Data 603: Statistical Modelling with Data

Multiple Linear Regression

Part IV: MULTIPLE REGRESSION DIAGNOSTICS

Residual Analysis: Checking the Regression Assumptions

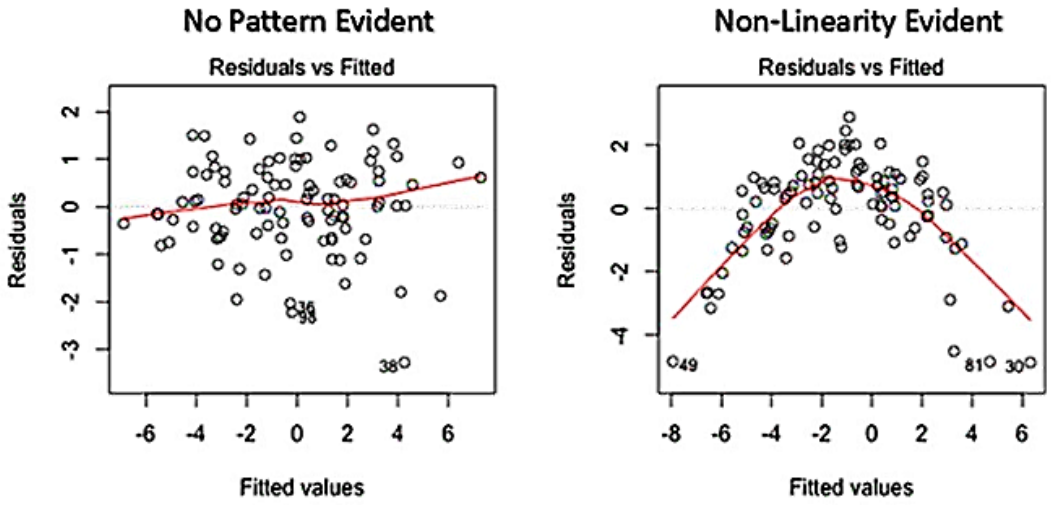
Most statistical tests rely upon certain assumptions about the variables used in the analysis. **When these assumptions are not met the results may not be trustworthy**. The assumptions and conditions for the multiple regression model sound nearly the same as for simple linear regression, but with more variables in the model.

1. Linearity Assumption

The linear regression model assumes that there is a straight-line relationship between the predictors and the response. If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect. In addition, the prediction accuracy of the model can be significantly reduced.

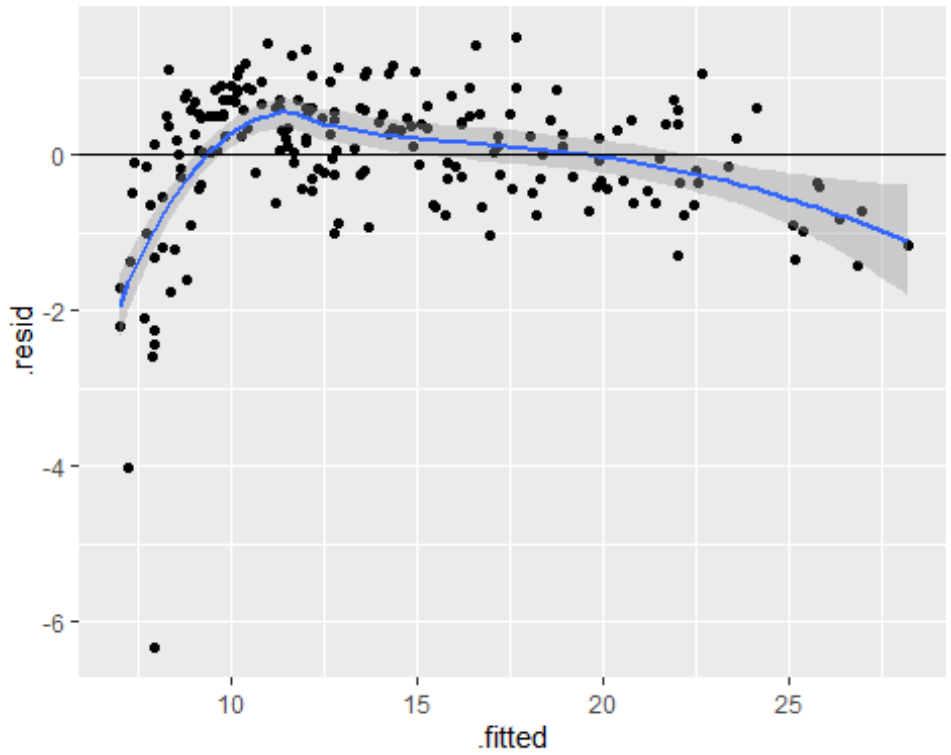
Residual plots are a useful graphical tool for identifying non-linearity. In the case of multiple regression model since there are multiple predictors, we instead plot the residuals versus predicted (or fitted) values $\widehat{y}_i$. Ideally, the residual plot will show no discernible pattern. The presence of a presence may indicate a problem with some aspect of the linear model.

If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to use non-linear transformations of the predictors, such as $\log(X)$, $\sqrt{X}$, and X^2, in the regression model.

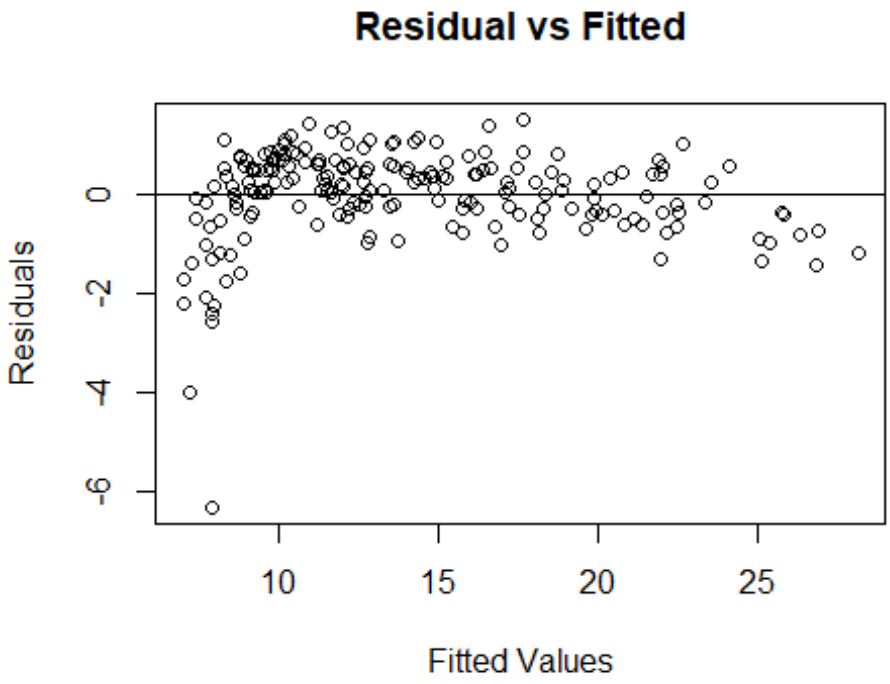


This scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values)

```
library(ggplot2)
Advertising=read.table("c:/Users/thuntida.ngamkham/OneDrive - University of
Calgary/dataset603/Advertising.txt", header = TRUE,  sep ="\t" )
model<-lm(sale~tv+radio+tv*radio, data=Advertising)
summary(model)

##
## Call:
## lm(formula = sale ~ tv + radio + tv * radio, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3366 -0.4028  0.1831  0.5948  1.5246
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
## tv          1.910e-02  1.504e-03  12.699   <2e-16 ***
## radio       2.886e-02  8.905e-03   3.241   0.0014 **
## tv:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9435 on 196 degrees of freedom
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16

ggplot(model, aes(x=.fitted, y=.resid)) +
  geom_point() + geom_smooth()+
  geom_hline(yintercept = 0)

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```r
#optional graph from plot()
plot(fitted(model), residuals(model),xlab="Fitted Values", ylab="Residuals")
abline(h=0,lty=1)
title("Residual vs Fitted")
```



R functions

ggplot(model, aes(x=…., y=…)): mapping 2 variables on a and y axis

.fitted : Fitted values of model

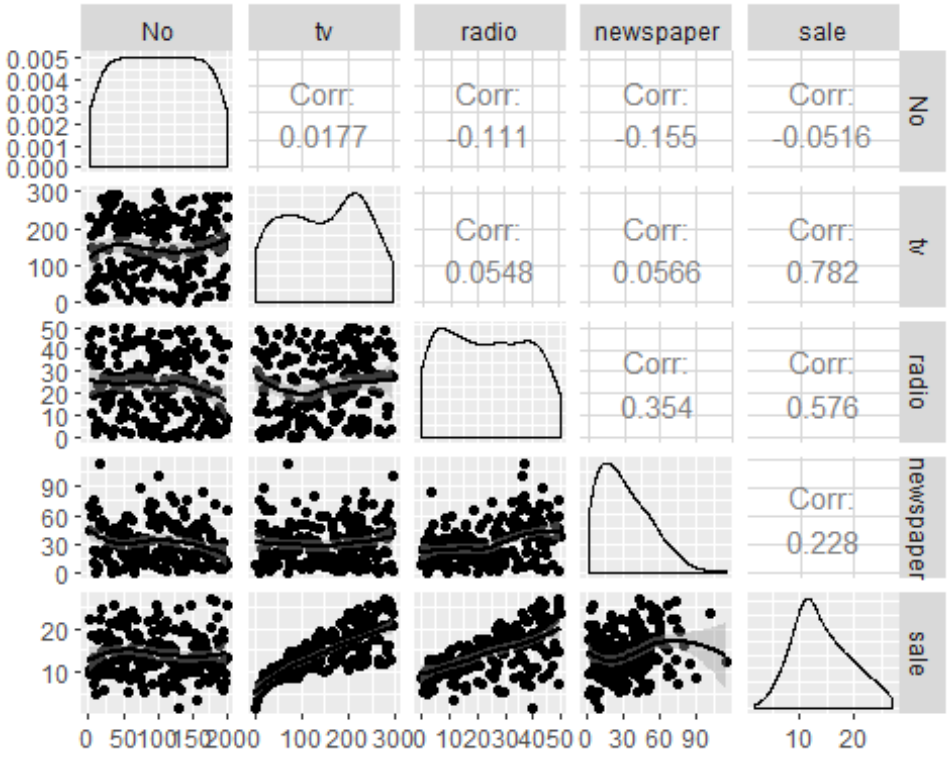
.resid : Residuals

geom_hline(yintercept = 0):add a horizontal line

geom_point() : add a layer of points to the plot

From the Advertising example, the output displays the residual plot that results from the model$\widehat{Sale} = 6.750 + 0.01910tv + 0.02886radio + 0.001086tv * radio$. There appears to be a little pattern in the residuals, suggesting that the quadratic term or logarithmic might improve the fit to the data.

```
library(ggplot2)
Advertising=read.table("c:/Users/thuntida.ngamkham/OneDrive - University of
Calgary/dataset603/Advertising.txt", header = TRUE,sep ="\t" )
library(GGally)
ggpairs(Advertising,lower = list(continuous = "smooth_loess", combo =
  "facethist", discrete = "facetbar", na = "na"))
```

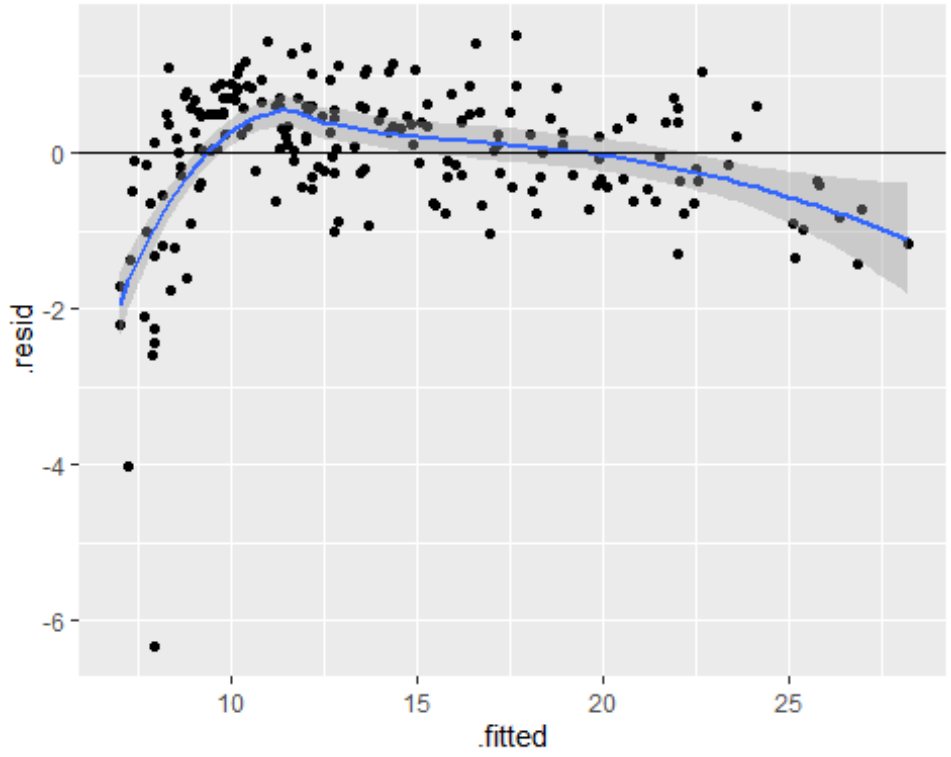



```
model<-lm(sale~tv+radio+tv*radio, data=Advertising)
quadmodel<-lm(sale~tv+I(tv^2)+radio+tv*radio, data=Advertising)
cubic<-lm(sale~tv+I(tv^2)+I(tv^3)+radio+tv*radio, data=Advertising)
summary(model)$adj.r.squared

## [1] 0.9672975

summary(quadmodel)$adj.r.squared

## [1] 0.985707

summary(cubic)$adj.r.squared

## [1] 0.99072

summary(cubic)

##
## Call:
## lm(formula = sale ~ tv + I(tv^2) + I(tv^3) + radio + tv * radio,
##     data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2184 -0.2106  0.0223  0.2454  1.1677
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  4.061e+00  1.871e-01  21.709  < 2e-16 ***
## tv           8.998e-02  4.193e-03  21.458  < 2e-16 ***
## I(tv^2)     -4.327e-04  3.180e-05 -13.604  < 2e-16 ***
## I(tv^3)      7.278e-07  7.058e-08  10.312  < 2e-16 ***
## radio        4.206e-02  4.801e-03   8.761 9.63e-16 ***
## tv:radio     1.044e-03  2.811e-05  37.129  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5026 on 194 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.9907
## F-statistic:  4250 on 5 and 194 DF,  p-value: < 2.2e-16
```

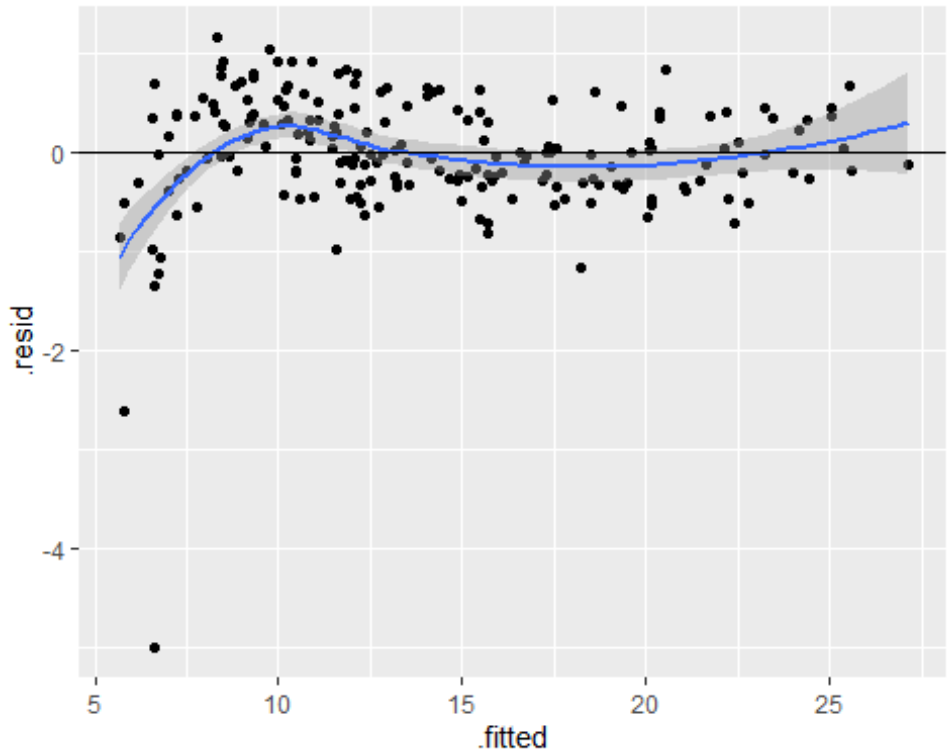
```
#residual vs fitted data plot for the simple  model
ggplot(model, aes(x=.fitted, y=.resid)) +
  geom_point() +geom_smooth()+
  geom_hline(yintercept = 0)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

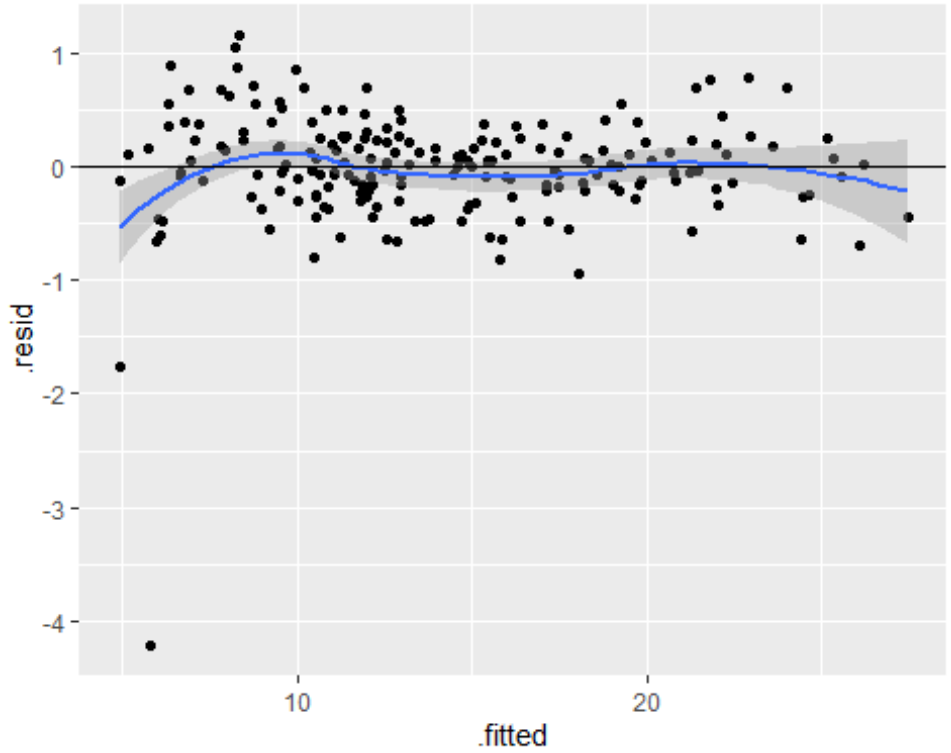



```
#residual vs fitted data plot for the quadratic  model
ggplot(quadmodel, aes(x=.fitted, y=.resid)) +
  geom_point() +geom_smooth()+
  geom_hline(yintercept = 0)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
#residual vs fitted data plot for the cubic model
ggplot(cubic, aes(x=.fitted, y=.resid)) +
  geom_point() +geom_smooth()+
  geom_hline(yintercept = 0)

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



From the output, there appears to be a little pattern for the quadratic regresssion model while the cubic model shows no pattern of the residuals at all. Moreover, the $R^2 - adj$ of the cubic model is 0.9907 indicates the variation in y that can be explained by this model is 99.02% with RMSE= 0.5026. Therefore, we can conclude that the cubic model is the best fit model to predict Y among the models we considered.

Inclass Practice Problem

From the clerical staff work hours, use residual plots to conduct a residual analysis of the data. Check Linearity Assumption. If a trend is detected, how would you like to transform the predictors in the model?

2. Independence Assumption

An important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \epsilon_3, \ldots, \epsilon_n$ are uncorrelated (must be mutually independent). What does this mean? For instance, if the errors are uncorrelated, then the fact that ϵ_i is positive provides little or no information about the sign of ϵ_{i+1}.

The assumption of independent errors is violated when successive errors are correlated. This typically occurs when the data for both dependent and independent variables are observed sequentially over a period of time-called **time-series data**

We can check displays of the regression residuals for evidence of patterns, trends or clumping, any of which would suggest a failure of independence. In the special case when

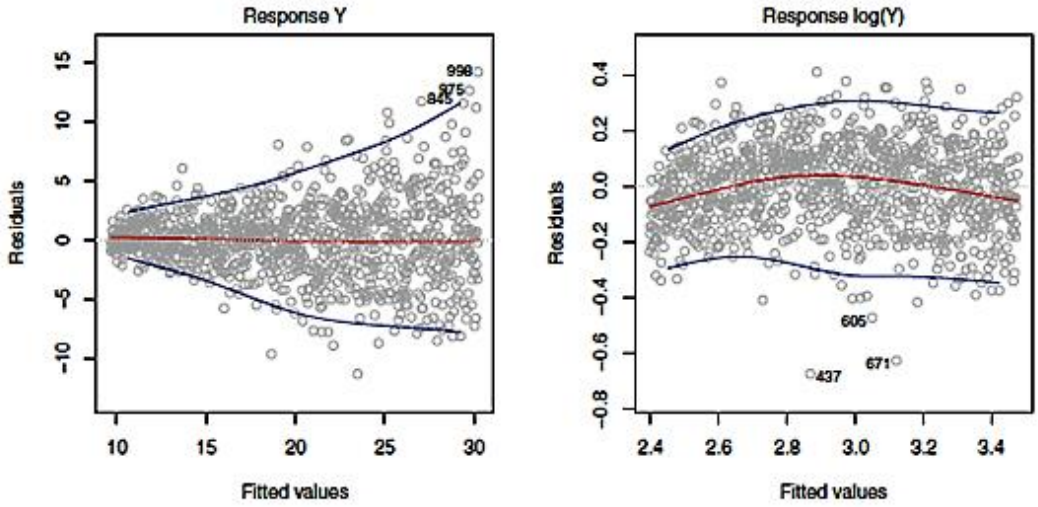
response Y is related to time (time series data), a common violation of the Independence Assumption is for the errors to be correlated. This violation can be check by plotting the residuals against the order of occurrence (time plot of the residuals and looking for pattern).

In the Advertising example, the subjects were not related to time, so we can pretty sure that their measurement are independent.

3. Equal Variance Assumption

Another important assumption of the linear regression model is that the error terms have a constant variance (homoscedasticity), $Var(\epsilon_i) = \sigma^2$. Unfortunately, it is often the case that the variances of the error terms are non-constant. For instance, the variances of the error terms may increase with the value of the response. One can identify non-constant variances in the errors, or **heteroscedasticity**

Heteroscedasticity means unequal scatter. In regression analysis, heteroscedasticity is a systematic change in the spread of the residuals over the range of measured values. An example is shown in the left-hand panel of the figure below, in which the magnitude of the residuals tends to increase with the fitted values. When faced with this problem, one possible solution is to transform the response Y using a concave function such as $\log(Y)$ or $\sqrt{X}$. Such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity. The right-hand panel of the figure below displays the residual plot after transforming the response.

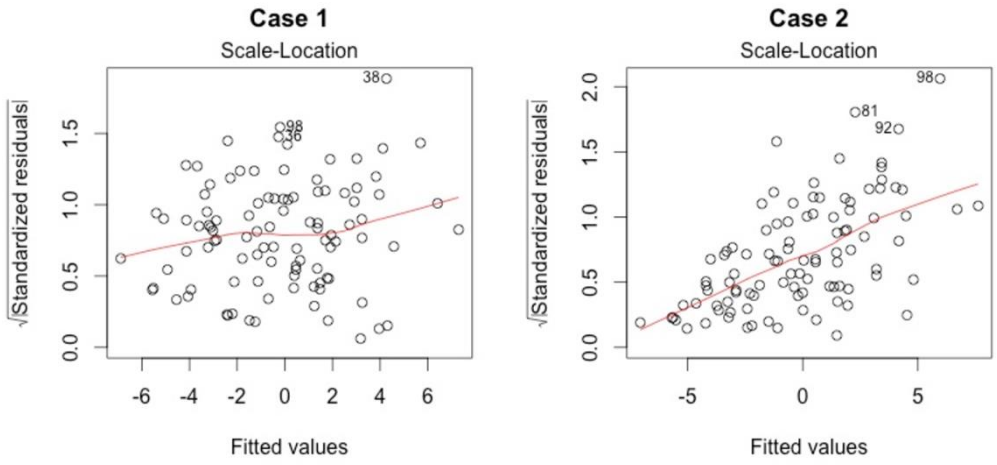


Residual plots.

In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns.

Left: The funnel shape indicates heteroscedasticity.

Right: The predictor has been log-transformed, and there is now no evidence of heteroscedasticity.



Residual plots.

A scale-location plot between fitted value and standardized residuals can also be checked for heteroscedasticity. It's also called Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. You can check the assumption of equal variance (homoscedasticity). It's good if you see a horizontal line with equally (randomly) spread points. From the figure above in Case 1, the residuals appear randomly spread. Whereas, in Case 2, the residuals begin to spread wider along the x-axis as it passes around 5. Because the residuals spread wider and wider, the red smooth line is not horizontal and shows a steep angle in Case 2.

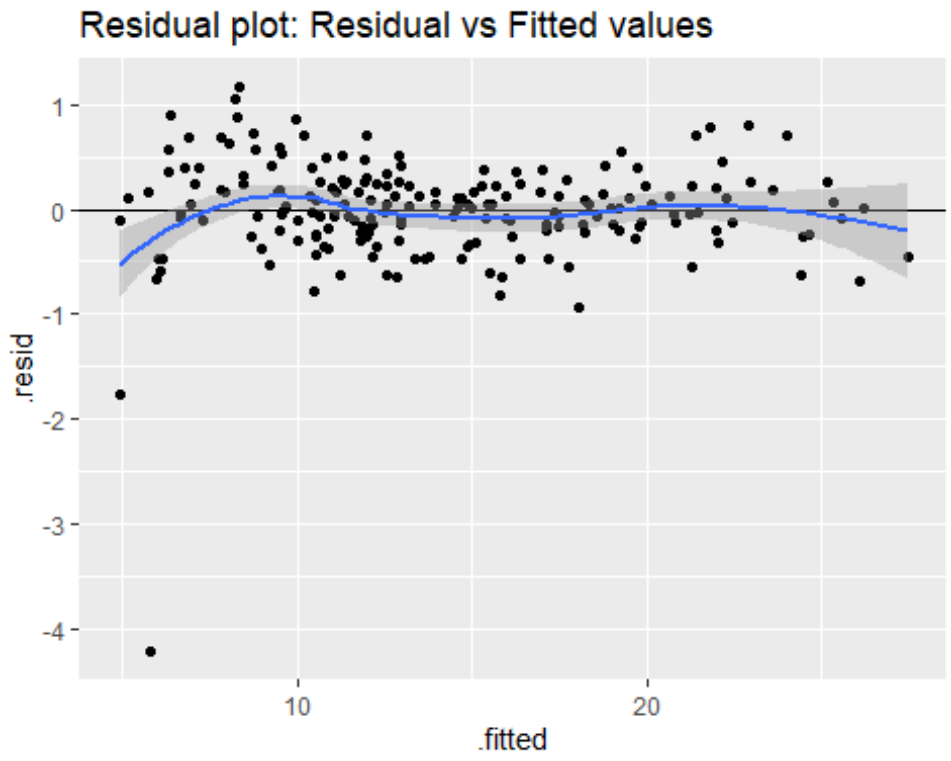
```
library(ggplot2)
Advertising=read.table("c:/Users/thuntida.ngamkham/OneDrive - University of
Calgary/dataset603/Advertising.txt", header = TRUE,sep ="\t" )
cubic<-lm(sale~tv+I(tv^2)+I(tv^3)+radio+tv*radio, data=Advertising)
summary(cubic)
```

```
##
## Call:
## lm(formula = sale ~ tv + I(tv^2) + I(tv^3) + radio + tv * radio,
##     data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2184 -0.2106  0.0223  0.2454  1.1677
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  4.061e+00  1.871e-01  21.709  < 2e-16 ***
## tv           8.998e-02  4.193e-03  21.458  < 2e-16 ***
## I(tv^2)     -4.327e-04  3.180e-05 -13.604  < 2e-16 ***
## I(tv^3)      7.278e-07  7.058e-08  10.312  < 2e-16 ***
## radio        4.206e-02  4.801e-03   8.761 9.63e-16 ***
## tv:radio     1.044e-03  2.811e-05  37.129  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5026 on 194 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.9907
## F-statistic:  4250 on 5 and 194 DF,  p-value: < 2.2e-16
```

```
#residuals plot
ggplot(cubic, aes(x=.fitted, y=.resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_smooth()+
  ggtitle("Residual plot: Residual vs Fitted values")
```

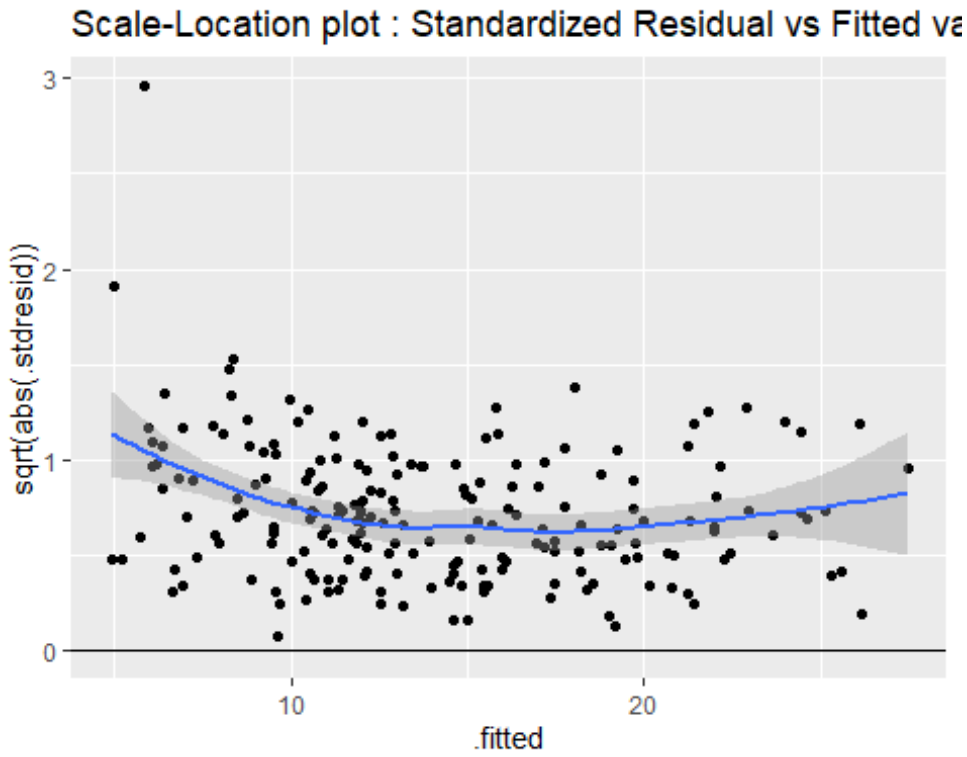
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



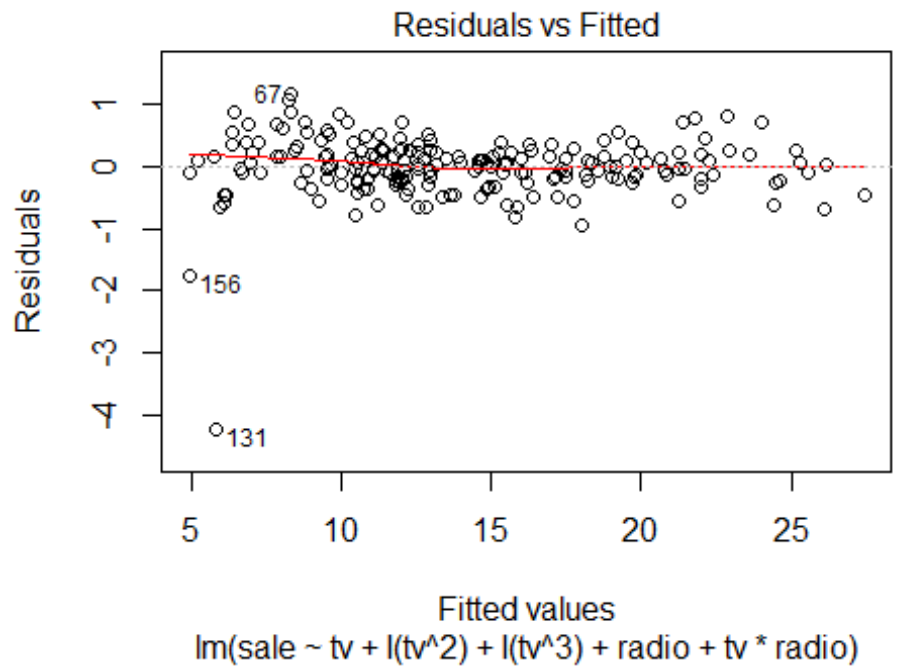

```
#a scale location plot
ggplot(cubic, aes(x=.fitted, y=sqrt(abs(.stdresid)))) +
  geom_point() +
  geom_hline(yintercept = 0) +
```

```
  geom_smooth()+
   ggtitle("Scale-Location plot : Standardized Residual vs Fitted values")
```

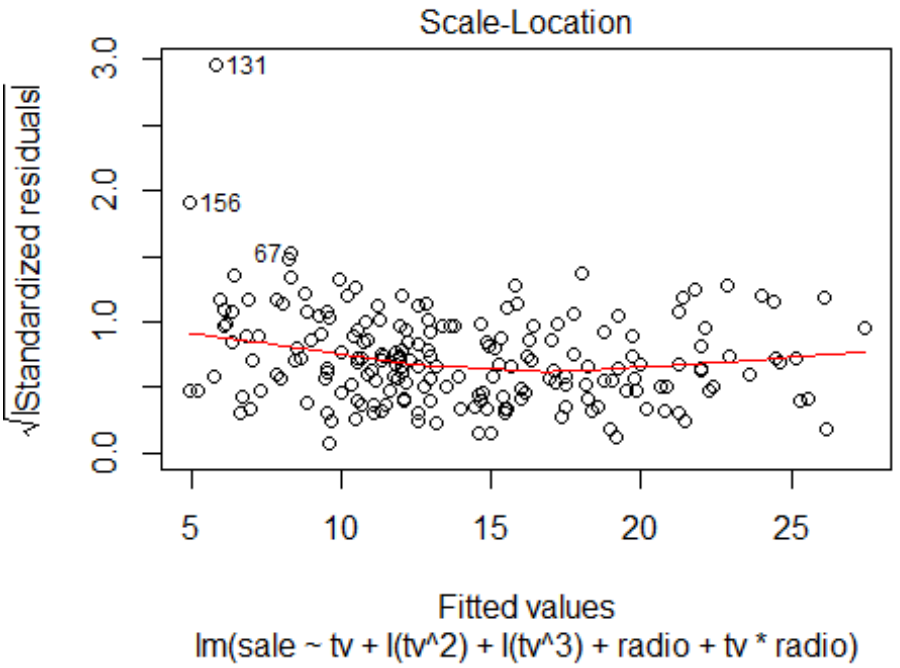
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
#optional graphs for residual plots and a scale location plot
plot(cubic, which=1) #residuals plot
```




```r
plot(cubic, which=3) #a scale location plot
```



R functions ggplot(model, aes(x=…., y=…)): mapping 2 variables on a and y axis .fitted : Fitted values of model .resid : Residuals .stdresid : Standardised Residuals geom_hline(yintercept = 0):add a horizontal line geom_point() : add a layer of points to the plot ggtitle() ; add a title to the plot

From the Advertising example, the output displays the residual plot and Scale-Location plot that result from the cubic model. In our case, the residuals tend to form a horizontal band-indicates that the plot does not provide evidence to suggest that heteroscedasticity exists.

A more formal, mathematical way of detecting heteroscedasticity is what is known as **the Breusch-Pagan test**. It involves using a variance function and using a $\chi^2 test$ to test

H_0: heteroscedasticity is not present (homoscedasticity)
H_a : heteroscedasticity is present
or
H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_p^2$
H_0: at least σ_i^2 is different from the others $i = 1,2,\ldots,p$

$\chi^2 = nR^2 \sim \chi^2_{p-1}$
where
n = sample size
R^2 = coefficient determination
p = number of regression coefficients

```
library(lmtest)
```

```
## Loading required package: zoo
```

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

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

```
Advertising=read.table("c:/Users/thuntida.ngamkham/OneDrive - University of
Calgary/dataset603/Advertising.txt", header = TRUE,  sep ="\t" )
cubic<-lm(sale~tv+I(tv^2)+I(tv^3)+radio+tv*radio, data=Advertising)
bptest(cubic)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  cubic
## BP = 22.934, df = 5, p-value = 0.0003476
```

```
morepower<-
lm(sale~tv+I(tv^2)+I(tv^3)+I(tv^4)+I(tv^5)+I(tv^6)+I(tv^7)+I(tv^8)+I(tv^9)+I(
```

```
tv^10)+I(tv^11)+radio+tv*radio, data=Advertising)
bptest(morepower)

##
##  studentized Breusch-Pagan test
##
## data:  morepower
## BP = 15.005, df = 13, p-value = 0.307
```

R function bptest(): to perform the Breusch-Pagan test

From the Advertising example, the output displays the Breusch-Pagan test that result from the cubic model. The p-value = 0.00034 <0.05, indicating that we do reject the null hypothesis. Therefore, the test provide evidence to suggest that heteroscedasticity does exist. However, a model with more power on tv (power of 11) shows evidence to suggest that heteroscedasticity does not exist.

Inclass Practice Problem

From the clerical staff work hours, use residual plots to conduct a residual analysis of the data. Check Equal Variance Assumption by graphs and the Breusch-Pagan test. If you detect a trend, how would you like to transform the predictors in the model?

R function

geom_smooth() : add the regression slope

4. Normality Assumption

The multiple linear regression analysis requires that the errors between observed and predicted values (i.e., the residuals of the regression) should be normally distributed. This assumption may be checked by looking at a histogram, a normal probability plot or a Q-Q-Plot.

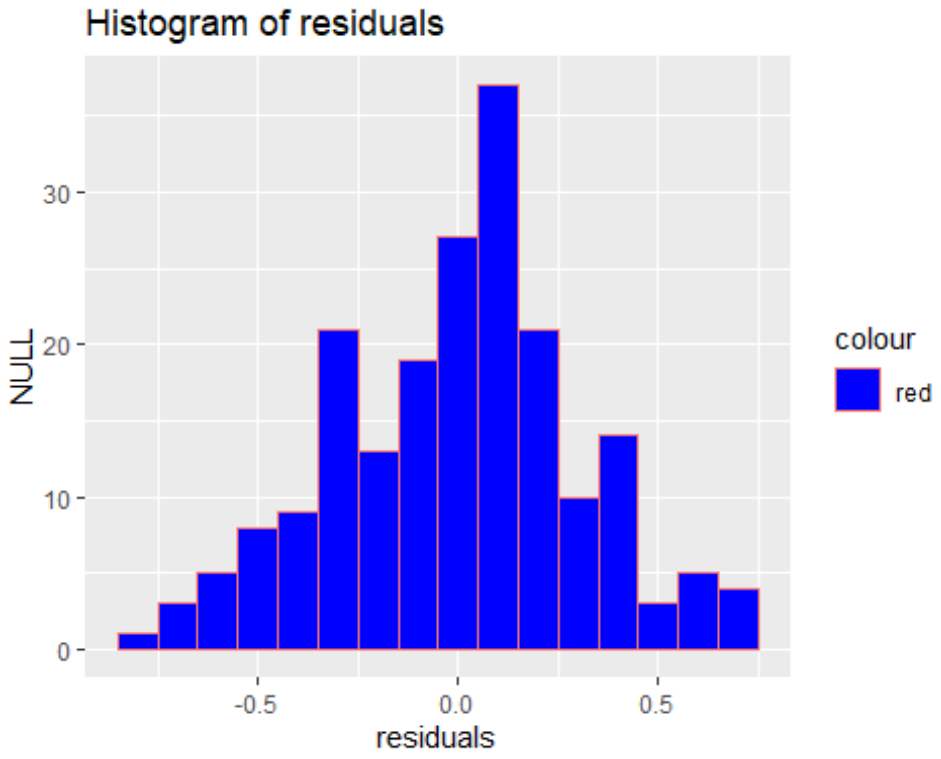
If the distribution is normal, the points on such a plot (Probability Plot or Q-Q-Plot) should fall close to the diagonal reference line. A bow-shaped pattern of deviations from the diagonal indicates that the residuals have excessive skewness. An S-shaped pattern of deviations indicates that the residuals have excessive kurtosis, i.e., there are either too many or two few large errors in both directions. Sometimes the problem is revealed to be that there are a few data points on one or both ends that deviate significantly from the reference line ("outliers"), in which case they should get close attention.

There are also a variety of statistical tests for normality, including the Kolmogorov-Smirnov test and the Shapiro-Wilk test.

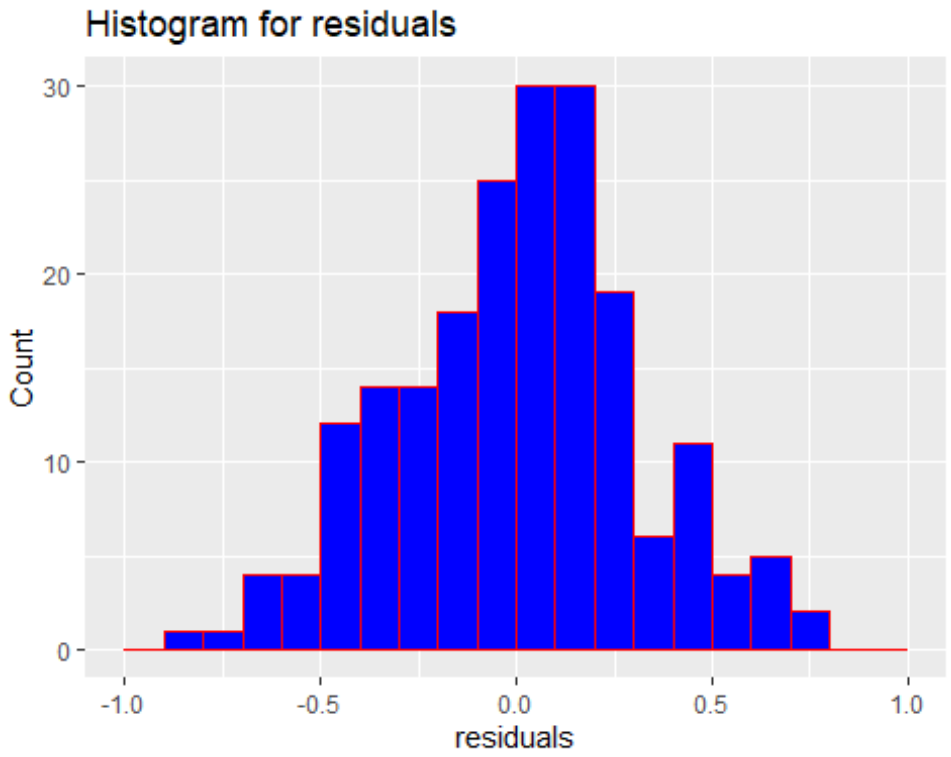
H_0: the sample data are significantly normally distributed
H_a: the sample data are not significantly normally distributed

```
library(ggplot2)
Advertising=read.table("c:/Users/thuntida.ngamkham/OneDrive - University of
```

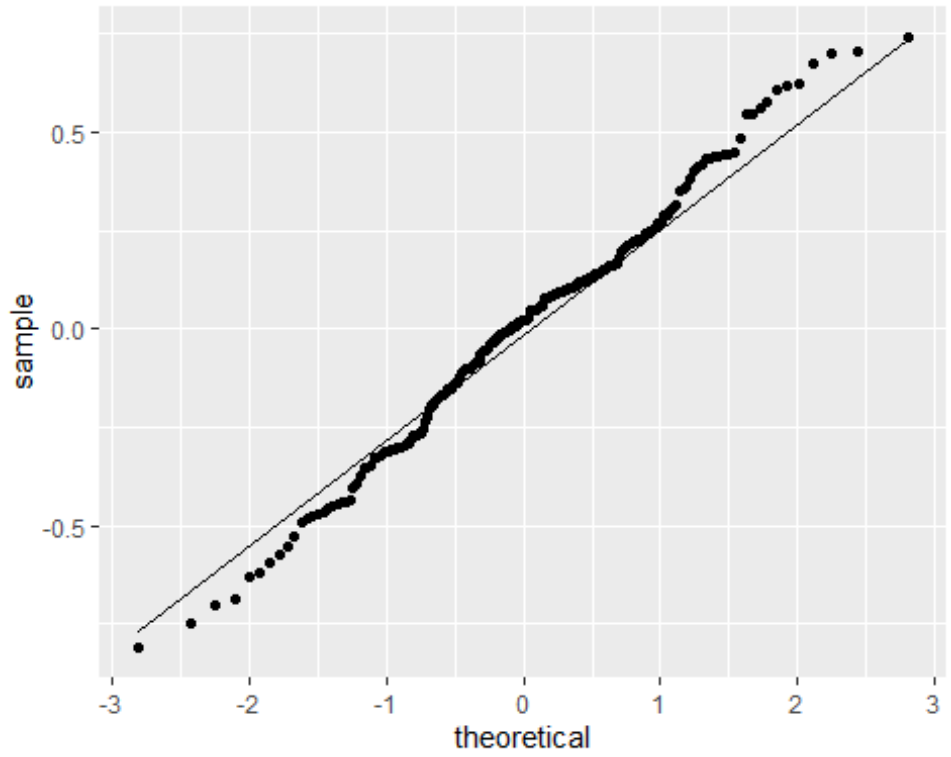
```
Calgary/dataset603/Advertising.txt", header = TRUE,  sep ="\t" )
#logmodel<-lm(sale~log(tv)+radio+tv*radio, data=Advertising)
morepower<-
lm(sale~tv+I(tv^2)+I(tv^3)+I(tv^4)+I(tv^5)+I(tv^6)+I(tv^7)+I(tv^8)+I(tv^9)+I(
tv^10)+I(tv^11)+radio+tv*radio, data=Advertising)
#option 1 (histogram)
qplot(residuals(morepower),
      geom="histogram",
      binwidth = 0.1,
      main = "Histogram of residuals",
      xlab = "residuals", color="red",
      fill=I("blue"))
```



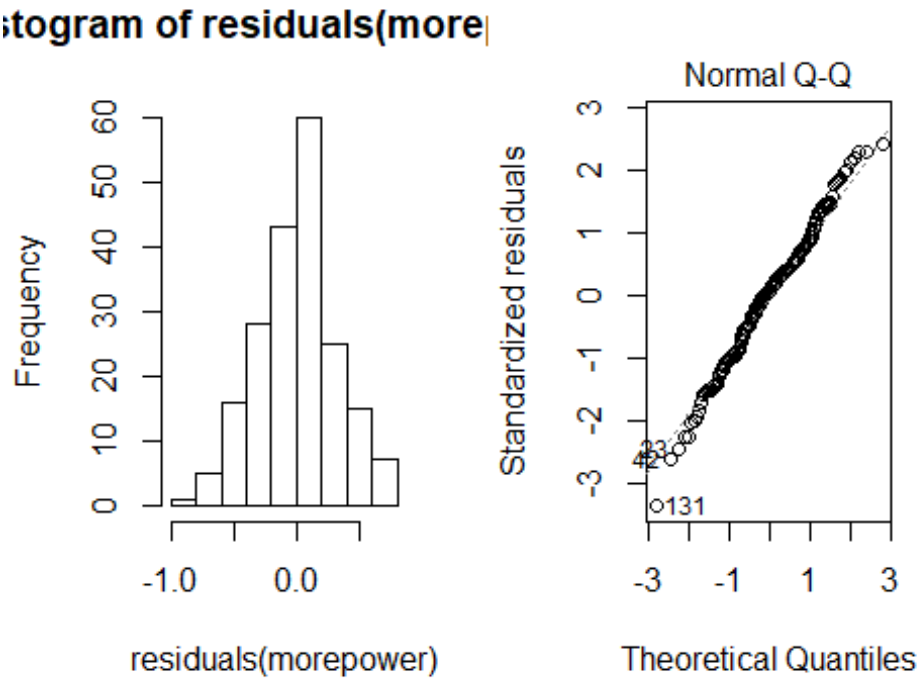

```
#option 2 (histogram)
ggplot(data=Advertising, aes(residuals(morepower))) +
  geom_histogram(breaks = seq(-1,1,by=0.1), col="red", fill="blue") +
  labs(title="Histogram for residuals") +
  labs(x="residuals", y="Count")
```




```r
#normal QQ plot
ggplot(Advertising, aes(sample=morepower$residuals)) +
  stat_qq() +
  stat_qq_line()
```




```r
#optional histogram
par(mfrow=c(1,2))
hist(residuals(morepower))
plot(morepower, which=2) #a Normal plot
```




```
#Testing for Normality
shapiro.test(residuals(morepower))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(morepower)
## W = 0.99171, p-value = 0.3129
```

R funtions stat_qq() : produce quantile-quantile plots stat_qq_line() : compute the slope and intercept of the line connecting the points at specified quartiles of the theoretical and sample distributions

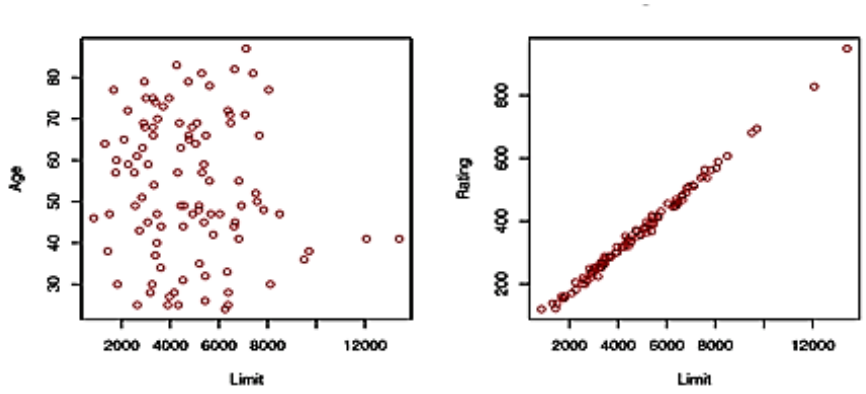
geom_qq_line and stat_qq_line compute the slope and intercept of the line connecting the points at specified quartiles of the theoretical and sample distributions. The outputs show that the residual data have normal distribution (from histogram and Q-Q plot). Moreover, Shapiro-Wilk normality test also confirms that the residuals are normally distributed as the p-value=0.3129 >0.05.

Inclass Practice Problem

From the clerical staff work hours, use residual plots to conduct a residual analysis of the data. Check Normality Assumption by graphs and the Shapiro-Wilk normality test. If you detect a trend, how would you like to transform the predictors in the model?

5. Multicollinearity

Often, two or more of the independent variables used in the model for $E(Y)$ provide redundant information. That is, the independent variables will be correlated with each other. For example, suppose we want to construct a model to predict the gasoline mileage rating, Y, of a truck as a function of its load, X_1, and the horsepower, X_2, of its engine. In general, you would expect heavier loads to require greater horsepower and to result in lower mileage ratings. Thus, although both X_1 and X_2 contribute information for the prediction of mileage rating, some of the information is overlapping, because X_1 and X_2 are (linearly) correlated. When the independent variables are (linearly) correlated, we say that multicollinearity exists. In practice, it is not uncommon to observe correlations among the independent variables. However, a few problems arise when serious multicollinearity is present in the regression analysis.



The scatter plot shows multicollonearity between Rating and Limit

In the left-hand panel of Figure 3, the two predictors limit and age appear to have no obvious relationship. In contrast, in the right-hand panel of Figure 3, the predictors limit and rating are very highly linearly correlated with each other, and we say that they are collinear.

What Problems Do Multicollinearity Cause?

Multicollinearity causes the following two basic types of problems:

1. The coefficient estimates can swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model.

2. Multicollinearity reduces the precision of the estimate coefficients, which weakens the statistical power of your regression model. You might not be able to trust the p-values to identify independent variables that are statistically significant.

Testing for Multicollinearity with Variance Inflation Factors (VIF)

If you can identify which variables are affected by multicollinearity and the strength of the correlation, you're well on your way to determining whether you need to fix it. Fortunately, there is a very simple test to assess multicollinearity in your regression model which is called **"The variance inflation factor (VIF)"**

The variance inflation factor (VIF)

VIF identifies correlation between independent variables and the strength of that correlation. It can be computed using the formula

$$VIF(\widehat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_j}}$$

where $R^2_{X_j|X\_j}$ is the R^2 from a regression of X_j onto all of the other predictors. If $R^2_{X_j|X\_j}$ is close to one, then collinearity is present, and so the VIF will be large.

Statistical software calculates a VIF for each independent variable. Value of VIFs start at 1 and have no upper limit and can be interpreted as following;

*VIFs=1 indicates that there is no collinearity between this independent variable and any others.

*1<= VIFs <=5 suggest that there is a moderate collinearity, but it is not severe enough to warrant corrective measures.

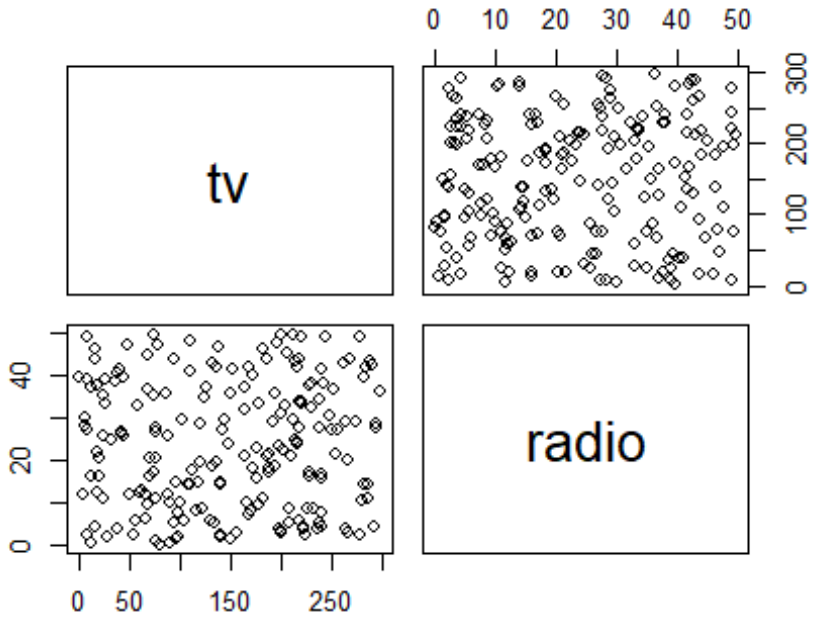
VIFs > 5 or 10 represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable.

We use VIFs to identify correlations between variables and determine the strength of the relationships.

Attention: *When the high VIFs are caused by the inclusion of powers or products of other variables, you can Safely Ignore Multicollinearity*

From Advertising data, check Multicollinearity Assumption by using scatter plotsand VIF

```
library(mctest) #for VIF
Advertising=read.table("c:/Users/thuntida.ngamkham/OneDrive - University of
Calgary/dataset603/Advertising.txt", header = TRUE,  sep ="\t" )
#logmodel<-lm(sale~log(tv)+radio+tv*radio, data=Advertising)
firstordermodel<-lm(sale~tv+radio, data=Advertising)
pairs(~tv+radio, data=Advertising)
```



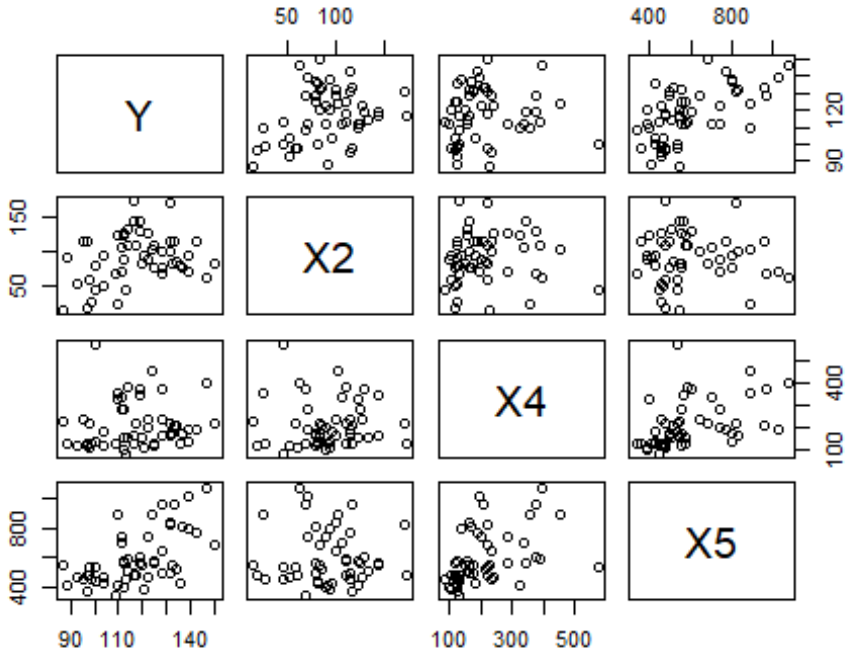

```
#Calculate VIF for multicollinearity model
#option 1
X<-cbind(Advertising$tv,Advertising$radio)
imcdiag(X,Advertising$sale, method="VIF")

##
## Call:
## imcdiag(x = X, y = Advertising$sale, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##      VIF detection
## V1 1.003         0
## V2 1.003         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================

#option 2
library(car)

## Loading required package: carData
```

```
vif(firstordermodel)

##       tv    radio
## 1.003013 1.003013
```

From the output, you can see that the VIF_{TV} =VIF_{Radio}=1.003013, which suggests that there is no correlation between these predictors.

From the clerical staff work hours, checking for Multicollinearity by scatter plots between independent predictors and VIF test. Note! consider only main effect perdictors

```
library(mctest) #for VIF
workhours=read.csv("c:/Users/thuntida.ngamkham/OneDrive - University of
Calgary/dataset603/CLERICAL.csv",
                   header = TRUE)
#improvemodel<-lm(Y~X2+I(X2^2)+X4+X5,data=workhours)
firstordermodel<-lm(Y~X2+X4+X5,data=workhours)
pairs(~Y+X2+X4+X5,data=workhours)
```




```
#Calculate VIF for multicollinearity model
#option 1
X<-cbind(workhours$X2,workhours$X4,workhours$X5)
imcdiag(X,workhours$Y, method="VIF")

##
## Call:
## imcdiag(x = X, y = workhours$Y, method = "VIF")
##
```

```
##
##  VIF Multicollinearity Diagnostics
##
##       VIF detection
## V1 1.0027         0
## V2 1.2469         0
## V3 1.2455         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

```
#option 2
library(car)
vif(firstordermodel)
```

```
##       X2       X4       X5
## 1.002657 1.246889 1.245509
```

```
#vif(improvemodel)
```

R functions imcdiag(x=,y=,method="VIF") : detects the existence of multicollinearity due to independent variables

Inclass practice Problem

From the credit card example, check for Multicollinearity by scatter plots between independent predictors and VIF test . Note! consider only main effect predictors

In the credit data example, a regression of balance on Age, Rating, and Limit indicates that the predictors have VIF values of 2.776906 , 230.869514, 1.039696, 479, 1.439007, and 1.009064. As we suspected, there is considerable multicollinearity in the data! When faced with the problem of multicollinearity.

There are two simple solutions.

The first solution is to drop one of the problematic variables from the regression model. This can usually be done without much compromise to the regression model, since the presence of multicollinearity implies that the information that this variable provides about the response is redundant in the presence of the other variables.

The second solution is to combine the collinear variables together into a single predictor. For instance, we might take the average of standardized versions of Limit and Rating in order to create a new variable that measures credit worthiness.

Inclass Practice Problem

From the credit card example, after dropping the Limit variable, find the best model and check for Multicollinearity by scatter plots between independent predictors and VIF test. Note! consider only main effect perdictors

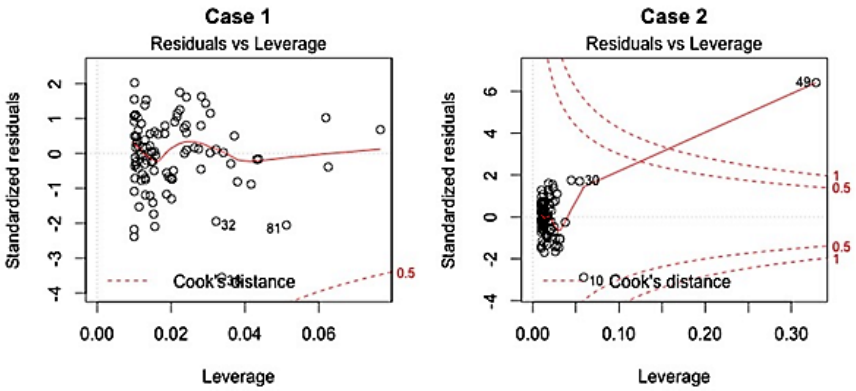
6. Outlier (The Effect on Individual Cases)

An outlying case is defined as a particular observation $(Y, X_1, X_2, \ldots X_p)$ that differs from the majority of the cases in the data set. There are several ways we can find and evaluate outlier or influential points.

1. Residuals vs Leverage plot

This plot helps us to find influential cases (i.e., subjects) if any. Not all outliers are influential in linear regression analysis (whatever outliers mean). Even though data have extreme values, they might not be influential to determine a regression line. That means, the results wouldn't be much different if we either include or exclude them from analysis. They follow the trend in the majority of cases and they don't really matter; they are not influential. On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values. They could be extreme cases against a regression line and can alter the results if we exclude them from analysis. Another way to put it is that they don't get along with the trend in the majority of the cases.

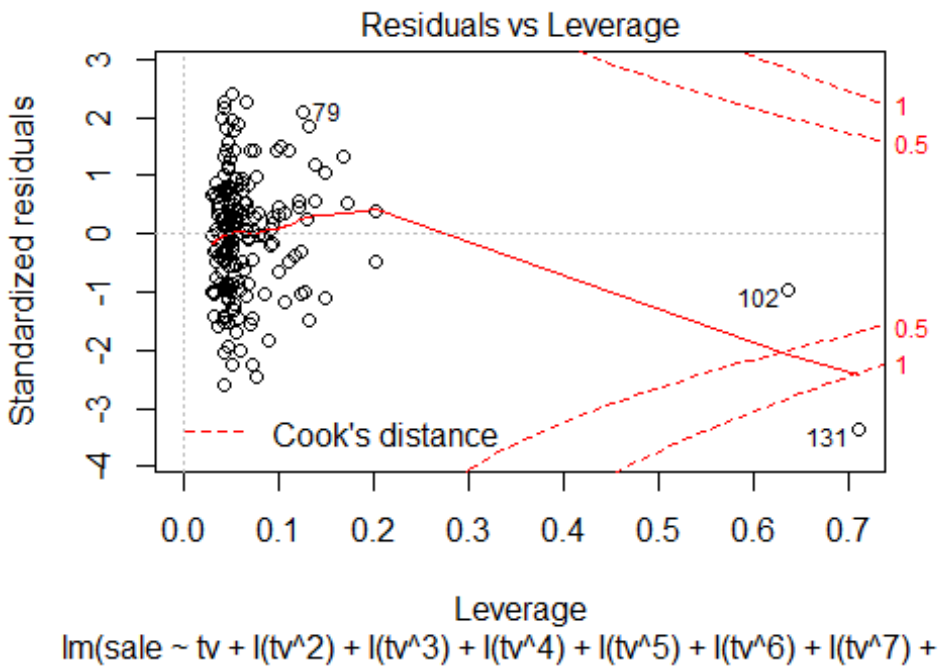
Unlike the other plots, this time patterns are not relevant. We watch out for outlying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line. Look for cases outside of a dashed line, Cook's distance. When cases are outside of the Cook's distance (meaning they have high Cook's distance scores), the cases are influential to the regression results. The regression results will be altered if we exclude those cases.



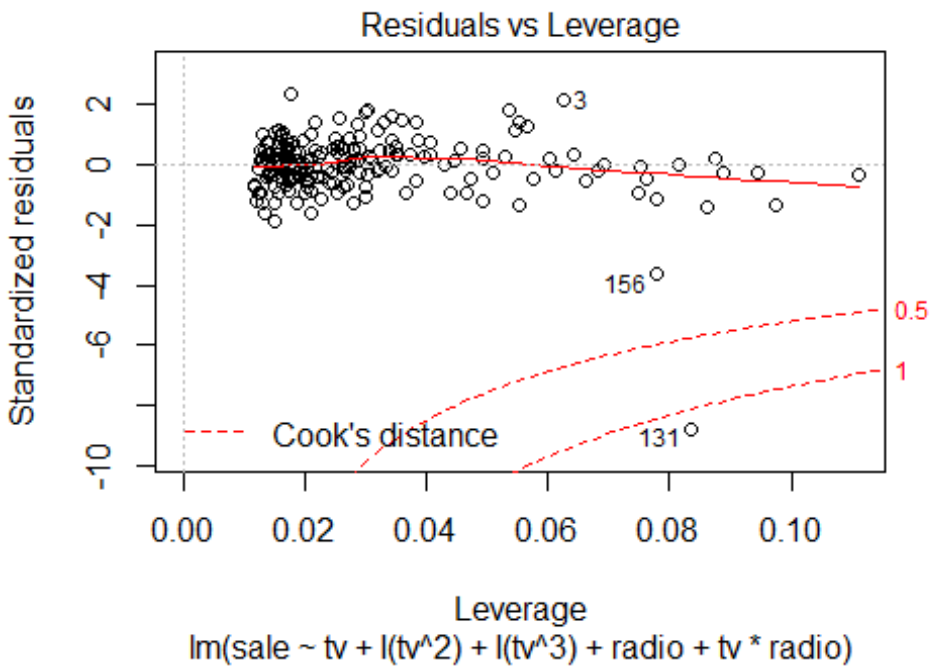
Residuals vs Leverage plot for detecting outliers or influential points

Case 1 is the typical look when there is no influential case, or cases. You can barely see Cook's distance lines (a red dashed line) because all cases are well inside of the Cook's distance lines. In Case 2, a case is far beyond the Cook's distance lines (the other residuals appear clustered on the left because the second plot is scaled to show larger area than the first plot). The plot identified the influential observation as #49.

```
Advertising=read.table("c:/Users/thuntida.ngamkham/OneDrive - University of
Calgary/dataset603/Advertising.txt", header = TRUE,  sep ="\t" )
morepower<-
lm(sale~tv+I(tv^2)+I(tv^3)+I(tv^4)+I(tv^5)+I(tv^6)+I(tv^7)+I(tv^8)+I(tv^9)+I(
tv^10)+I(tv^11)+radio+tv*radio, data=Advertising)
plot(morepower,which=5)
```




```
cubic<-lm(sale~tv+I(tv^2)+I(tv^3)+radio+tv*radio, data=Advertising)
plot(cubic,which=5)
```



From the Adertising example (both cubic and morepower models), you can see that data point 131 is an influential case.

2. Cook's Distance

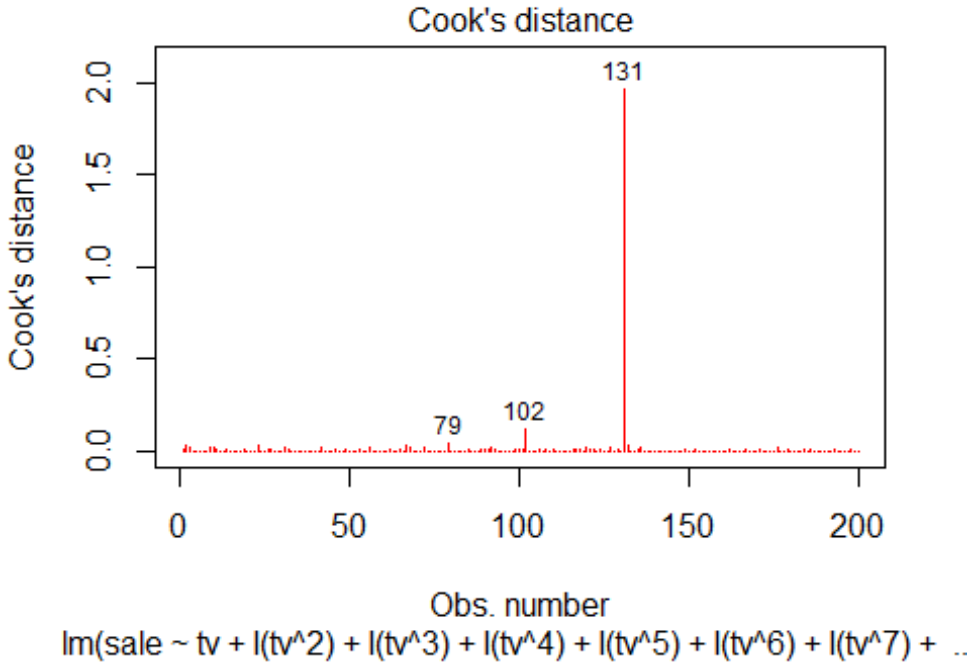
A measure of the overall influence an outlying observation has on the estimated coefficients was proposed by R. D. Cook (1979). The **Cook's distance** D_i measures the effect of deleting a given observation, and is interpreted for the ith observation as follows:

A large value of D_i indicates that the observed Y_i value has strong influence on the estimated coefficients (since the residual, the leverage, or both will be large). A general rule of thumb is that observations with a Cook's D of more than 3 times the mean, μ, is a possible outlier. An alternative interpretation is to investigate any point over $4/n$, where n is the number of observations. Other authors suggest that any "large" D_i should be investigated. How large is "too large"? The consensus seems to be that a D_i value of more that 1 indicates an influential value, but you may want to look at values above 0.5. Like the other numerical measures of influence, options for calculating Cook's distance are available in most statistical software packages.

```
Advertising=read.table("c:/Users/thuntida.ngamkham/OneDrive - University of
Calgary/dataset603/Advertising.txt", header = TRUE,  sep ="\t" )
morepower<-
lm(sale~tv+I(tv^2)+I(tv^3)+I(tv^4)+I(tv^5)+I(tv^6)+I(tv^7)+I(tv^8)+I(tv^9)+I(
tv^10)+I(tv^11)+radio+tv*radio, data=Advertising)
Advertising[cooks.distance(morepower)>0.5,] #have Cook statistics larger than
0.5
```

```
##      No  tv radio newspaper sale
## 131 131 0.7  39.6       8.7  1.6
```

```
plot(morepower,pch=18,col="red",which=c(4))
```



3. Leverage points

Points that fall horizontally far from the line are points of high leverage; these points can strongly influence the slope of the least squares line. If one of these high leverage points does appear to actually invoke its influence on the slope of the line then we call it an influential point. Usually we can say a point is influential if, had we plotted the line without it, the influential point would have been unusually far from the least squares line.

Leverage values for multiple regression models are extremely difficult to calculate without the aid of a computer. Fortunately, most of the statistical software packages have options that give the leverage associated with each observation.

A good rule of thumb to identify an observation y_i as influential if its leverage value h_i is

$$h_i > \frac{2p}{n}$$

where h_i is the leverage for the ith observation

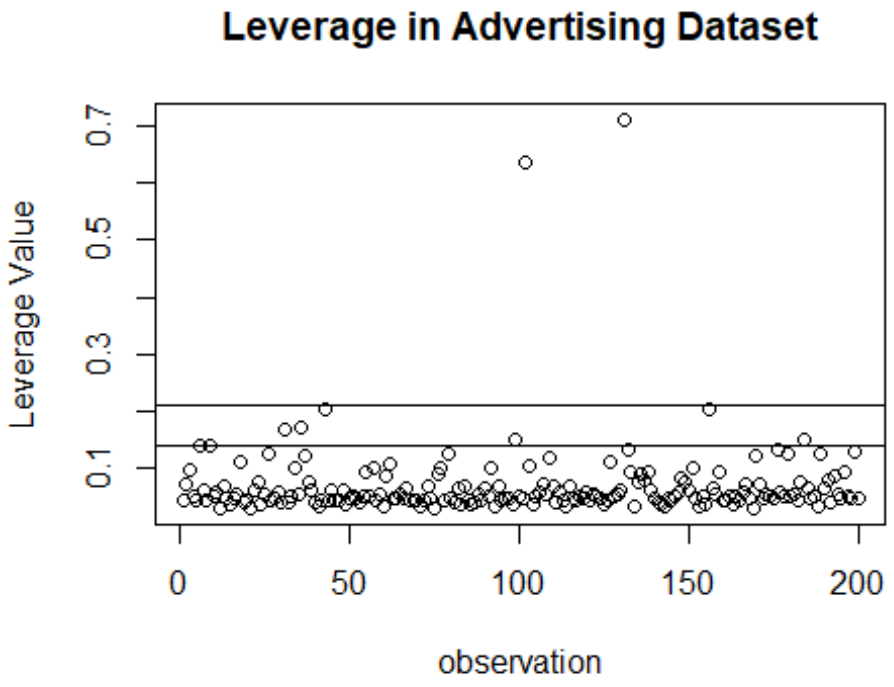
p = the number of predictors

n= the number of the sample size

```
morepower<-
lm(sale~tv+I(tv^2)+I(tv^3)+I(tv^4)+I(tv^5)+I(tv^6)+I(tv^7)+I(tv^8)+I(tv^9)+I(
tv^10)+I(tv^11)+radio+tv*radio, data=Advertising)
lev=hatvalues(morepower)
p = length(coef(morepower))
n = nrow(Advertising)
outlier = lev[lev>(2*p/n)]
print(outlier)
```

```
##        31        36        43        99       102       131       156
## 0.1681832 0.1730329 0.2027613 0.1493568 0.6369562 0.7100702 0.2026995
##       184
## 0.1496396
```

```
plot(rownames(Advertising),lev, main = "Leverage in Advertising Dataset",
xlab="observation",
    ylab = "Leverage Value")
abline(h = 2 *p/n, lty = 1)
abline(h = 3 *p/n, lty = 1)
```



Big question now is, once we identify an outlier, or influential observation, **what do we do with it?**

For a good understanding of the regression model, if we have some outliers or influential points, we may want to….

-See what happens when we exclude these from the model as an outlier has occured due to an error in data collection or recoding, then the solution is to simply remove the observation.

-Investigate these cases separately as it may happen that we mistyped.

Inclass Practice Problem

From the clerical staff work hours, using residual plots to conduct a residual analysis of the data. Check any potential outliers.

Inclass Practice Problem

Check assumptions for the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_4 + \beta_6 X_5 + \beta_7 X_3 * X_4 + \epsilon$ to predict executive salary (Y)

How to deal with Heteroscedasticity?

1. Log-transformation

As the simple solver, log-transformation can be one of the candidates. When log() takes the numbers, the difference between big and small numbers relatively becomes small.

2. Box-Cox transformations.

This will help the 'transformed' data to have equal variance, and, as usually happens, will also make the transformed data to follow a normal distribution.

3. Weight Least Squares Regression

If one wants to correct for heteroskedasticity by using a fully efficient estimator rather than accepting inefficient OLS and correcting the standard errors, the appropriate estimator is weight least squares, which is an application of the more general concept of generalized least squares.

Important points about OLS Regression

1. There must be linear relationship between independent and dependent variables
2. Multiple regression suffers from multicollinearity and heteroskedasticity.
3. Linear Regression is very sensitive to Outliers. It can terribly affect the regression line and eventually the forecasted values.
4. Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable