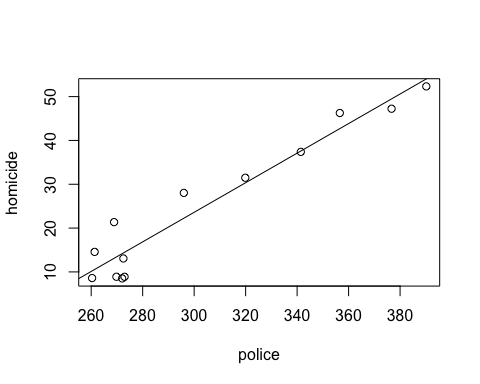
Applied Statistics Mini Project (Spring 2017)

## Math 7343 Mini Project

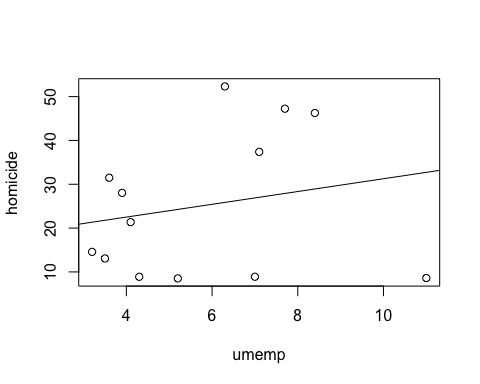
### Visual Analysis

library(readxl)  
detroit <- read\_excel("~/Documents/Graduate-Math-Projects/detroit.xlsx")  
plot(homicide ~ police, data = detroit)  
abline(lm(homicide ~ police, data = detroit))



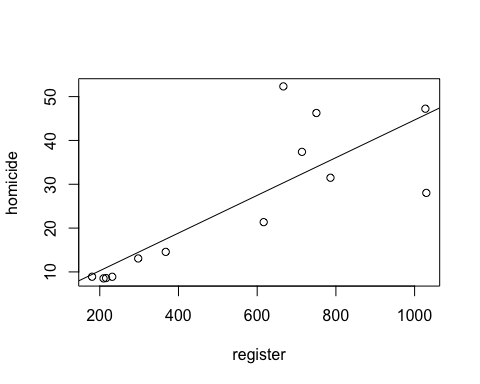
This graph shows a linear relationship that shows a strong, positive, and highly correlated relationship between between the two variables. This means that when there are more police, there tends to be more homicides.

plot(homicide ~ umemp, data = detroit)  
abline(lm(homicide ~ umemp, data = detroit))



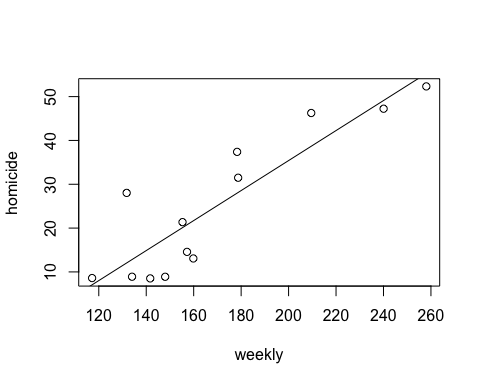
This graph shows a slight linear relationship that is pretty weak, positive and does not have a highly correlated relationship between the two variables. This means we cannot determine whether unemployment affects the amount of homicides.

plot(homicide ~ register, data = detroit)  
abline(lm(homicide ~ register, data = detroit))



The graph shows a linear relationship that is strong, positive and has a moderately correlated relationship between the two variables. This means as the amount of handgun registers increases, the amount of homicides will probably increase.

plot(homicide ~ weekly, data = detroit)  
abline(lm(homicide ~ weekly, data = detroit))



The graph shows a linear relationship that is strong, positive and has a moderately correlated relationship between the two variables. This means as the amount of weekly earnings increases, the amount of homicides will probably increase.

### Linear Regression Models

reg.fit <- lm(homicide ~ police, data = detroit)  
reg.fit1 <- lm(homicide ~ umemp, data = detroit)  
reg.fit2 <- lm(homicide ~ register, data = detroit)  
reg.fit3 <- lm(homicide ~ weekly, data = detroit)

summary(reg.fit)

##   
## Call:  
## lm(formula = homicide ~ police, data = detroit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.649 -2.244 -1.258 3.559 8.254   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -77.63026 8.63094 -8.994 2.11e-06 \*\*\*  
## police 0.33745 0.02804 12.035 1.13e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.547 on 11 degrees of freedom  
## Multiple R-squared: 0.9294, Adjusted R-squared: 0.923   
## F-statistic: 144.8 on 1 and 11 DF, p-value: 1.129e-07

summary(reg.fit1)

##   
## Call:  
## lm(formula = homicide ~ umemp, data = detroit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -24.128 -14.059 -1.297 10.354 26.462   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 16.673 12.735 1.309 0.217  
## umemp 1.460 2.047 0.713 0.491  
##   
## Residual standard error: 16.73 on 11 degrees of freedom  
## Multiple R-squared: 0.04416, Adjusted R-squared: -0.04274   
## F-statistic: 0.5082 on 1 and 11 DF, p-value: 0.4908

summary(reg.fit2)

##   
## Call:  
## lm(formula = homicide ~ register, data = detroit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.914 -2.901 -2.154 1.398 22.007   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.662063 5.708193 0.291 0.776337   
## register 0.043003 0.009175 4.687 0.000664 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.886 on 11 degrees of freedom  
## Multiple R-squared: 0.6663, Adjusted R-squared: 0.636   
## F-statistic: 21.97 on 1 and 11 DF, p-value: 0.0006642

summary(reg.fit3)

##   
## Call:  
## lm(formula = homicide ~ weekly, data = detroit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.709 -6.182 -1.877 3.361 15.987   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -33.05875 9.33585 -3.541 0.00462 \*\*   
## weekly 0.34233 0.05341 6.410 5.01e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.865 on 11 degrees of freedom  
## Multiple R-squared: 0.7888, Adjusted R-squared: 0.7696   
## F-statistic: 41.09 on 1 and 11 DF, p-value: 5.011e-05

The variables of police, handgun registration, and weekly earnings are considered to be statistically significant as their p-values – 1.13 multiplied by 10^(-7), 6.6 multiplied by 10^(-4), and 5.01 multiplied by 10^(-5) respectively – are less than the 0.05 significance level. The unemployment variable has a p-value of 0.4908, which is a lot larger than the 0.05 significance level. The p-values represent the probability that the effect we see comes from chance, so the three identified as significant have very low probabilities of being the result of random chance, while the unemployment rate is about even odds that the effect is the result of random chance.

The adjusted R-squared values for police, unemployment, handgun registration, and weekly earnings are 0.923, -0.0427, 0.636, and 0.769 respectively. This means that the police variable explains the largest percentage of variation in the homicide rate. We came to this conclusion since R-squared is used to provide the proportion of variation in the response that is attributable to the explanatory variable, it is the best value to use to determine which explanatory variable explains the largest proportion of the response variance.

### Multiple Linear Regression Model

reg.fit4 <- lm(formula = homicide ~ police + umemp + register + weekly, data = detroit)  
library(MASS)  
step <- stepAIC(reg.fit4, direction = "both")

## Start: AIC=34.53  
## homicide ~ police + umemp + register + weekly  
##   
## Df Sum of Sq RSS AIC  
## - umemp 1 4.178 89.973 33.149  
## <none> 85.795 34.531  
## - weekly 1 18.311 104.106 35.046  
## - police 1 41.330 127.125 37.643  
## - register 1 119.890 205.685 43.898  
##   
## Step: AIC=33.15  
## homicide ~ police + register + weekly  
##   
## Df Sum of Sq RSS AIC  
## - weekly 1 14.298 104.272 33.067  
## <none> 89.973 33.149  
## + umemp 1 4.178 85.795 34.531  
## - police 1 105.274 195.248 41.221  
## - register 1 137.422 227.395 43.203  
##   
## Step: AIC=33.07  
## homicide ~ police + register  
##   
## Df Sum of Sq RSS AIC  
## <none> 104.27 33.067  
## + weekly 1 14.30 89.97 33.149  
## + umemp 1 0.17 104.11 35.046  
## - register 1 123.14 227.41 41.204  
## - police 1 970.76 1075.03 61.397

By using the stepwise selection method, we were able to determine that the best multiple regression model has police and registered handguns per 100,000 as the explanatory values for homicide rates.

summary(lm(homicide ~ police + register, data = detroit))

##   
## Call:  
## lm(formula = homicide ~ police + register, data = detroit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.3398 -1.8011 -0.1346 2.3278 4.8227   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -64.96371 7.15240 -9.083 3.81e-06 \*\*\*  
## police 0.26993 0.02797 9.649 2.20e-06 \*\*\*  
## register 0.01447 0.00421 3.437 0.00637 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.229 on 10 degrees of freedom  
## Multiple R-squared: 0.9676, Adjusted R-squared: 0.9612   
## F-statistic: 149.5 on 2 and 10 DF, p-value: 3.551e-08

It can be assumed that the chance of such a relationship being the result of pure chance is miniscule and these two explanatory variables account for 96.12% of the variables in the response variable by looking at the p-value of 3.551\*10-8 and the adjusted R-squared value of 0.9612.

The appropriate model for this relationship is: y = -64.963 + 0.26993x + 0.01447z

If the number of registered handguns is held constant, then the model indicates that each additional police officer is correlated with an additional .26993 homicides, while if the number of police officers is held constant, then the model indicates that each additional handgun is correlated with an additional .0145 homicides. Therefore, the number of police officers explains the larger percentage of variation in the homicide rate.

### Prediction

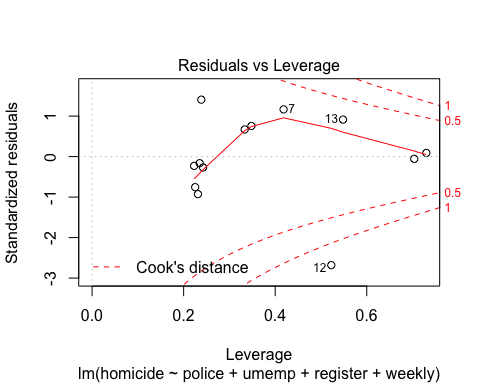
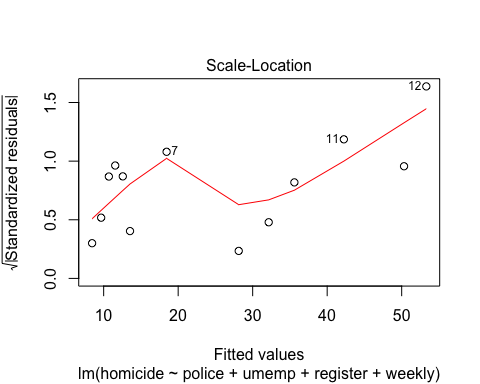
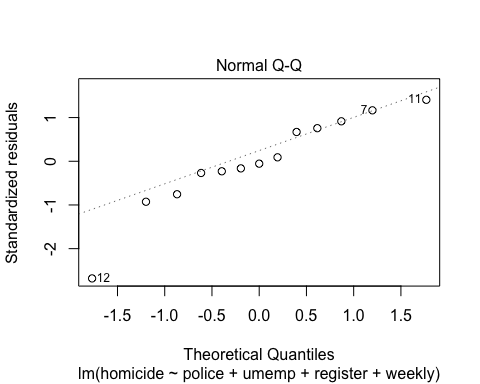
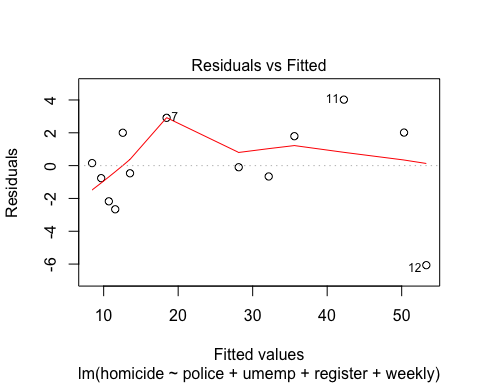
predict(reg.fit4, newdata = detroit[10,], interval = "prediction")

## fit lwr upr  
## 10 35.59766 26.87764 44.31768

The 95% confidence interval is 26.87764 to 44.31768 homicides per 100,00 people in the year 1970.

### Regression Plots

plot(reg.fit4)



The Residual vs Fitted plot shows some outliers, but for the most part, the values lie near zero which indicates that the model is not biased in one direction or the other. Residuals are the difference between the actual values and the predicted values from the linear model, and if they were very large or found mostly on one side of zero, it would cast doubt on the validity of the model. This model is considered a good predictive one since the residuals are small values that are balanced on either side of zero.

The Normal Q-Q plot still show one outlier, but is considered a better model as all the other values are found right along the line that has a slope of 1. This confirms the normality of the data given in the problem. The Normal Q-Q plot plots the standardized quantiles against the theoretical quantiles. That is, it plots the values actually predicted by your model against the values one would expect if it was perfectly accurate and perfectly normal.

The Scale-Location plot shows that the residuals appear randomly spread. This confirms the assumption that there is equal variance in the data.

The Residuals vs Leverage plot shows that there are no influential cases in the data. Even though the Cook's distance lines arevisible, most of the data points are well inside these distance lines - with the exception of 1 outlier.