# Statistical Diagnostics

# Impact of Multicollinearity

- Multicollinearity can hinder our ability to use the *t* statistics and related *p*-values to assess the importance of the independent variables
  - Even when the multicollinearity itself is not severe
- With multicollinearity, the *t* statistic and *p*-value measure the additional importance of the independent variable $x_j$ over the combined importance of the other independent variables
- When two variables are multicollinear, they contribute redundant information
- This causes the resulting t statistic to be smaller than it would be if the variable were used alone

Conditional number of $\mathbf{A} = \dfrac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}$

Example: When $k$ is large,

$$y = \beta_0 + \beta_1 x + \ldots + \beta_k x^k + \varepsilon$$

and suppose that $x$ is taken 1, 2, …, 10.

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 4 & \cdots & 2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 10 & 100 & \cdots & 10^k \end{pmatrix}$$

Remark: You can use R command **kappa()** to get the conditional number of matrix

The conditional number of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is

| *k* | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| *Cond. Num* | 187.11 | 40847 | 11583988 | 4398349265 | 2.15170E+12 |

# Variance Inflation Factors (VIF)

The **variance inflation factor** for the j$^{th}$ independent (or predictor) variable $x_j$ is

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

where $R_j^2$ is the multiple coefficient of determination for the regression model relating $x_j$ to the other predictors – $x_1, \ldots, x_{j-1}, x_{j+1}, x_k$

$$x_j = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{j-1} x_{j-1} + \beta_{j+1} x_{j+1} + \ldots + \beta_k x_k + \varepsilon$$

Notes:

$\mathrm{VIF}_j = 1$ implies $x_j$ not related to other predictors

$\max(\mathrm{VIF}_j) > 10$ suggest severe multicollinearity

$\mathrm{mean}(\mathrm{VIF}_j)$ substantially greater than 1 suggests severe multicollinearity

# Example : Hospital Manpower Data

**Correlation Matrix:**

**cor(manpower)**

```
        X1   X2   X3   X4   X5   Y
X1    1.00 0.91 1.00 0.93 0.67 0.99
X2    0.91 1.00 0.91 0.91 0.45 0.95
X3    1.00 0.91 1.00 0.93 0.67 0.99
X4    0.93 0.91 0.93 1.00 0.46 0.94
X5    0.67 0.45 0.67 0.46 1.00 0.58
Y     0.99 0.95 0.99 0.94 0.58 1.00
```

**There is probably a multicollinearity problem.**

## Example : Hospital Manpower Data

Install the **car package. (by install.packages("car"), This package is available at** http://cran.r-project.org)

### Variance Inflation Factors (VIF)

```
lm.manpower1<-lm(Y~.,data=manpower)
vif(lm.manpower1)
       X1          X2          X3           X4          X5
9553.086254  7.941355  8896.349112  22.924111  4.305850
```

```
lm.manpower<-lm(Y~X2+X3+X5,data=manpower)
vif(lm.manpower)

 X2         X3        X5
7.737331 11.269342  2.492901
```

# Exercise

- Write you own R function to calculate the VIF of the multiple linear model.
- Try the data manpower. The linear model is Y~X2+X3+X5. Compare your result with that obtained by command ***vif*** in package “car” to test you code.
- Use the data Real-Estate.txt to try your R function again.

# *How to solve the multicollinearity problem?*

- Use the ***Principle Component Regression***

- ***Ridge Regression***

$$\text{Original}: \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\text{Ridge Regression}: \quad \hat{\boldsymbol{\beta}}_{\boldsymbol{r}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} \qquad \lambda \in [0, \infty)$$

- Use of ***Orthogonal Polynomials*** for polynomial regression models.

- ***Selection of Variables***

# Statistical diagnostics

- Is the model correct?
    - Are there any outliers?
        - Is the variance constant?
            - Is the error normally distributed?

# The Residuals

*studentized residuals*:

$$\boxed{r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}}$$

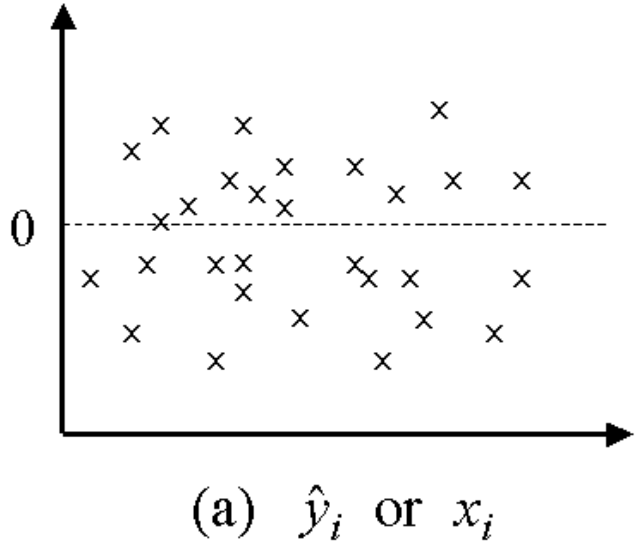
```
manpower1=read.table(file.choose(),header=T)
lm1<-lm(Y~X2+X3+X5,data=manpower1)
reg1<-summary(lm1)
hat1=lm.influence(lm1)$hat
re.s=reg1$residuals/(reg1$sigma*sqrt(1-hat1))
```

# Residual Plots

- Residuals versus each independent variable
- Residuals versus predicted y's
- Residuals in time order
  (if the response is a time series)
- Histogram of residuals
- Normal plot of the residuals

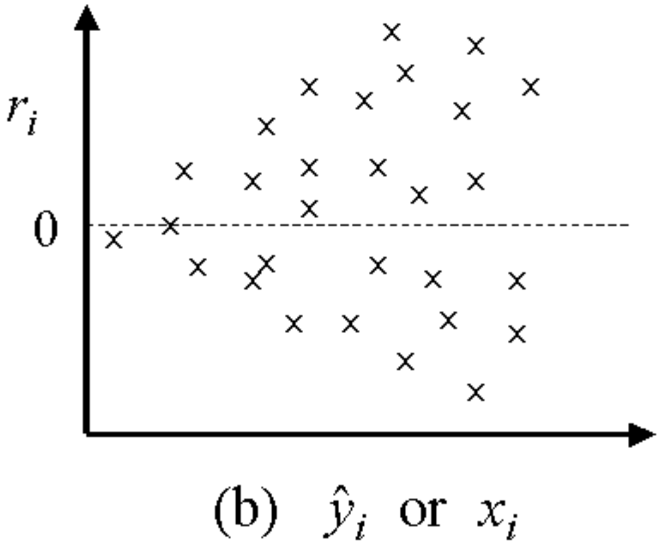
# Plotting of Residues

Residuals versus each independent variable or residuals versus predicted y's



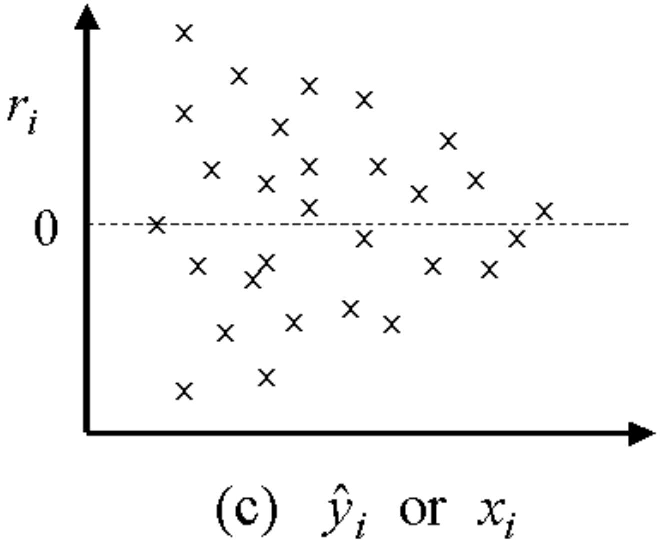
A null plot:

(normal one)



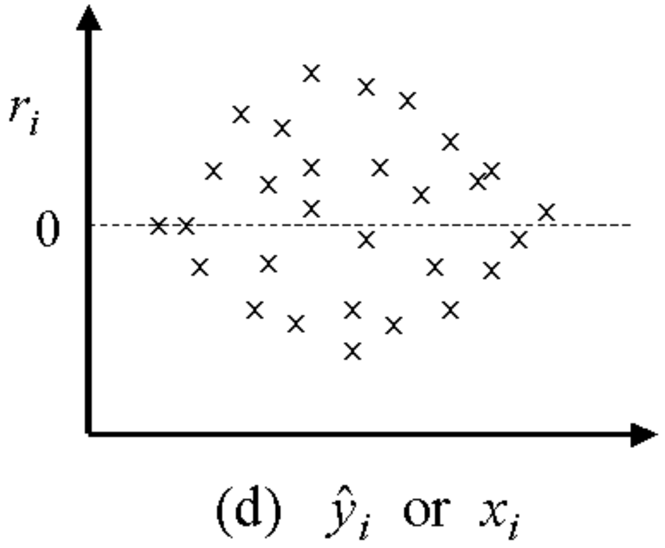
## Right-opening megaphone:

It suggests variance increasing with the quality plotted on the x-axis. This will often occur if an intrinsically positive response varies over a wide range, say from near zero into the thousands.



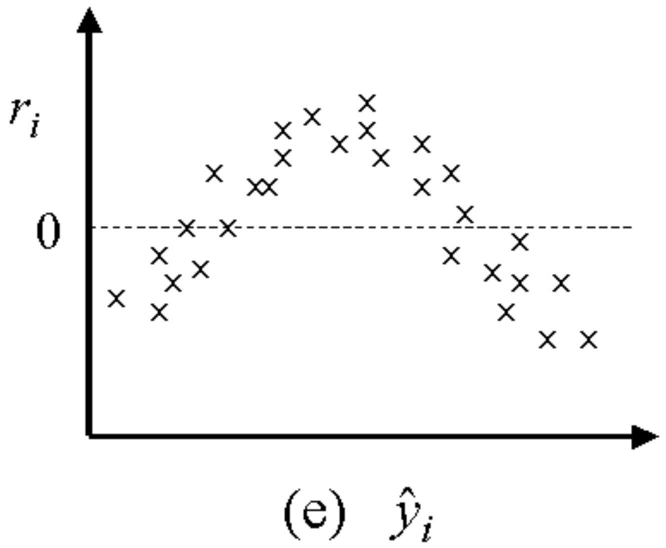
## Left-opening megaphone:

It suggests variance decreasing with the quality plotted on the x-axis.



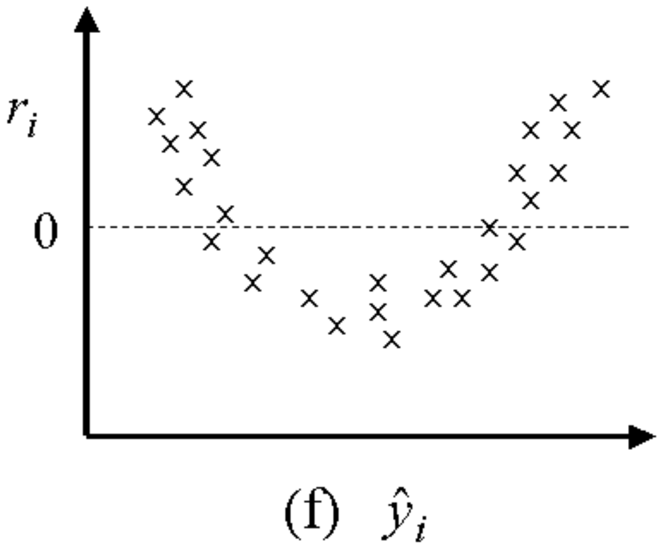
## Double outward bow:

It can occur if the response is constrained to lie between a minimum and a maximum value, for example, a percentage between 0 and 100. Large and small percentages are often less variable than the percentages near 50%.



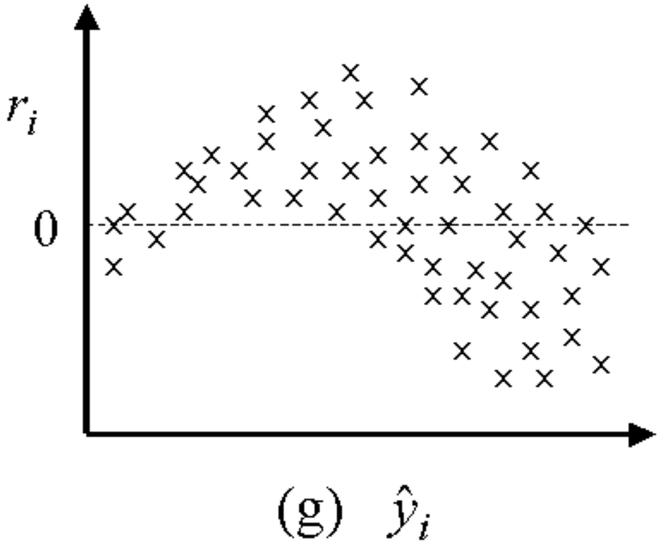
Non-linearity:

This will often call for transformation of the data, either the response or the predictors, or use of nonlinear models.

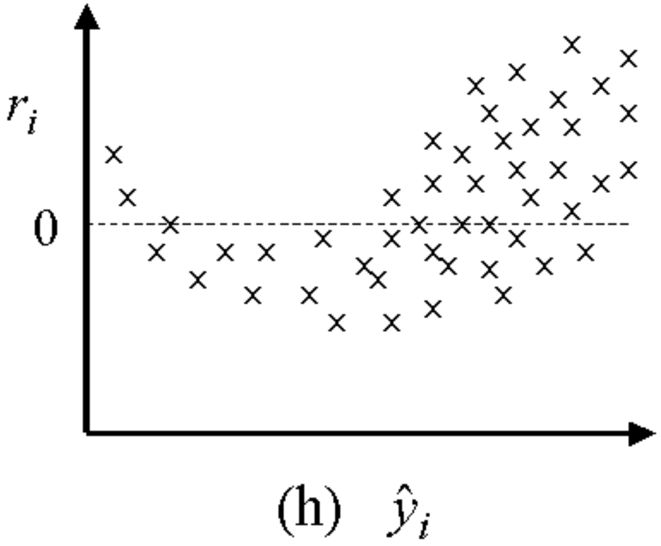


Non-linearity

This will often call for transformation of the data, either the response or the predictors, or use of nonlinear models



Non-linearity and
Non-constant variance:



Non-linearity and
Non-constant variance

# Diagnostic Plots

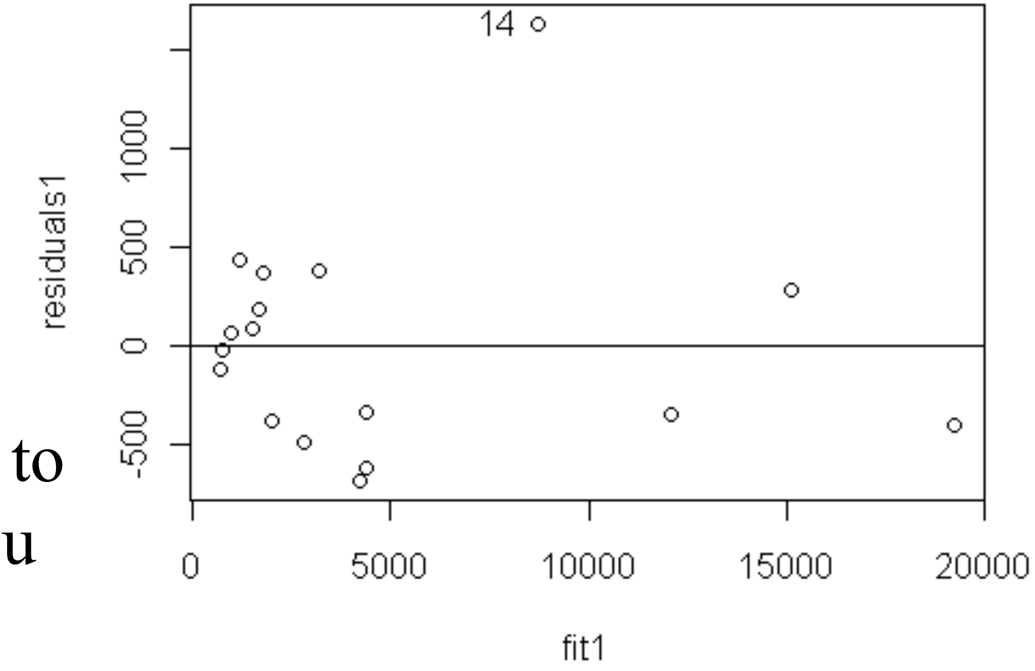
a) Residual against estimated response plots :

$$e_j \quad \text{against} \quad \hat{y}_j$$

If the model is correct, we expect to see a plot with random pattern such that the variance of $e_{y|X}$ at different values of $\hat{y}_j$'s are about constant.

# Residual against estimated response plots :

```
lm.manpower<-lm(Y~X2+X3+X5,data=manpower)
fit1<-fitted(lm.manpower)
residuals1<-resid(lm.manpower)
plot(fit1,residuals1)
abline(h=0)
```



Use the identify() function to identify the observation you are interested.

```
identify(fit1,residuals1,row.names(manpower))
```

When non-constant variance is diagnosed, but variances are unknown, we could consider the following two approaches to dealing with non-constant variance:

- Weighted least square

$$\mathrm{Q} = \sum_{i=1}^{n} (y_i - \beta x_i)^2 w_i \rightarrow \min$$

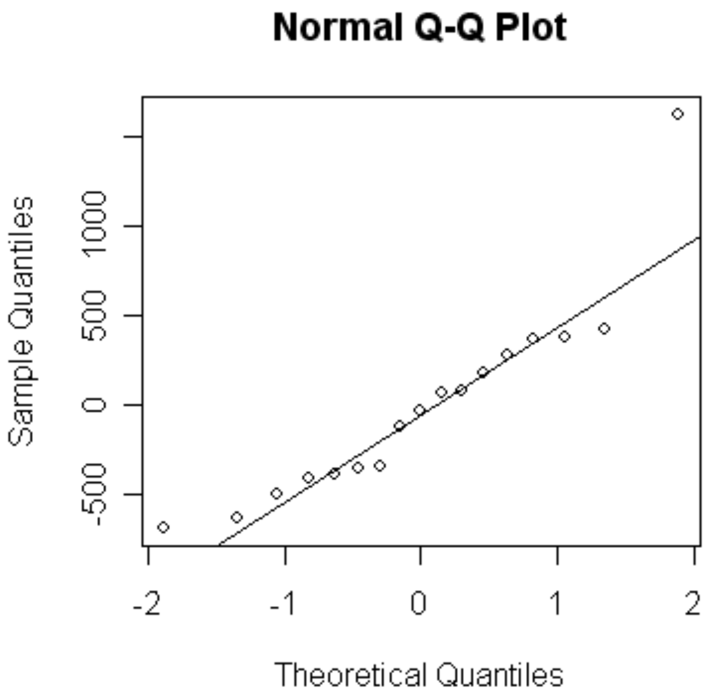
- Variance stabilizing transformations

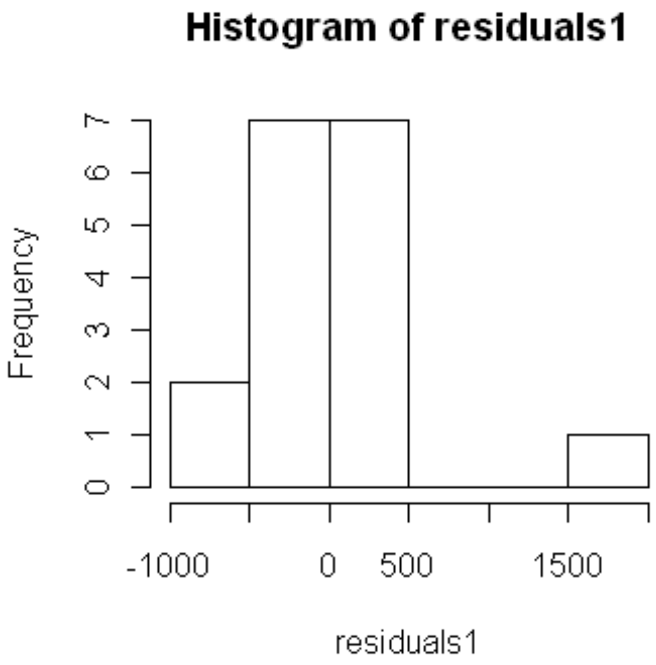
  Transform y to h(y), for example, $\sqrt{y},\ \log y$

# Diagnostic Plots

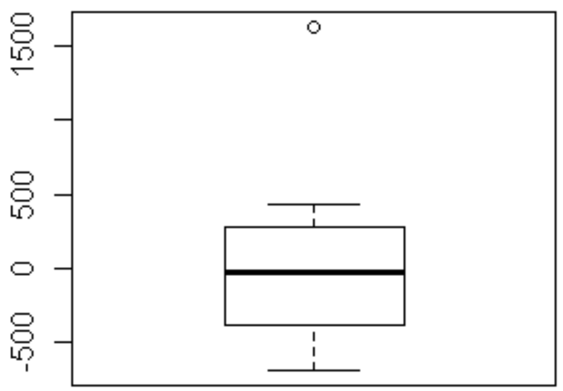
## b) Normal residual plots

```
residuals1<-resid(lm.manpower)
qqnorm(residuals1)
qqline(residuals1)
```





```
hist(residuals1)
```



```
boxplot(residuals1)
```

# The consequences of non-normality are:

1. The least squares estimates may not be optimal.

2. The test and confidence intervals are invalid. However, it has been shown that only really long tailed distribution cause a problem. Mild- non-normality can safely be ignored and the larger the sample size the less troublesome the non-normality.

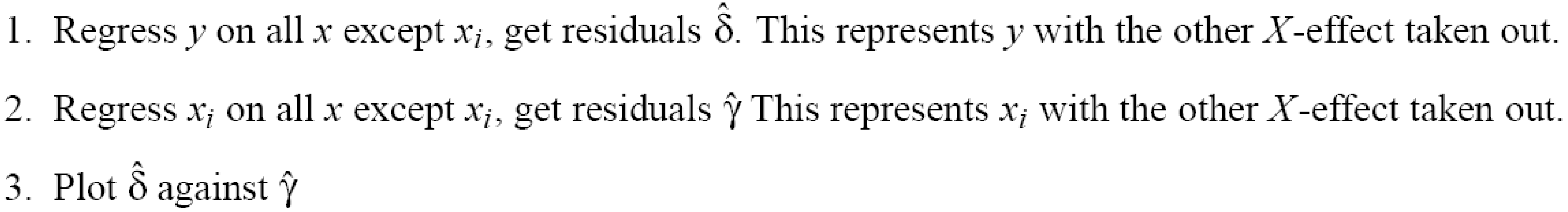
# When non-normality is diagnosed, what to do?

1. A transformation of the response may solve the problem.
2. Other changes in the model may help.
3. Accept non-normality and base the inference on the assumption of another distribution. Alternatively use robust methods which give less weight to outlying points. That is appropriate for long tailed distribution.
4. For short-tailed distribution, the consequence of non-normality are not serious and can reasonably be ignored.

# Diagnostic Plots

***C. Partial regression residual plots*** are designed to show the relationship between *y* and each $x_j$ , after the effects of all other predictors have been removed.

Procedure:



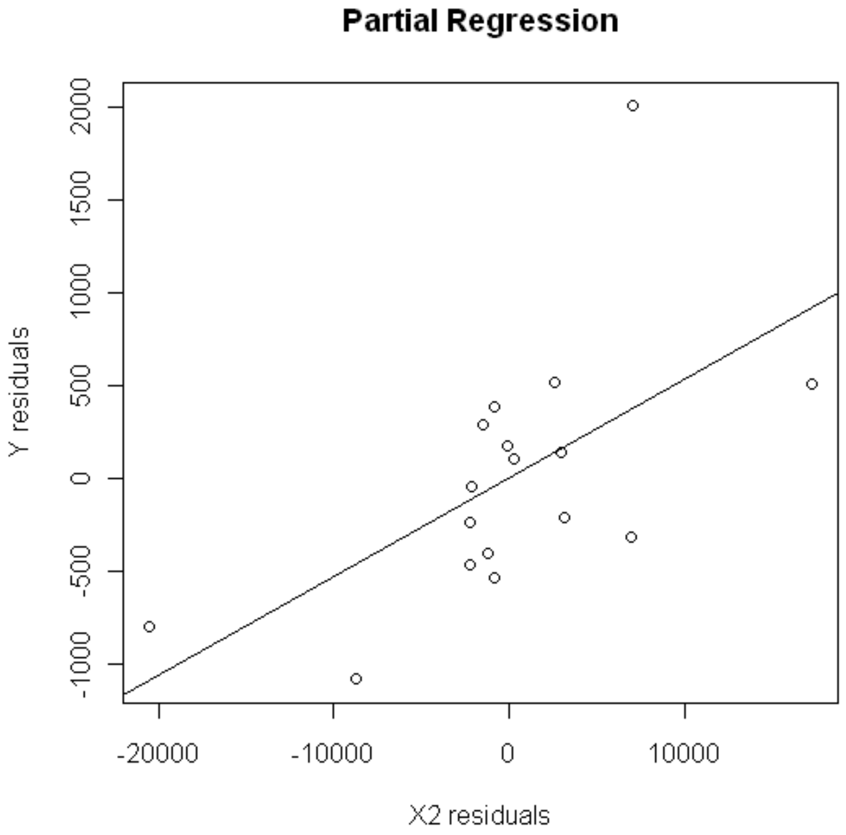
1. Regress $y$ on all $x$ except $x_i$, get residuals $\hat{\delta}$. This represents $y$ with the other $X$-effect taken out.
2. Regress $x_i$ on all $x$ except $x_i$, get residuals $\hat{\gamma}$ This represents $x_i$ with the other $X$-effect taken out.
3. Plot $\hat{\delta}$ against $\hat{\gamma}$

Variation in the plot shows how strong is the term $\hat{\beta}_j x_j$ in the regression.

eg. If y and $x_j$ are related in a linear manner, then the plot should show a linear trend with slope $\hat{\beta}_j$ and intercept 0.

# Partial regression residual plots

```
d<-lm(Y~X3+X5,data=manpower)$res
m<-lm(X2~X3+X5,data=manpower)$res
plot(m,d,xlab='X2 residuals',ylab='Y residuals',main='Partial
Regression')
abline(0,lm.manpower$coef[2])
```



```
lm(d~m)$coef
  (Intercept)             m
-2.867662e-14  5.298733e-02
```

# Diagnostic Plots

d) Component - plus - resident plots :

$$\mathbf{e}_{\mathrm{y|X}} + \mathbf{x}_j b_j \quad \text{against} \quad \mathbf{x}_j$$

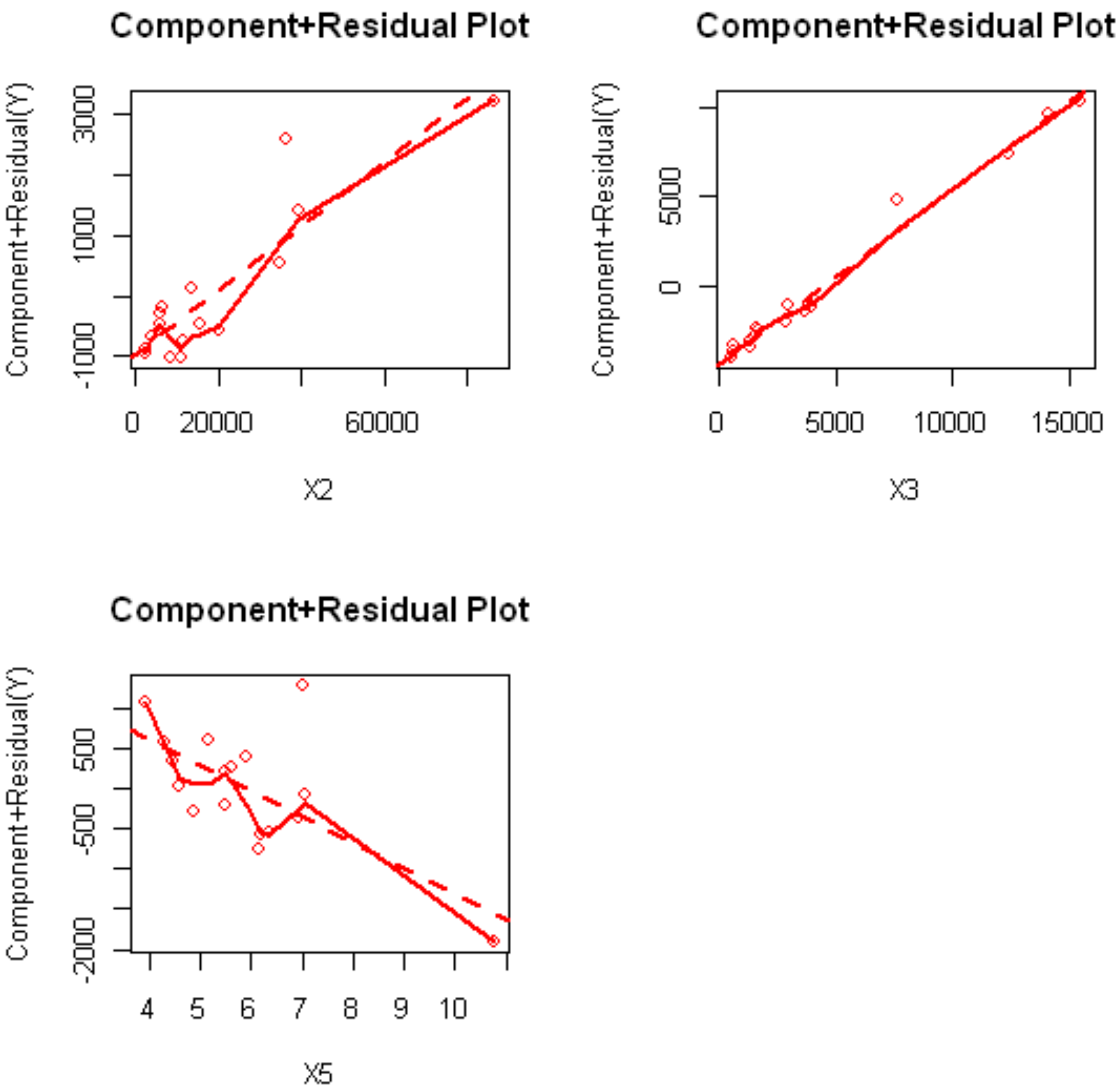
Procedure :

Plot $\mathrm{e}_{\mathrm{y|X}} + \mathrm{x}_{\mathrm{j}}\mathrm{b}_{\mathrm{j}}$ against $\mathrm{x}_{\mathrm{j}}$

To see how y and $\mathrm{x}_{\mathrm{j}}$ are related.

In some cases this can result in a more effective detection on non - linearity in the regressor $\mathrm{x}_{\mathrm{j}}$.

```
library(car)
crPlots(lm.manpower)
```



# Detection of Outliers

- **An outliers test:**

$$t_i = \frac{y_i - \hat{y}_i}{s_{-i}\sqrt{1 - h_{ii}}}$$

where

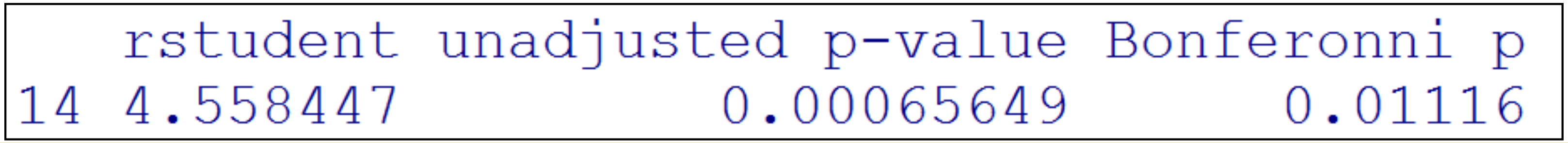
$$s_{-i}^2 = \frac{(n-p)s^2 - e_i^2 \big/ (1 - h_{ii})}{n - p - 1}$$

If the hypothesis is true, $t \sim t_{n\text{-}p\text{-}1}$

```
library(car)
outlierTest(lm.manpower)
```

or

```
rstudent(lm.manpower)
```



```
   rstudent unadjusted p-value Bonferonni p
14 4.558447         0.00065649      0.01116
```

# How to handle outliers?

1. Check original source of data. Remove the outlier if it appears to be an error.
2. Provide two data analysis. (with and without outliers)
3. Use weighted least squares estimation. Put less weight on outliers.

# Influence of cases

- For a regression model with *p* parameters: Any observation with $h_{ii}>2p/n$ has potential for exerting strong influence on the results. This does not apply for data set with $2p/n>1$.
- Pay attention on observation (s) that have $h_{ii}>2n/p$ or $|t_i|>2$.

$$t_i = \frac{y_i - \hat{y}_i}{s_{-i}\sqrt{1-h_{ii}}}$$

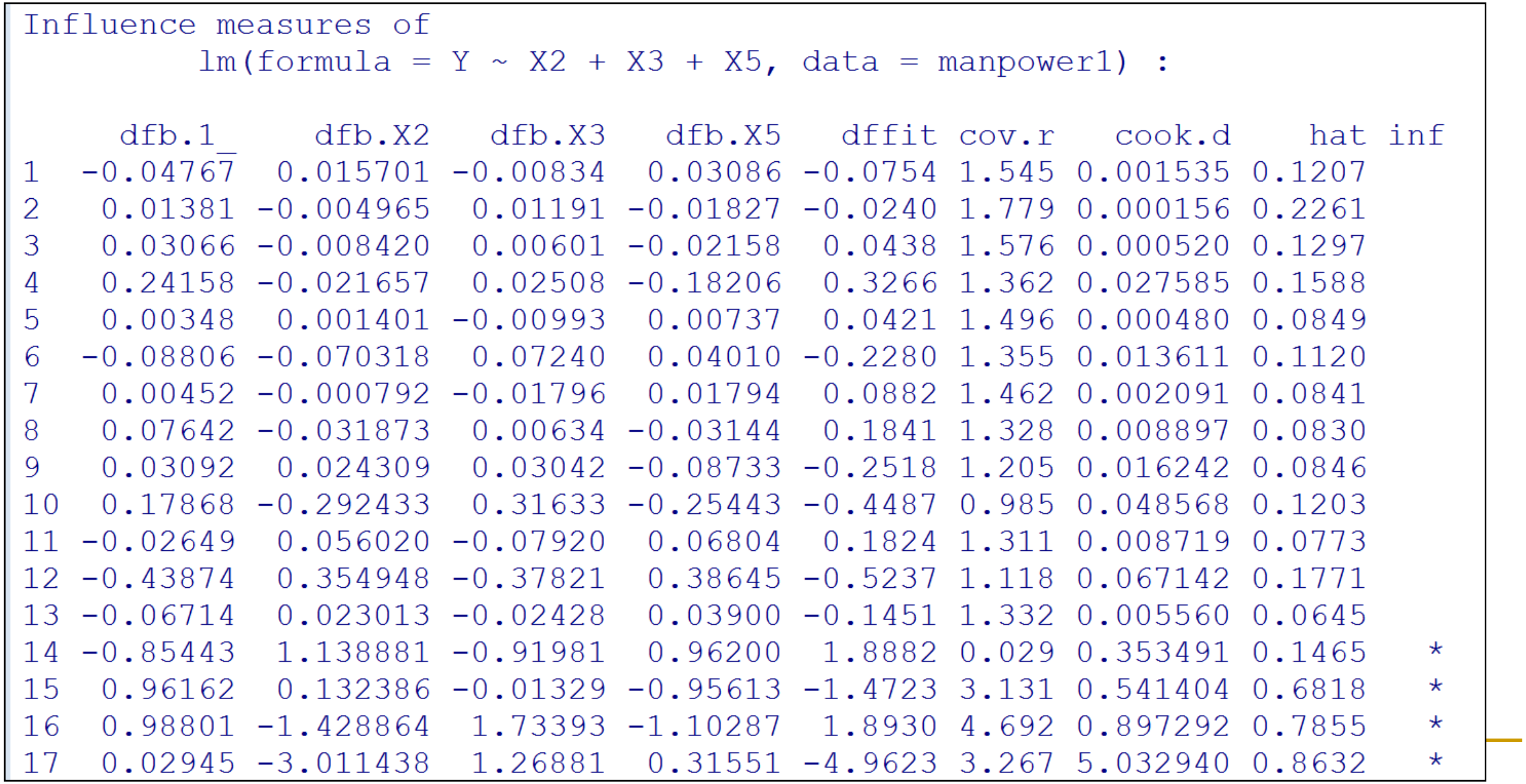
```
ti=rstudent(lm1)
which(abs(ti)>2)
```

Hat matrix diagonal $h_{ii}$:

```
lm1<-lm(Y~X2+X3+X5,data=manpower1)
x<-model.matrix(lm1)
hat1=hat(x)
which(hat1>(2*4)/17)
or
lm.influence(lm1)$hat
```

# Influence case (DFFITS, DFBETA, Cook distance)

influence.measures(lm1)



```
Influence measures of
         lm(formula = Y ~ X2 + X3 + X5, data = manpower1) :

     dfb.1_    dfb.X2   dfb.X3   dfb.X5   dffit cov.r   cook.d    hat inf
1  -0.04767  0.015701 -0.00834  0.03086 -0.0754 1.545 0.001535 0.1207
2   0.01381 -0.004965  0.01191 -0.01827 -0.0240 1.779 0.000156 0.2261
3   0.03066 -0.008420  0.00601 -0.02158  0.0438 1.576 0.000520 0.1297
4   0.24158 -0.021657  0.02508 -0.18206  0.3266 1.362 0.027585 0.1588
5   0.00348  0.001401 -0.00993  0.00737  0.0421 1.496 0.000480 0.0849
6  -0.08806 -0.070318  0.07240  0.04010 -0.2280 1.355 0.013611 0.1120
7   0.00452 -0.000792 -0.01796  0.01794  0.0882 1.462 0.002091 0.0841
8   0.07642 -0.031873  0.00634 -0.03144  0.1841 1.328 0.008897 0.0830
9   0.03092  0.024309  0.03042 -0.08733 -0.2518 1.205 0.016242 0.0846
10  0.17868 -0.292433  0.31633 -0.25443 -0.4487 0.985 0.048568 0.1203
11 -0.02649  0.056020 -0.07920  0.06804  0.1824 1.311 0.008719 0.0773
12 -0.43874  0.354948 -0.37821  0.38645 -0.5237 1.118 0.067142 0.1771
13 -0.06714  0.023013 -0.02428  0.03900 -0.1451 1.332 0.005560 0.0645
14 -0.85443  1.138881 -0.91981  0.96200  1.8882 0.029 0.353491 0.1465   *
15  0.96162  0.132386 -0.01329 -0.95613 -1.4723 3.131 0.541404 0.6818   *
16  0.98801 -1.428864  1.73393 -1.10287  1.8930 4.692 0.897292 0.7855   *
17  0.02945 -3.011438  1.26881  0.31551 -4.9623 3.267 5.032940 0.8632   *
```