Diagnostics and transformations for linear regression

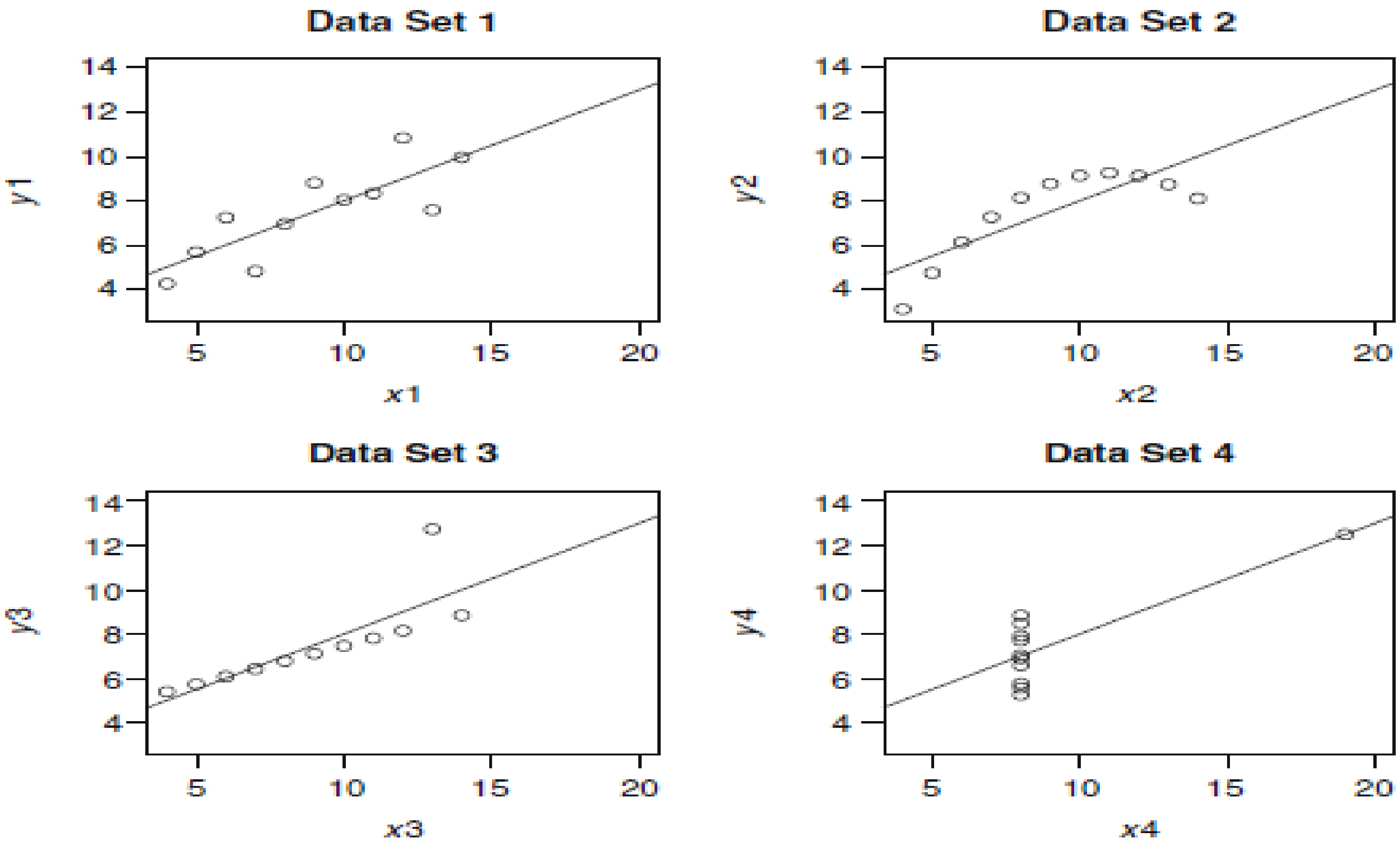
Table 3.1 Anscombe's four data sets

Case	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.1	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.1	5.39	12.5
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

Sheather, S.J., (2009). *A Modern Approach to Regression with R*, Springer.

Regression output from R

```
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)   3.0001      1.1247    2.667    0.02573    *
x1            0.5001      0.1179    4.241    0.00217    **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*'  0.05 '.' 0.1  ' ' 1

Residual standard error: 1.237  on  9  degrees  of  freedom
Multiple R-Squared:  0.6665,  Adjusted R-squared:  0.6295
F-statistic:  17.99  on  1  and  9  DF,   p-value:  0.002170

Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)    3.001       1.125    2.667    0.02576    *
x2             0.500       0.118    4.239    0.00218    **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*'  0.05 '.' 0.1  ' ' 1

Residual standard error: 1.237  on  9  degrees  of  freedom
Multiple R-Squared:  0.6662,  Adjusted R-squared:  0.6292
F-statistic:  17.97  on  1  and  9  DF,   p-value:  0.002179

Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)   3.0025      1.1245    2.670    0.02562    *
x3            0.4997      0.1179    4.239    0.00218    **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*'  0.05 '.' 0.1  ' ' 1

Residual  standard  error:  1.236  on  9  degrees  of  freedom
Multiple  R-Squared:  0.6663,  Adjusted  R-squared:  0.6292
F-statistic:  17.97  on  1  and  9  DF,   p-value:  0.002176

Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)   3.0017      1.1239    2.671    0.02559    *
x4            0.4999      0.1178    4.243    0.00216    **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*'  0.05 '.' 0.1  ' ' 1

Residual  standard  error:  1.236  on  9  degrees  of  freedom
Multiple  R-Squared:  0.6667,  Adjusted  R-squared:  0.6297
F-statistic:  18  on  1  and  9  DF,   p-value:  0.002165
```

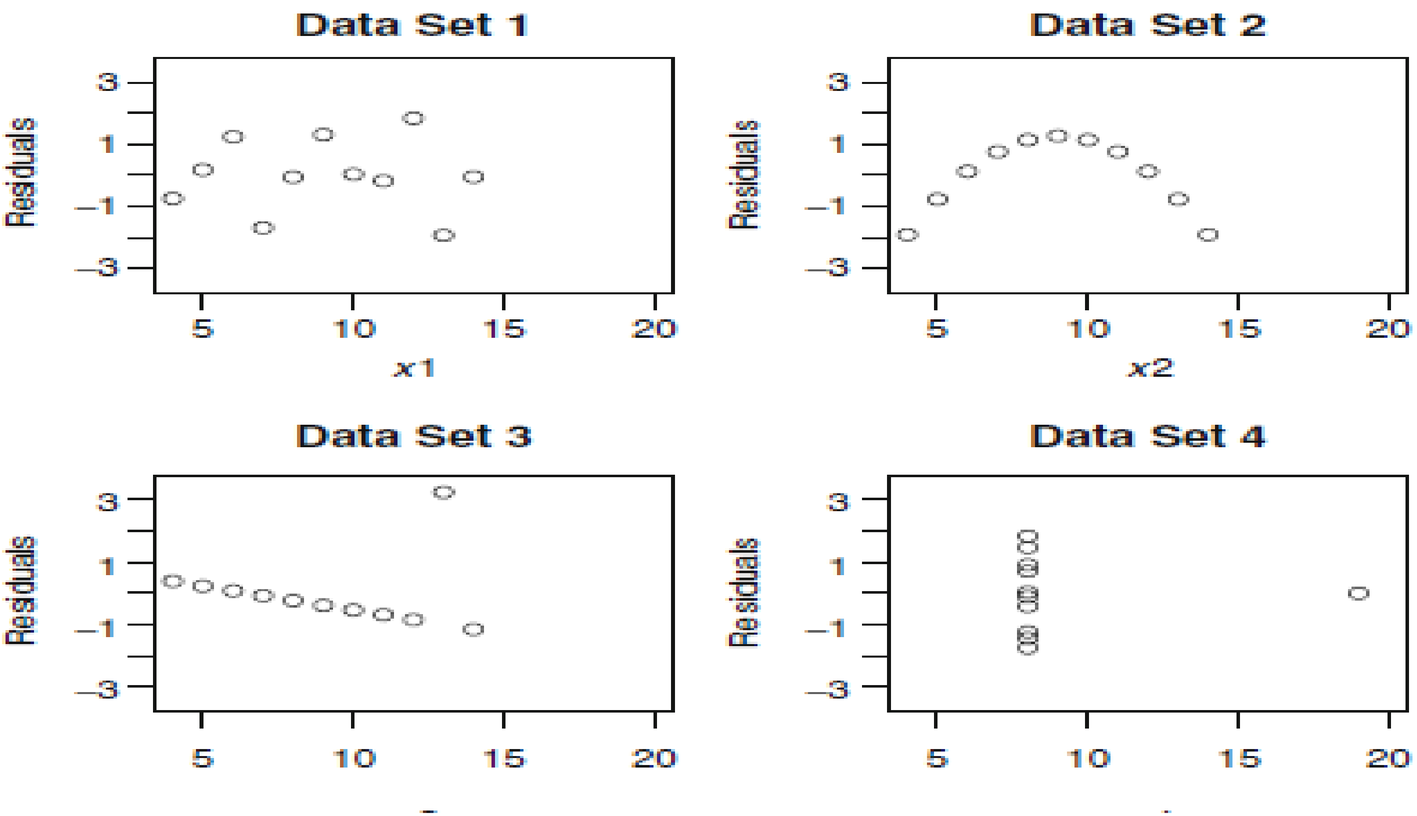


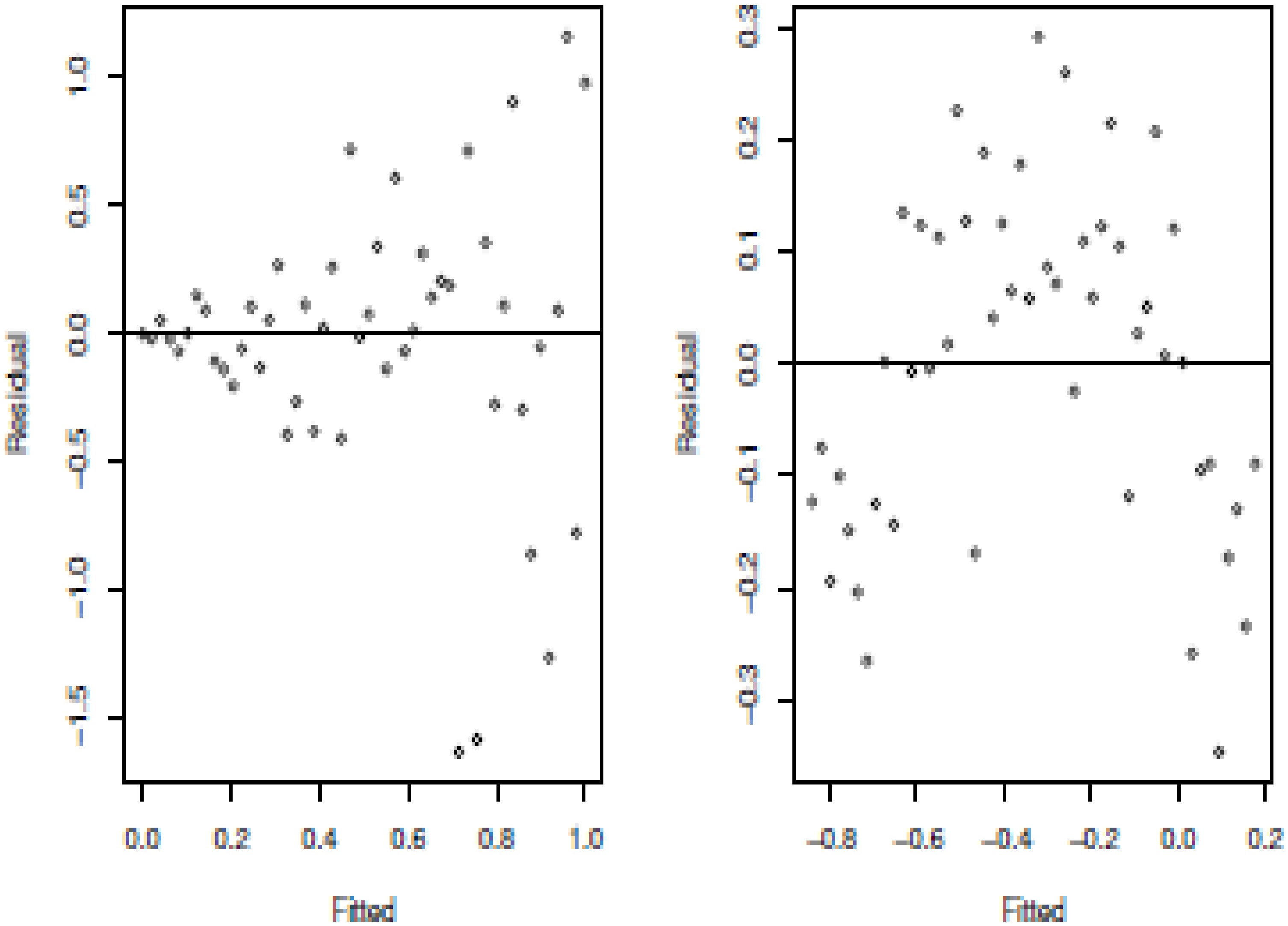
Also: http://www.youtube.com/watch?v=sfH43temzQY

Tools to check the appropriateness of the fitted model

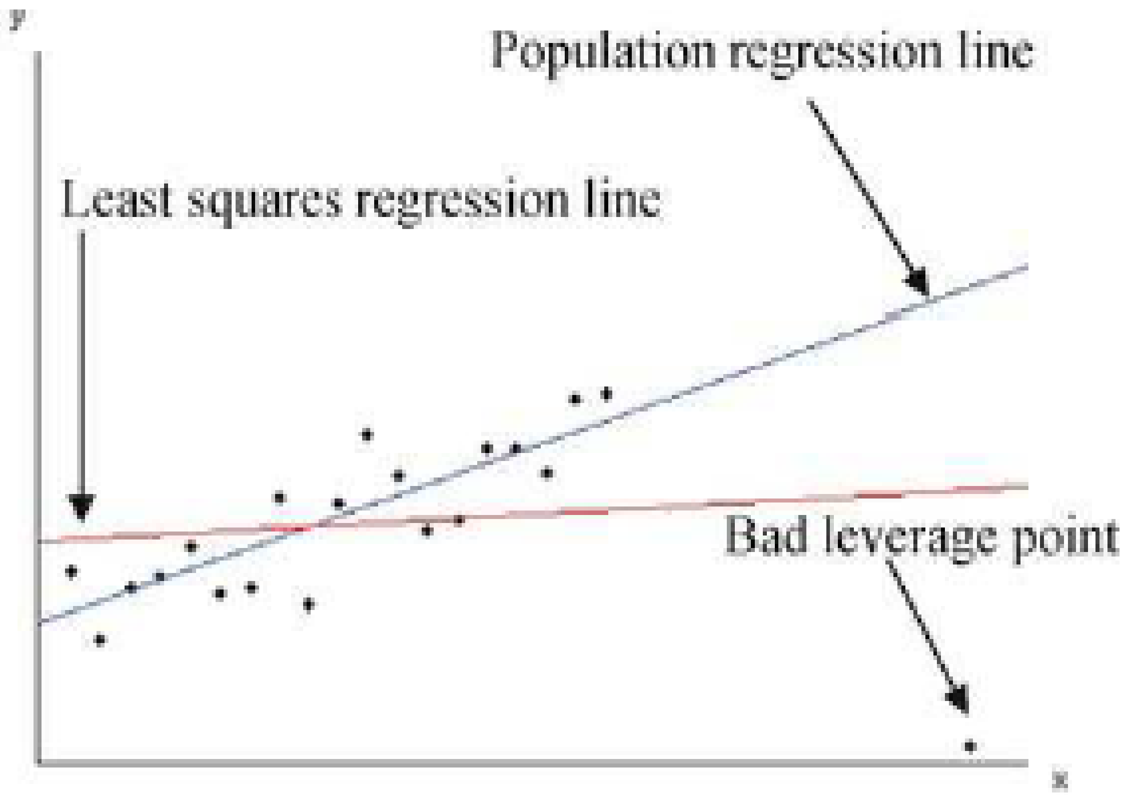
- Residuals
- Outlier
- Leverage
- Influence

(standardized) Residuals

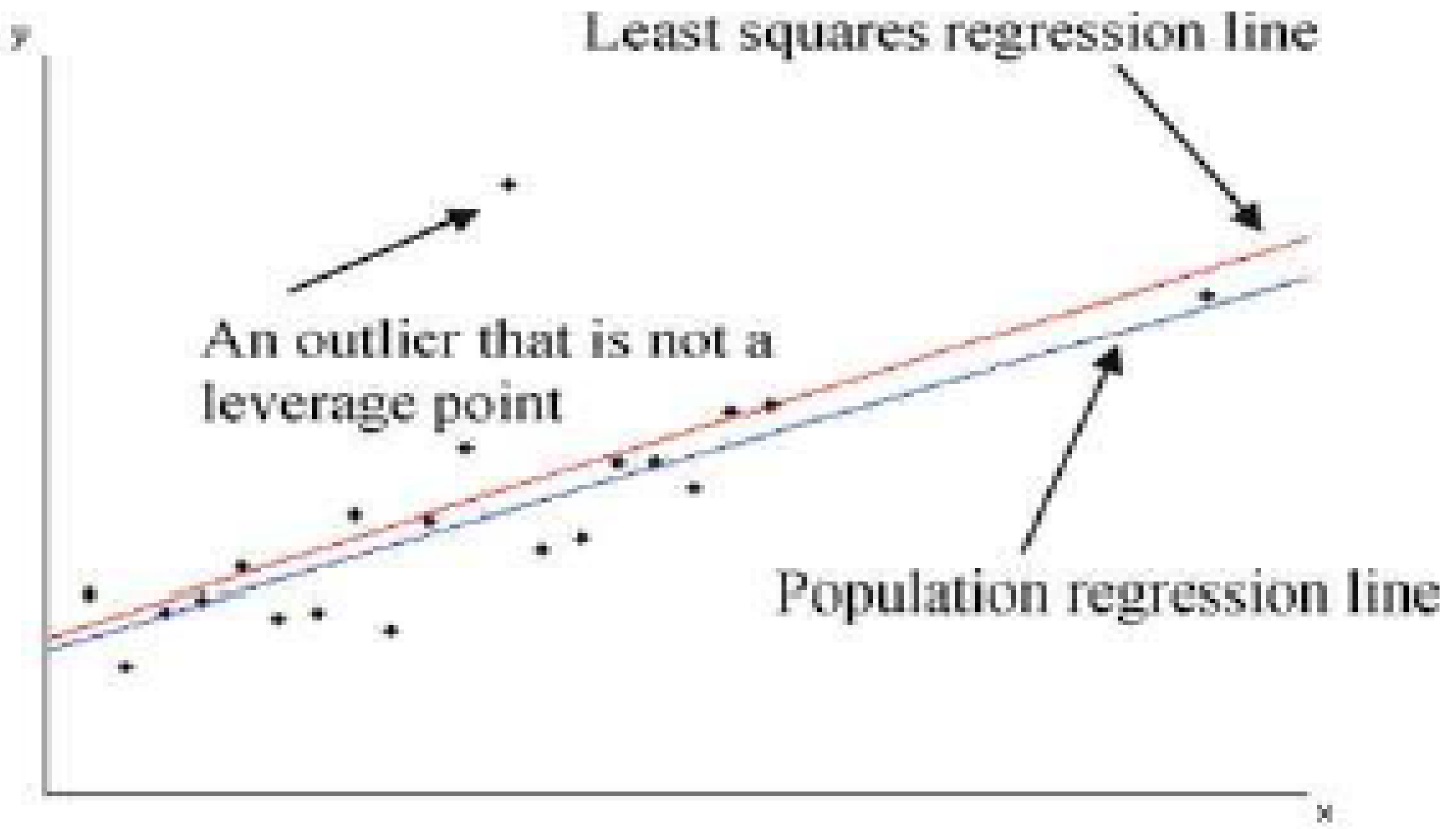



Residual
Fitted
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.2
0.4
0.6
0.8
1.0
0.3
0.2
0.1
−0.1
−0.2
−0.3
−0.8
−0.6
−0.4
−0.2

Leverage Points



Leverage point $\neq$ outlier



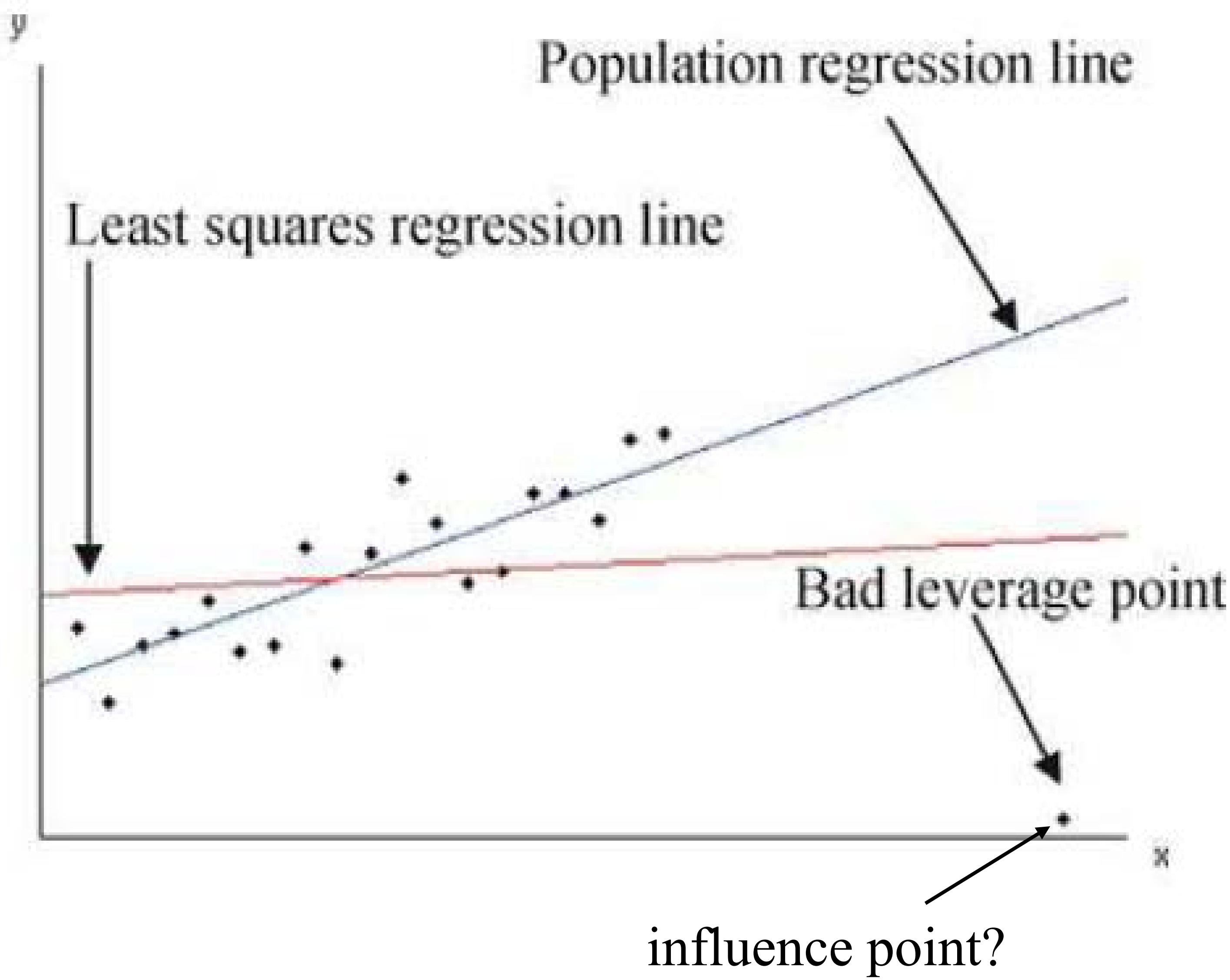
Rule for identifying leverage points

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} > 2 * \text{average}\left(h_{ii}\right) = \frac{2\left(p+1\right)}{n}$$

p : the number of covariates

Strategies for dealing with “bad” leverage points

- Remove invalid data points (not be routinely deleted)
- Fit a different regression model


y
Population regression line
Least squares regression line
Bad leverage point
x
influence point?

Standardized residuals

$$Var\left(\hat{e}_i\right) = \left(1 - h_{ii}\right)\sigma^2$$

$$r_i = \frac{\hat{e}_i}{\sqrt{\dfrac{\sum_{i=1}^{n} \hat{e}_i^{\,2}}{n - (p+1)}} \cdot \sqrt{1 - h_{ii}}} = \frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}}$$

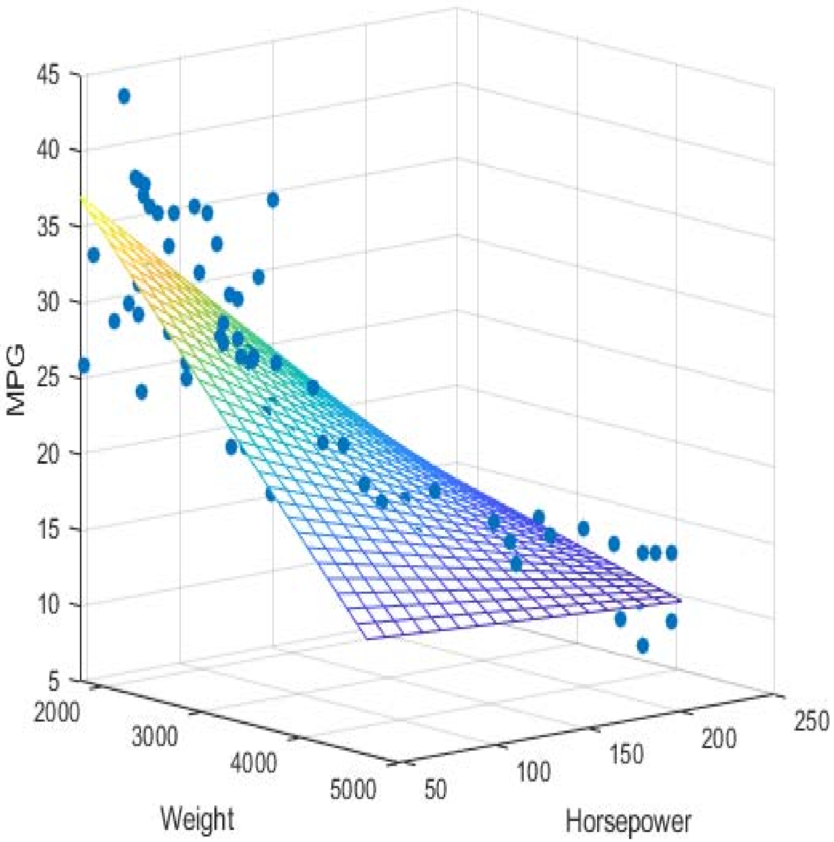
Studentized (deleted) residuals

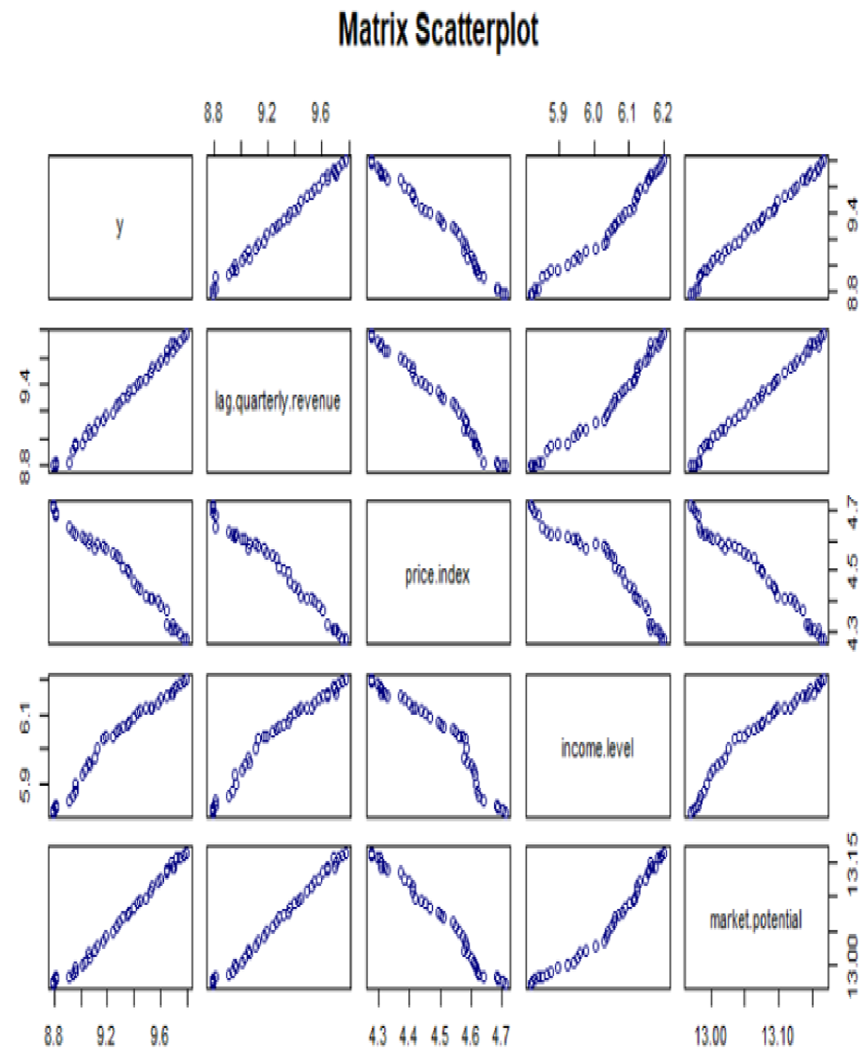
Let $\tilde{r}_i = \dfrac{\hat{e}_i}{s_{(-i)}\sqrt{1-h_{ii}}}$ where $s_{(-i)}^2 = \dfrac{(n-p-1)s^2 - \dfrac{\hat{e}_i^2}{1-h_{ii}}}{n-p-2}$.

$|\tilde{r}_i| \sim t_{n-p-2,\frac{\alpha}{2n}}$.

Multiple covariates case?

- outlier
- “good” or “bad” leverage points
- influence points





- https://www.mathworks.com/help/stats/regress.html
- https://www.educba.com/multiple-linear-regression-in-r/

Cook distance

- Cook (1977)

$$Cook_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_{j(-i)} - \hat{y}_j \right)^2}{(p+1)s^2} = \frac{r_i^2}{p+1} \cdot \frac{h_{ii}}{1-h_{ii}}.$$

- $\hat{y}_{j(-i)}$: the least square estimate obtained by deleting the i-th observation from data

Some relative influence measures using Cook's distance

- Cook and Weisberg (1982),

$$Cook_i > 1.$$

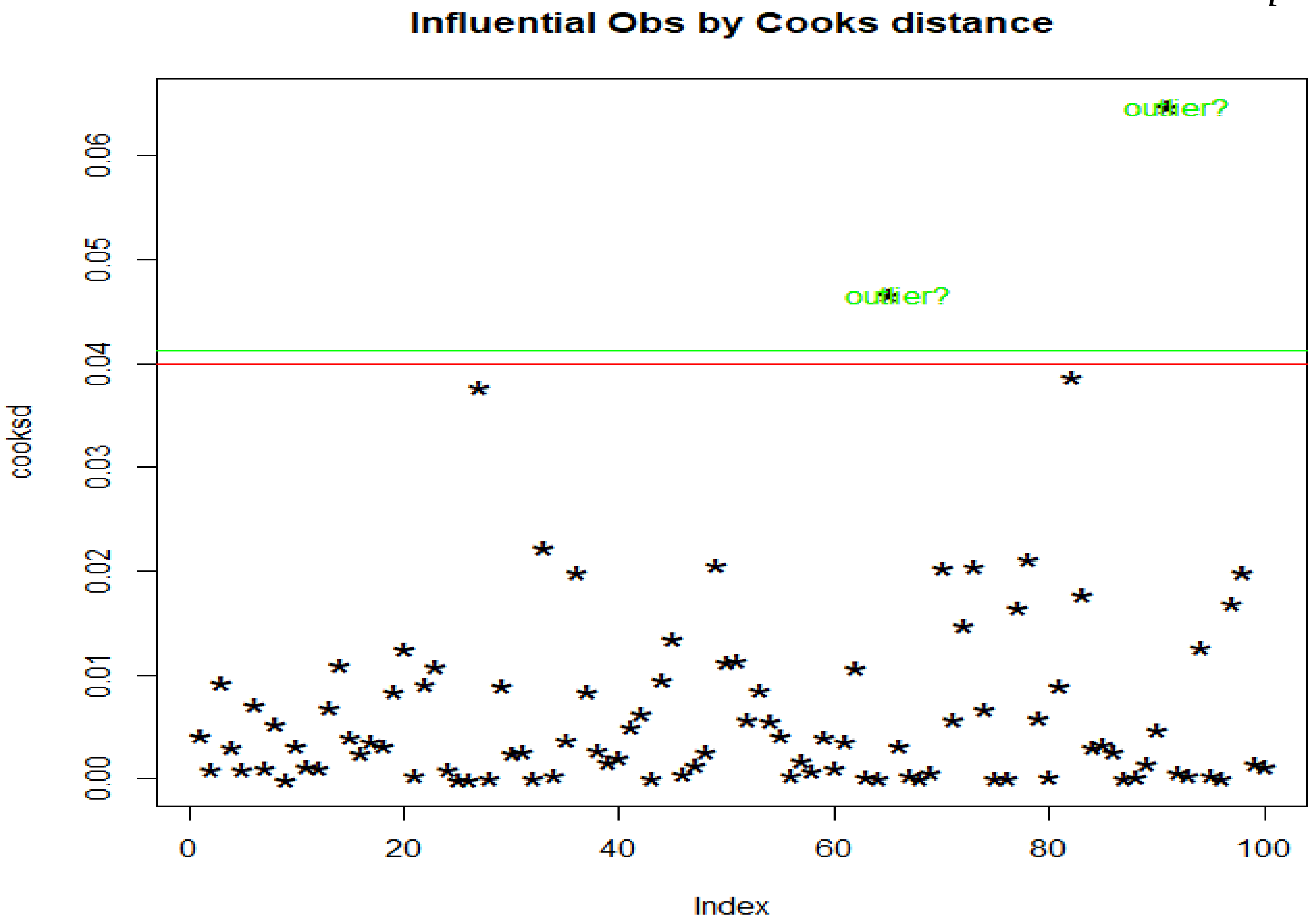
- Bollen and Jackman (1990),

$$Cook_i > \frac{4}{n}.$$

- Fox (2002),

$$Cook_i > \frac{4}{n-(p+1)}.$$

$$Cook_i > \frac{4}{n}.$$



```
set.seed(1082)
n=100 ; x1 = rexp(n) ; x2 =rlnorm(n, meanlog=log(2), sdlog = 1 ) # mean(x2)~3.4 , var(x2)~12.5
y = 3+ 2*x1 + 1*x2 + rnorm(n)
cooksd = cooks.distance(lm(y~x))

plot(cooksd, pch="*", cex=2, main="Influential Obs by Cooks distance") # plot cook's distance
abline(h = 1, col="blue")  # add cutoff line: simple rule
abline(h = 4/n, col="red")  # add cutoff line: cook and weisberg 1982
abline(h = 4/(n-(2+1)), col="green")  # add cutoff line: bollen and jackman 1990

text( x=1:length(cooksd)+1, y=cooksd, labels=ifelse(cooksd>4/n, "outlier?",""), col="red" )  # add labels using cook and weisberg 1982
```

DFFITS (difference between the fitted values)

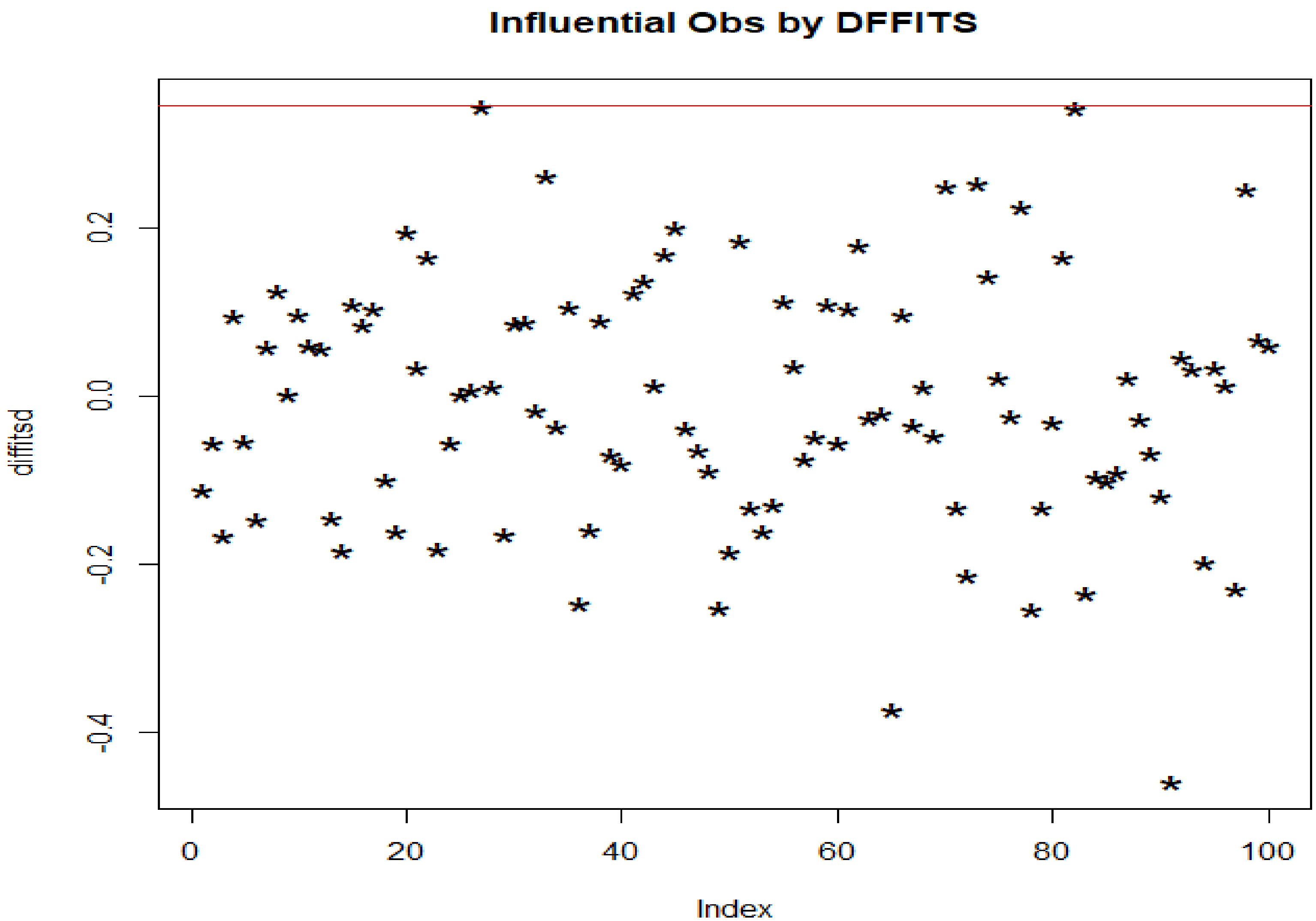
- Belsley, Kuh, and Welsch (1980, 2004)

$$\mathrm{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(-i)}}{s_{(-i)}\sqrt{h_{ii}}} = \tilde{r}_i \sqrt{\frac{1-h_{ii}}{h_{ii}}}$$

Some relative influences measure using DFFITS

$$\left|\mathrm{DFFITS}_i\right| > 1 = \left|\frac{\hat{y}_i - \hat{y}_{i(-i)}}{s_{(-i)}\sqrt{h_{ii}}}\right| = \left|\tilde{r}_i\right|\sqrt{\frac{1-h_{ii}}{h_{ii}}}$$

As large n, $\mathrm{DFFITS}_i > 2\sqrt{\frac{p+1}{n-(p+1)}} \approx 2\sqrt{\frac{p+1}{n}}$


Influential Obs by DFFITS
diffitsd
0.2
0.0
-0.2
-0.4
0
20
40
60
80
100
Index

```
set.seed(1082)
n=100 ; x1 = rexp(n) ; x2 =rlnorm(n, meanlog=log(2), sdlog
= 1 ) # mean(x2)~3.4 , var(x2)~12.5
y = 3+ 2*x1 + 1*x2 + rnorm(n)

diffitsd = dffits(lm(y~ x1 + x2))
plot(diffitsd, pch="*", cex=2, main="Influential Obs by
DFFITS")
abline(h = 2*sqrt((2+1)/n), col="red")
text( x=1:length(diffitsd)+1, y=diffitsd,
labels=ifelse(abs(diffitsd)>2*sqrt((2+1)/n), "outlier?",""),
col="red" )
```

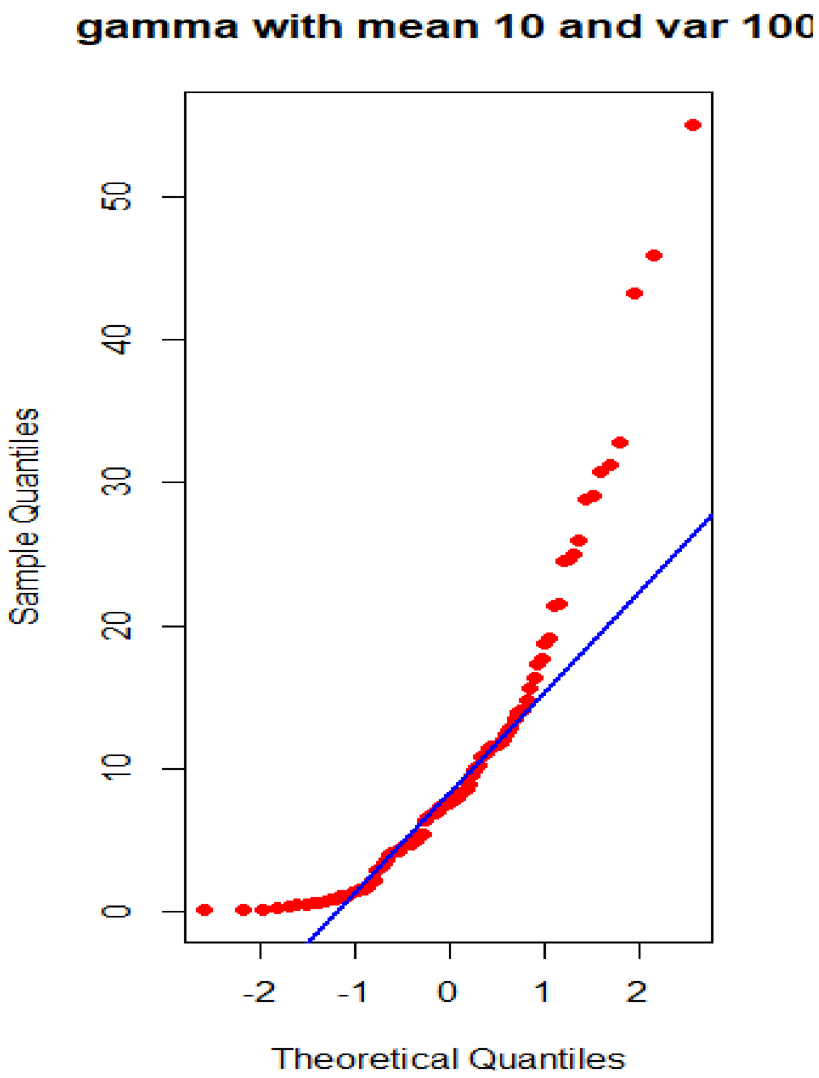
Predicted residual sums of squares (PRESS)

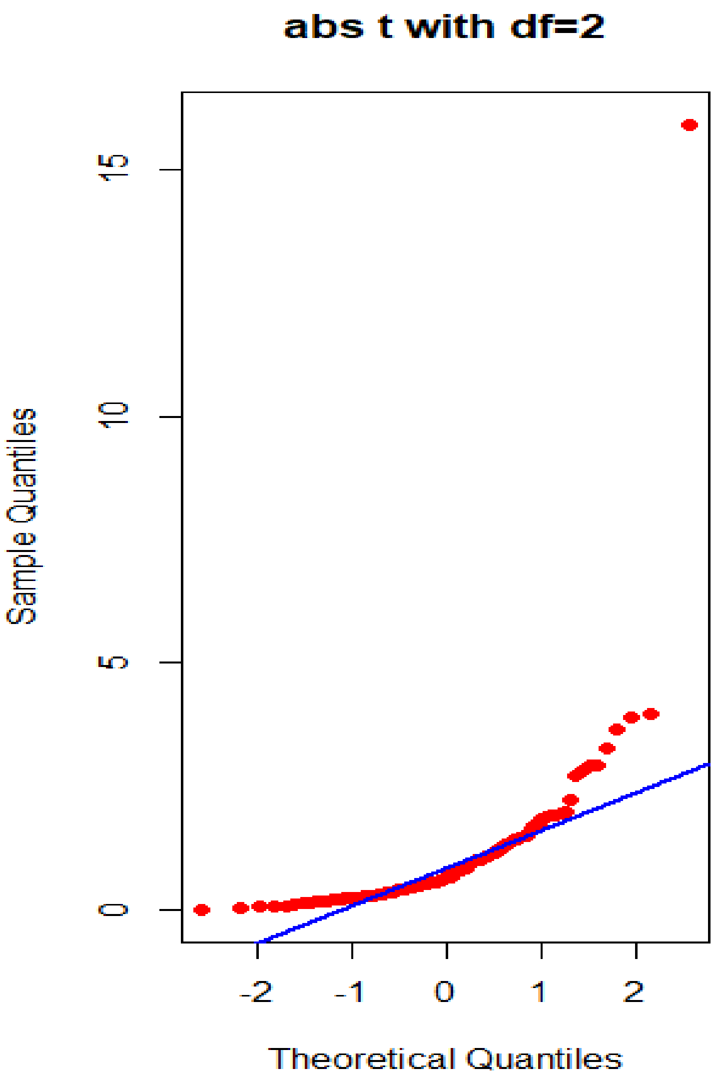
$$PRESS = \sum_{i=1}^{n} \left(y_i - \hat{y}_{i(-i)} \right)^2$$

PRESS(package: “qpcR”)
https://www.rdocumentation.org/packages/qpcR/versions/1.4-1/topics/PRESS

Transformations

- Overcome problems due to nonconstant variance
- Estimate percentage effects
- Overcome problems due to nonlinearity


gamma with mean 10 and var 100
Sample Quantiles
Theoretical Quantiles
0
10
20
30
40
50
-2
-1
0
1
2


abs t with df=2
Sample Quantiles
Theoretical Quantiles
0
5
10
15
-2
-1
0
1
2

```
par(mfrow=c(1, 2)) ; set.seed(1082)

sim = rgamma(100, 1, 1/10) ;
qqnorm(sim, pch=19, col = "red", main =
"gamma with mean 10 and var 100") ;
qqline(sim, col="blue",lwd =2 )

sim = abs( rt( 100, df=2 ) ) ;
qqnorm(sim, pch=19, col = "red", main = c("abs
t with df=2" )) ; qqline(sim, col="blue",lwd =2 )
```

Tests of normality in R

- Kolmogorov-Smirnov (Lilliefors) test (lillie.test, package: “nortest”)
- Anderson-Darling test (ad.test, package: “nortest”)
- Shapiro-Wilk test (shapiro.test)
- Jarque-Bera test (jarque.bera.test, package “tseries”)
- …etc

Box-Cox transformations

- (power transformation), on positive responses,

$$\Psi_S(y;\lambda)=\begin{cases}\dfrac{y^{\lambda}-1}{\lambda} & \text{if } \lambda\neq 0\\ \log(y) & \text{if } \lambda=0\end{cases},$$

and $\Psi_S(y;\lambda)$ is continuous on λ.

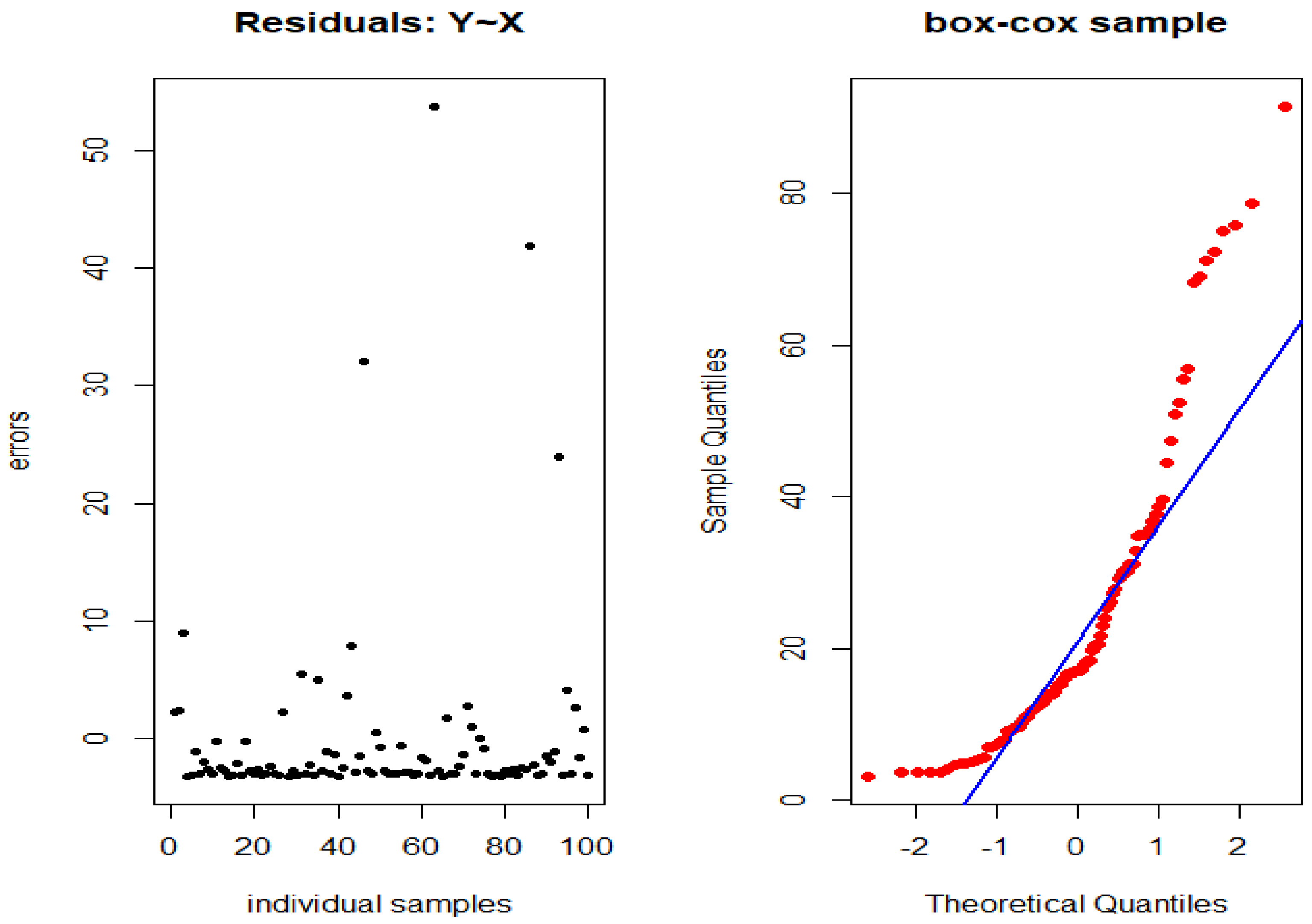
Maximize the log-likelihood function

$$\Psi\left(Y_i;\lambda\right) = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

$$\hat{\lambda} = \arg\max_{\lambda}\left\{\log L\left(\beta_0,\beta_1,\sigma^2,\lambda\right)\right\}$$

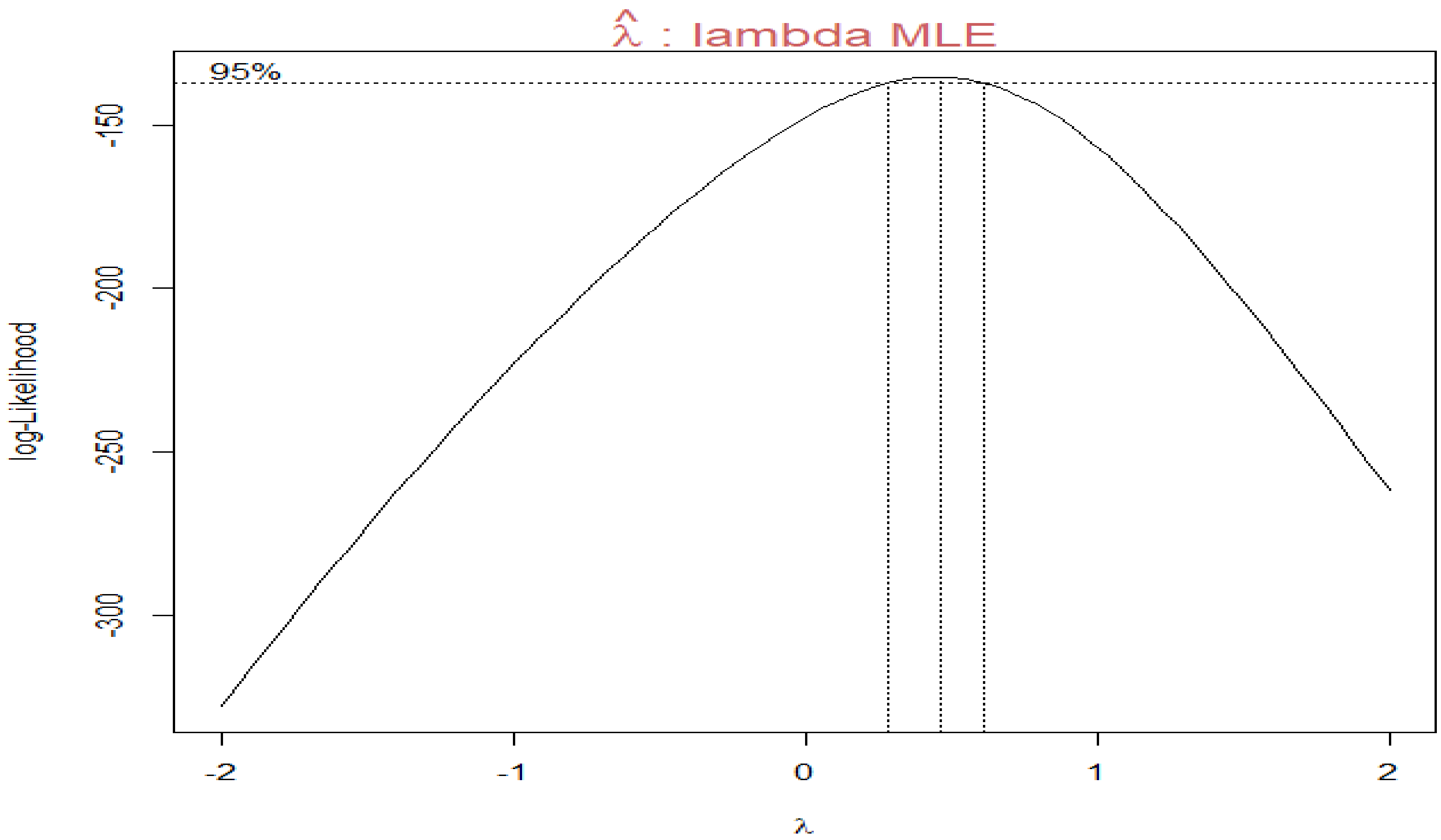
$$= \arg\max_{\lambda}\left\{\frac{-n}{2}\log\left(2\pi\right) - n\log\left(\sigma\right)\right.$$

$$\left. - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\Psi\left(Y_i;\lambda\right) - \left(\beta_0 + \beta_1 x_i\right)\right)^2 + \left(\lambda-1\right)\sum_{i=1}^{n}\log\left(y_i\right)\right\}$$

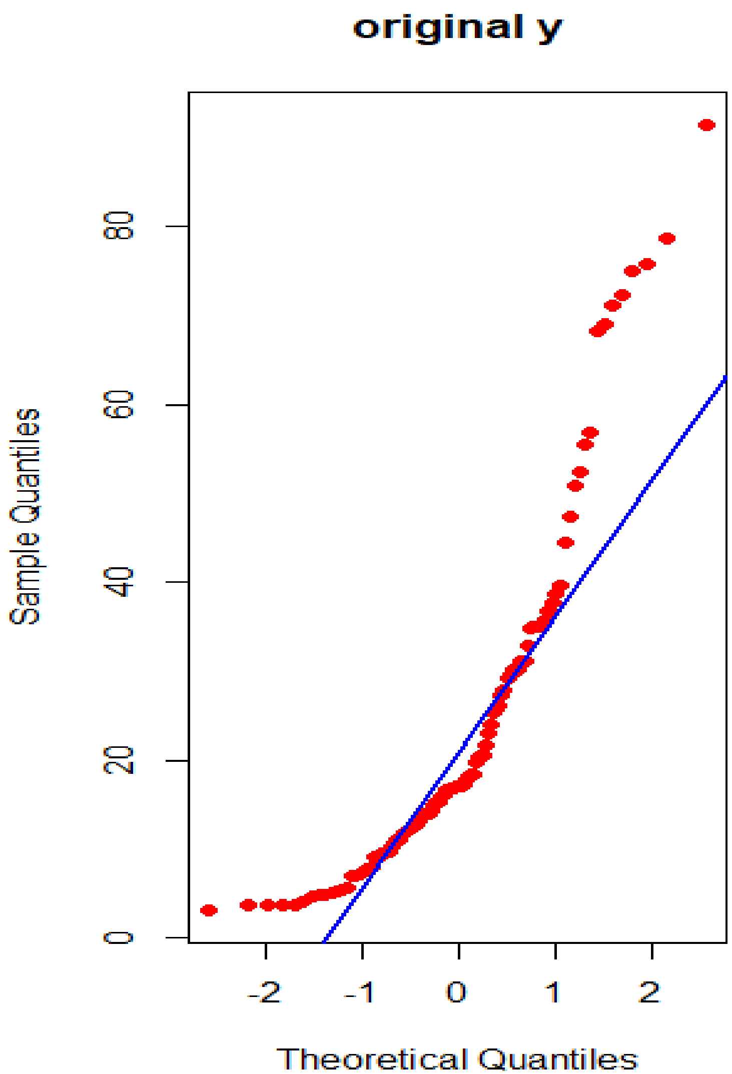


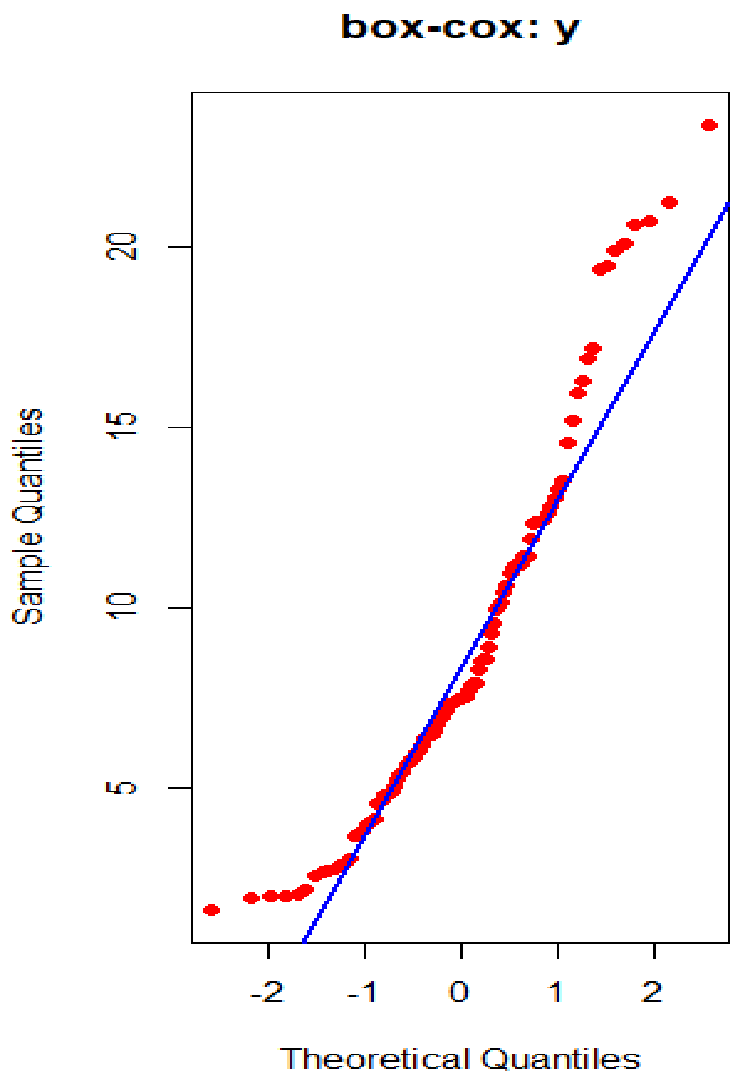
Residuals: Y~X
errors
0
10
20
30
40
50
0
20
40
60
80
100
individual samples
box-cox sample
Sample Quantiles
0
20
40
60
80
-2
-1
0
1
2
Theoretical Quantiles

```
par(mfrow=c(1, 2)) ; set.seed(108)
x=rgamma(100, 1, 1/10)
y = 3 + 2*x  + abs( rt( 100, df=2 ) )^2
error = lm( y ~ x )$residuals
plot( error, xlab = "individual samples", ylab
= "errors", main = "Residuals: Y~X",
      xlim = c(0, 100), pch=20, col= "black" )
qqnorm(y, pch=19, col = "red", main =
c("box-cox sample" )) ; qqline(y,
col="blue",lwd =2 )
```



library(MASS) ; boxcox(lm(y~x), lambda=seq(-2, 2, by=0.1))
mtext(expression(paste(hat(lambda), " : lambda MLE")), cex = 1.5, font=4, col=rgb(0.7,0.1,0.1,0.7))


original y
Sample Quantiles
Theoretical Quantiles
0
20
40
60
80
-2
-1
0
1
2


box-cox: y
Sample Quantiles
Theoretical Quantiles
5
10
15
20
-2
-1
0
1
2

```
lambda.t = 0.6
y.t = (y^lambda.t-1)/lambda.t
error.t = lm( y.t ~ x )$residuals
par(mfrow=c(1, 2)) ; set.seed(108)
qqnorm(y, pch=19, col = "red", main =
c("original y" )) ; qqline(y, col="blue",lwd =2 )
qqnorm(y.t, pch=19, col = "red", main =
c("box-cox: y" )) ; qqline(y.t, col="blue",lwd
=2 )
```

Modified Box-Cox transformations

- Box and Cox (1964)

$$\Psi_M(y;\lambda) = \begin{cases} \dfrac{y^{\lambda}-1}{GM(\mathbf{y})^{\lambda-1}\lambda} & \text{if } \lambda \neq 0 \\ GM(\mathbf{y})\log(y) & \text{if } \lambda = 0 \end{cases},$$

where $GM(\mathbf{y}) = \left(\prod_{i=1}^{n} y_i\right)^{1/n}$. $\hat{\lambda} = \arg\min_{\lambda} SSE.$

Box-Cox transformations

- (power transformation), on responses,

$$\Psi(y;\lambda,c)=\begin{cases}\dfrac{(y+c)^{\lambda}-1}{\lambda} & \text{if } \lambda\neq 0\\ \log(y+c) & \text{if } \lambda=0\end{cases}.$$

Logarithm or other transformations

$$\log(Y_i) = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i.$$

$$\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

$$\sqrt{Y_i} = \beta_0 + \beta_1 x_i + \varepsilon_i \text{....etc}$$