Applied Statistical Regression
AS 2013 – Week 08

Marcel Dettling

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

ETH Zürich, November, 2012

Applied Statistical Regression
AS 2013 – Week 08

Residual Analysis – Model Diagnostics

Why do it? And what is it good for?

a) To make sure that estimates and inference are valid

- $E[E_i] = 0$
- $Var(E_i) = \sigma_E^2$
- $Cov(E_i, E_j) = 0$
- $E_i \sim N(0, \sigma_E^2 I),\ i.i.d$

b) Identifying unusual observations

Often, there are just a few observations which "are not in accordance" with a model. However, these few can have strong impact on model choice, estimates and fit.

Applied Statistical Regression
AS 2013 – Week 08

Residual Analysis – Model Diagnostics

Why do it? And what is it good for?

c) Improving the model

- Transformations of predictors and response
- Identifying further predictors or interaction terms
- Applying more general regression models

- There are both model diagnostic graphics, as well as numerical summaries. The latter require little intuition and can be easier to interpret.

- However, the graphical methods are far more powerful and flexible, and are thus to be preferred!

Applied Statistical Regression
AS 2013 – Week 08

Residuals vs. Errors

All requirements that we made were for the errors E_i. However, they cannot be observed in practice. All that we are left with are the residuals r_i.

But:

- the residuals r_i are only estimates of the errors E_i, and while they share some properties, others are different.
- in particular, even if the errors E_i are uncorrelated with constant variance, the residuals r_i are not: they are correlated and have non-constant variance.
- does residual analysis make sense?

Applied Statistical Regression
AS 2013 – Week 08

Standardized/Studentized Residuals

Does residual analysis make sense?

- the effect of correlation and non-constant variance in the residuals can usually be neglected. Thus, residual analysis using raw residuals r_i is both useful and sensible.

- The residuals can be corrected, such that they have constant variance. We then speak of standardized, resp. studentized residuals.

$$\tilde{r}_i = \frac{r_i}{\hat{\sigma}_E \cdot \sqrt{1-h_{ii}}}, \text{ where } Var(\tilde{r}_i) = 1 \text{ and } Cor(\tilde{r}_i, \tilde{r}_j) \text{ is small.}$$

- R uses these $\tilde{r}_i$ for the Normal Plot, the Scale-Location-Plot and the Leverage-Plot.

Toolbox for Model Diagnostics

There are 4 "standard plots" in R:

- Residuals vs. Fitted, i.e. Tukey-Anscombe-Plot
- Normal Plot
- Scale-Location-Plot
- Leverage-Plot

Some further tricks and ideas:

- Residuals vs. predictors
- Partial residual plots
- Residuals vs. other, arbitrary variables
- Important: Residuals vs. time/sequence

Applied Statistical Regression
AS 2013 – Week 08

Example in Model Diagnostics

Under the life-cycle savings hypothesis, the savings ratio (aggregate personal saving divided by disposable income) is explained by the following variables:

```
lm(sr ~ pop15 + pop75 + dpi + ddpi, data=LifeCycleSavings)
```

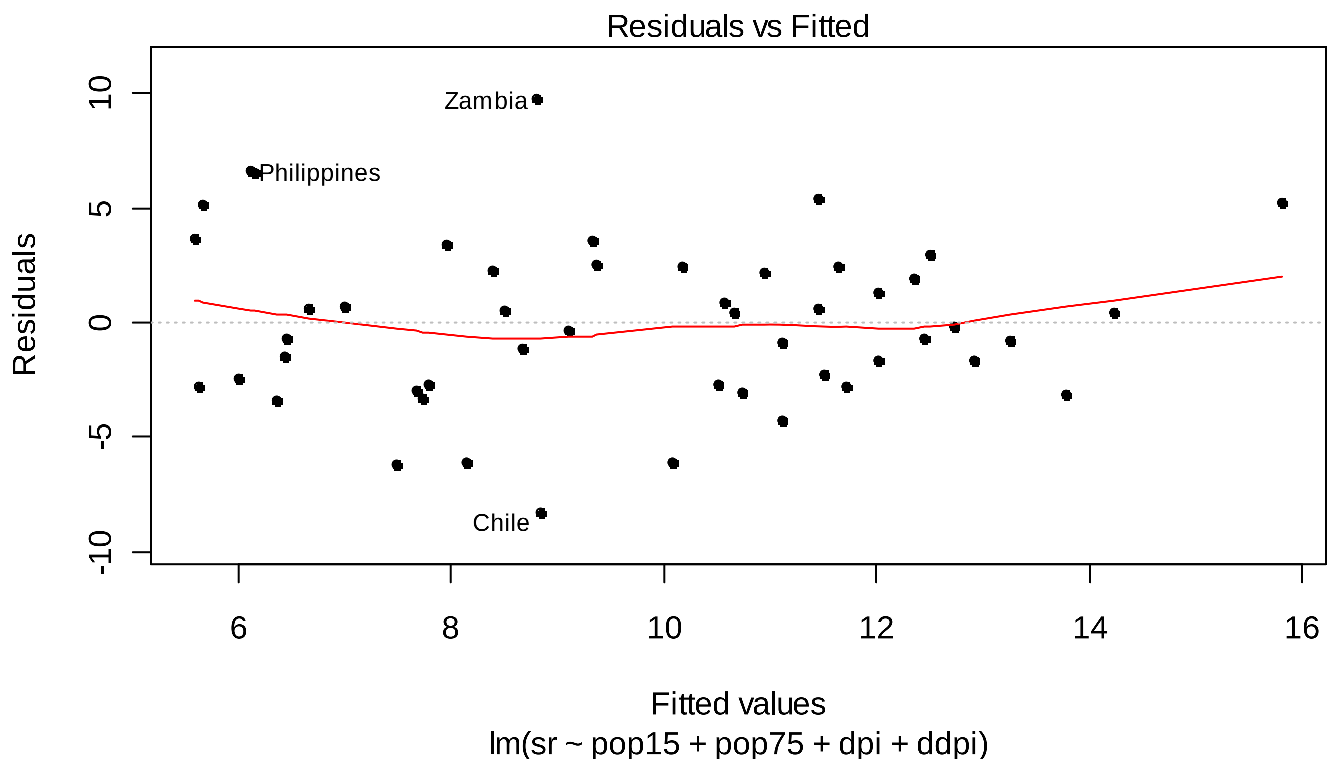
`pop15`: percentage of population < 15 years of age
`pop75`: percentage of population > 75 years of age
`dpi`: per-capita disposable income
`ddpi`: percentage rate of change in disposable income

The data are averaged over the decade 1960–1970 to remove the business cycle or other short-term fluctuations.

Applied Statistical Regression
AS 2013 – Week 08

Tukey-Anscombe-Plot

Plot the residuals r_i versus the fitted values $\hat{y}_i$



Applied Statistical Regression
AS 2013 – Week 08

Tukey-Anscombe-Plot

Is useful for:

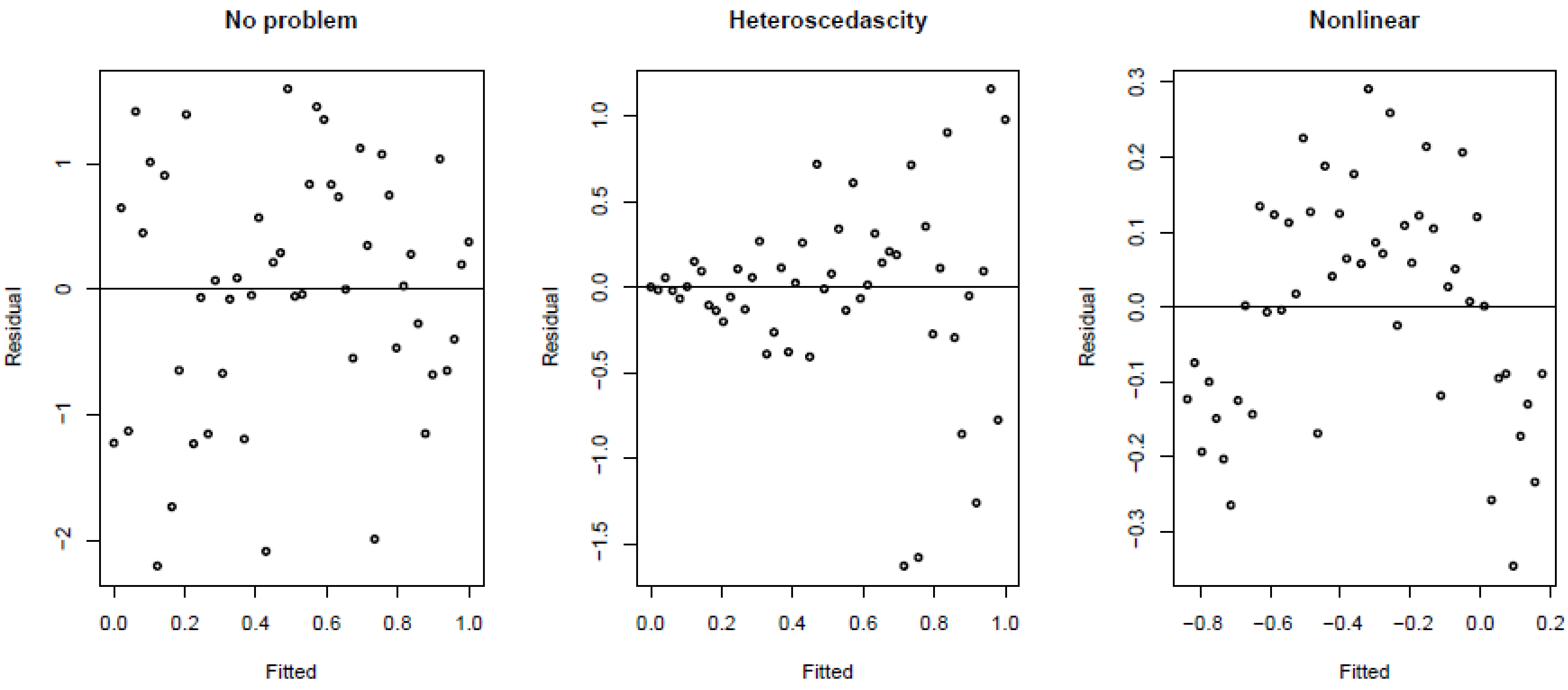
- finding structural model deficiencies, i.e. $E[E_i] \neq 0$
- if that is the case, the response/predictor relation could be nonlinear, or some predictors could be missing
- it is also possible to detect non-constant variance (→ then, the smoother does not deviate from 0)

When is the plot OK?

- the residuals scatter around the x-axis without any structure
- the smoother line is horizontal, with no systematic deviation
- there are no outliers

Applied Statistical Regression
AS 2013 – Week 08

Tukey-Anscombe-Plot



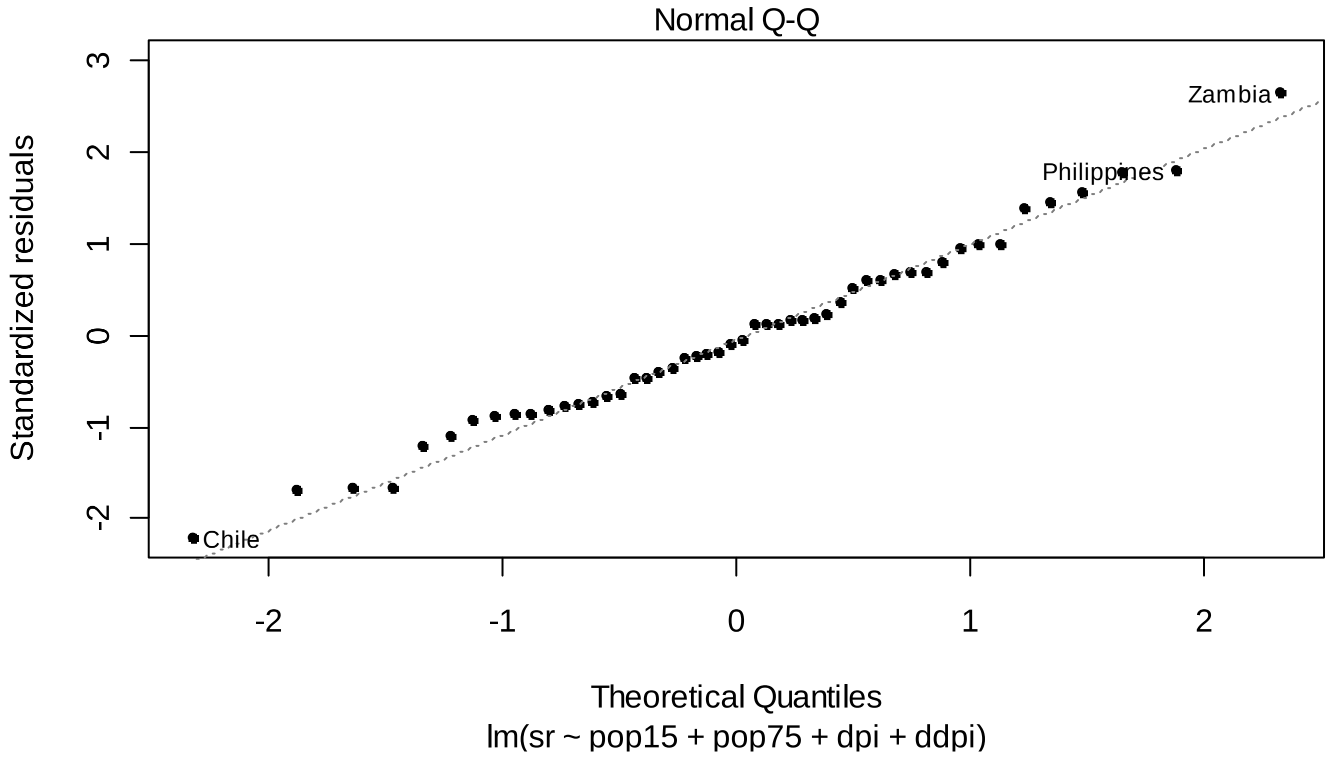
Tukey-Anscombe-Plot

When the Tukey-Anscombe-Plot is not OK:

- If structural deficencies are present ($E[E_i] \neq 0$, often also called "non-linearities"), the following is recommended:
  - "fit a better model", by doing transformations on the response and/or the predictors
  - sometimes it also means that some important predictors are missing. These can be completely novel variables, or also terms of higher order
- Non-constant variance: transformations usually help!

Normal Plot

Plot the residuals $\tilde{r}_i$ versus `qnorm(i/(n+1),0,1)`



Applied Statistical Regression
AS 2013 – Week 08

Normal Plot

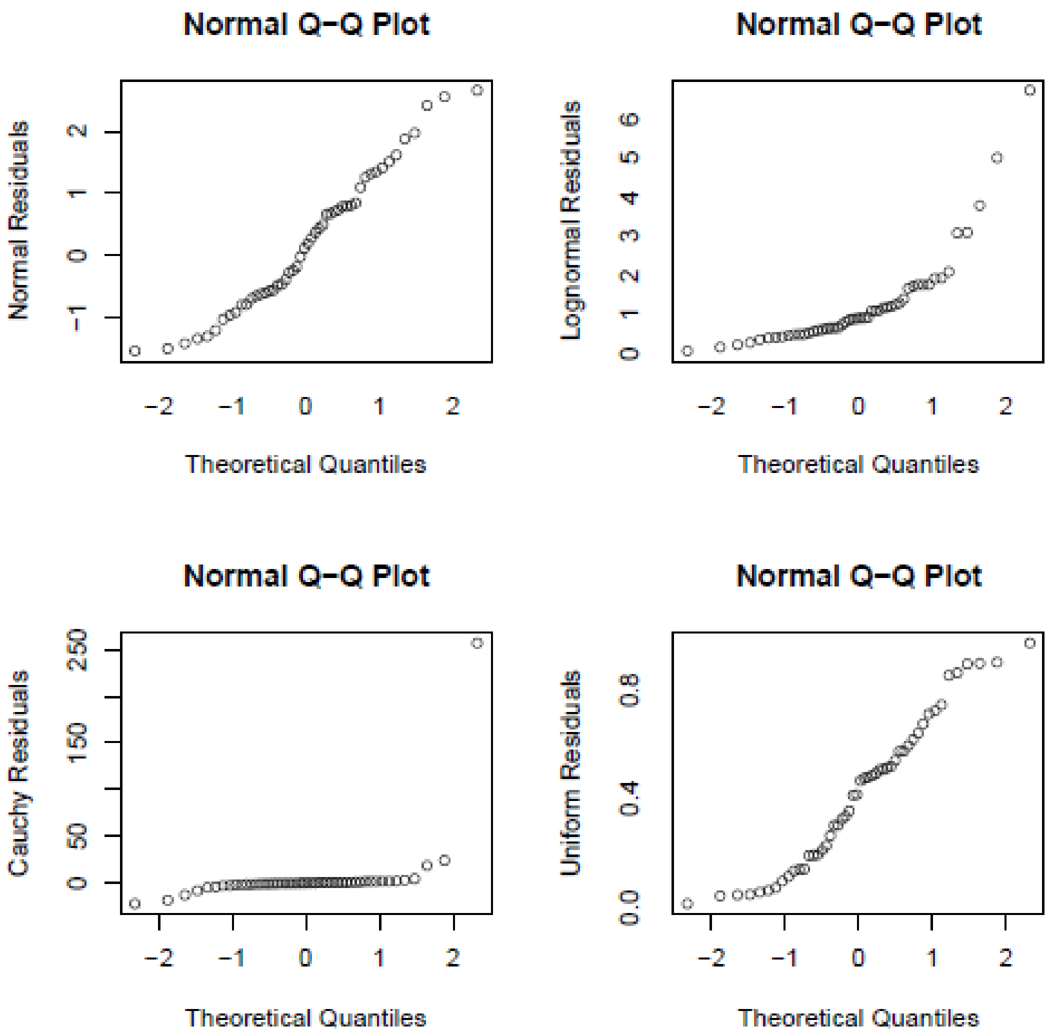
Is useful for:

- for identifying non-Gaussian errors: $E_i \overset{!}{\sim} N(0, \sigma_E^2 I)$

When is the plot OK?

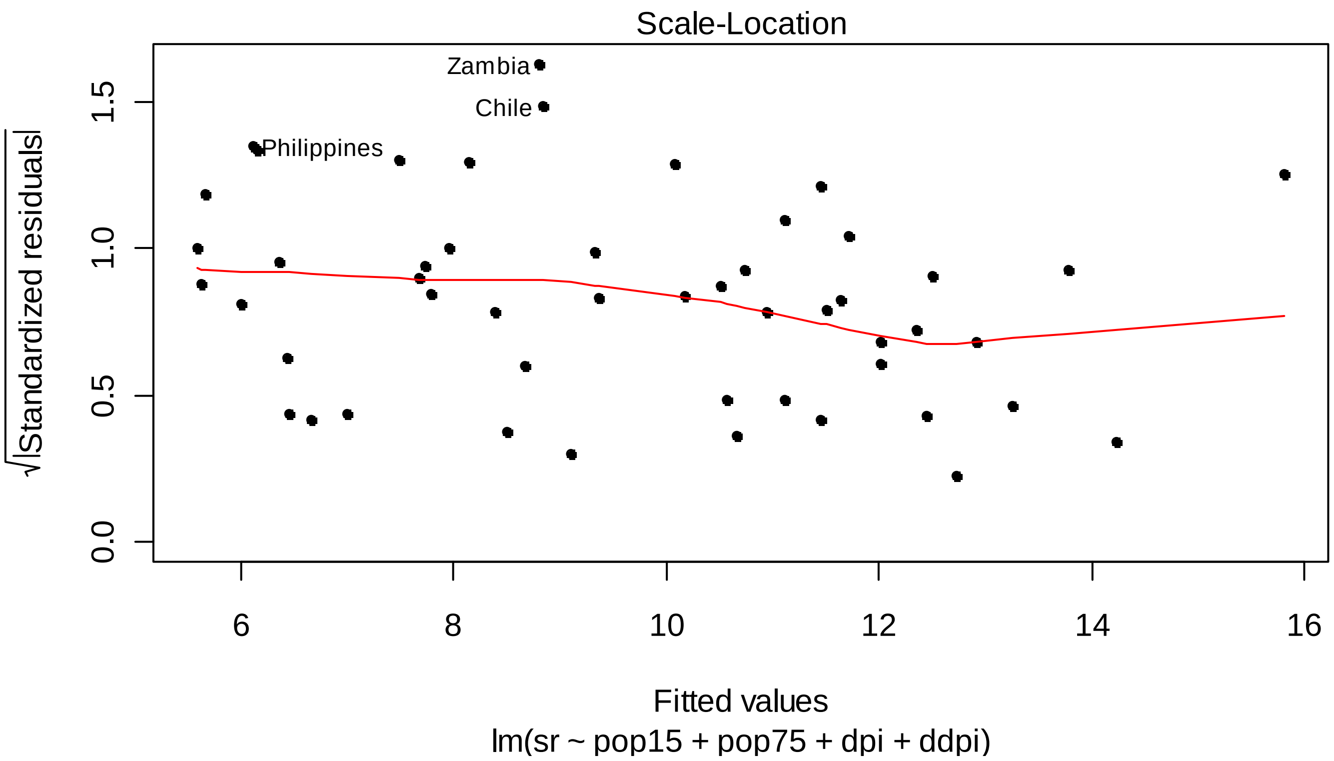
- the residuals $\tilde{r}_i$ must not show any systematic deviation from line which leads to the 1st and 3rd quartile.
- a few data points that are slightly "off the line" near the ends are always encountered and usually tolerable
- skewed residuals need correction: they usually tell that the model structure is not correct. Transformations may help.
- long-tailed, but symmetrical residuals are not optimal either, but often tolerable. Alternative: robust regression!

Normal Plot


Normal Q−Q Plot
Normal Residuals
Theoretical Quantiles
−2 −1 0 1 2
−1 0 1 2
Normal Q−Q Plot
Lognormal Residuals
Theoretical Quantiles
−2 −1 0 1 2
0 1 2 3 4 5 6
Normal Q−Q Plot
Cauchy Residuals
Theoretical Quantiles
−2 −1 0 1 2
0 50 150 250
Normal Q−Q Plot
Uniform Residuals
Theoretical Quantiles
−2 −1 0 1 2
0.0 0.4 0.8

Scale-Location-Plot

Plot $\sqrt{|\tilde{r}_i|}$ **versus** $\hat{y}_i$



Scale-Location-Plot

Is useful for:

- identifying non-constant variance: $Var(E_i) \neq \sigma_E^2$
- if that is the case, the model has structural deficencies, i.e. the fitted relation is not correct. Use a transformation!
- there are cases where we expect non-constant variance and do not want to use a transformation. This can the be tackled by applying weighted regression.

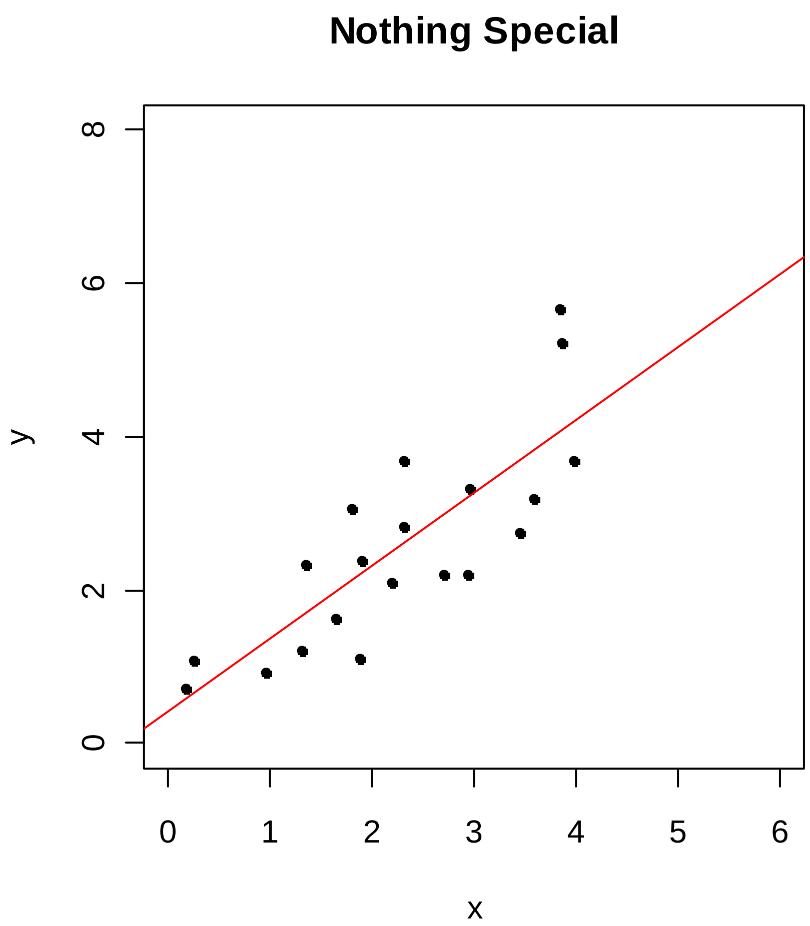
When is the plot OK?

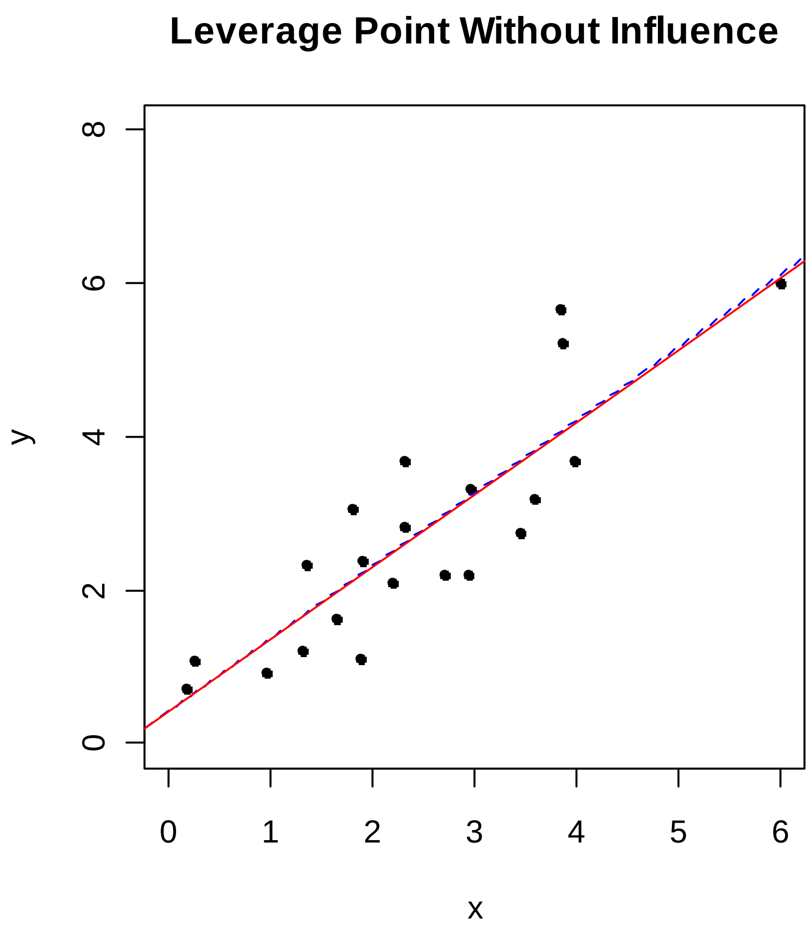
- the smoother line runs horizontally along the x-axis, without any systematic deviations.

Unusual Observations

- There can be observations which do not fit well with a particular model. These are called ***outliers***.
- There can be data points which have strong impact on the fitting of the model. These are called ***influential observations***.
- A data point can fall under **none, one or both** the above definitions – there is no other option.
- A ***leverage point*** is an observation that lies at a "different spot" in predictor space. This is potentially dangerous, because it can have strong influence on the fit.

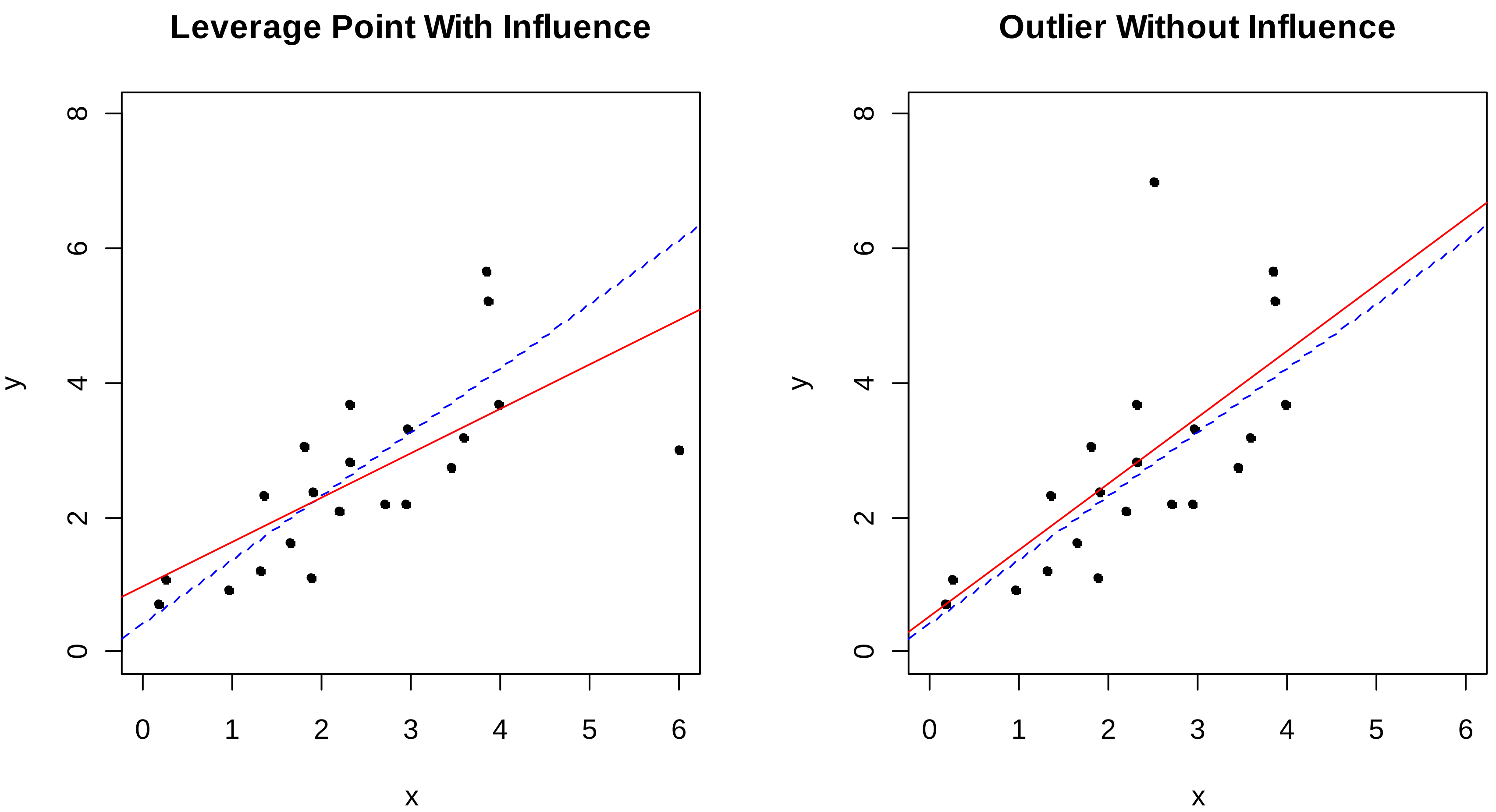
Unusual Observations





Applied Statistical Regression
AS 2013 – Week 08

Unusual Observations



How to Find Unusual Observations?

1) **Poor man's approach**

   Repeat the analysis n-times, where the i-th observation is left out. Then, the change is recorded.

2) **Leverage**

   If y_i changes by Δy_i, then $h_{ii}\Delta y_i$ is the change in $\hat{y}_i$.

   High leverage for a data point ($h_{ii} > 2(p+1)/n$) means that it forces the regression fit to adapt to it.

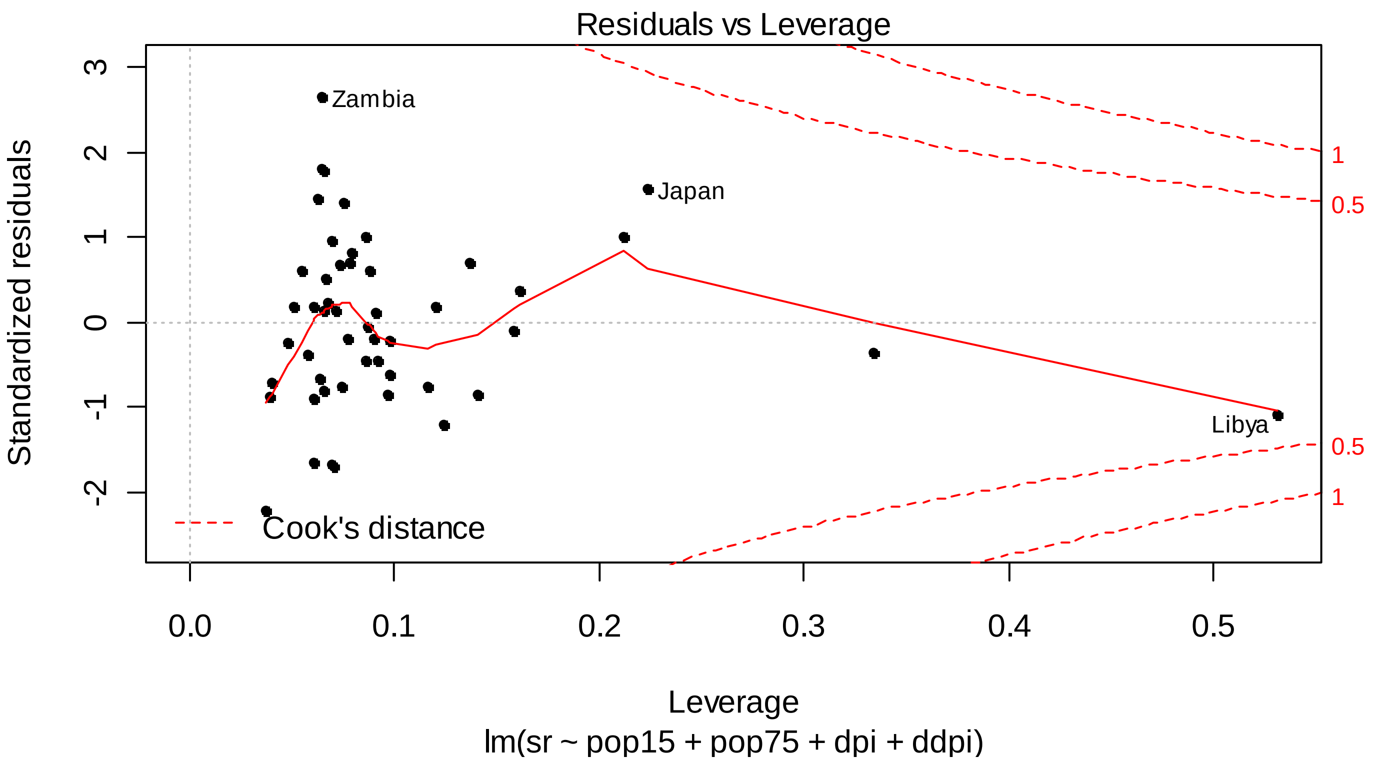
3) **Cook's Distance**

$$D_i = \frac{\sum(\hat{y}_j - y_{j(i)})^2}{(p+1)\sigma_E^2} = \frac{h_{ii}}{1-h_{ii}} \cdot \frac{r_i^{*2}}{(p+1)}$$

   Be careful if Cook's Distance > 1.

Leverage-Plot

Plot the residuals $\tilde{r}_i$ versus the leverage h_{ii}



Leverage-Plot

Is useful for:

- identifying outliers, leverage points and influential observation at the same time.

When is the plot OK?

- no extreme outliers in y-direction, no matter where
- high leverage, here $h_{ii} > 2(p+1)/n = 2(4+1)/50 = 0.2$ is always potentially dangerous, especially if it is in conjunction with large residuals!
- This is visualized by the Cook's Distance lines in the plot: >0.5 requires attention, >1 requires much attention!

Leverage-Plot

What to do with unusual observations:

- First check the data for gross errors, misprints, typos, etc.
- Unusual observations are also often a problem if the input is not suitable, i.e. if predictors are extremely skewed, because first-aid-transformations were not done. Variable transformations often help in this situation.
- Simply omitting these data points is not a very good idea. Unusual observations are often very informative and tell much about the benefits and limits of a model.