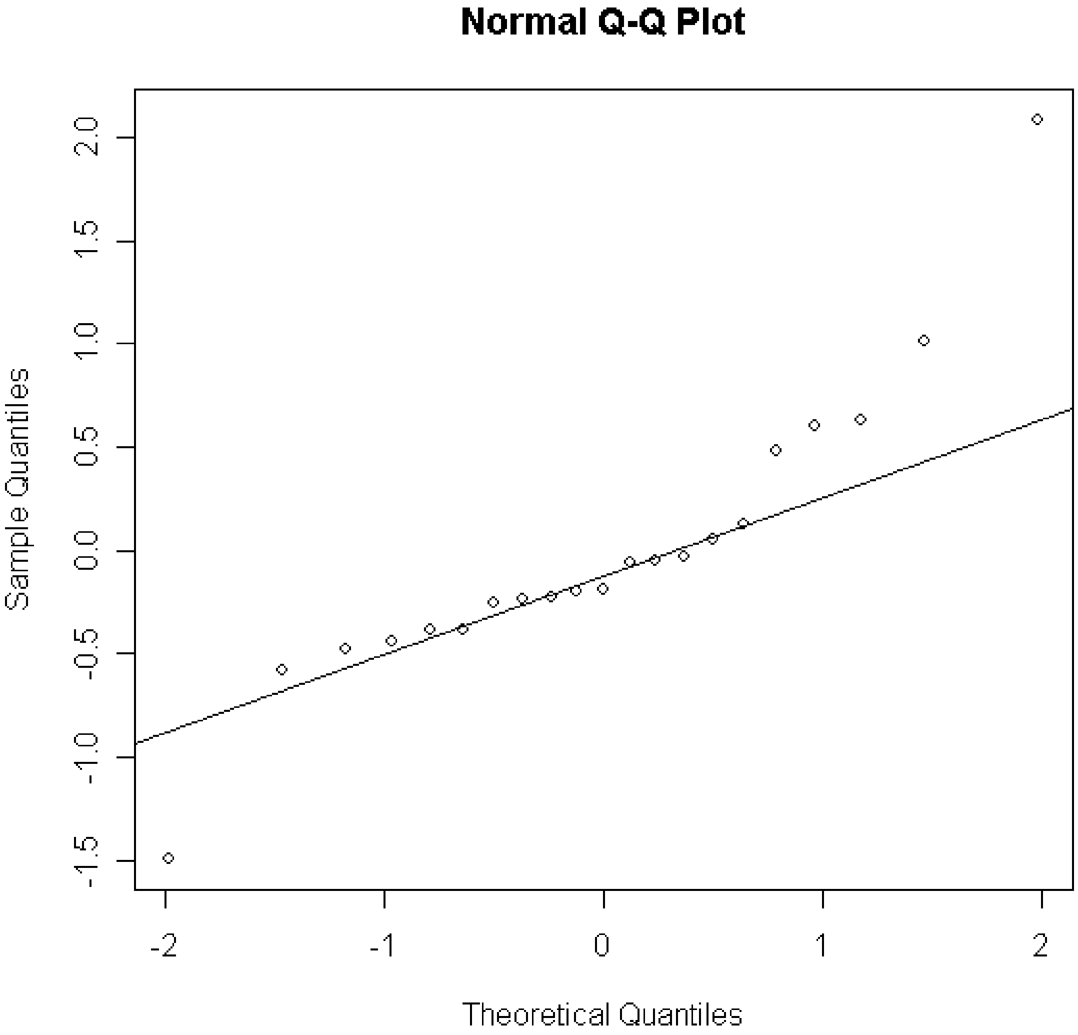
Each curve looks odd, and not like a particular function. School could be transformed with a cubic, so we'll try that.

```
> model2 = lm(Arrests~Abuse + poly(School,3)+Sped)
> summary(model2)
…
Multiple R-Squared: 0.8101,     Adjusted R-squared: 0.7467
…
```

This is better than the 0.7358 for model1, but not by much.

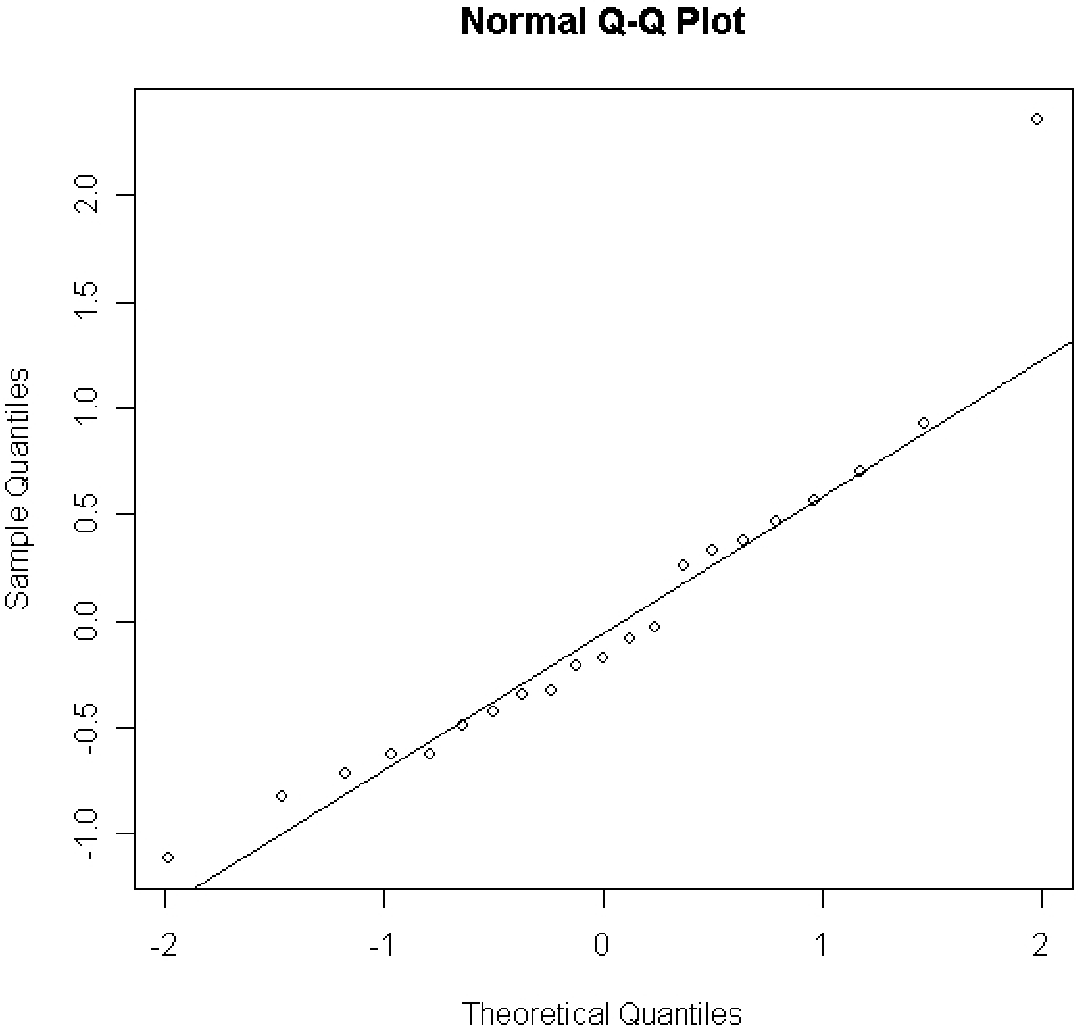
Now we check for the normality of the residuals.

```
> qqnorm(residuals(model2))
> qqline(residuals(model2))
```



This is not a very normal set of residuals, so we check the other model to see if its residuals are more normally distributed.

```
> qqnorm(residuals(model1))
> qqline(residuals(model1))
```

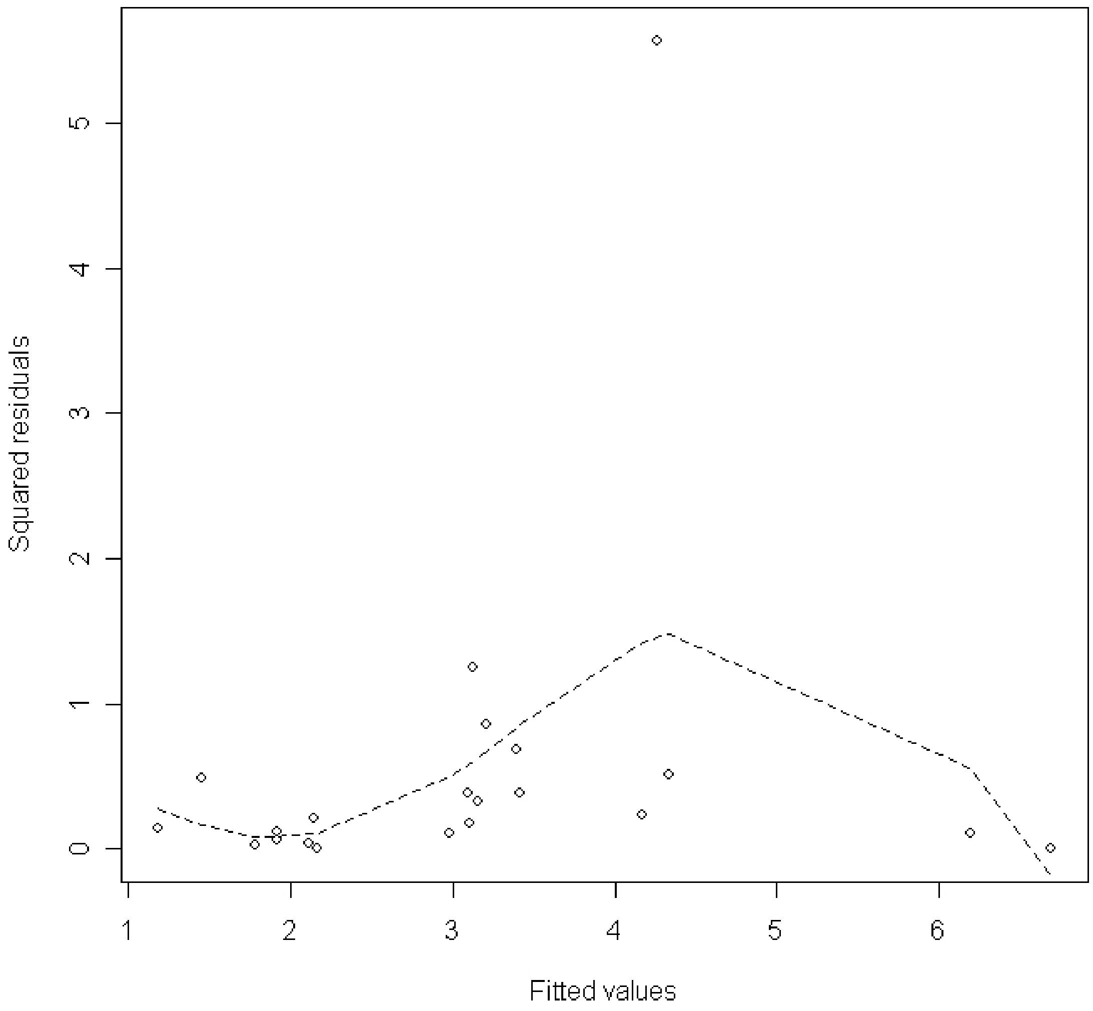


These are indeed a nice fit to the normal distribution, so we will stick with the simpler linear model after all. The residuals need to be normally distributed for our model to be good.
Next use durbin-watson to test for independence.

```
> durbin.watson(model1)
 lag Autocorrelation D-W Statistic p-value
   1      -0.1451842      1.981533 0.472
 Alternative hypothesis: rho != 0
```

This statistic is fine.

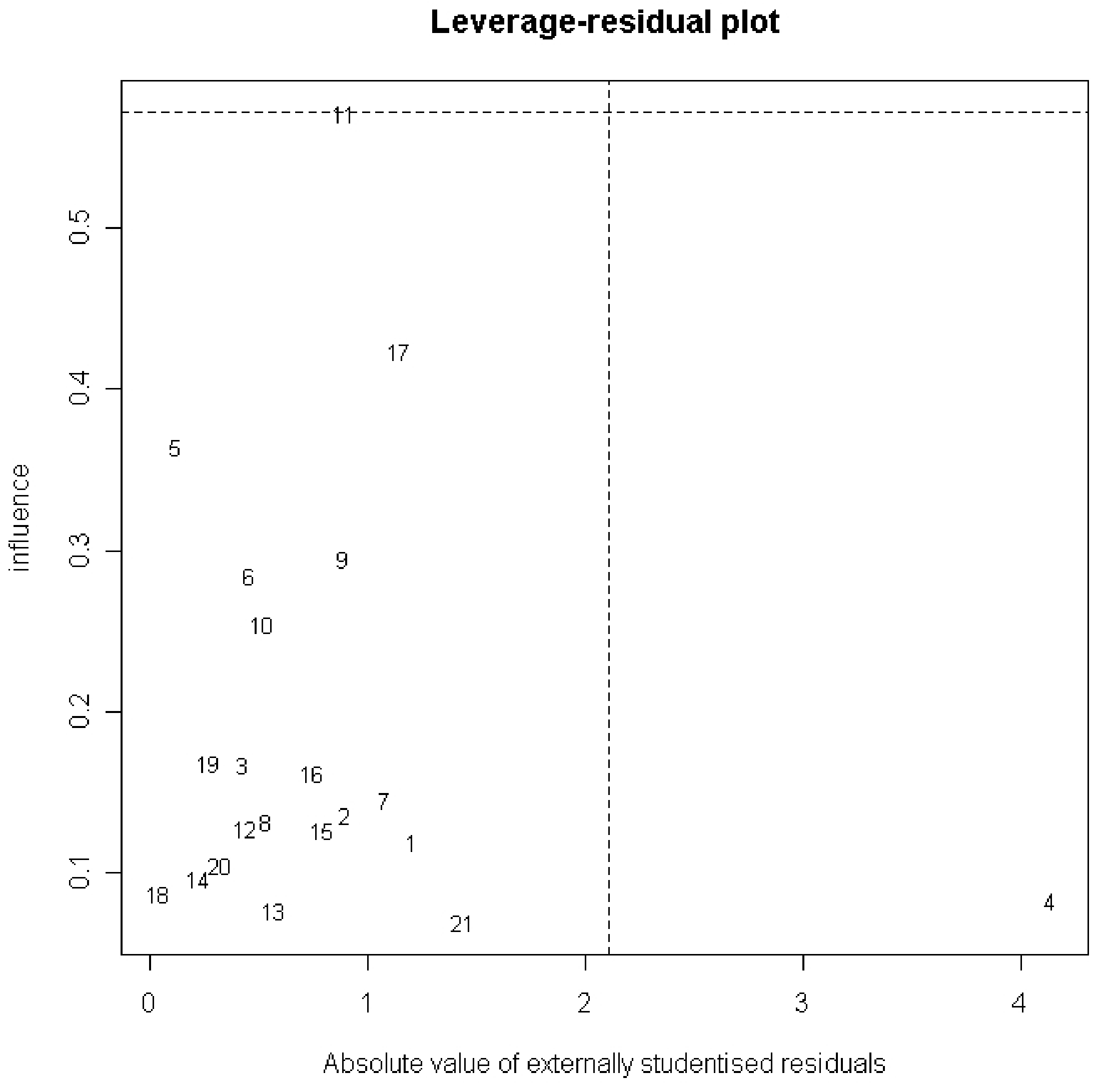
```
> funnel(model1)
```



The variance seems to be close to 0 with the exception of one outlier that throws it way off. We need to eliminate this outlier to have a relatively constant variance.

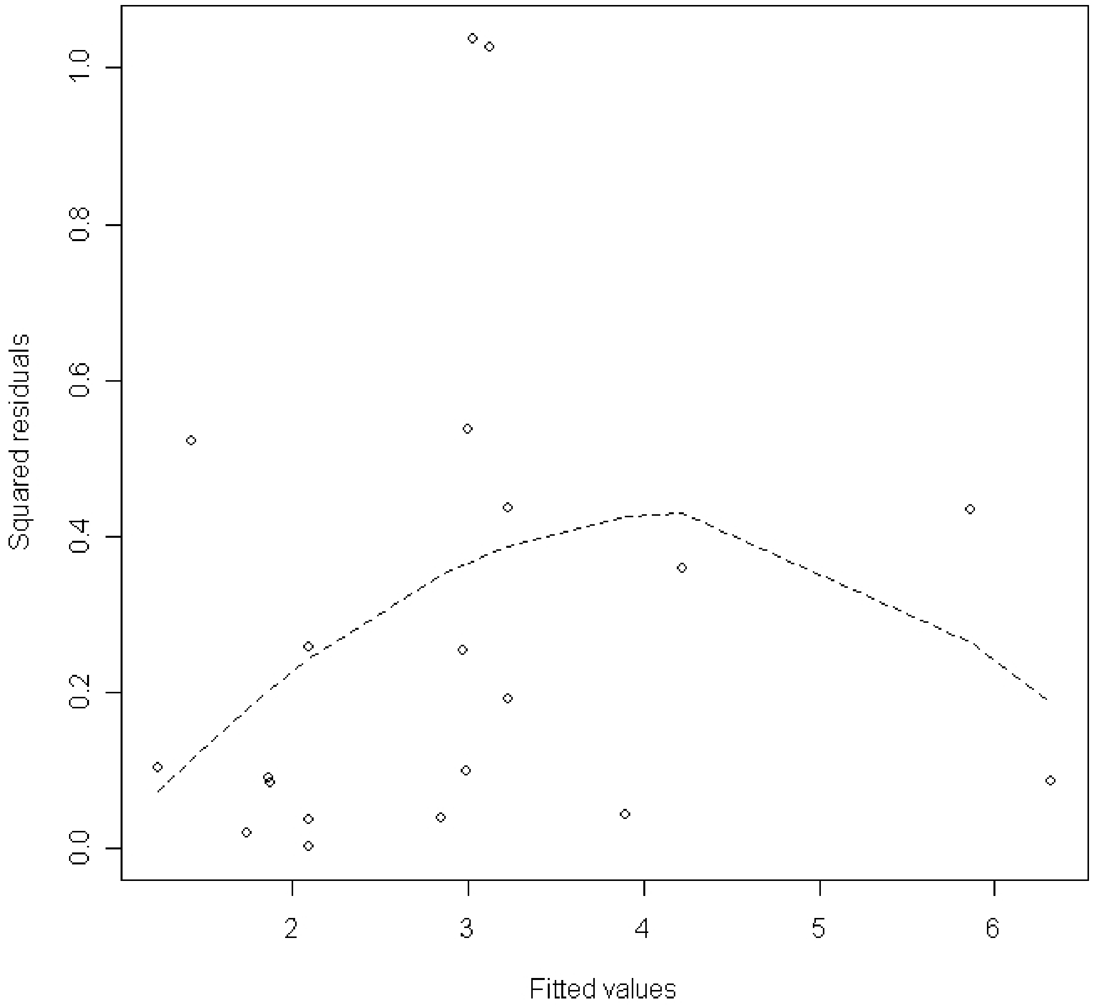
To eliminate this outlier we should find it with an lrplot

```
> lrplot(model1)
```



Clearly value 4 is both very non-influential and has a very high externally studentised residual. We can remove it without any large change to the model.

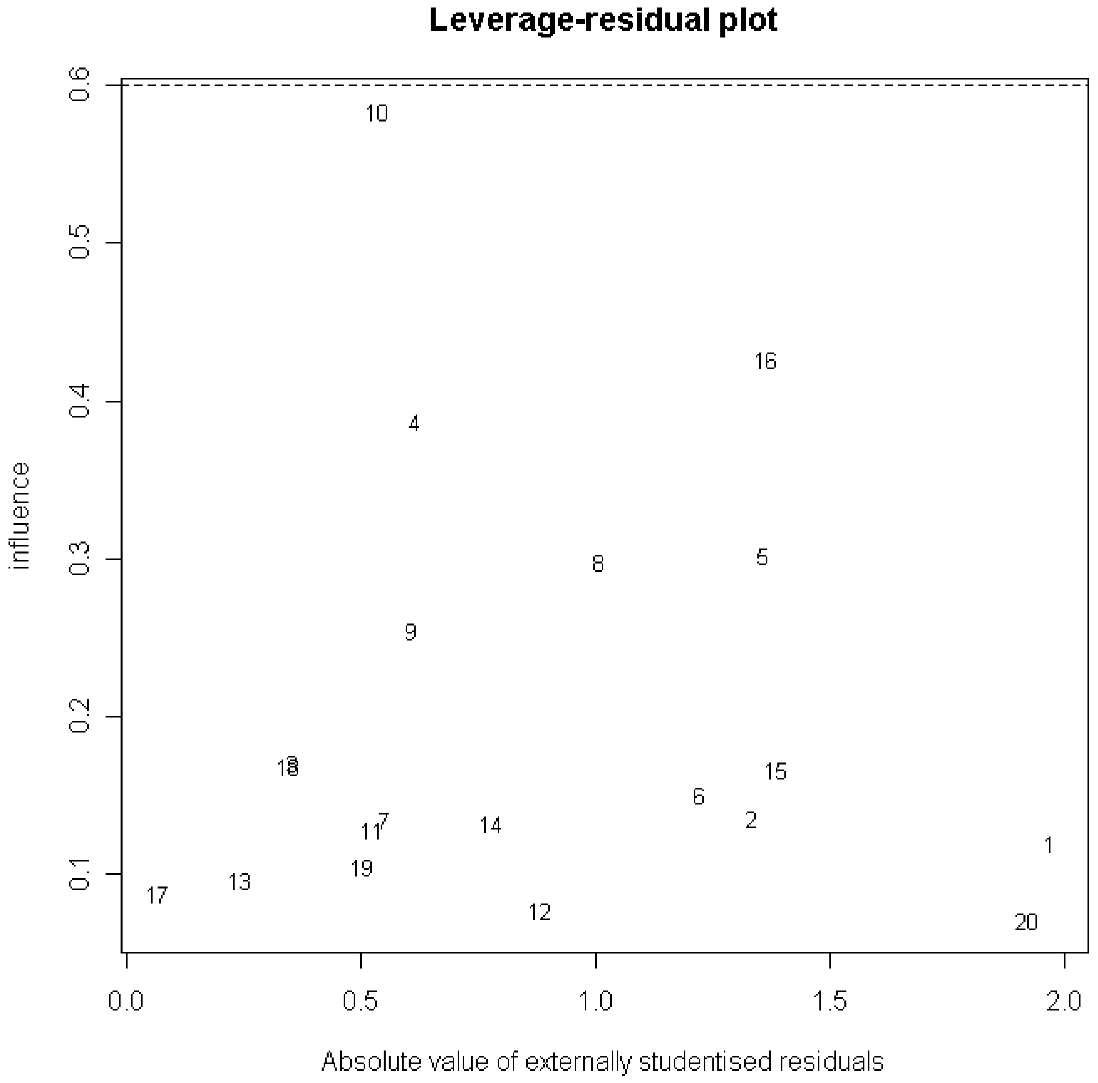
```
> v = c(1,2,3,5:21)
> data2 = data[v,1:8]
> detach(data)
> attach(data2)
> model3 = lm(Arrests~Sped+Abuse+School)
> funnel(model3)
```



This looks much better as the residuals are now low and more linear.

Check for any other outliers:

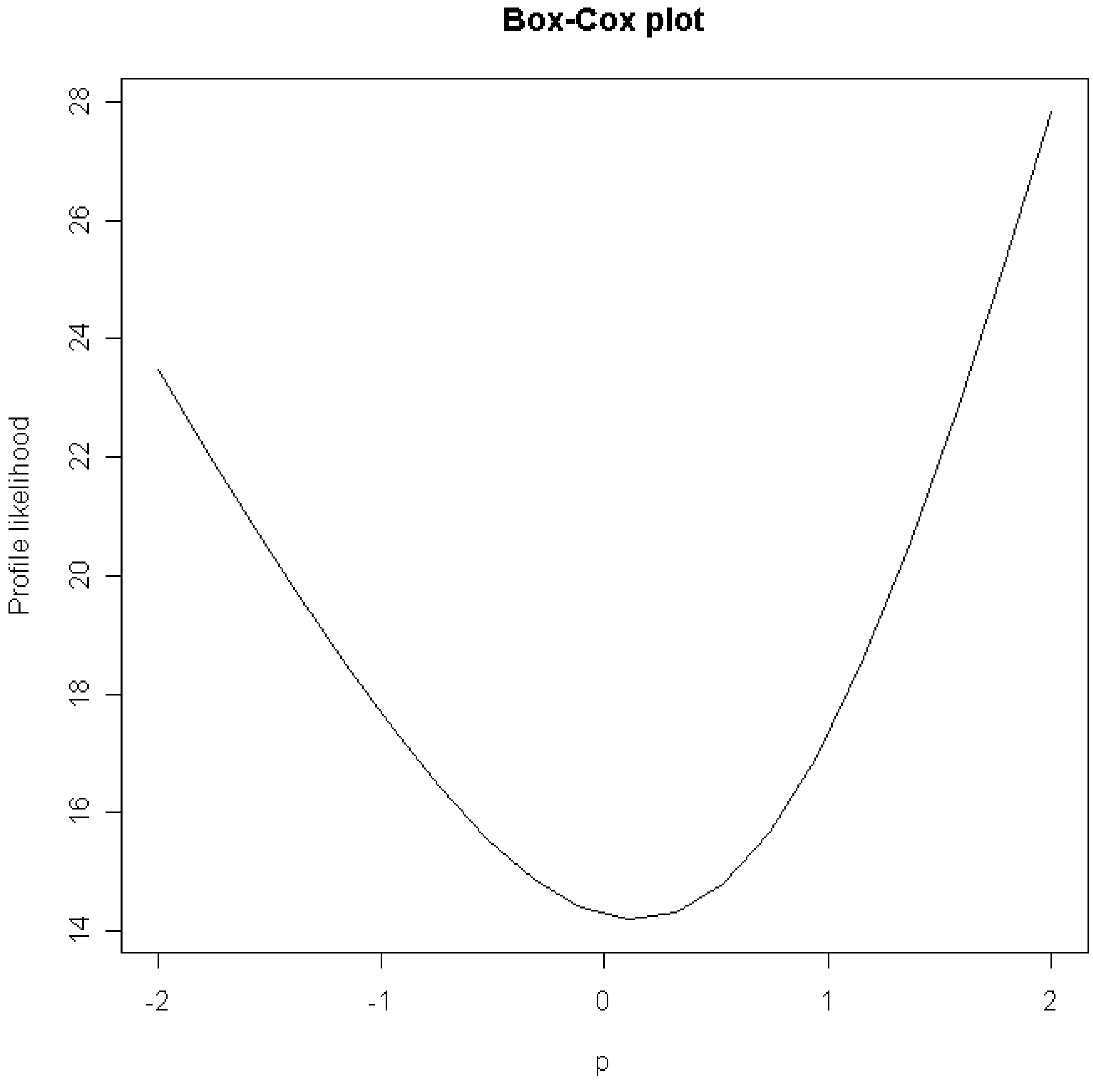
```
> lrplot(model3)
```



Looks great.

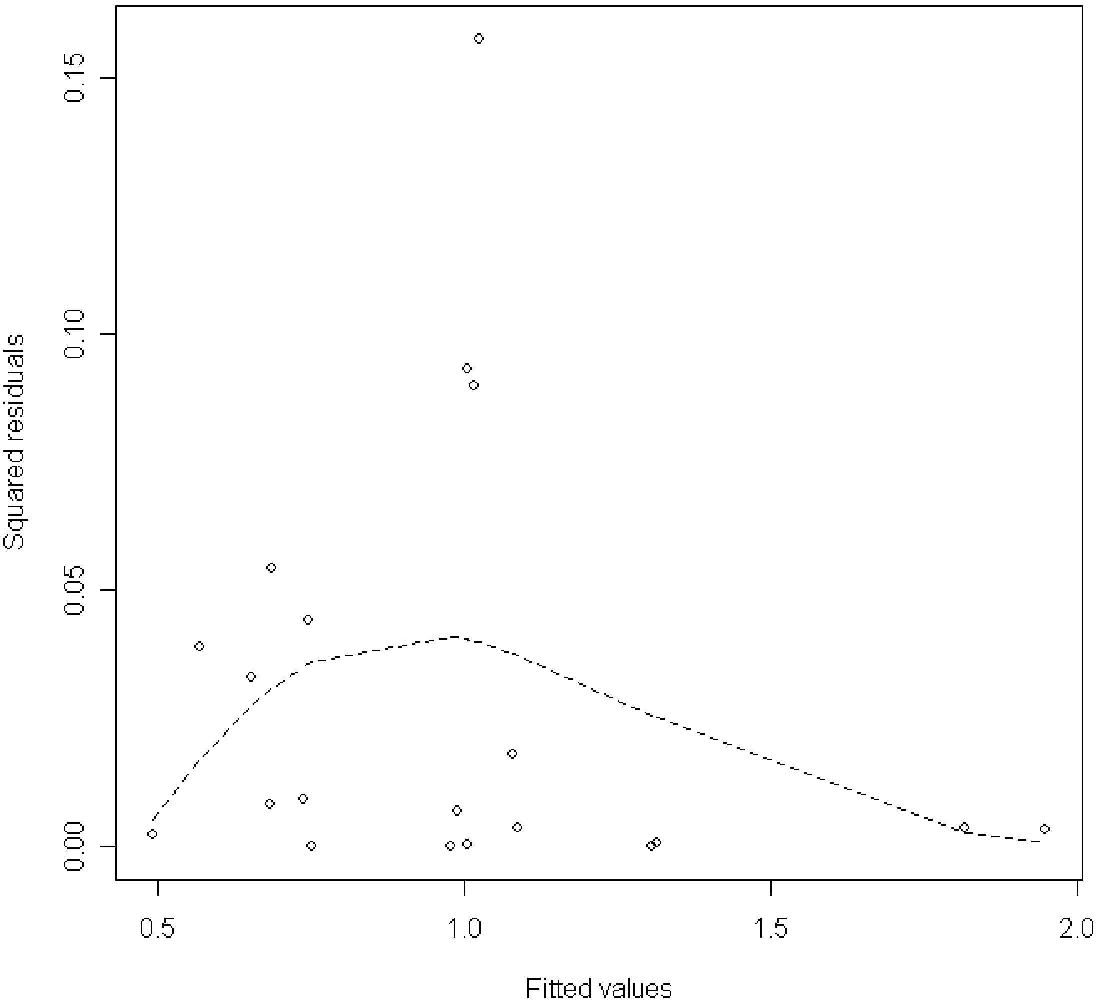
We might be able to use a box cox plot to get a better fit.

```
> boxcoxplot(Arrests~Sped+Abuse+School, data2)
```



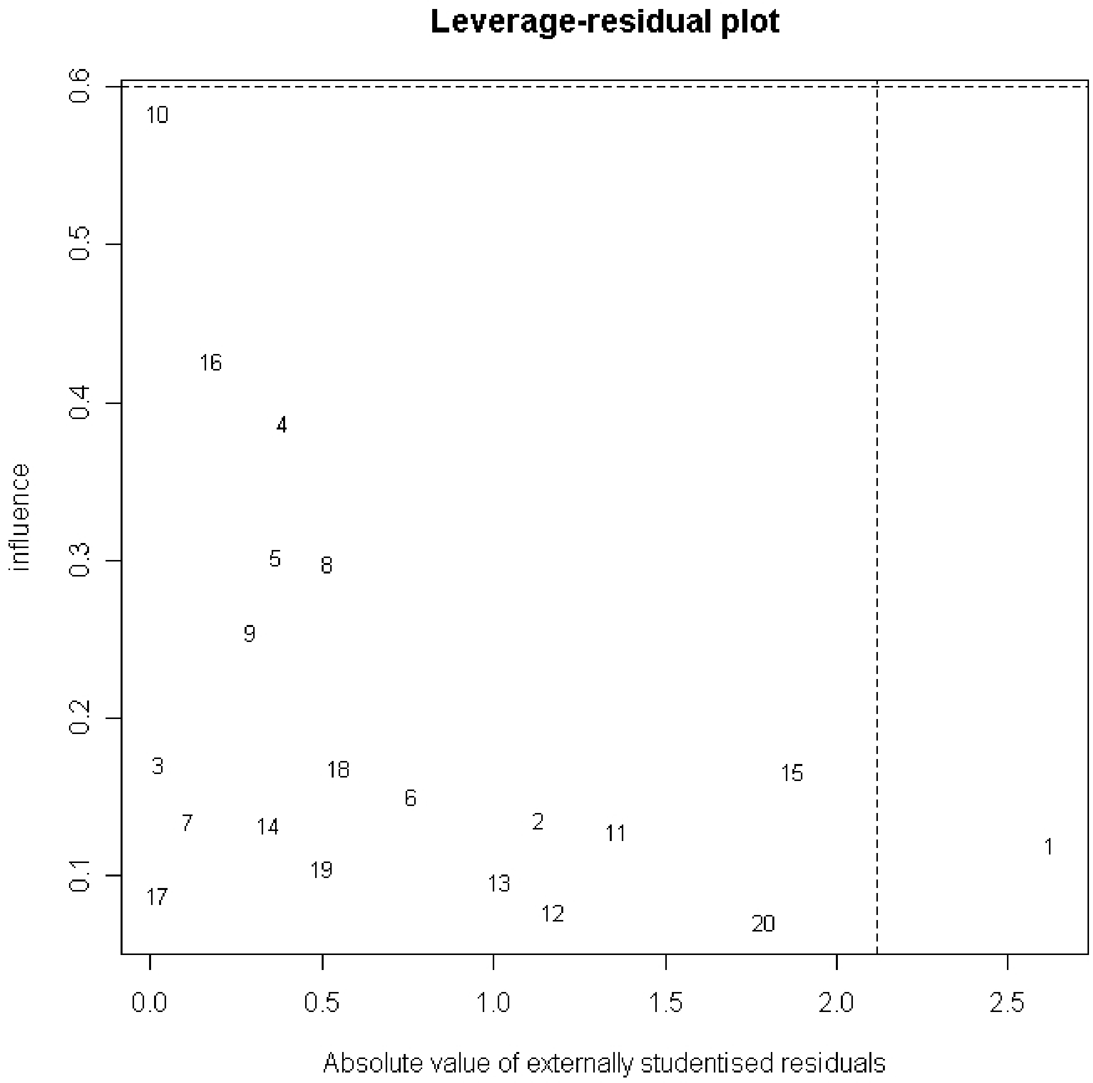
As you can see the lowest point is very close to 0, so we should try a logarithmic transform.

```
> model4 = lm(log(Arrests)~Sped+Abuse+School)
> funnel(model4)
```



more linear, but seems like there is one large outlier again, do another lrplot to try and find it, to remove it. The adj R^2 value goes up to 0.795

```
> lrplot(model4)
```



lets remove data point 1

```
> data3 = data2[2:20,1:8]
> model 5 = lm(log(Arrests)~Sped+Abuse+School, data=data3)
```

now the lrplot is nice, and the funnel still has the two points up above, but we cant remove to many counties or we'll have a model that's only accurate for so little.
And now the adj R^2 value is up to 0.8494

The model seems good, now to find which variables are most influential with anova:

```
> anova(model5)
Analysis of Variance Table

Response: log(Arrests)
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Sped       1 1.80173 1.80173  69.385 5.218e-07 ***
Abuse      1 0.64287 0.64287  24.757 0.0001660 ***
School     1 0.26970 0.26970  10.386 0.0056924 ** 
Residuals 15 0.38951 0.02597
```

The anova shows that sped is the most influential variable and that school is no all that influential.

```
> anova(lm(log(Arrests)~Abuse+School+Sped, data = data3))
Analysis of Variance Table

Response: log(Arrests)
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Abuse      1 2.28545 2.28545 88.0123 1.150e-07 ***
School     1 0.16450 0.16450  6.3348  0.023702 *  
Sped       1 0.26436 0.26436 10.1804  0.006079 ** 
Residuals 15 0.38951 0.02597
```

When you put abuse before sped, it accounts for much of the SSQ, which makes sense because they have a decently high correlation. When you put school first however it still has a small number, as it is not that influential and not very linearly related to the other two.

Finally we have found a decent model for arrests, and which variables are more influential; the conclusion is on the first page.

The model is log(Arrests) = 0.956301 + 0.086790(sped) +0.236819(abuse)-0.028434(school)

Abuse has the largest coefficient, and also had much of the SSQ accounted for it.

```
> summary(model5)

Call:
lm(formula = log(Arrests) ~ Sped + Abuse + School, data = 
data3)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.2651261 -0.0667853  0.0007264  0.0810304  0.2951937 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.956301   0.579752   1.649  0.11982    
Sped         0.086790   0.027201   3.191  0.00608 ** 
Abuse        0.236819   0.043972   5.386 7.57e-05 ***
School      -0.028434   0.008823  -3.223  0.00569 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1 

Residual standard error: 0.1611 on 15 degrees of freedom
Multiple R-Squared: 0.8745,     Adjusted R-squared: 0.8494 
F-statistic: 34.84 on 3 and 15 DF,  p-value: 5.308e-07
```