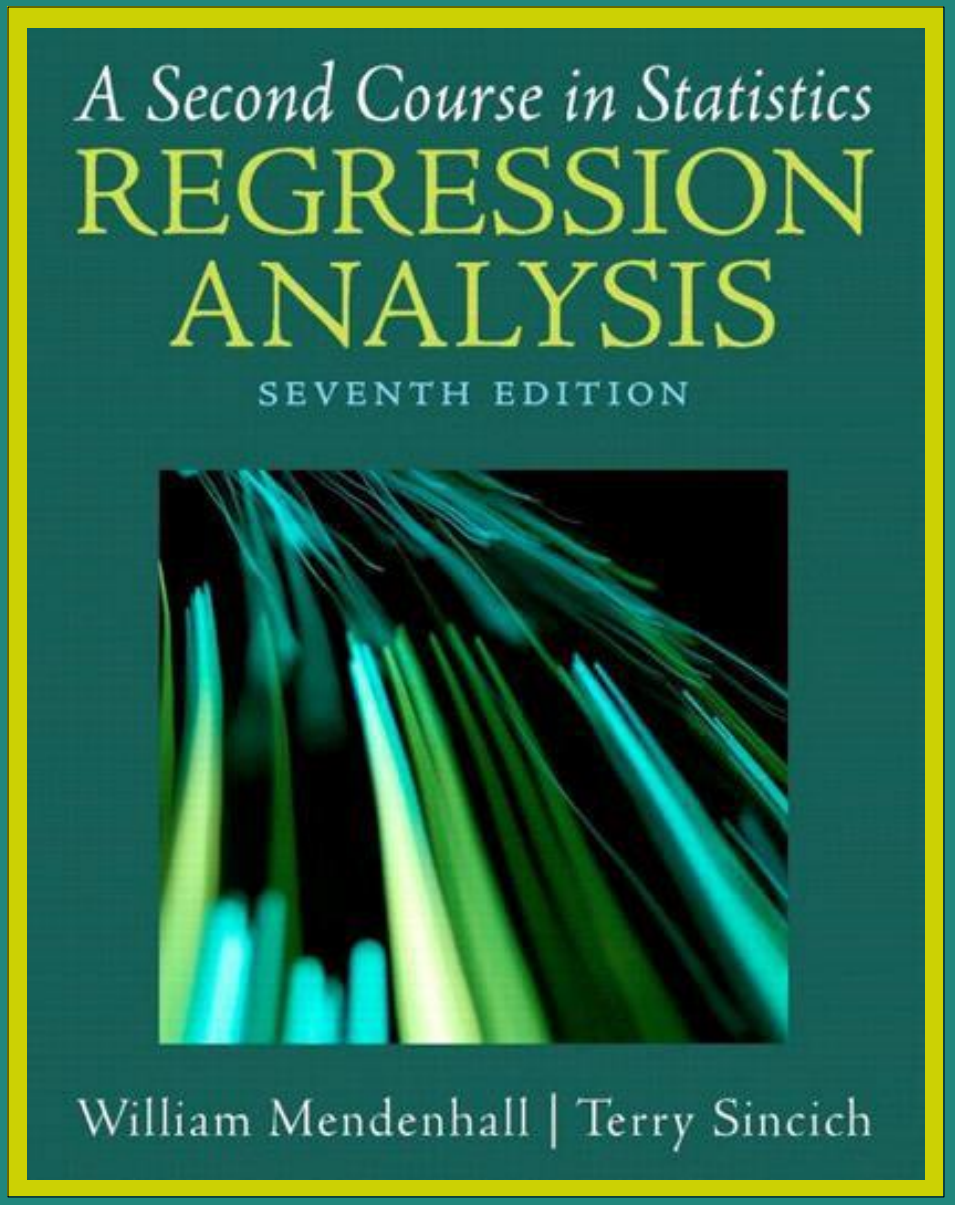
Chapter 8

Residual Analysis

Lecture 1

Dr Brenda Vo



PEARSON

STAT210/410 Study Plan

Topic	Weeks covered	Readings	Assessment
Topic 1: Simple Linear regression (SLR)	Wk 1	Chapter 3	Online Quiz due 9th March
Topic 2: Multiple Linear Regression (MLR)	Wk2 & 3	Chapter 4	Written Assessment A2 due 23rd March
Topic 3: Model building	Wk 4	Chapter 5	
Topic 4: Variable Screening and regression pitfalls	Wk 5	Chapters 6, 7	
Topic 5: Residual Analysis	Wk 6	Chapter 8	Written Assessment A3 due 13th April
Topic 6 Generalised Linear Models (GLMs)	Wk 9 & 10	Chapter 9	
Topic 7: Principles of Experimental Design	Wk 11	Chapter 11	Written Assessment A4 due 11th May
Topic 8: ANOVA, contrasts	Wk 12 & 13	Chapter 12	
STAT410 ONLY			
ART: Nonparametric Regression		Section 9.9	Written Assessment ART due 18th May

Chapter 8 outline

Lecture 1:

- Review assumptions of linear model
- Detect lack of fit
- Power transformations

Lecture 2:

- Detect outliers and influential points

Assumptions of the linear model

$$\varepsilon \sim N(0, \sigma^2)$$

The residuals

- are normally distributed,
- have a mean of 0
- have constant variance
- are independent

- Also check for *outliers* and *influential* points

Violated Assumptions

Issues….

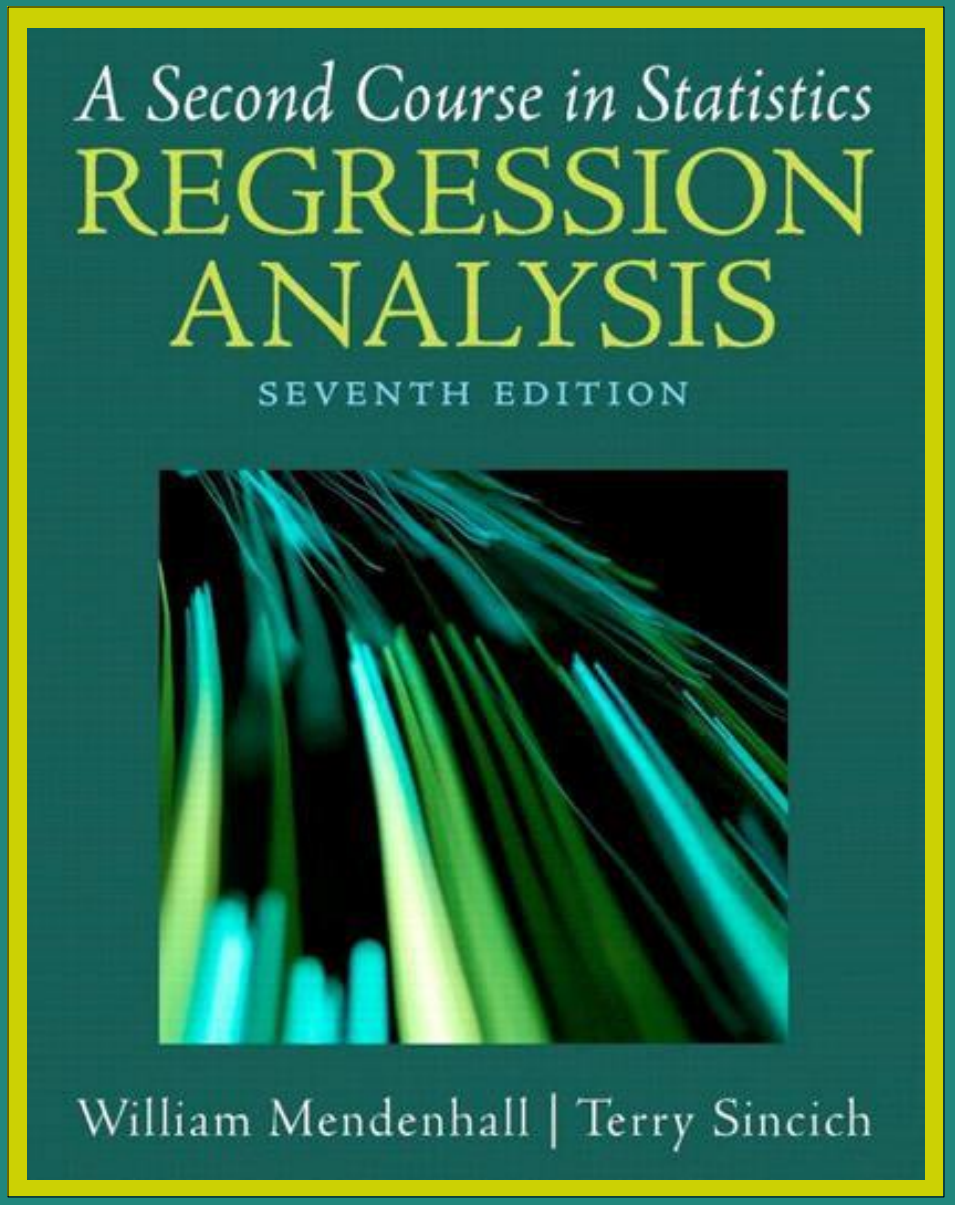
- Using model for estimation or prediction may be invalid
- Assumptions are reasonably “robust”

Solutions…..

- Transform response and/or predictor variables
- Weighted regression (constant variance)
- Generalized Linear Models - assume a non-normal error distribution

Section 8.3

Detecting Lack of Fit



PEARSON

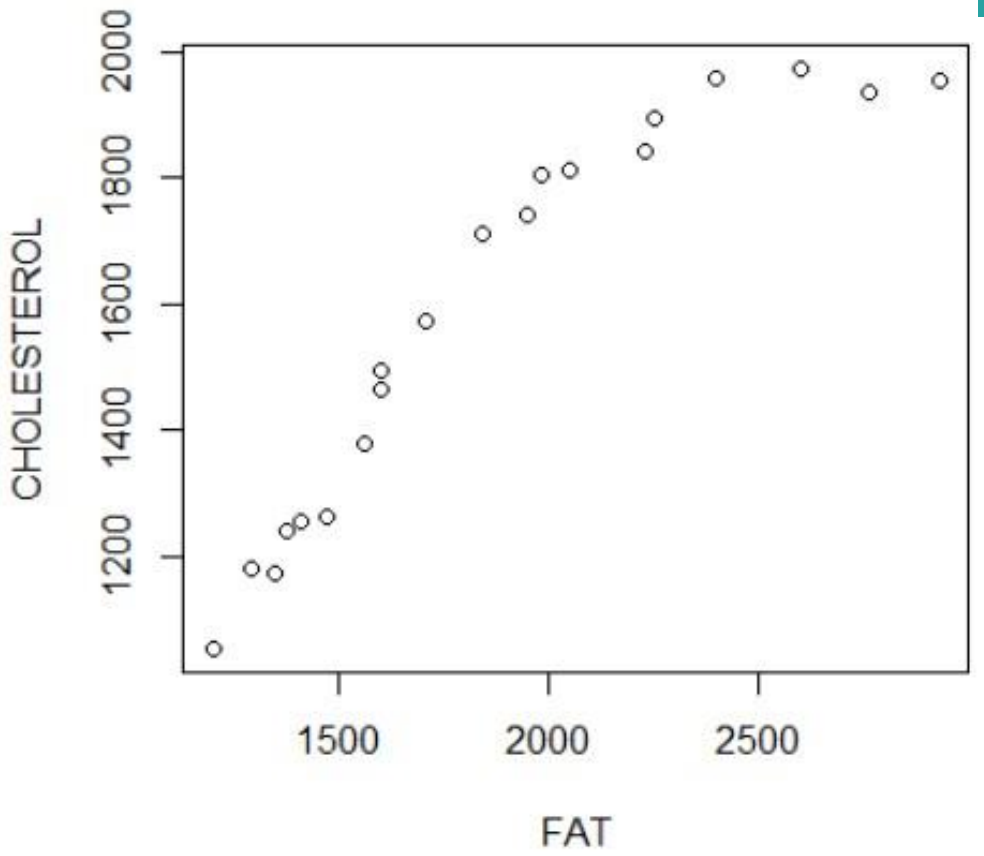
Detecting Lack of Fit

- Plot the residuals versus each of the independent variables x_1, x_2 … x_k
- Plot the residuals against the fitted value
- In each plot, look for trends, dramatic changes in variability.

Scatterplot

Table 8.1 Data for 20 Olympic athletes

Athlete	Fat intake x, milligrams	Cholesterol y, milligrams/liter
1	1,290	1,182
2	1,350	1,172
3	1,470	1,264
4	1,600	1,493
5	1,710	1,571
6	1,840	1,711
7	1,980	1,804
8	2,230	1,840
9	2,400	1,956
10	2,930	1,954
11	1,200	1,055
12	1,375	1,241
13	1,410	1,254
14	1,560	1,377
15	1,600	1,465
16	1,950	1,741
17	2,050	1,810
18	2,250	1,893
19	2,600	1,972
20	2,760	1,935



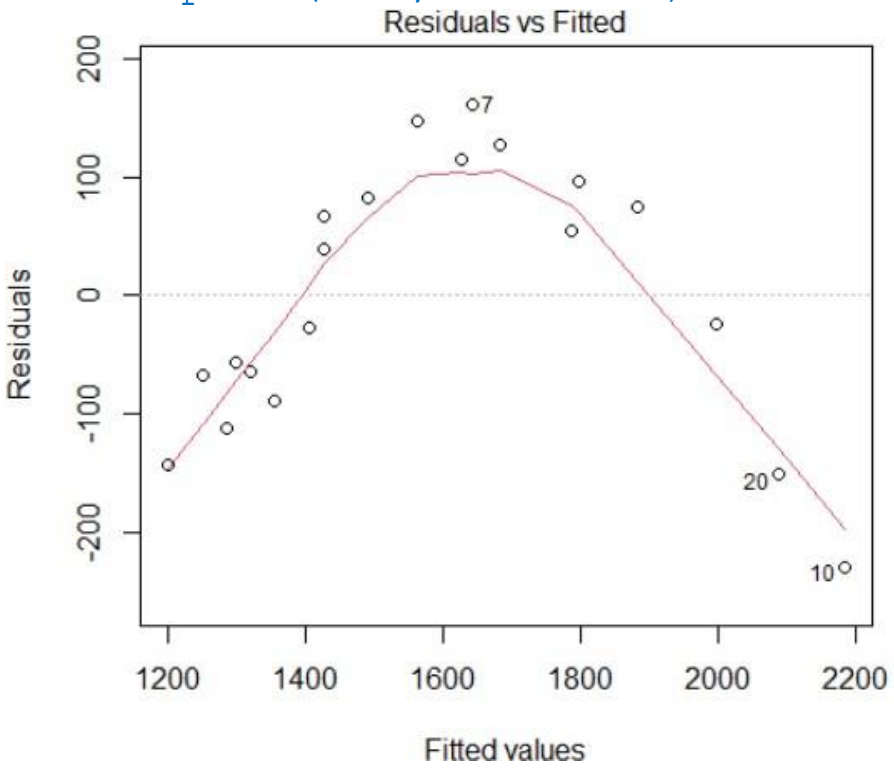

```
oly<-read.table("OLYMPIC.txt",header=T)
plot(CHOLESTEROL~FAT, data=oly)
```

R printout for first-order model

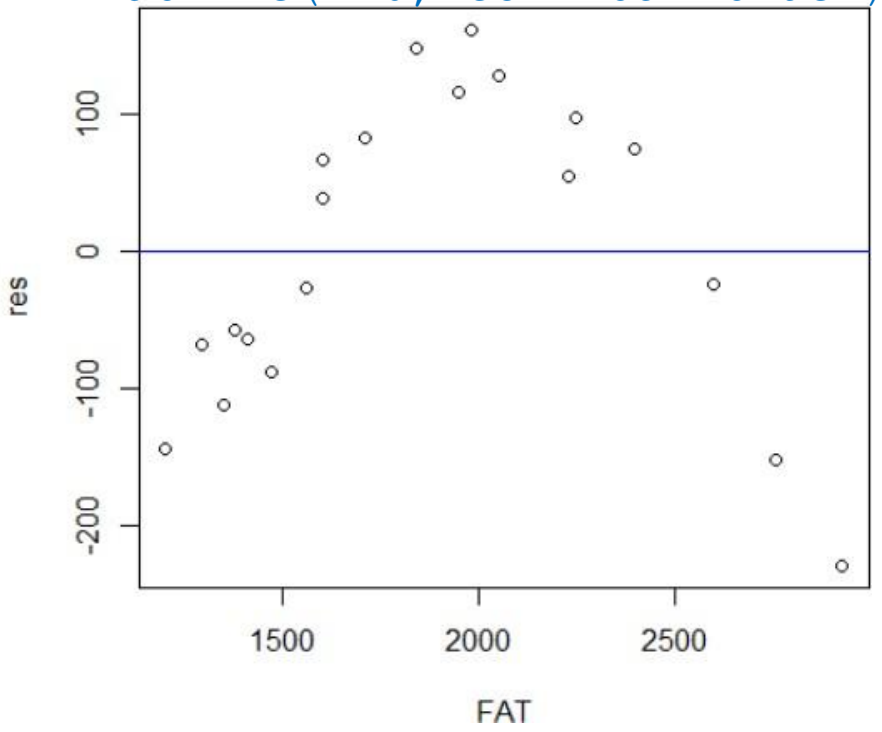
```
mod<-lm(CHOLESTEROL~FAT)
anova(mod)
summary(mod)
############################
Analysis of Variance Table

Response: CHOLESTEROL
          Df  Sum Sq Mean Sq F value  Pr(>F)
FAT        1 1617913 1617913     122 1.8e-09
Residuals 18  237980   13221

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 515.7050    99.9789    5.16  6.6e-05
FAT           0.5692     0.0515   11.06  1.8e-09

Residual standard error: 115 on 18 degrees of freedom
Multiple R-squared:  0.872,       Adjusted R-squared:  0.865
F-statistic:  122 on 1 and 18 DF,  p-value: 1.85e-09
```

Plots of residuals

```
plot(mod, which=1)
```




```
plot(mod$residuals~oly$FAT)
abline(h=0, col="darkblue")
```



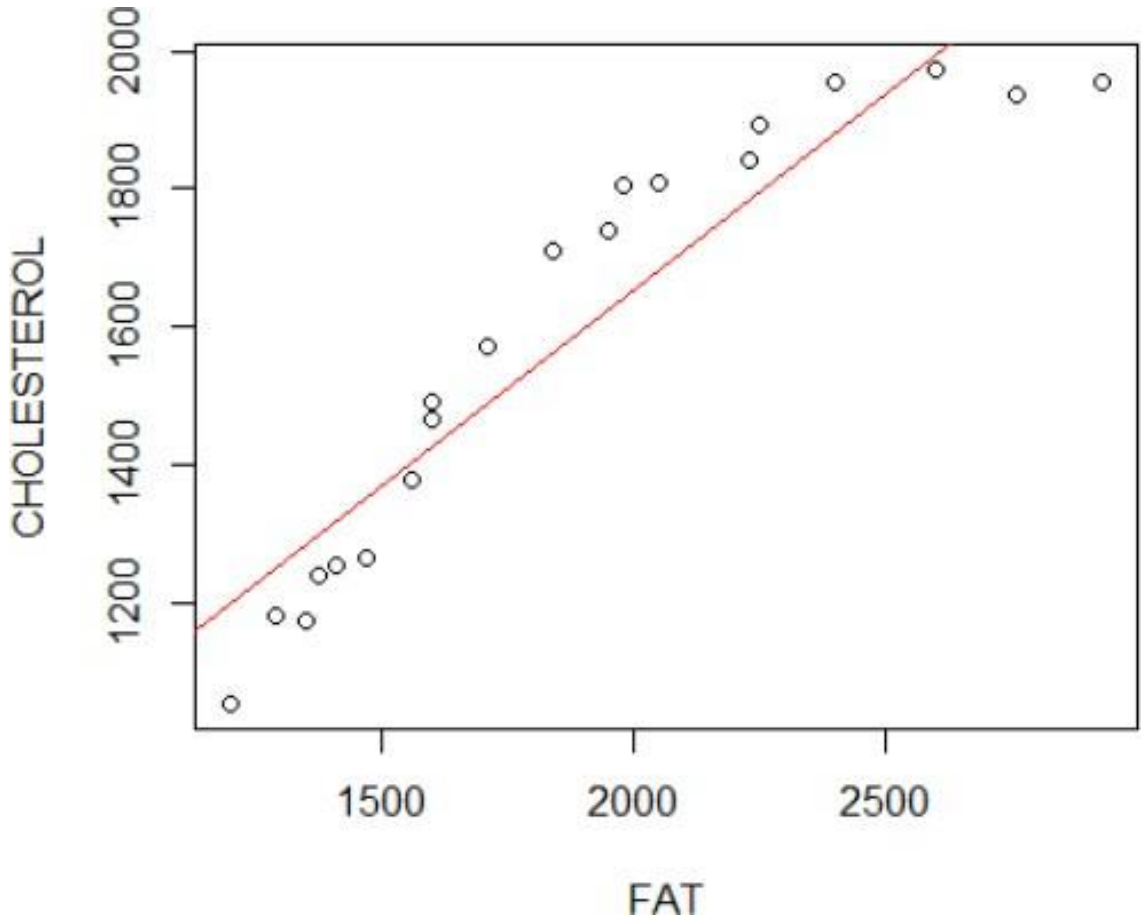
Residuals vs fitted value of y

Residuals vs x values

Observed, predicted and residuals

```
res<-mod$residuals
preds<-predict(mod)
cbind(FAT, CHOLESTEROL, preds, res)
#####################################
    FAT CHOLESTEROL preds     res

1   1290         1182  1250  -67.96
2   1350         1172  1284 -112.11
3   1470         1264  1352  -88.41
4   1600         1493  1426   66.59
5   1710         1571  1489   81.98
.     ..           ..    ..      ..
17  2050         1810  1683  127.46
18  2250         1893  1796   96.62
19  2600         1972  1996  -23.60
20  2760         1935  2087 -151.67
```

Scatterplot with SLR line

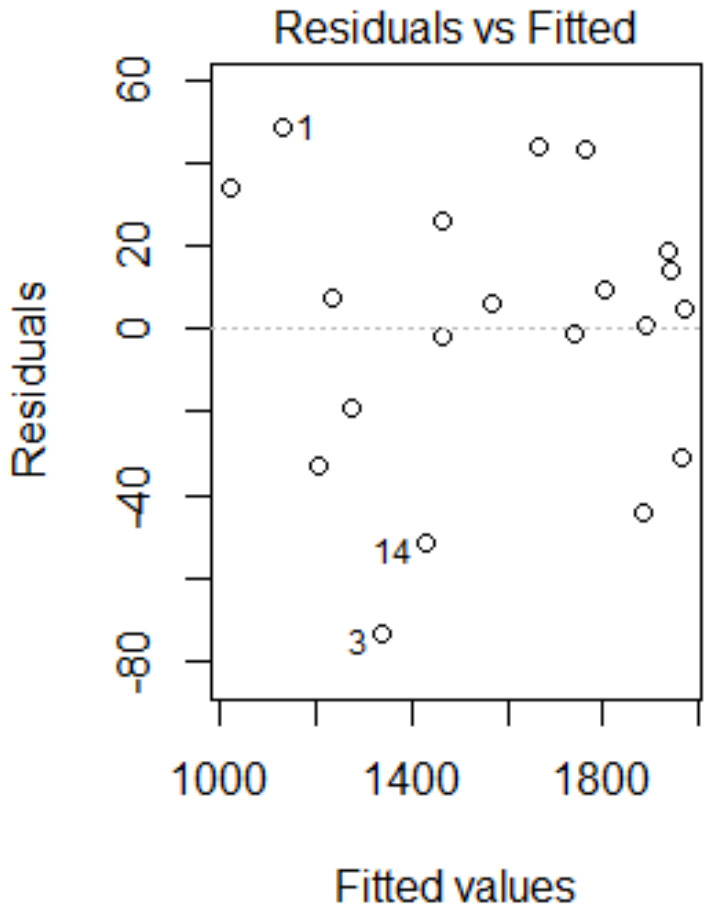


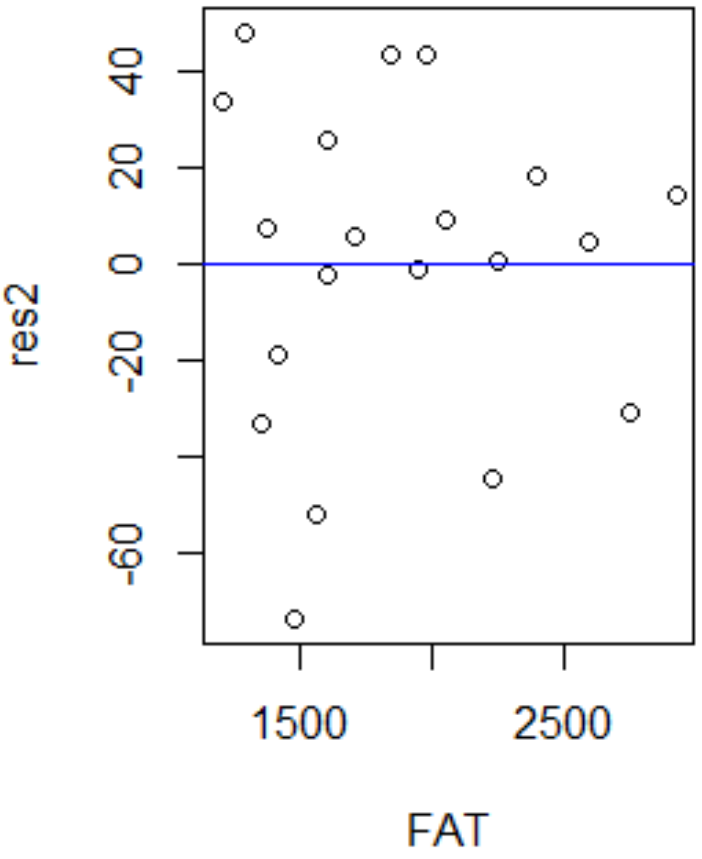
Quadratic Model

```
mod2<-lm(CHOLESTEROL~FAT + I(FAT^2),data = oly)
anova(mod2)
summary(mod2)
```

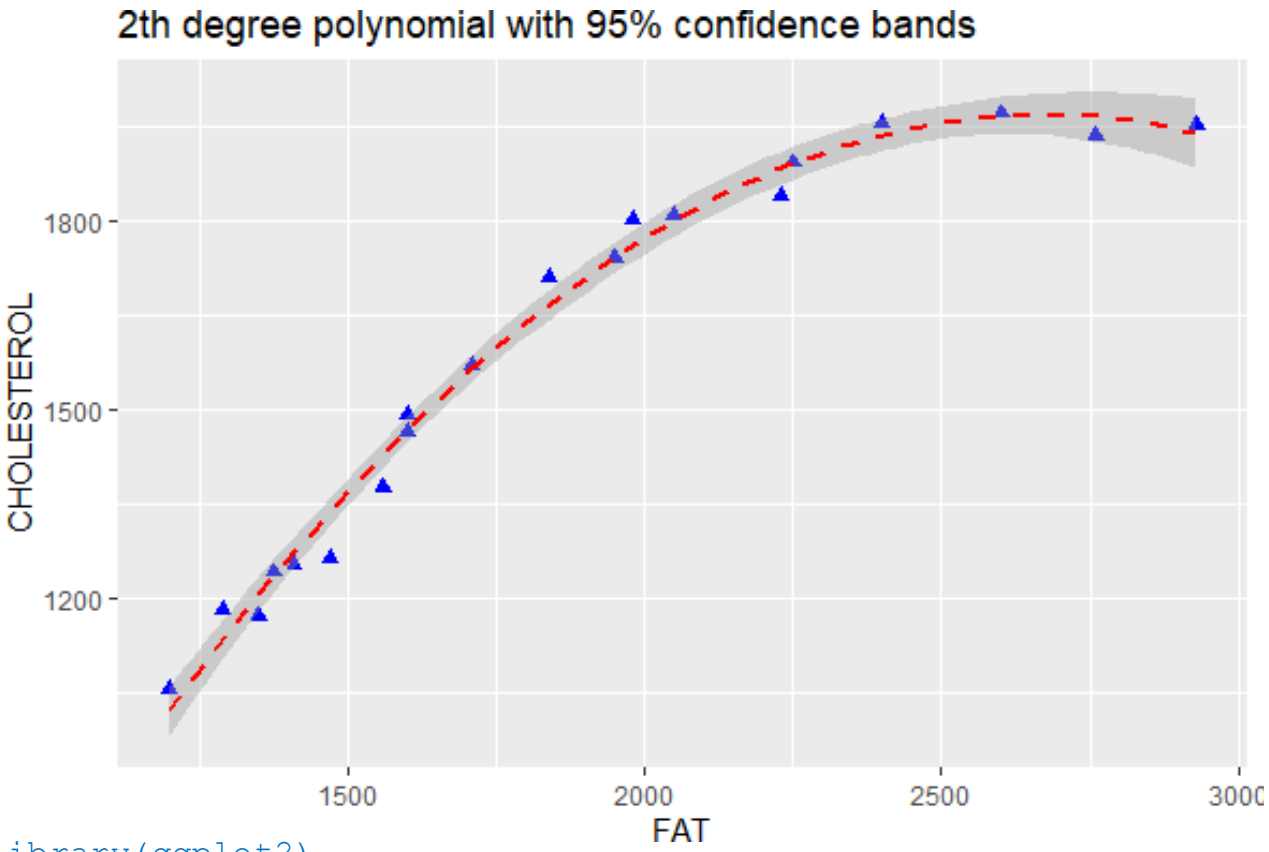
```
Analysis of Variance Table
Response: CHOLESTEROL
           Df  Sum Sq Mean Sq F value  Pr(>F)
FAT         1 1617913 1617913    1300 < 2e-16
I(FAT^2)    1  216817  216817     174 2.3e-10
Residuals 17   21163    1245

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.16e+03   1.31e+02   -8.88  8.6e-08
FAT          2.34e+00   1.35e-01   17.31  3.1e-12
I(FAT^2)    -4.39e-04   3.33e-05  -13.20  2.3e-10

Residual standard error: 35.3 on 17 degrees of freedom
Multiple R-squared:  0.989,   Adjusted R-squared:  0.987
F-statistic:  737 on 2 and 17 DF,  p-value: <2e-16
```

Plots of residuals: Quadratic model





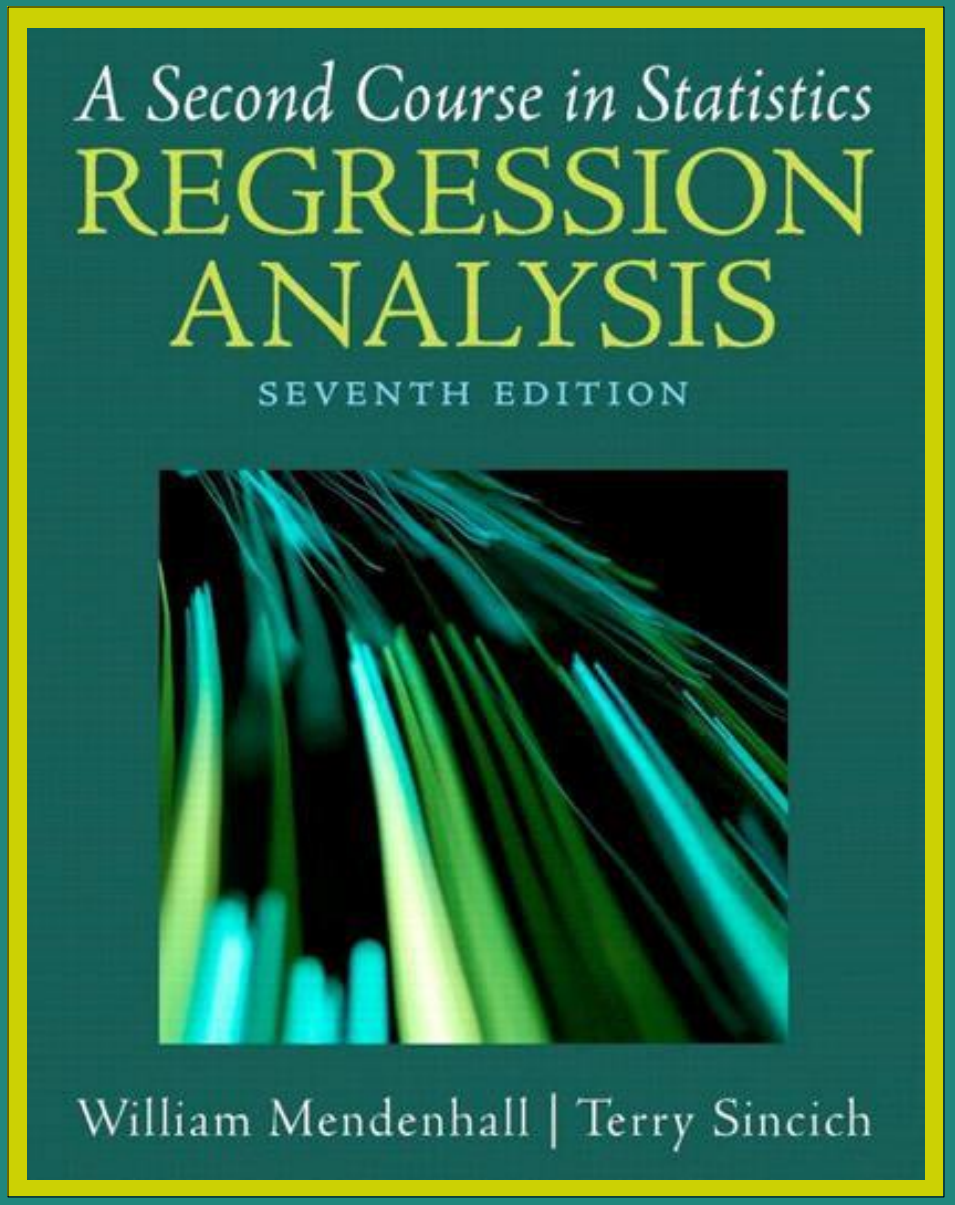
Quadratic model




```
library(ggplot2)
par(mfrow =c(1,1))
ggplot(data=oly, aes(x=FAT, y=CHOLESTEROL)) +
  geom_point(pch=17, color="blue", size=2) +
  geom_smooth( method = "lm", formula = y ~ poly(x, 2), color="red",
linetype=2) +
  labs(title="2th degree polynomial with 95% confidence bands",
       x="FAT", y="CHOLESTEROL")
```

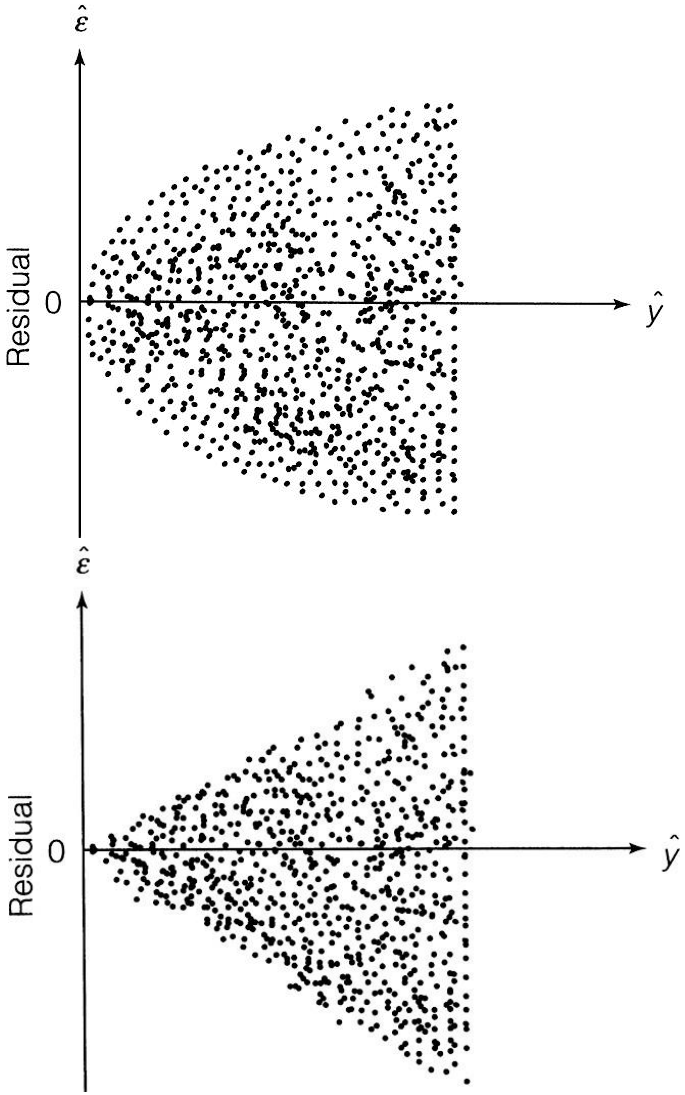
Section 8.4

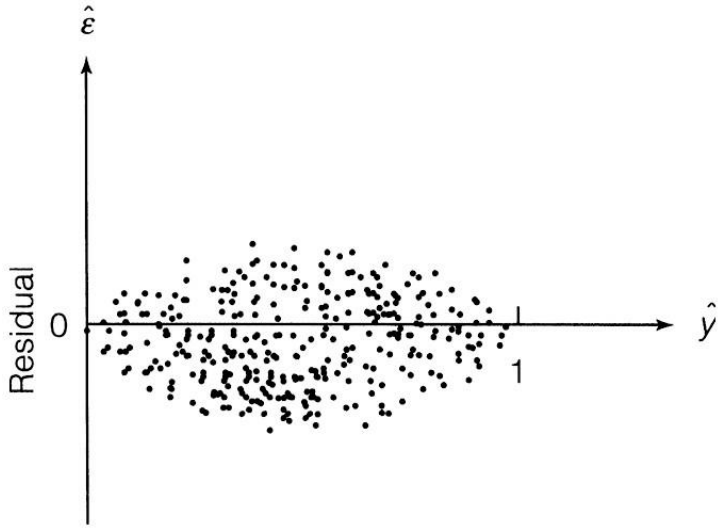
Detecting Unequal Variances

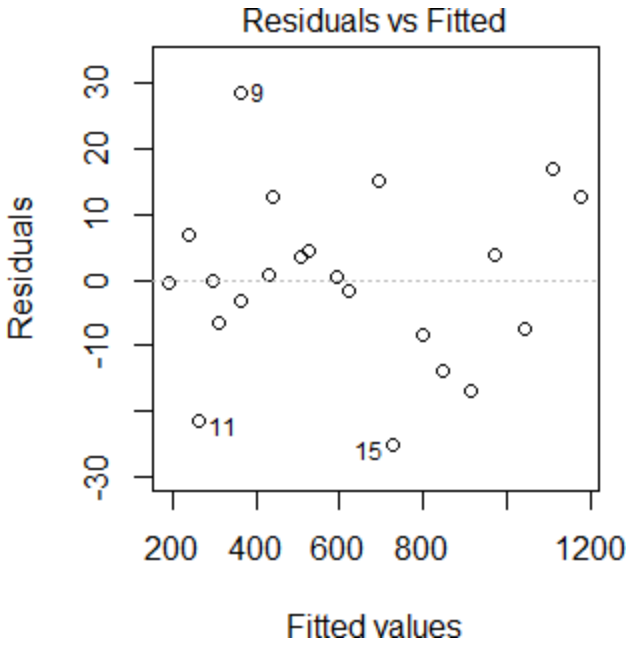


PEARSON

plot of residuals vs fitted values

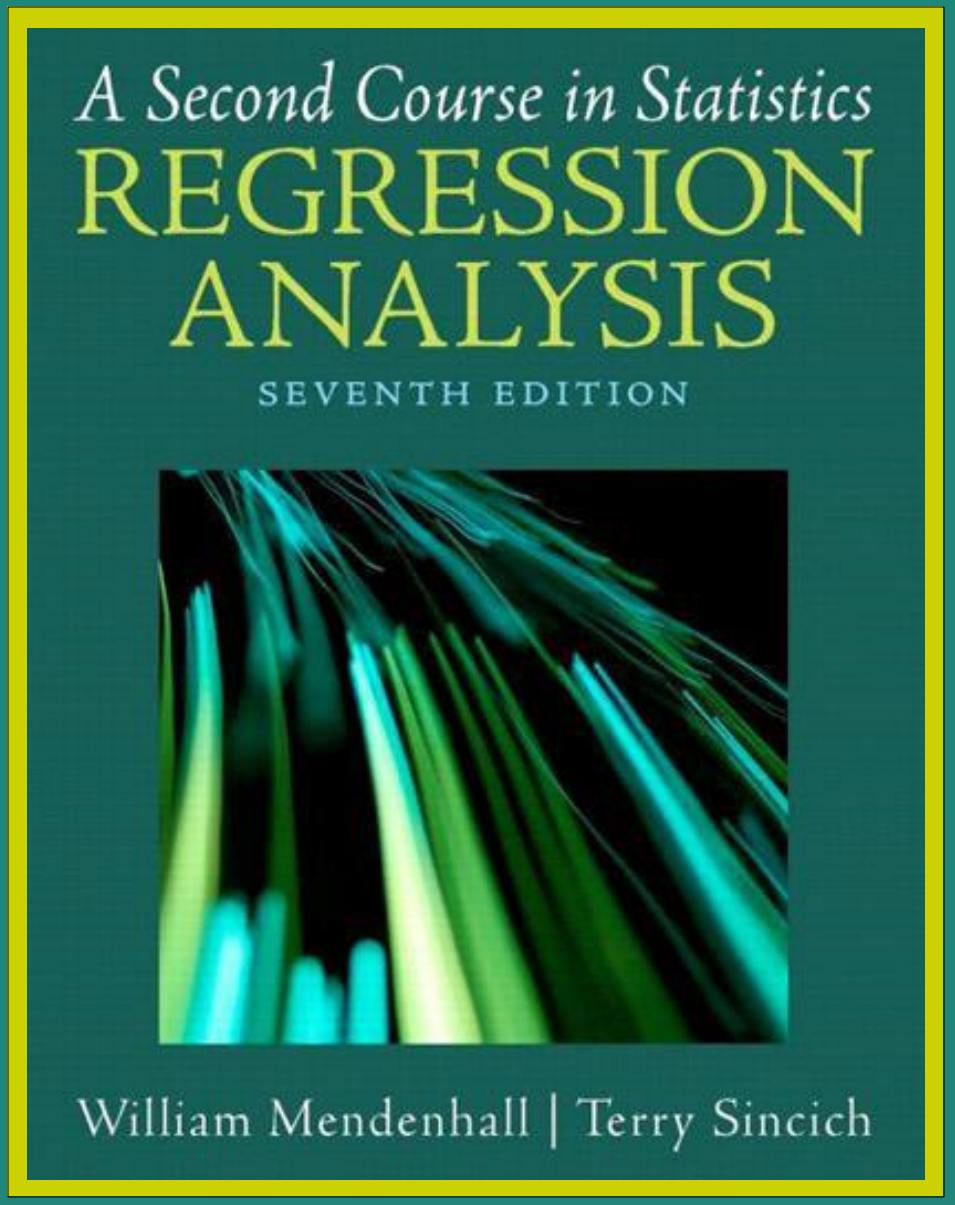






Section 8.5

Checking the Normality Assumption



PEARSON

Assumption of normally distributed residuals

- This is the least restrictive of the assumptions.
- Regression is robust to departures from normality
- Inferences derived from the analysis remain valid even if the assumption is not exactly satisfied
- Moderate departures have little impact on CIs for the model parameters

Formal test

```
shapiro.test(model$residuals)
```

Limitations:

- Formal tests only useful for relatively small samples.
- Tests have low ***power*** *i.e.* the probability of detecting a non-normal error distribution when it exists is low

Power Transformations

How can we address the problem of violated model assumptions?

One approach is to transform the response variable.

Example

The following table gives the counts of numbers of poppies grown under six different treatment regimes. We will analyse the data and check the assumptions.

	1	2	3	4	mean	sd
A	538	422	377	315	413	94
B	438	442	319	380	395	58
C	77	61	157	52	87	48
D	115	57	100	45	79	34
E	17	31	87	16	38	34
F	18	26	77	20	35	28

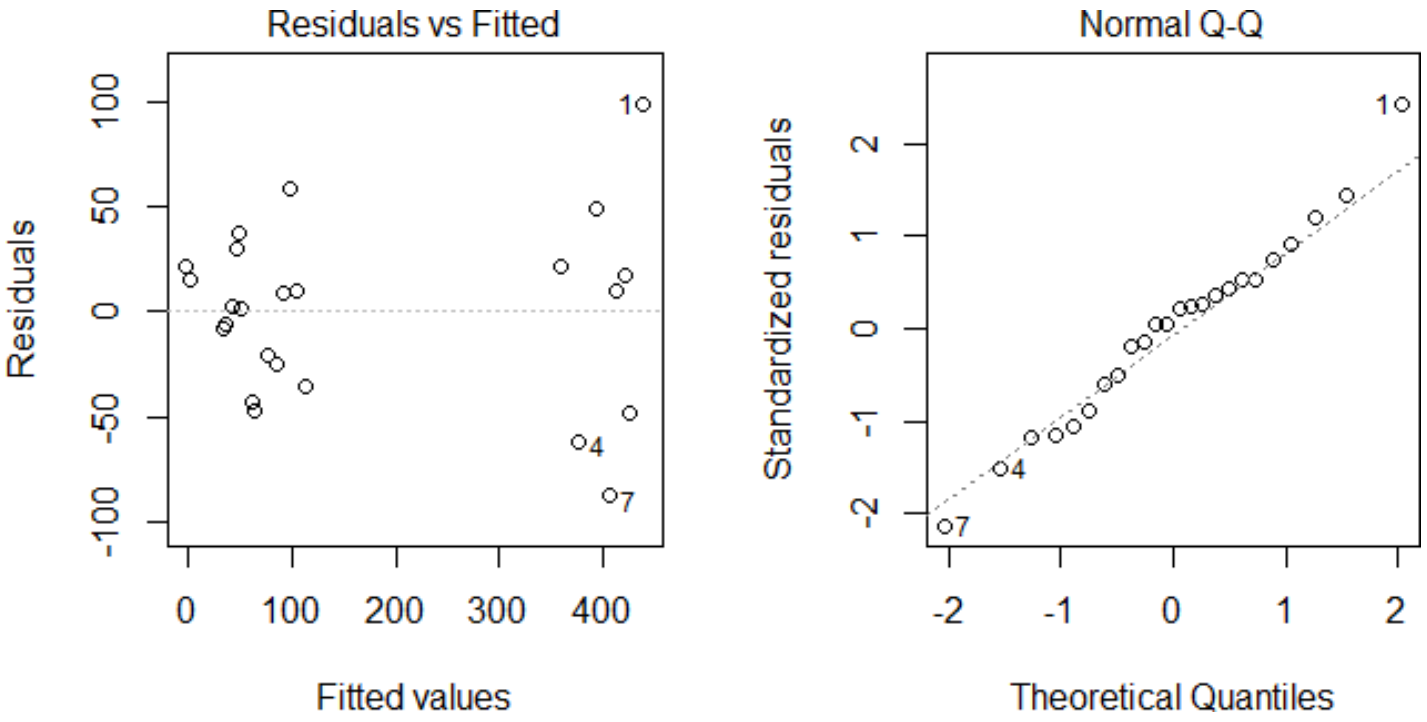
Example

```
## Create data frame

poppies <- expand.grid(Block = 1:4,
                       Treatment = LETTERS[1:6])
poppies$Block <- factor(poppies$Block)
poppies$count <- c(538, 422, 377, 315,
                   438, 442, 319, 380,
                   77, 61, 157, 52,
                   115, 57, 100, 45,
                   17, 31, 87, 16,
                   18, 26, 77, 20)
## Fit a MLR model
mod1 <- lm(count ~ Block + Treatment, data = poppies)
summary(mod1)
```

Example

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    439.04      31.54  13.919 5.55e-10 ***
Block2         -27.33      29.74  -0.919   0.3726
Block3         -14.33      29.74  -0.482   0.6368
Block4         -62.50      29.74  -2.102   0.0529 .
TreatmentB     -18.25      36.42  -0.501   0.6236
TreatmentC    -326.25      36.42  -8.957 2.08e-07 ***
TreatmentD    -333.75      36.42  -9.163 1.56e-07 ***
TreatmentE    -375.25      36.42 -10.303 3.37e-08 ***
TreatmentF    -377.75      36.42 -10.371 3.09e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51.51 on 15 degrees of freedom
Multiple R-squared:  0.9426,  Adjusted R-squared:  0.912
F-statistic: 30.81 on 8 and 15 DF,  p-value: 5.927e-08
```

Example




```
par(mfrow =c(1,2))
plot(mod1,which=1:2,add.smooth = F)
```

Transformations

- Transform the data to a different scale and carry out the analysis in the new scale.
- Transformations often can be used to solve problems of non-homogeneity of variance and non-normal error distribution.
- Power transformations can make a skewed distribution more symmetric and stabilize the spread of the data.

Power Transformations

The most commonly used power transformations ($Y' = Y^{\lambda}$) in *order of increasing strength* are:

Transformation	Y'	λ	Back Transform
No transform	Y	1	
Square root	$\sqrt{Y}$	0.5	Y^2
Log (Natural log)	$\log_e Y$	0 (by definition)	e^Y
Reciprocal	$1/Y$	-1	$1/Y$

Choosing a transformation

The most appropriate transformation is that for which the data

- Shows no sign of a variance-mean relationship
- Shows least difference between the group variances as expressed by

$$F_{max} = \frac{largest\ group\ variance}{smallest\ group\ variance}$$

Comparing transforms

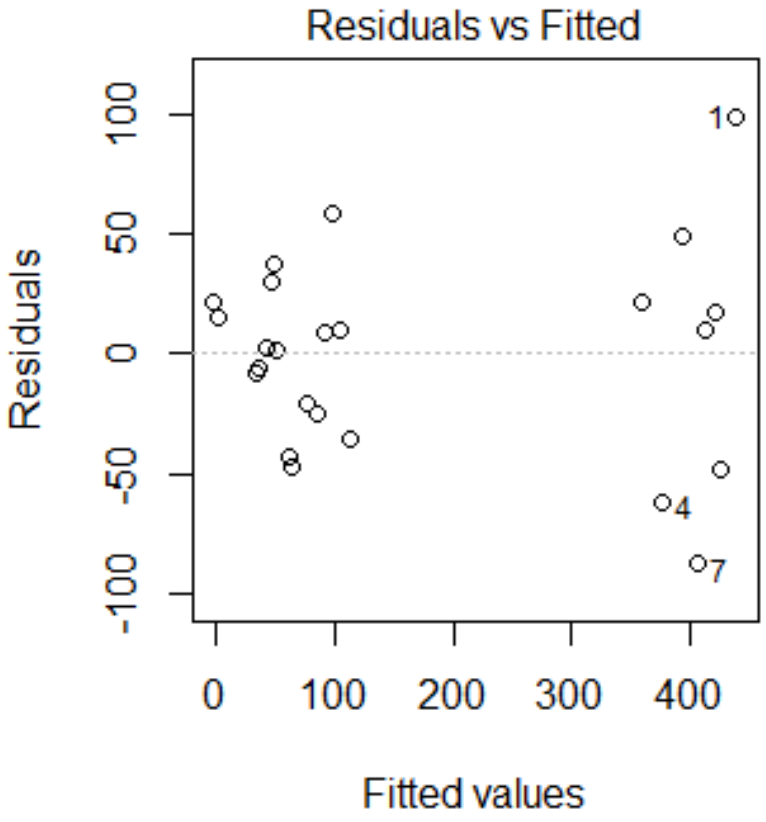
Treat	Y	var	$Y^{1/2}$	var	$\log_e(Y)$	var
A	413.00	8869	20.23	5.24	6.00	0.05
B	394.75	3353	19.83	2.21	5.97	0.02
C	86.75	2300	9.08	5.70	4.37	0.24
D	79.25	1126	8.75	3.69	4.30	0.20
E	37.75	1125	5.75	6.18	3.38	0.62
F	35.25	786	5.65	4.48	3.37	0.44
F_{max}		11.31		2.81		31.00

Q What does this table suggest?

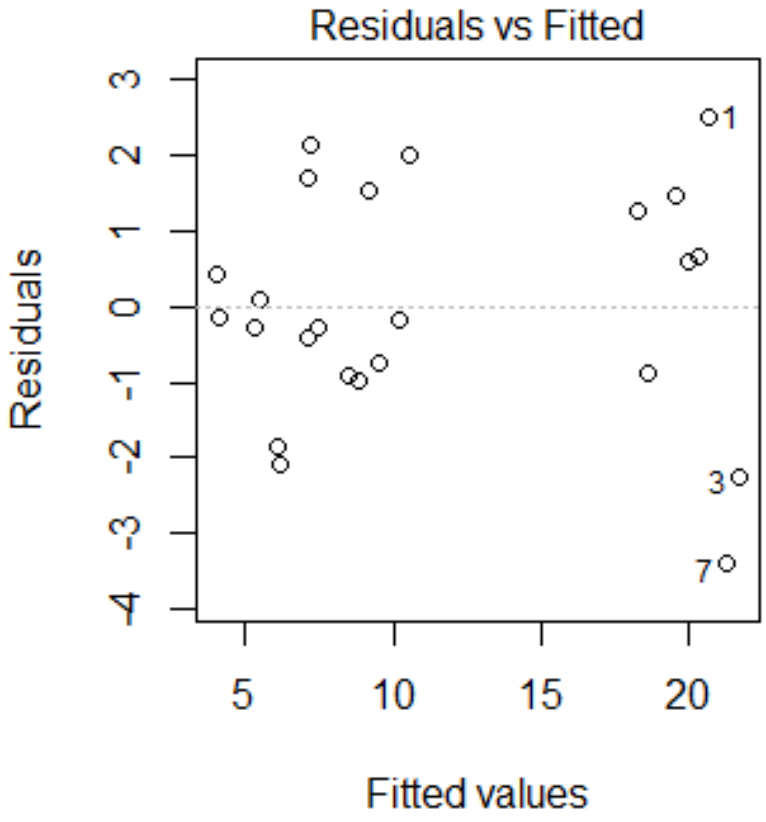
Square root transform

```
mod2<-lm(sqrt(count)~ Block + Treatment, data = poppies)
anova(mod2)
#####
Analysis of Variance Table
Response: sqrt(count)
           Df Sum Sq Mean Sq F value     Pr(>F)
Block       3  29.62   9.872  2.8014    0.07578 .
Treatment   5 904.62 180.925 51.3405 6.628e-09 ***
Residuals 15  52.86   3.524
```

Compare diagnostics



count ~ Block + Treatment



sqrt(count) ~ Block + Treatment

Square root transform: Block 1 means

sqrt(count) ~Treatment + Block

```
mod2a<-lm(sqrt(count)~Treatment+Block-1,data=poppies)
confint(mod2a)
```

```
            Estimate Std. Error 2.5 % 97.5 %
TreatmentA    20.677       1.15 18.23 23.127
TreatmentB    20.278       1.15 17.83 22.728
TreatmentC     9.533       1.15  7.08 11.983
TreatmentD     9.197       1.15  6.75 11.647
TreatmentE     6.206       1.15  3.76  8.656
TreatmentF     6.098       1.15  3.65  8.549
```

Back transformations

- Tests, inference and CIs are made in the transformed scale.
- To express the results in the original scale of measurement make a second transformation

  ➔ transforming the results back to the original scale.

- This process is known as back transformation

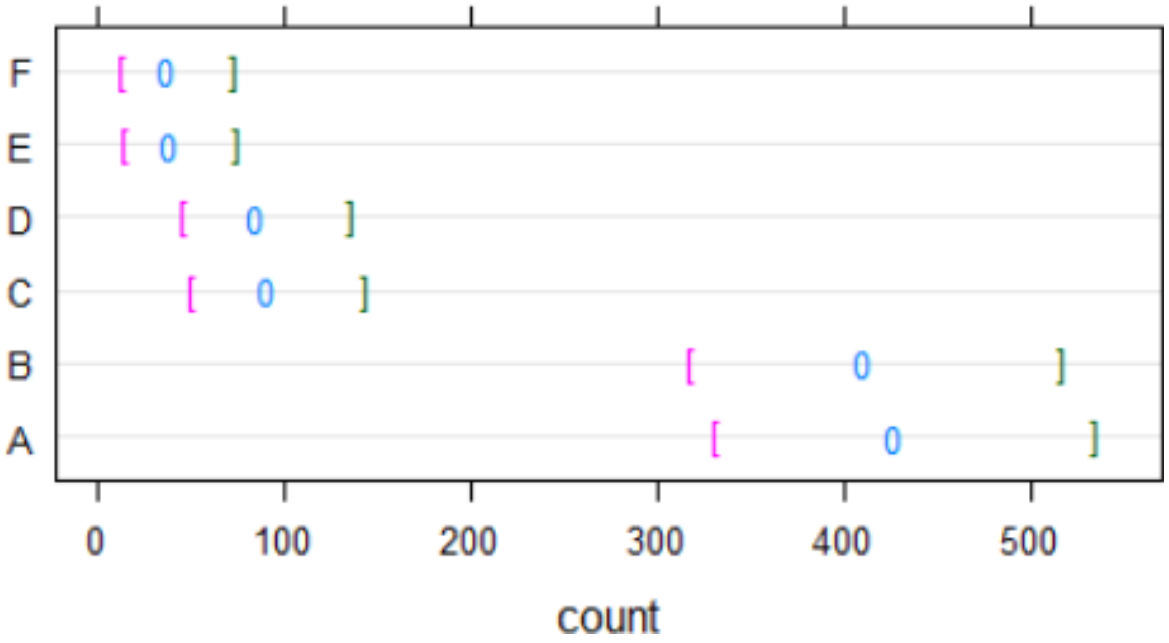
Back transformations

```
## back transforming
pred.df <- data.frame(Block=1,Treatment=LETTERS[1:6])
pred.df$Block <- factor(pred.df$Block)
preds <- predict(mod2,newdata=pred.df,interval="confidence")
pred.df <- cbind(pred.df,preds)

#square results (Backtransform of sqrt)
pred.df$backT_fit <- pred.df$fit^2
pred.df$backT_lwr <- pred.df$lwr^2
pred.df$backT_upr <- pred.df$upr^2
```

Backtransformations

```
> pred.df
```

	Block	Treatment	fit	lwr	upr	backT_fit	backT_lwr	backT_upr
1	1	A	20.68	18.23	23.13	427.5	332.2	534.9
2	1	B	20.28	17.83	22.73	411.2	317.8	516.6
3	1	C	9.53	7.08	11.98	90.9	50.2	143.6
4	1	D	9.20	6.75	11.65	84.6	45.5	135.6
5	1	E	6.21	3.76	8.66	38.5	14.1	74.9
6	1	F	6.10	3.65	8.55	37.2	13.3	73.1



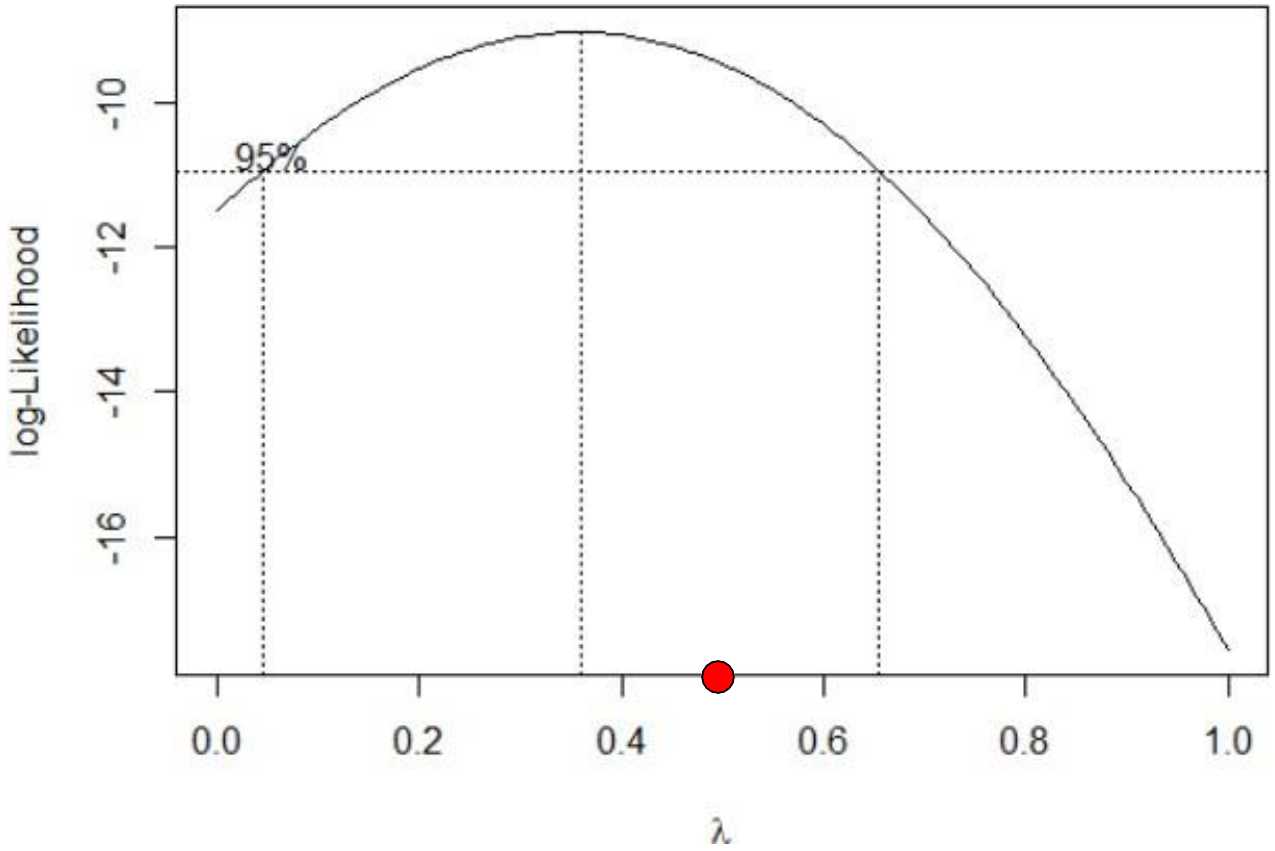
The Box – Cox procedure

The Box-Cox procedure offers a method of finding a suitable value of λ for a power transformation of the type Y^{λ}.

Example: Estimate the value of λ needed to transform the poppy count data

```
### Box-Cox procedure

install.packages("MASS")
library(MASS)

par(mfrow =c(1,1))
boxcox(count~Block+Treatment, data = poppies,
       lambda=seq(from=0,to=1, by=0.01))
```

The Box – Cox procedure



$\lambda = 0.5$ lies inside the 95% CI

Lecture 1 recap:

- Review assumptions of linear model
- Detect lack of fit using residual plots
- Power transformations – box-cox procedure

Lecture 2:

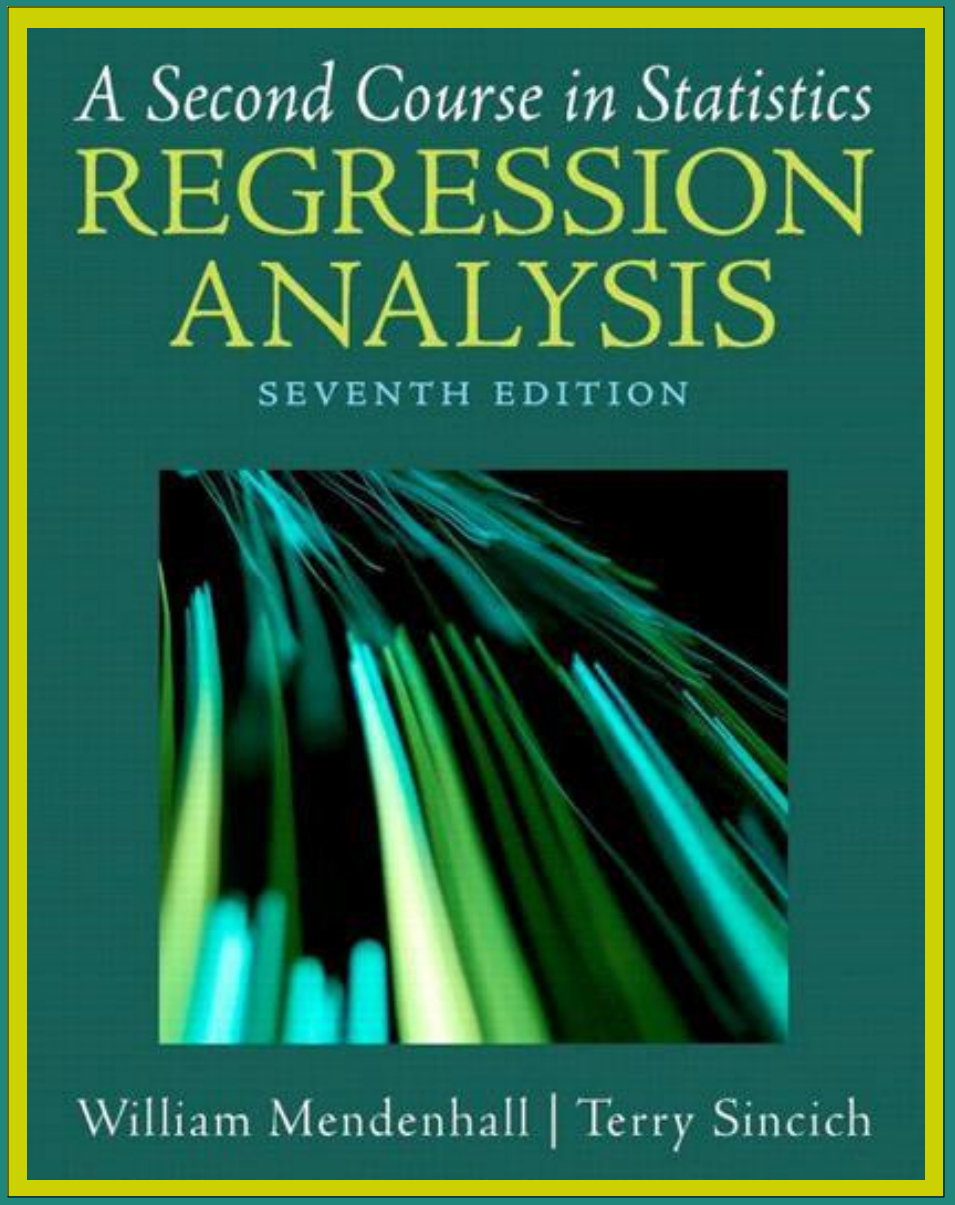
- Detect outliers and influential points

Chapter 8

Residual Analysis

(Lecture 2)

Dr Brenda Vo



PEARSON

Chapter 8 outline

Lecture 1:

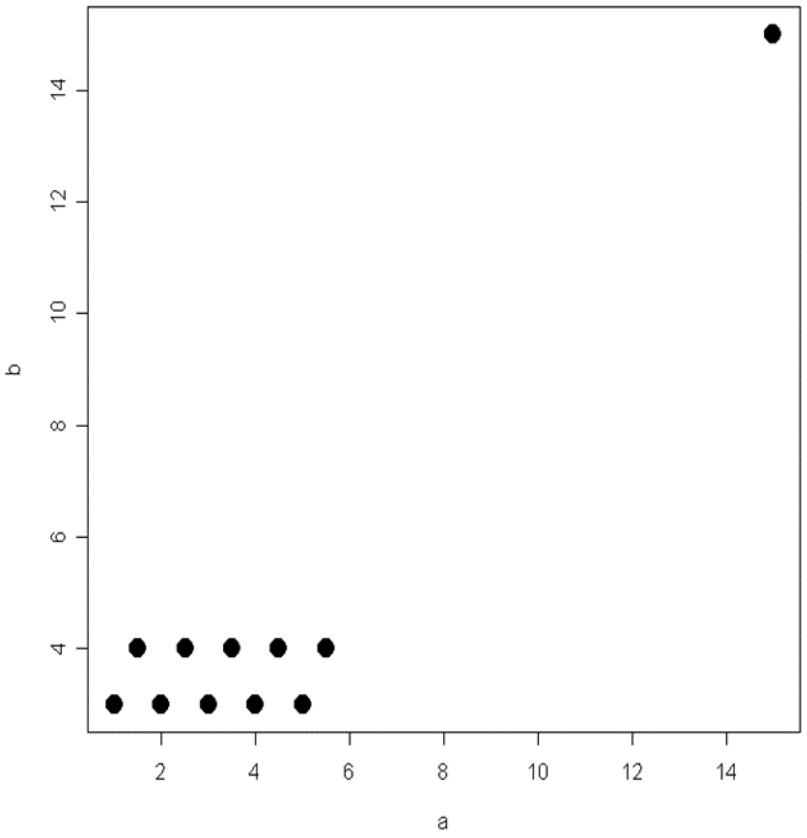
- Review assumptions of linear model
- Detect lack of fit
- Power transformations

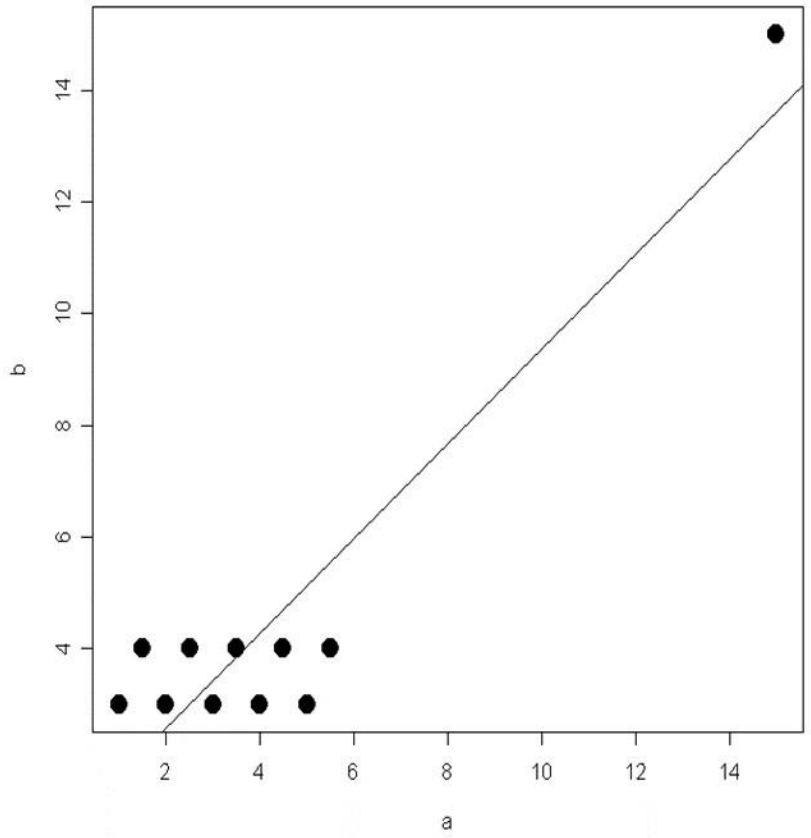
Lecture 2:

- Detect outliers and influential points (section 8.6)

Q: How would you describe the relationship between a and b?

Q: What would the line of best fit look like?





Standardised Residuals

The standardised residual for the i^{th} observation

$$z_i = \frac{\hat{\epsilon}_i}{s} = \frac{y_i - \hat{y}_i}{s}$$

Note that this is simply the z-score for the residual. Z~N(0,1)

An observation whose absolute value of the standardised residual is

- greater than 2 is a potential outlier.
- greater than 3 is an outlier

Influential points, leverage and Cook's distance

Influential points

- have a large influence on the regression analysis
- their inclusion or exclusion changes the model, the values of the regression parameters, the predictions etc.

Identify influential points using

- Leverage
- Cook's distance

Life Cycle Savings

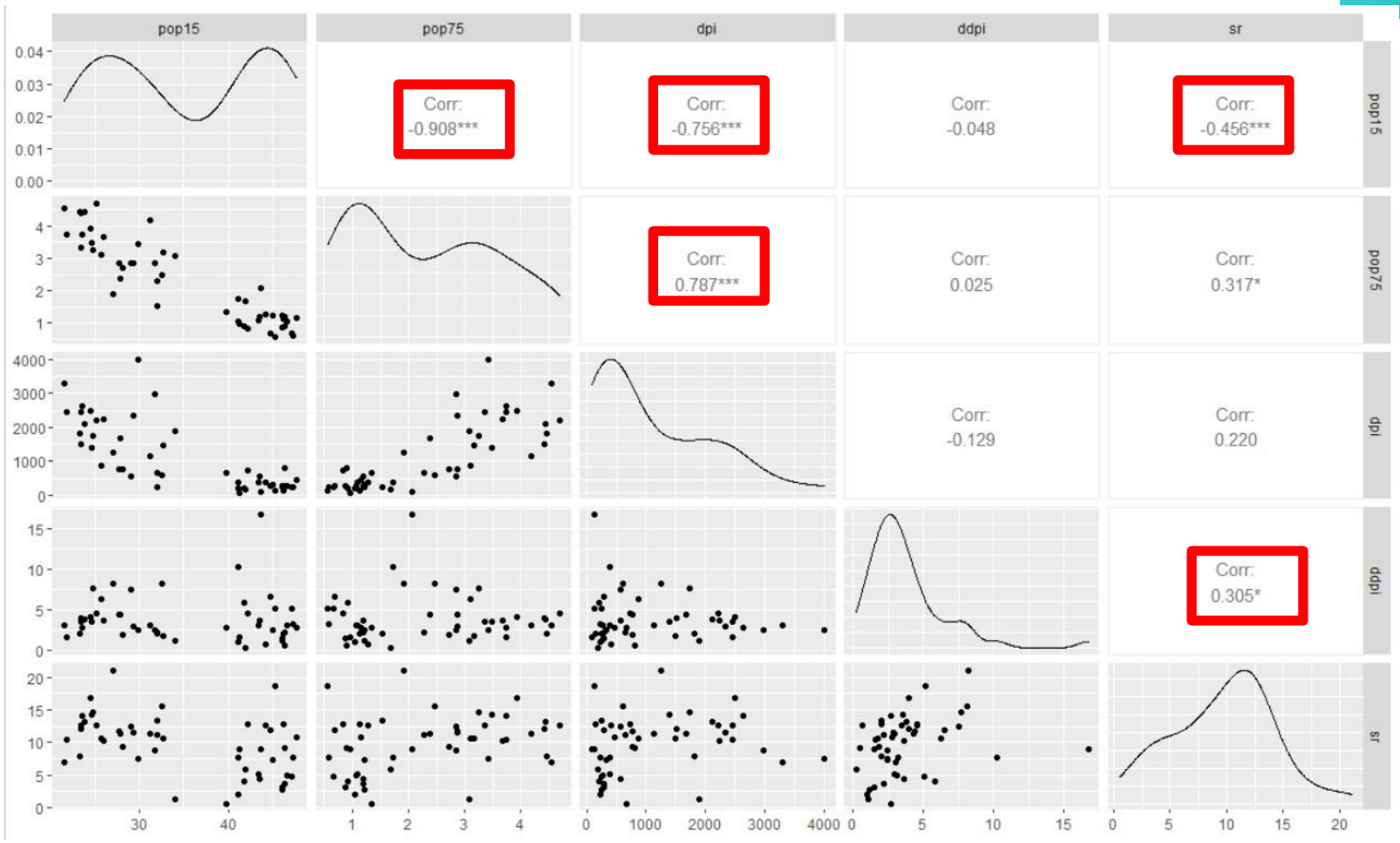
Data on the savings ratio 1960–1970:
50 observations (countries) on 5 variables.

sr	numeric	aggregate personal savings
pop15	numeric	% of population under 15
pop75	numeric	% of population over 75
dpi	numeric	real per-capita disposable income
ddpi	numeric	% growth rate of dpi

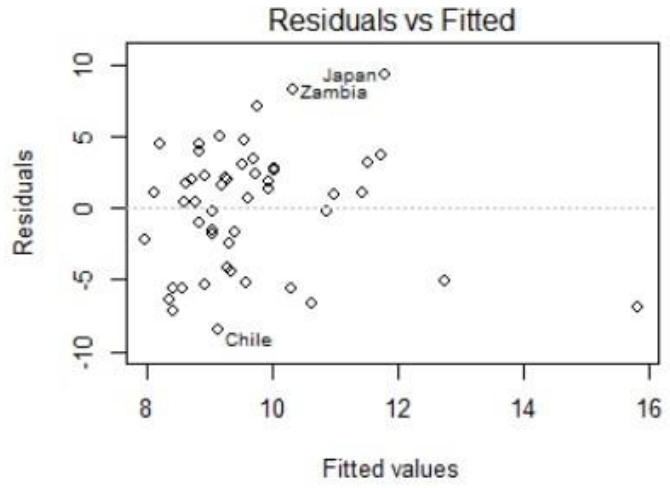
Data set available in R: **`data("LifeCycleSavings")`**

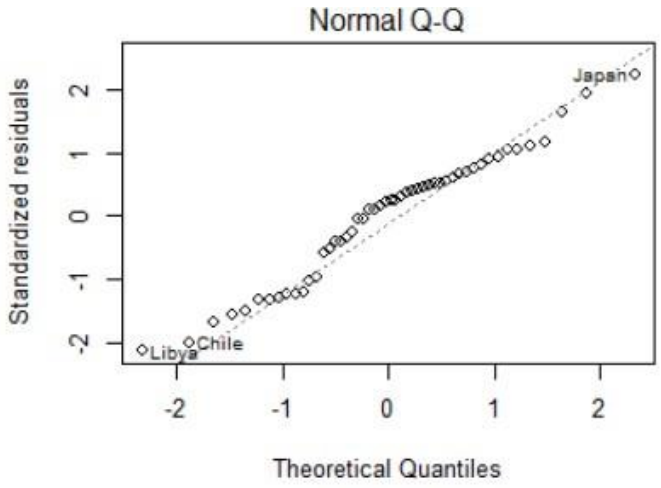
```
library(Ggally)
ggpairs(LifeCycleSavings[,c(2,3,4,5,1))
```

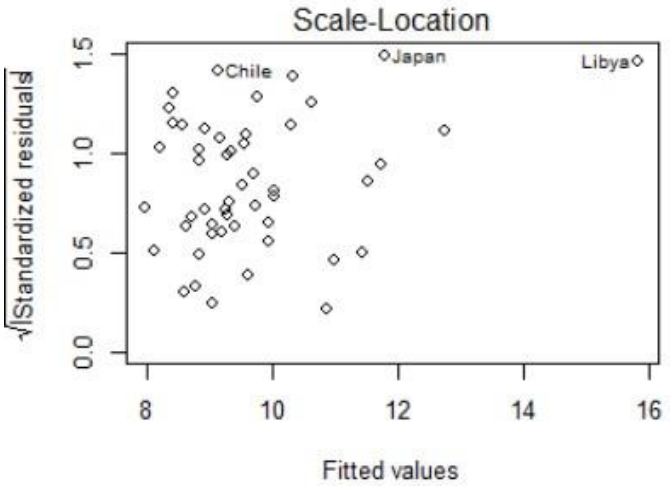
LifeCycleSavings pairs plot

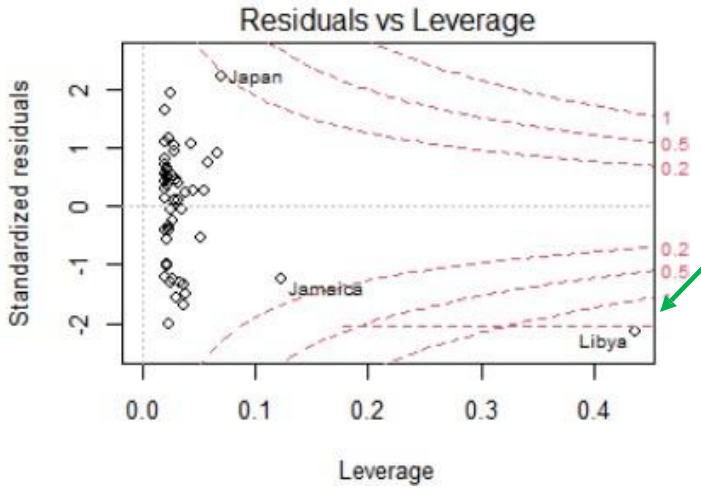


```
mod1 <- lm(sr ~ ddpi, data = LifeCycleSavings)
summary(mod1)
par(mfrow = c(2, 2))
plot(mod1, add.smooth=F, cook.levels=c(0.2,0.5,1.0))
```



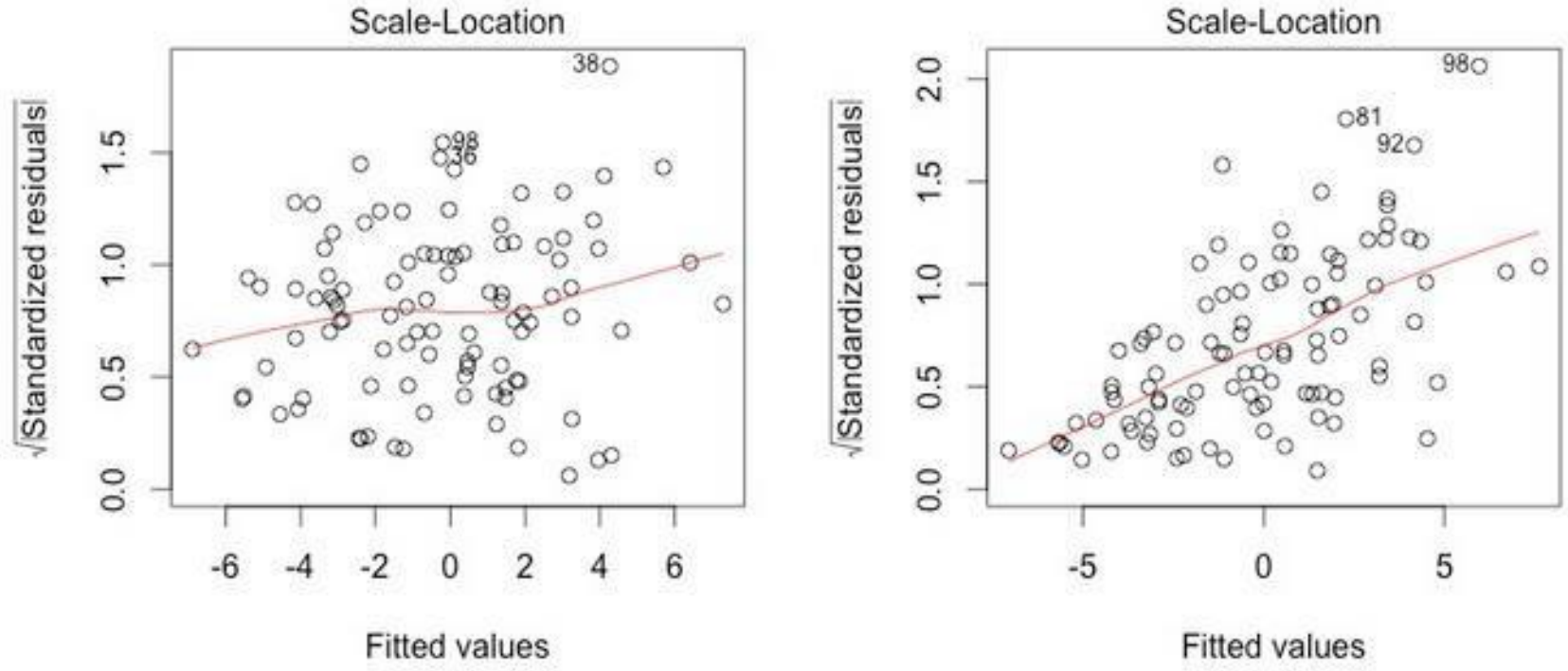






Influential point

Scale-Location plot



- Check the assumption of equal variance (homoscedasticity)
- Check if residuals are spread equally along the x-axis.

Influential Points: Leverage

- The leverage for i^{th} observation is denoted as h_i
- The leverage h_i measures the influence of y_i on its predicted value $\hat{y}_i$

Rule of Thumb for Detecting Influence with Leverage

The observed value of y_i is influential if

$$h_i > \frac{2(k+1)}{n}$$

where h_i is the leverage for the ith observation and k = the number of β's in the model (excluding β_0).

For Life Savings example

$$\frac{2(k+1)}{n} = \frac{2\,(1+1)}{50} = 0.08$$

Libya has $h_i \approx 0.42 > 0.08$ → Libya is an influential point

Influential Points: Cook's Distance

Cook's distance, D_i, of an observation i is a measure of *the overall influence* of that observation on the estimated regression parameters β

$$D_i = \frac{(y_i - \hat{y}_{(i)})^2}{(k+1)MSE}\left[\frac{h_i}{(1-h_i)^2}\right] \qquad i = 1, \ldots, n$$

D_i depends on

- The residual $(y_i - \hat{y}_{(i)})$
- The leverage h_i

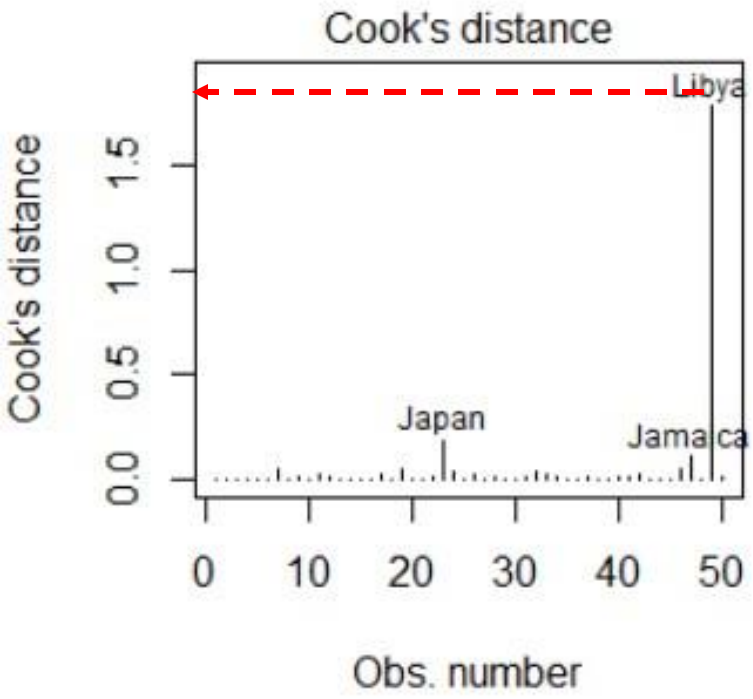
➢ Large value of D_i indicates that the observed value i^{th} has strong influence on the estimated β coefficients.

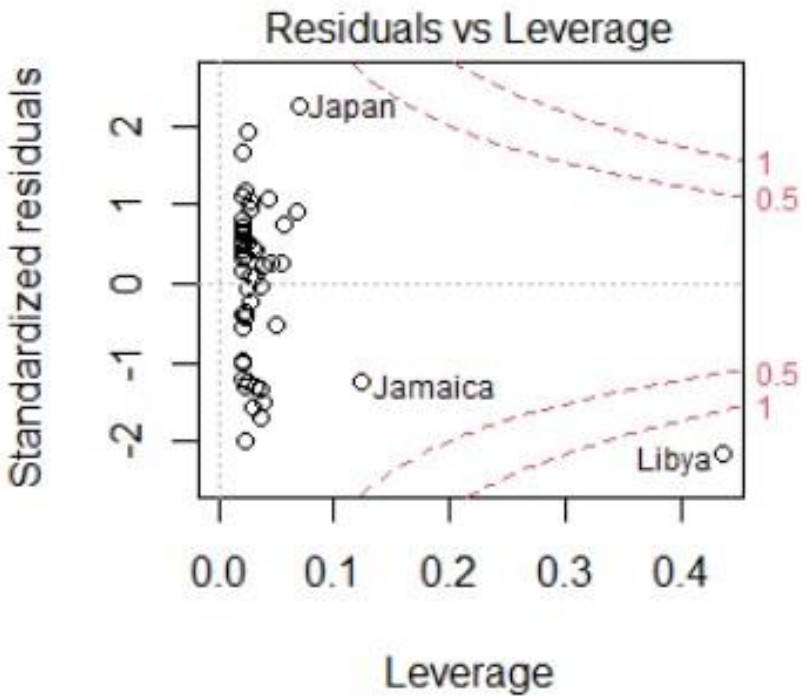
Influential Points: Cook's Distance

- $D_i \sim F_{k+1,n-(k+1)}$ where k is the number of estimated parameters
- If D_i lies at or *above the 50th percentile* of the *F distribution*, the ith observation is considered to have a major influence
- if D_i is below the *20th* percentile then it has little influence

Cook's distance (D_i) and leverage (h_i)

```
par(mfrow = c(1, 2))
plot(mod1,  which=4:5, add.smooth = F)
```





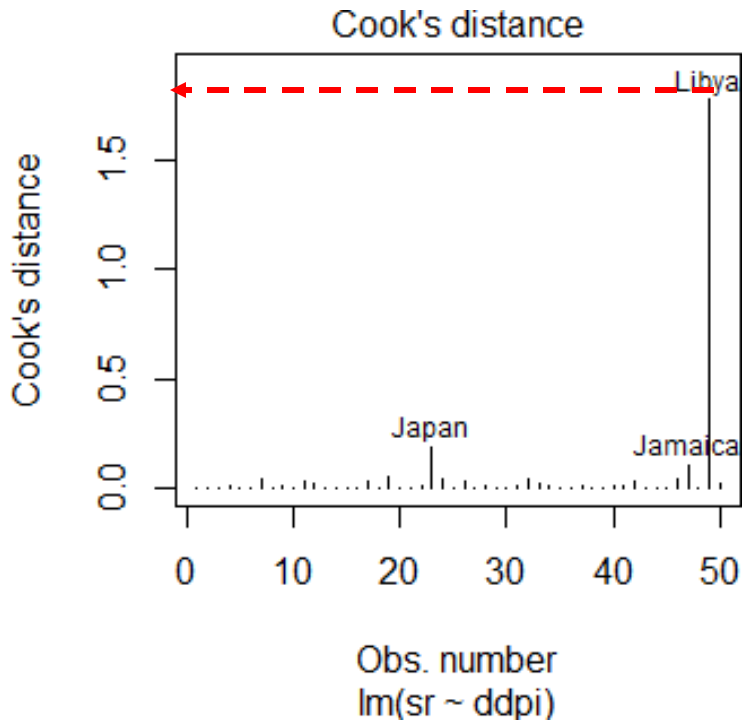
Libya has a Cook's distance that lies outside the red lines
–> highly influential

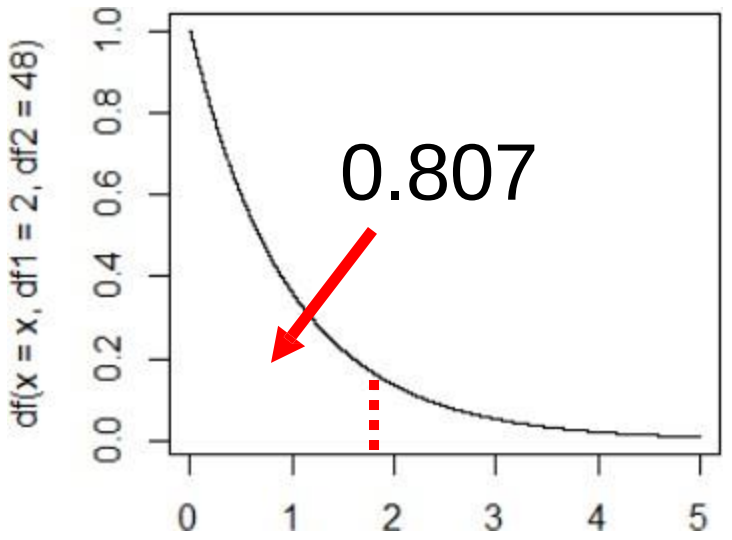
Cook's distance (D_i)

Calculate F percentile

```
> pf(1.7, 2, 48)
[1] 0.807
```

i.e the Cook's distance of 1.7 lies above the 80^{th} of the F distribution.

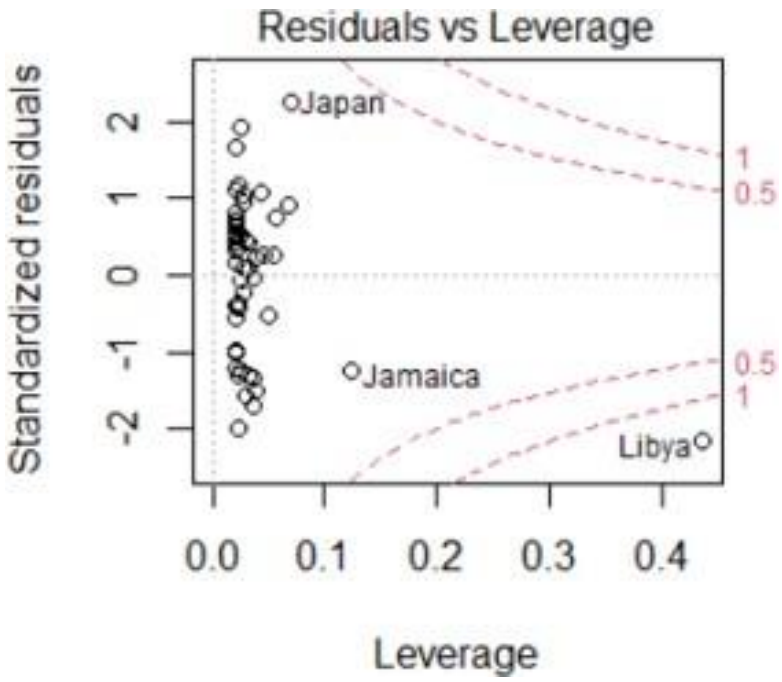




F distribution with 2 and 48 df

Residuals vs leverage plot

- Patterns are not relevant
- Look at observations at the upper right corner and at the lower right corner
- Look for cases outside of the red dashed lines
  - ➢ High Cook's distance scores

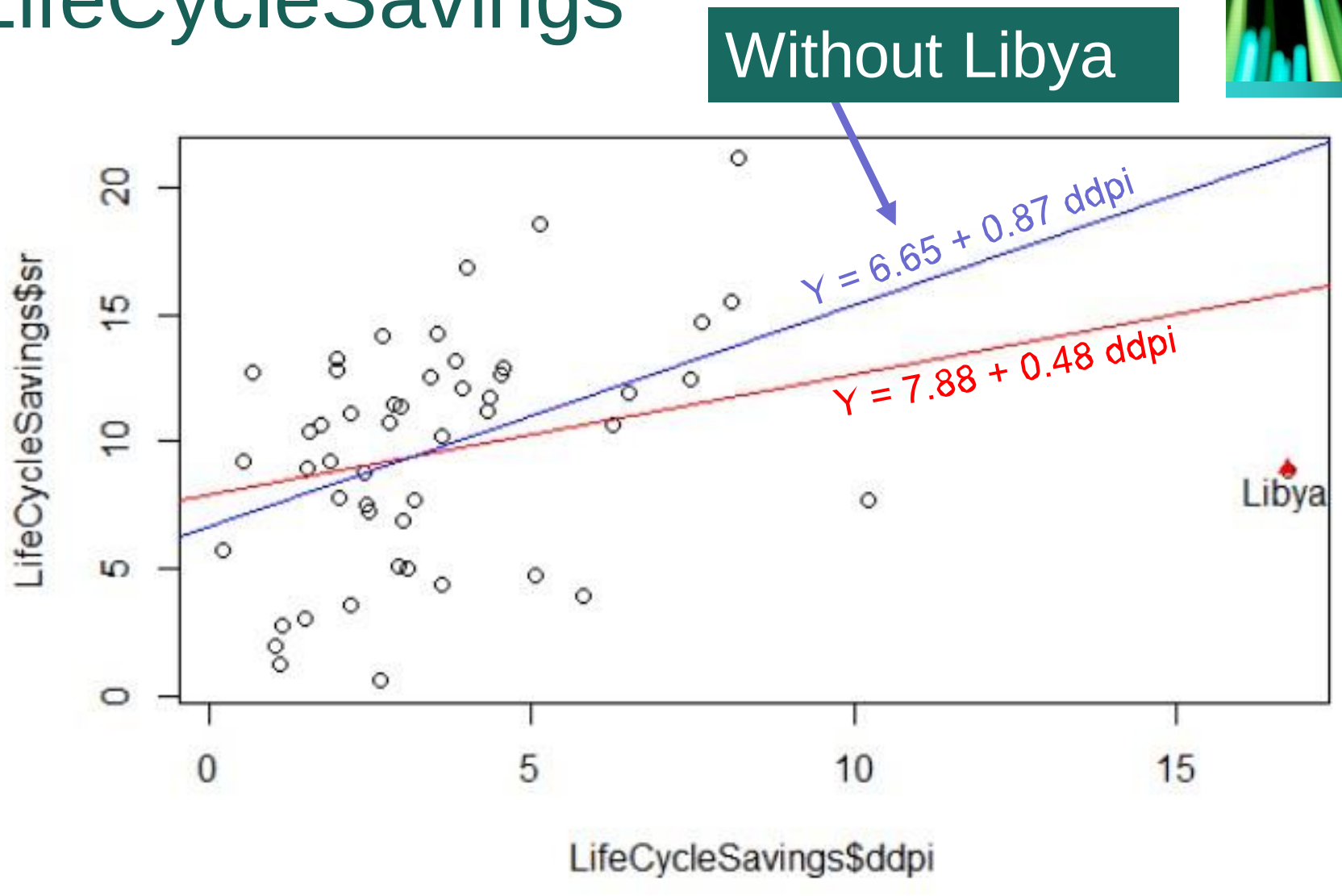


Cook's distance

```
install.packages("car")
library(car)
sort(cooks.distance(mod1))
```

```
      Greece       Canada        India      Uruguay       Norway
    4.19e-05     4.28e-05     1.36e-04     1.75e-04     2.33e-04
United Kingdom      China      Finland United States   Costa Rica
    8.31e-04     9.54e-04     9.87e-04     1.52e-03     1.53e-03

   Guatamala      Tunisia        Korea     Paraguay        Chile
    2.94e-02     3.39e-02     3.87e-02     4.51e-02     4.73e-02
      Zambia      Iceland      Jamaica        Japan        Libya
    4.74e-02     5.52e-02     1.10e-01     1.86e-01     1.77e+00
```

LifeCycleSavings



Review

- What are the four indicators of multicollinearity?
- State the assumptions of the linear model.
- List the 5 diagnostic (residuals) plots and explain what information they provide about the model.
- Why would you consider transforming the response variable?
- List the common power transformations discussed in lectures. What value of lambda corresponds to each?

Review

- What are the four indicators of multicollinearity?

1. Significant pairwise correlations for the I.V.s
2. Non-significant t-tests for many individual parameters
3. Opposite sign to expected for estimated parameters
4. Variance Inflation Factor, VIF>10

Review

- State the assumptions of the linear model.

  Residuals, $\varepsilon \sim N(0, \sigma^2)$, where σ^2 is constant, and residuals are independent.

- List the 5 diagnostic (residuals) plots and explain what information they provide about the model.

  1. Residuals vs fitted: check assumption of constant variance, and residuals are randomly centred around 0
  2. Normal QQ plot: check assumption of normality and also for potential outliers.
  3. Scale location plot: check constant variance assumption when n is small.
  4. Residuals vs leverage plot: check for influential points.
  5. Cook's distance plot: check for influential points.

Review

- Why would you consider transforming the response variable?
  - If the model assumptions have been violated.

- List the common power transformations discussed in lectures. What value of lambda corresponds to each?
  1. Square root transformation, $\lambda = 0.5$
  2. Log transformation, $\lambda = 0$, by definition
  3. Reciprocal, $\lambda = -1$