Applied Statistical Regression
HS 2011 – Week 07

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 8, 2011

Applied Statistical Regression
HS 2011 – Week 07

Residual Analysis – Model Diagnostics

Why do it? And what is it good for?

a) To make sure that estimates and inference are valid

- $E[\varepsilon_i] = 0$
- $Var(\varepsilon_i) = \sigma_\varepsilon^2$
- $Cov(\varepsilon_i, \varepsilon_j) = 0$
- $\varepsilon_i \sim N(0, \sigma_\varepsilon^2 I), \ i.i.d$

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
HS 2011 – Week 07

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
HS 2011 – Week 07

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
HS 2011 – Week 07

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.

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

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

Applied Statistical Regression
HS 2011 – Week 07

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
HS 2011 – Week 07

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)
```

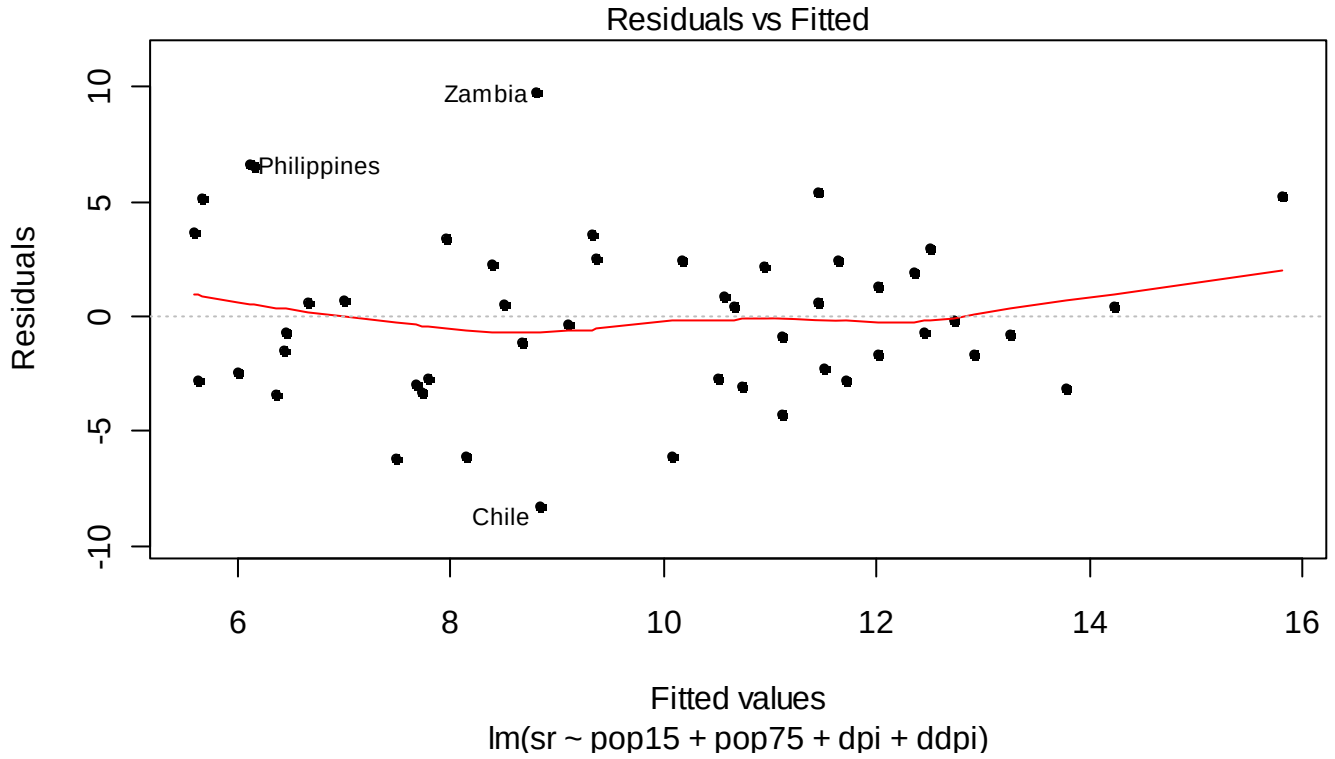
`pop15`: percentage of population < 15 years of age
`pop75`: percentage of population > 75 years of age
`dpi`: per-capita disposable income
`ddpi`: percentage rate of change in disposable income

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

Applied Statistical Regression
HS 2011 – Week 07

Tukey-Anscombe-Plot

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



Applied Statistical Regression
HS 2011 – Week 07

Tukey-Anscombe-Plot

Is useful for:

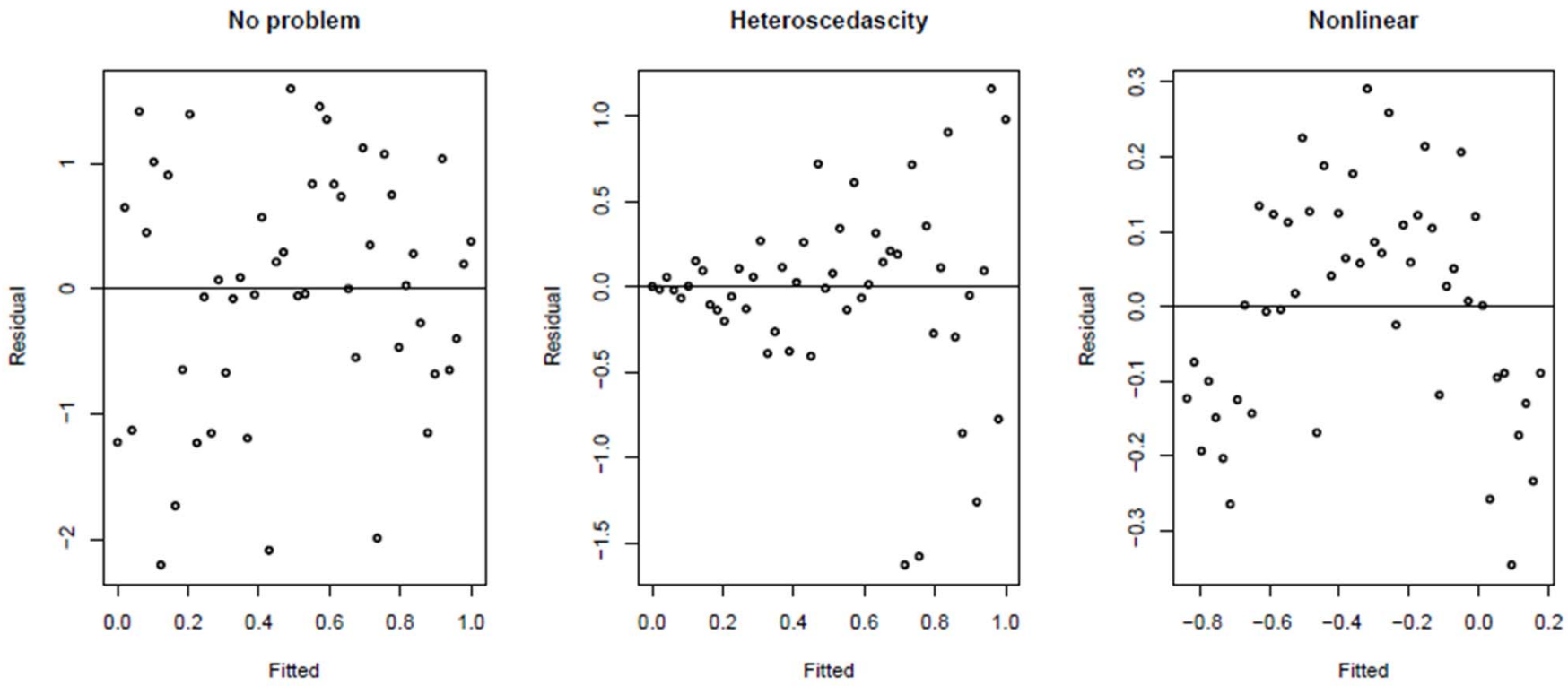
- finding structural model deficiencies, i.e. $E[E_i] \neq 0$
- if that is the case, the response/predictor relation could be nonlinear, or some predictors could be missing
- it is also possible to detect non-constant variance (→ then, the smoother does not deviate from 0)

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
HS 2011 – Week 07

Tukey-Anscombe-Plot



Tukey-Anscombe-Plot

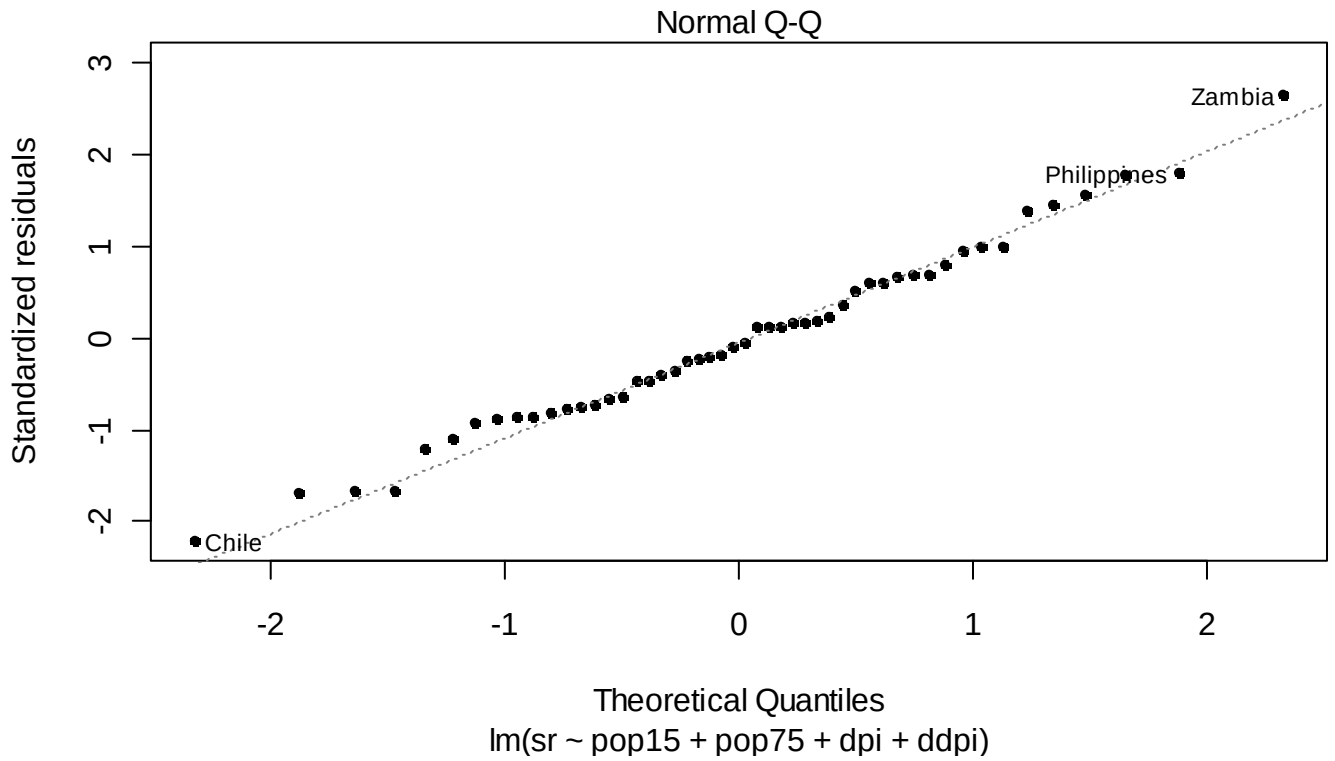
When the Tukey-Anscombe-Plot is not OK:

- If structural deficencies are present ($E[\varepsilon_i] \neq 0$, 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!

Applied Statistical Regression
HS 2011 – Week 07

Normal Plot

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



Applied Statistical Regression
HS 2011 – Week 07

Normal Plot

Is useful for:

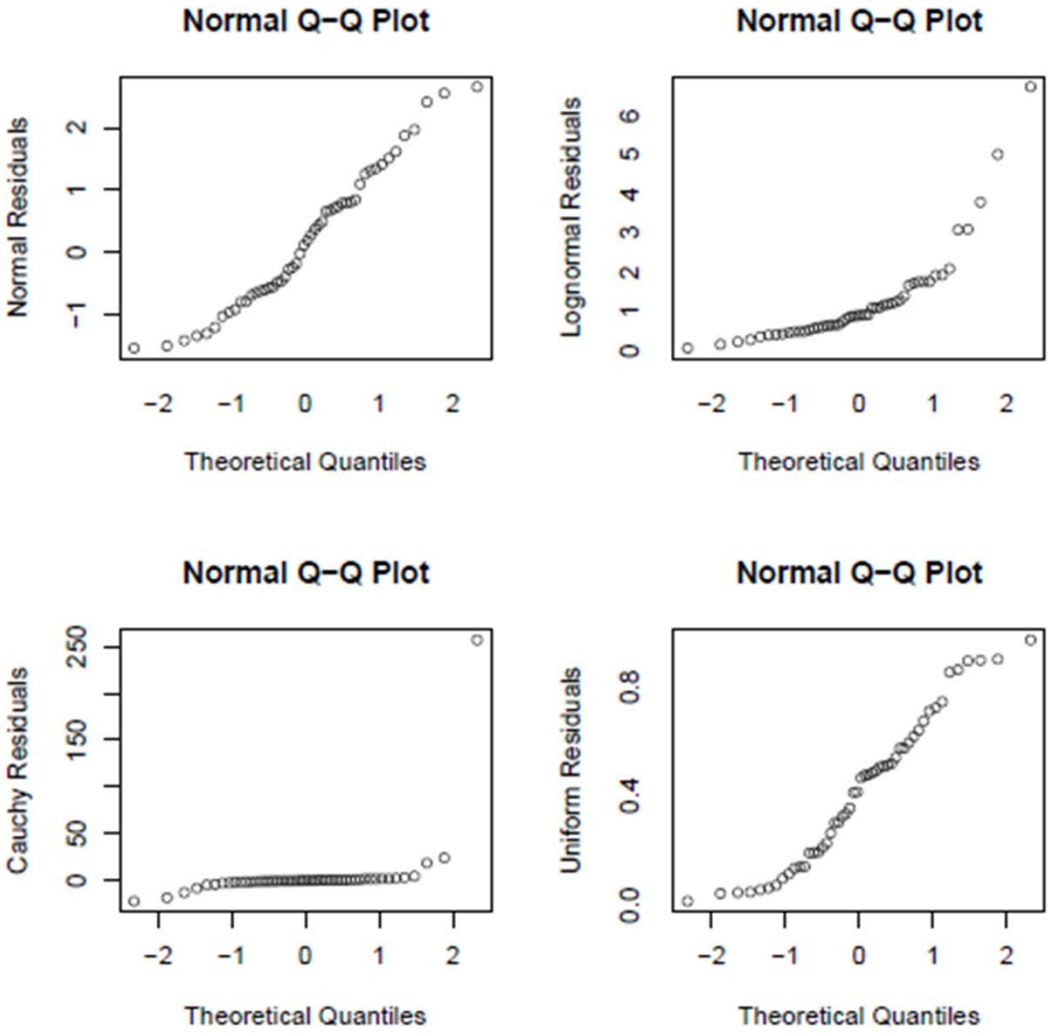
- for identifying non-Gaussian errors: $E_i \sim^{!} N(0, \sigma_E^2 I)$

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!

Applied Statistical Regression
HS 2011 – Week 07

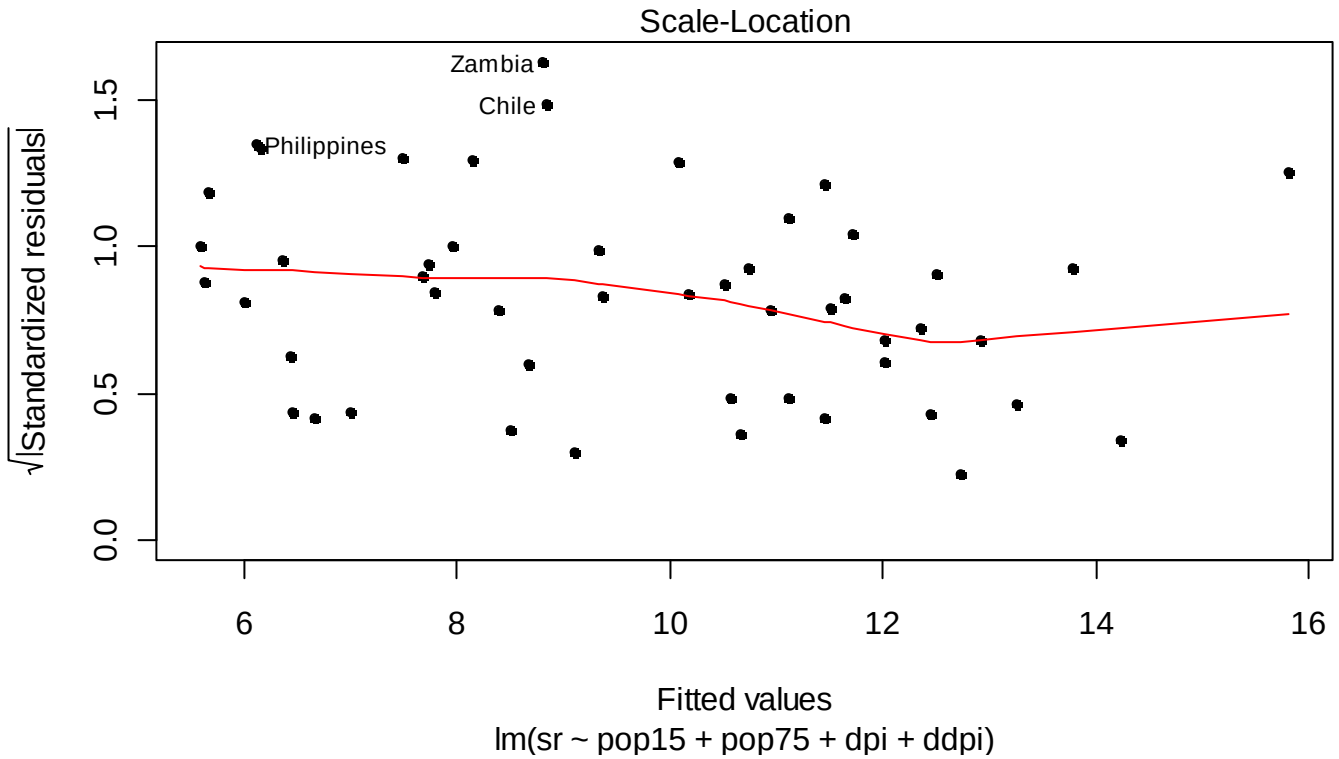
Normal Plot



Applied Statistical Regression
HS 2011 – Week 07

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?

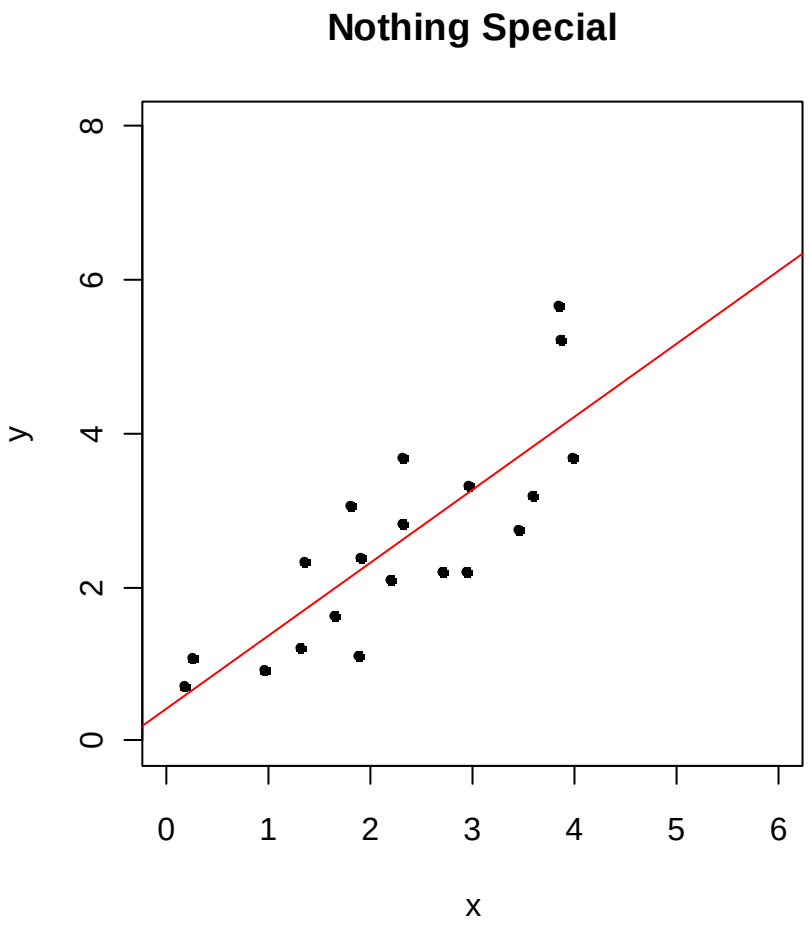
- the smoother line runs horizontally along the x-axis, without any systematic deviations.

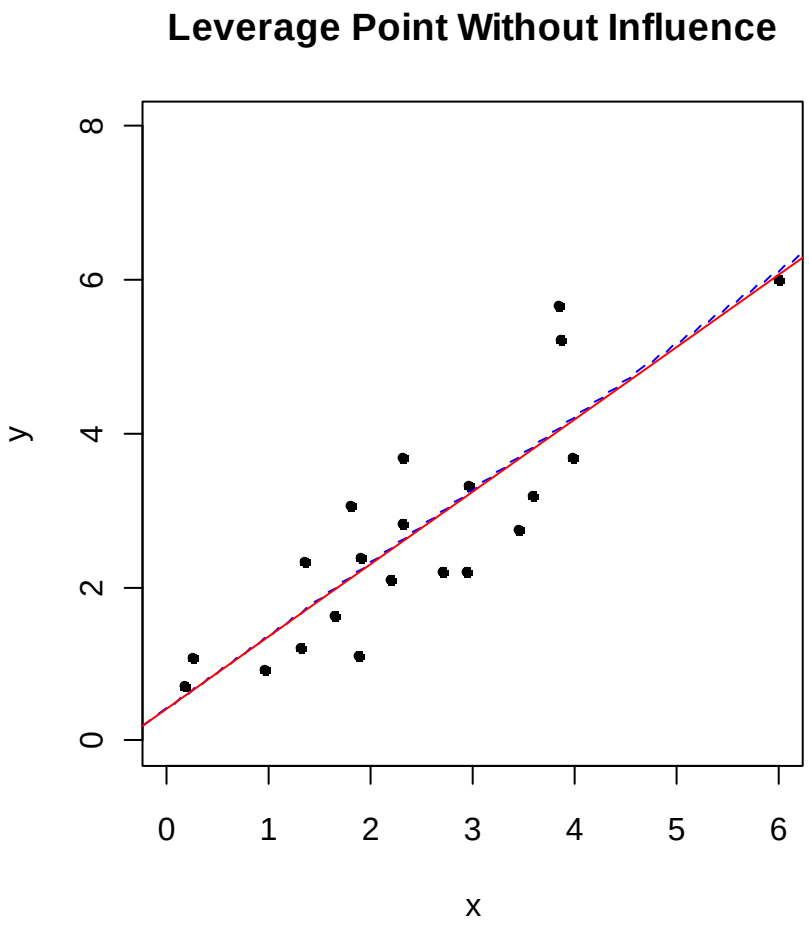
Unusual Observations

- 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.

Applied Statistical Regression
HS 2011 – Week 07

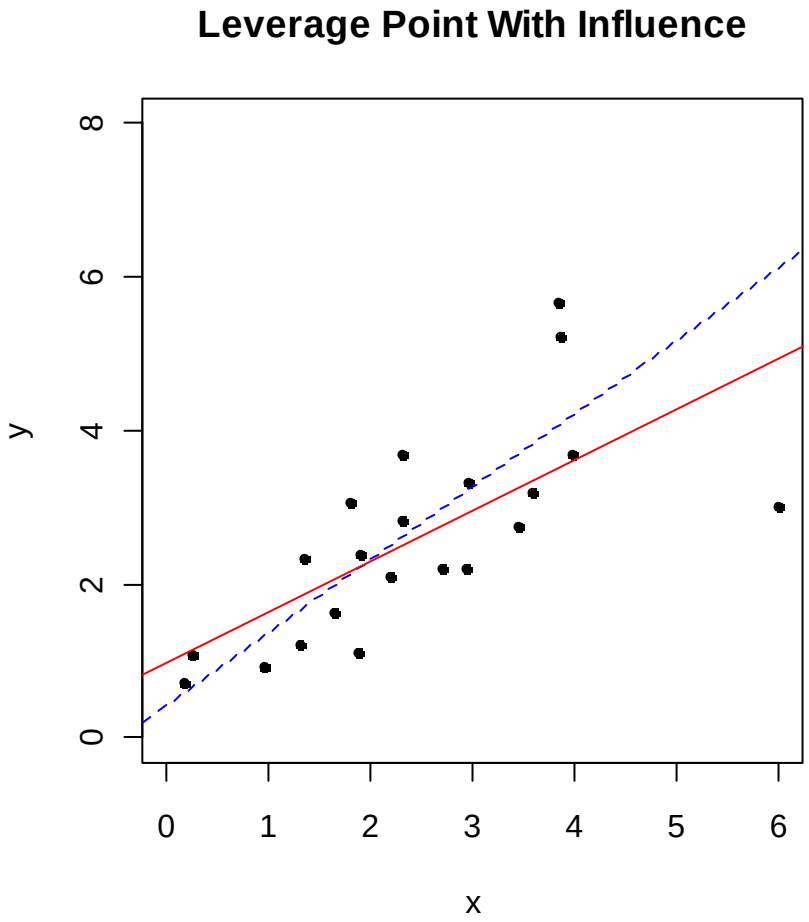
Unusual Observations

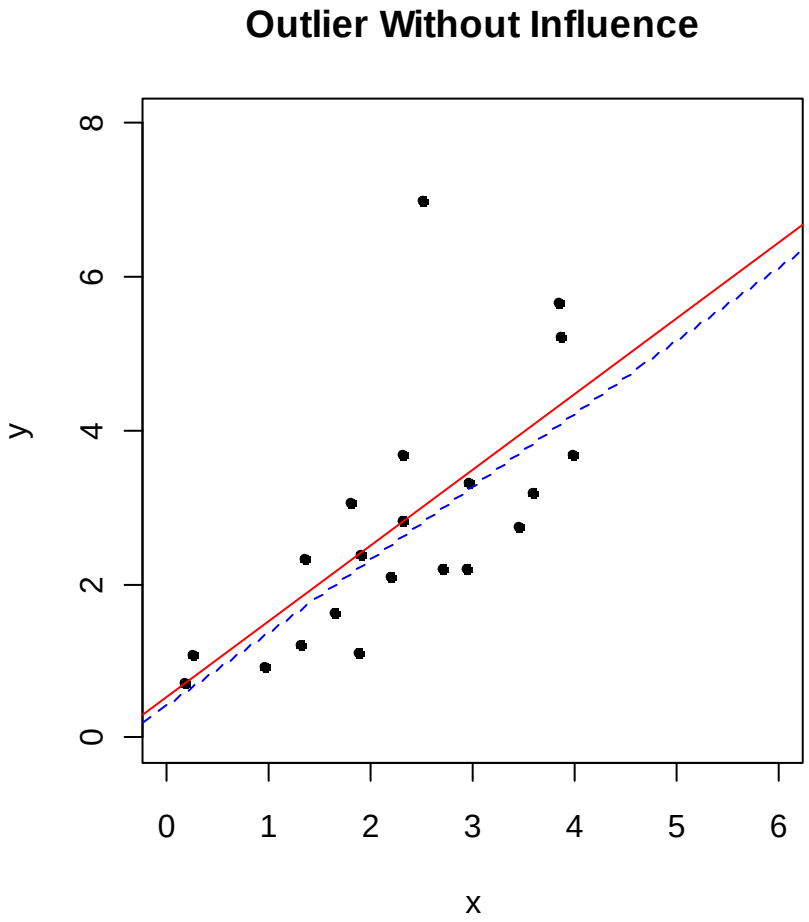




Applied Statistical Regression
HS 2011 – Week 07

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

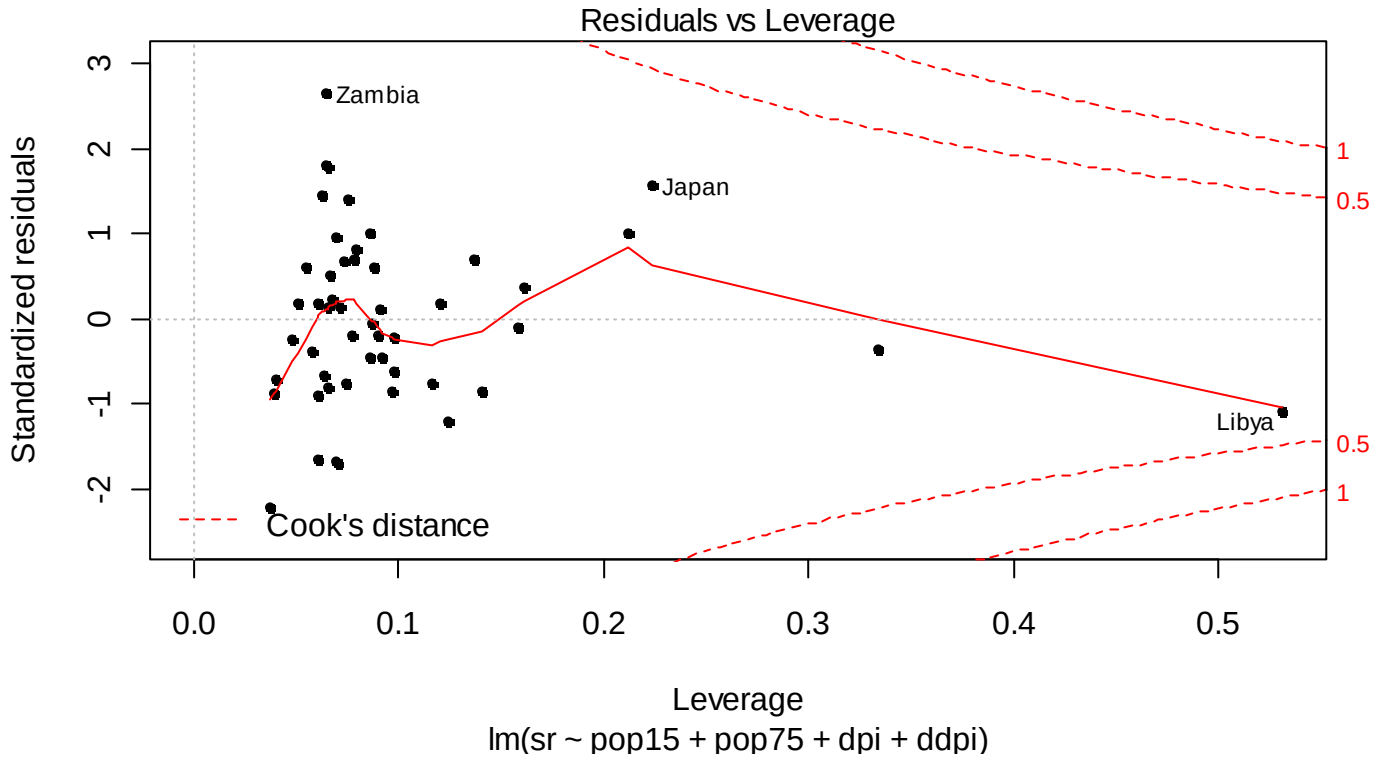
$$D_i = \frac{\sum(\hat{y}_j - y_{j(i)})^2}{(p+1)\sigma_\varepsilon^2} = \frac{h_{ii}}{1-h_{ii}} \cdot \frac{r_i^{*2}}{(p+1)}$$

Be careful if Cook's Distance > 1.

Applied Statistical Regression
HS 2011 – Week 07

Leverage-Plot

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



Applied Statistical Regression
HS 2011 – Week 07

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.

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

Residuals vs. (Potential) Predictors

General Remark:

We are allowed to plot the residuals versus any arbitrary variable we wish. This includes:

- predictors that were used in fitting the model
- potential predictors which were not (yet) used in the model
- in particular also the time/sequence of the observations

All these plots have one thing in common:

All these residual plots must not show any structure. If they do, the model has some deficiencies, and can be improved!

Residuals vs. (Potential) Predictors

Example:

This dataset deals with the prestige of Canadian occupations. There are 102 different observations and 6 columns:

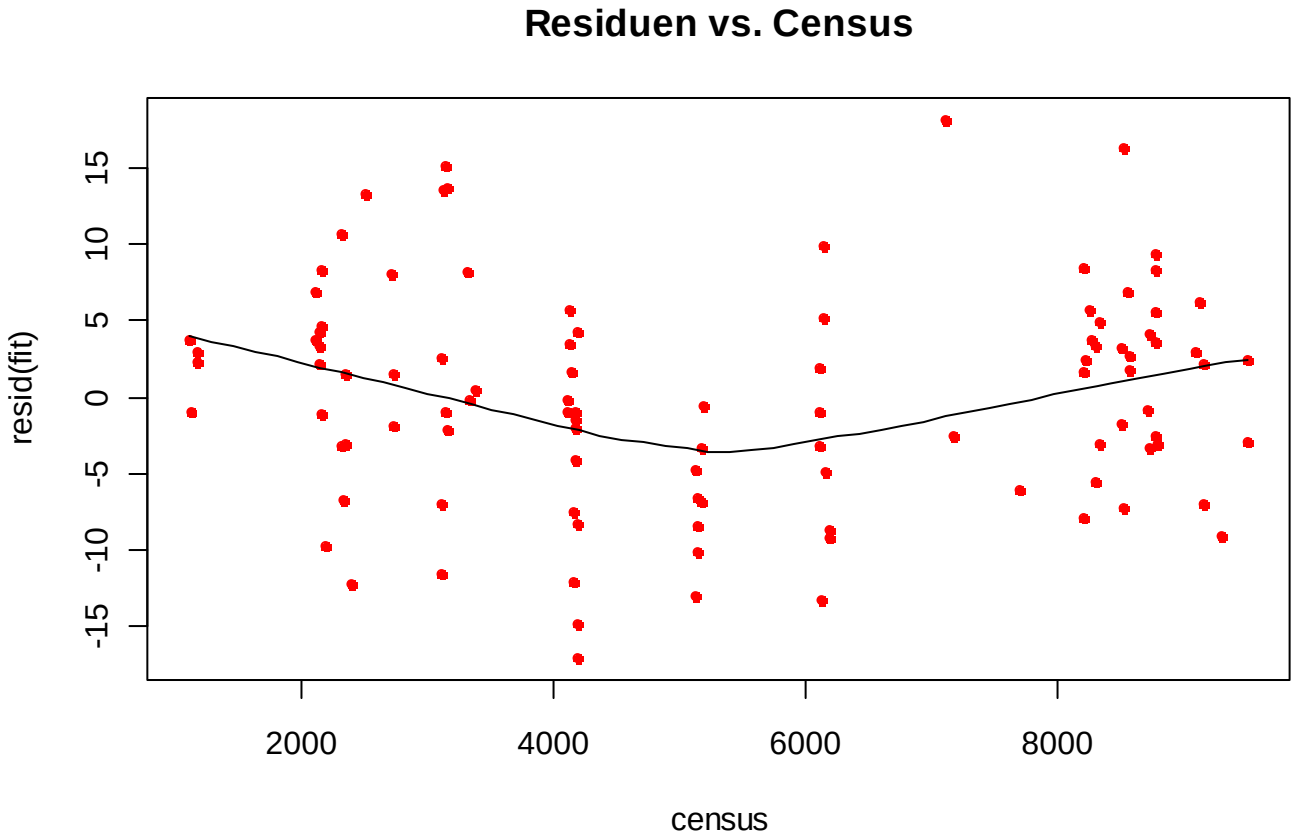
	educ	income	women	prest	cens	type
gov.administrators	13.11	12351	11.16	68.8	1113	prof
general.managers	12.26	25879	4.02	69.1	1130	prof
accountants	12.77	9271	15.70	63.4	1171	prof

We start with fitting the model: `prestige ~ income + education`, but do not take into account any of the remaining predictors.

Applied Statistical Regression
HS 2011 – Week 07

Residuals vs. Potential Predictors

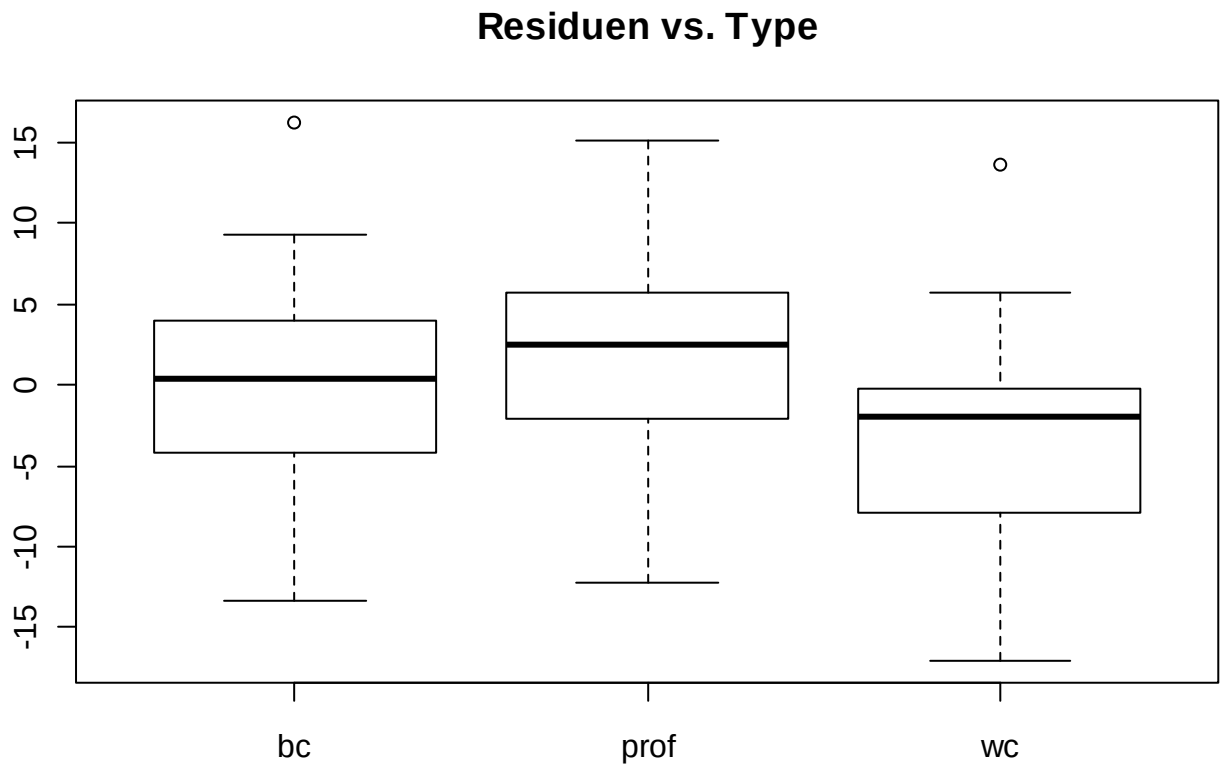
```
> scatter.smooth(census, resid(fit), col="red", pch=20)
```



Applied Statistical Regression
HS 2011 – Week 07

Residuals vs. Potential Predictors

```
> boxplot(resid(fit) ~ type)
```



Motivation for Partial Residual Plots

Problem:

We sometimes want to learn about the relation between a predictor and the response, and also visualize it. Is it also of importance whether it is directly linear.

How can we infer this?

- we can plot y versus predictor x_k
- however, the problem is that all the other predictors also influence the response and thus blur our impression
- thus, we require a plot which shows the "isolated" influence of predictor x_k on the response y

Applied Statistical Regression
HS 2011 – Week 07

Partial Residual Plots

Idea:

We remove the estimated effect of all the other predictors from the response and plot this versus the predictor x_k.

$$y - \sum_{k \neq j} x_j \hat{\beta}_j = \hat{y} + r - \sum_{k \neq j} x_j \hat{\beta}_j = x_k \hat{\beta}_k + r$$

We then plot these so-called partial residuals versus the predictor x_k. We require the relation to be linear!

Partial residual plots in R:

- `library(car); crPlots(...)`
- `library(faraway); prplot(...)`

Partial Residual Plots: Example

We try to predict the prestige of a number of 102 different profession with a set of 2 predictors:

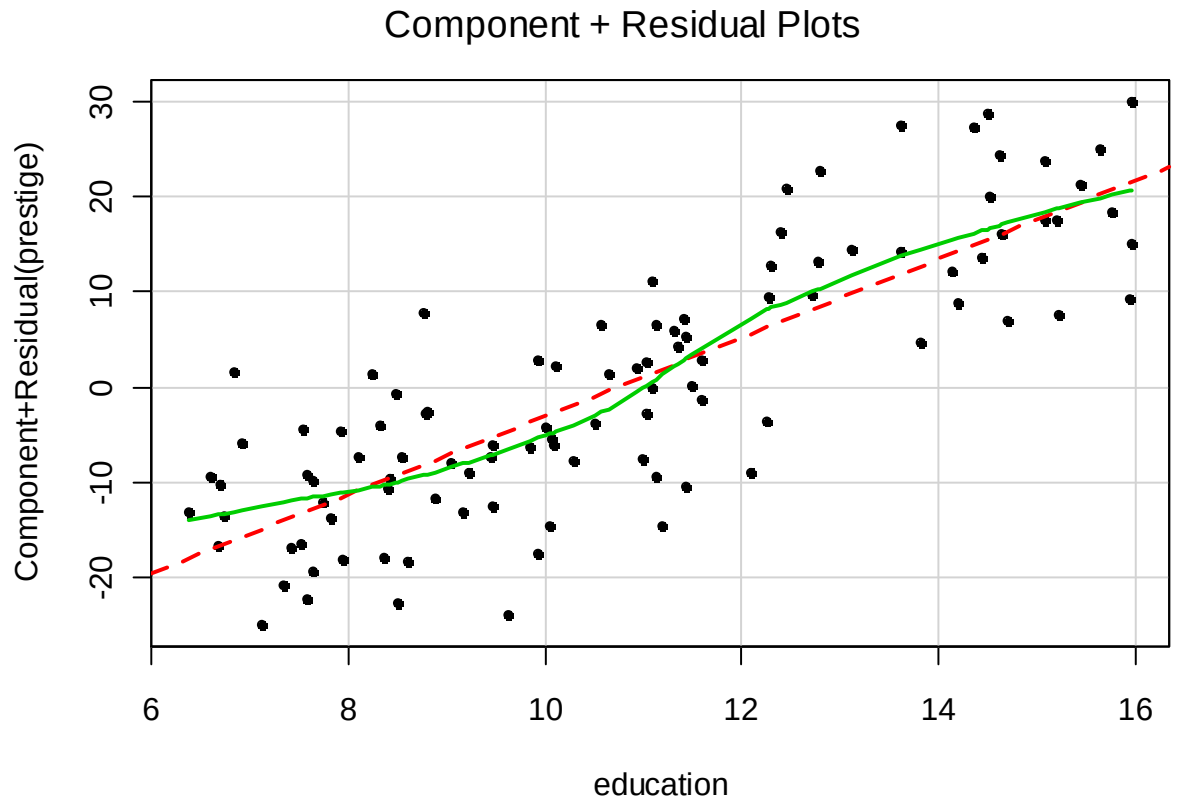
prestige ~ education + income

```
> data(Prestige)
> head(Prestige)
                    education income women prestige census type
gov.administrators      13.11  12351 11.16     68.8   1113 prof
general.managers        12.26  25879  4.02     69.1   1130 prof
accountants             12.77   9271 15.70     63.4   1171 prof
purchasing.officers     11.42   8865  9.11     56.8   1175 prof
chemists                14.62   8403 11.68     73.5   2111 prof
...
```

Applied Statistical Regression
HS 2011 – Week 07

Partial Residual Plots: Example

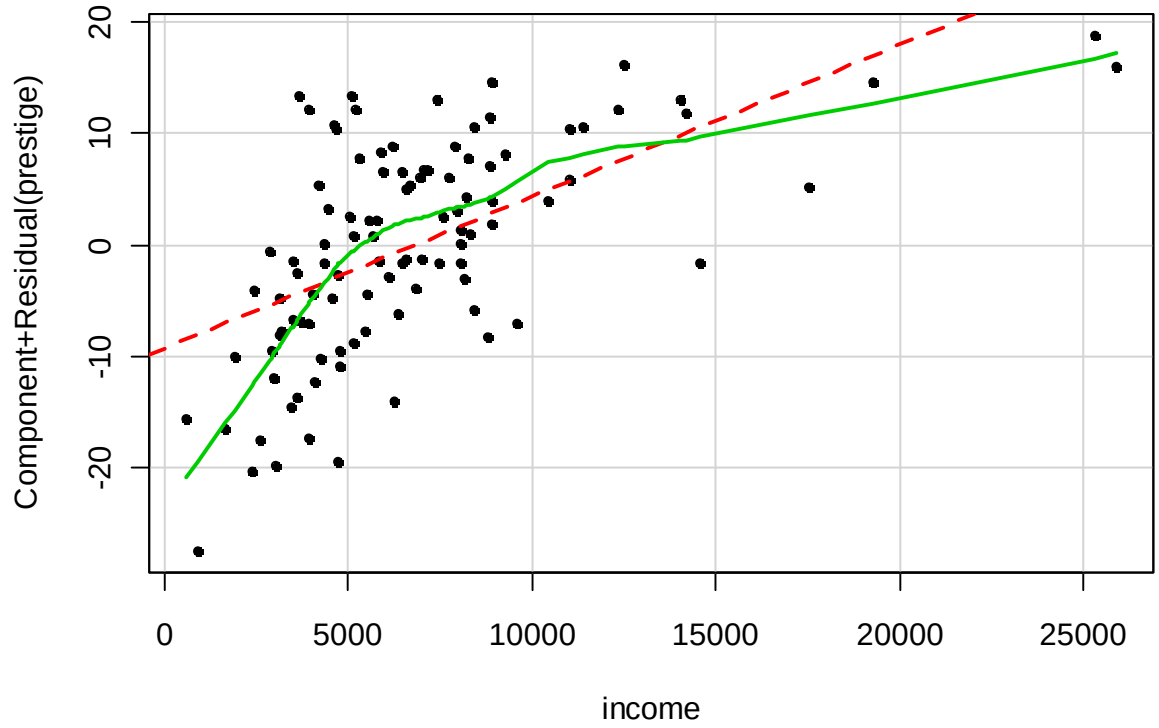
```
library(car); data(Prestige)
fit <- lm(prestige ~ education + income, data=Prestige)
crPlots(fit, layout=c(1,1))
```



Applied Statistical Regression
HS 2011 – Week 07

Partial Residual Plots: Example

```
library(car); data(Prestige)
fit <- lm(prestige ~ education + income, data=Prestige)
crPlots(fit, layout=c(1,1))
```



Evident non-linear influence of income on prestige.

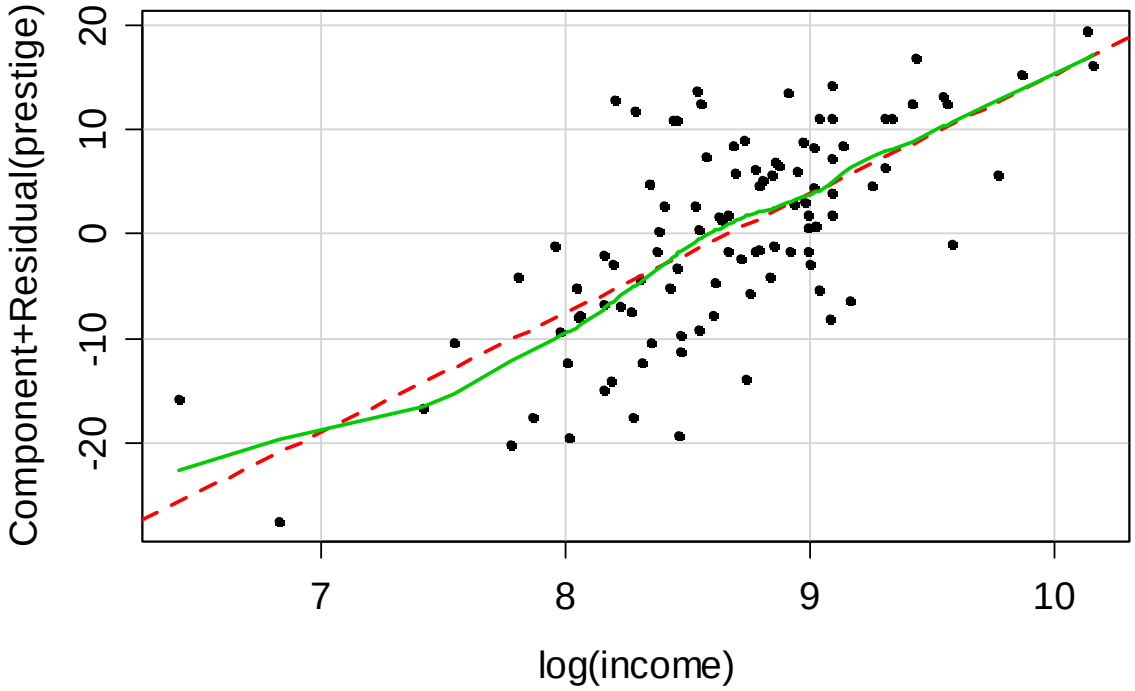
→ not a good fit!

→ correction needed

Applied Statistical Regression
HS 2011 – Week 07

Partial Residual Plots: Example

```
library(car); data(Prestige)
fit <- lm(prestige ~ education + log(income), Prestige)
crPlots(fit, layout=c(1,1))
```



After a log-trsf of predictor 'income', things are fine

Applied Statistical Regression
HS 2011 – Week 07

Partial Residual Plots

Summary:

Partial residual plots show the marginal relation between a predictor x_k and the response y.

When is the plot OK?

If the red line with the actual fit, and the green line of the smoother do not show systematic differences.

What to do if the plot is not OK?

- apply a transformation
- use Generalized Additive Models (GAM, tbd later)

Applied Statistical Regression
HS 2011 – Week 07

Checking for Correlated Errors

Background:

For LS-fitting we require uncorrelated errors. For data which have timely or spatial structure, this condition happens to be violated quite often.

Example:

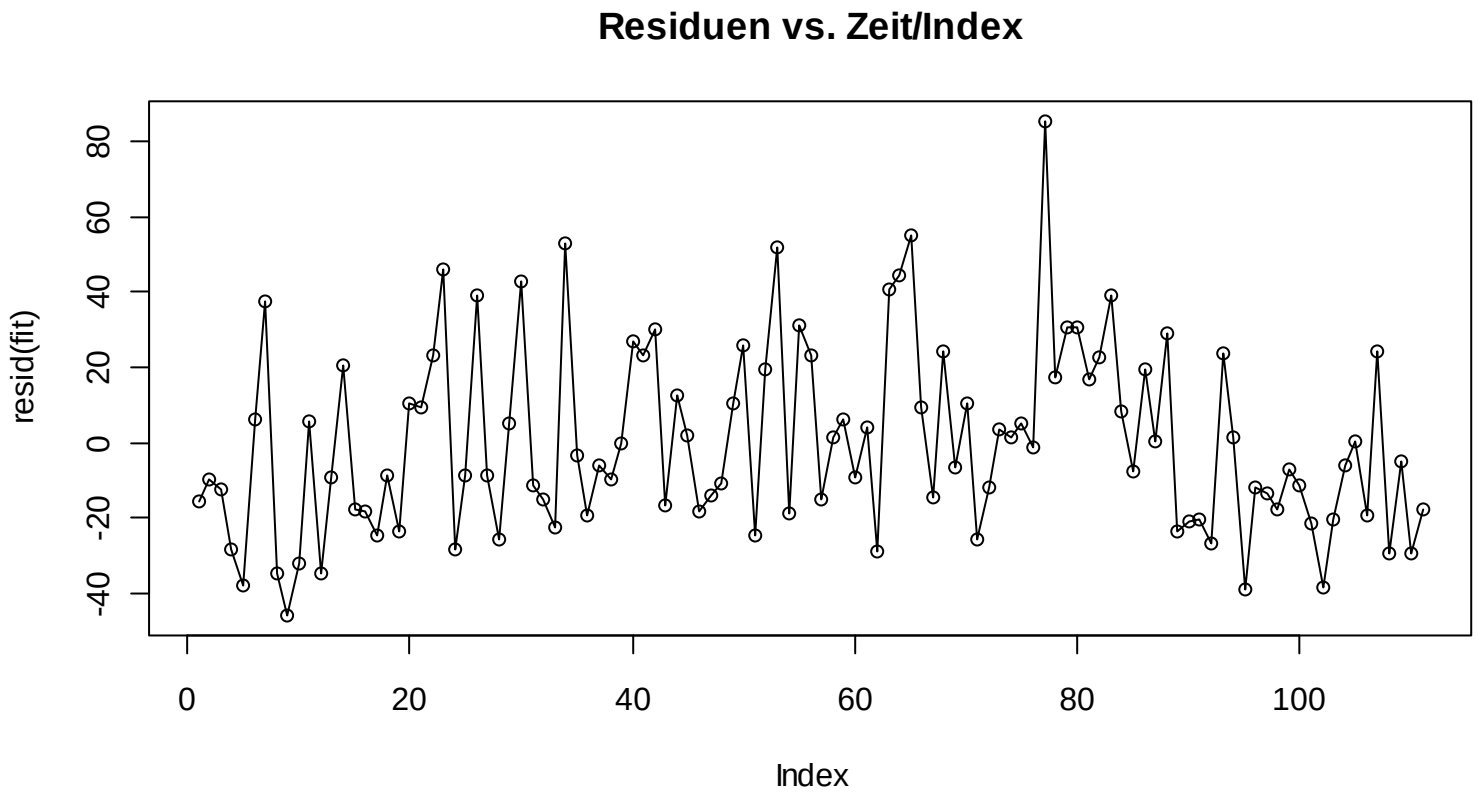
- library(faraway); data(airquality)
- Ozone ~ Solar.R + Wind
- Measurements from 153 consecutive days in New York
- data have a timely sequence

→ **to be handled with care!**

Applied Statistical Regression
HS 2011 – Week 07

Residuals vs. Time/Index

```
> plot(resid(fit)); lines(resid(fit))
```



Alternative: Durbin-Watson-Test

The Durbin-Watson-Test checks if consecutive observations show a sequential correlation:

Test statistic: $$DW = \frac{\sum_{i=2}^{n} (r_i - r_{i-1})^2}{\sum_{i=1}^{n} r_i^2}$$

- under the null hypothesis "no correlation", the test statistic has a χ^2- distribution. The p-value can be computed.
- the DW-test is somewhat problematic, because it will only detect simple correlation structure. When more complex dependency exists, it has very low power.

Applied Statistical Regression
HS 2011 – Week 07

Durbin-Watson-Test

R-Hints:

```
library(lmtest)

> dwtest(Ozone ~ Solar.R + Wind, data=airquality)

        Durbin-Watson test

data:  Ozone ~ Solar.R + Wind
DW = 1.6127, p-value = 0.01851
alternative hypothesis: true autocorrelation is greater than 0
```

The null hypothesis is rejected. We conclude that the residuals are correlated. For more details, see the exercises...

Residuals vs. Time/Index

When is the plot OK?

- There is no systematic structure present
- There are no long sequences of pos./neg. residuals
- There is no back-and-forth between pos./neg. residuals

What to do if the plot is not OK?

1) Search for and add the "forgotten" predictors
2) Using the generalized least squares method (GLS)
   → to be discussed in Applied Time Series Analysis
3) Estimated coefficients and fitted values are not biased, but confidence intervals and tests are: be careful!

Applied Statistical Regression
HS 2011 – Week 07

Further Strategies for Problem Solving

Where are we?

- We know the model assumptions and the standard plots for diagnostics. And we also know how we can identify problems in these plots.
- So far, we discussed how "non-linear" relations (i.e. missing transformations in response/predictors) can be recognized, or how we can identify missing predictors.
- Now, we will be discussing two specific model violations, which cannot be dealt with using transformations: these are **non-constant variance** and **long-tailed errors**.

Applied Statistical Regression
HS 2011 – Week 07

Weighted Regression

When to use?

Weighted regression is used when symmetrically distributed errors have zero expectation, but, according to the Scale-Location-Plot, have non-constant variance.

Important:

If non-constant variance is observed together with non-optimal model structure, and/or skewly distributed errors, then weighted regression is not the right tool. In that case, better search for a response/predictor transformation.

Applied Statistical Regression
HS 2011 – Week 07

Weighted Regression: Model

The model is:

$$Y = X\beta + \varepsilon \text{, wobei } \varepsilon \sim N(0, \sigma_\varepsilon^2 \Sigma)$$

→ For the non-weighted ordinary least squares regression, the error covariance matrix is the identity: $\Sigma = I$

→ We still assume uncorrelated errors, but no longer do we assume uncorrelated errors. The covariance matrix can thus be:

$$\Sigma = diag\left(\frac{1}{w_1}, \frac{1}{w_2}, ..., \frac{1}{w_n}\right) \neq I$$

Applied Statistical Regression
HS 2011 – Week 07

Weighted Regression: And Now?

In a weighted least squares problem, the regression coefficients are estimated by minimizing a weighted sum of squares:

$$\sum_{i=1}^{n} w_i r_i^2$$

If the design matrix has full rank, this minimization problem has an explicit and unique solution. Moreover:

- Observations with small variance (i.e. where one is "sure" about the position of the data point) obtain large weight in the regression fit, and vice versa.

Applied Statistical Regression
HS 2011 – Week 07

Where Are the Weights from?

1) If the response Y_i is the mean from several independent observations, but not the same number of every data point. Then use: $w_i = n_i$.

   Example: Regression where daily cost in a mental hospital is explained with some socio-demographic predictors. The response variable is:

   "Total cost for the stay" / "Length of stay in days"

   The bigger the number of days that were used for assessing the cost, the more precise (=lower variance) the average cost is determined.

Where are the weights from?

2) One knows or can easily see that the variance in the residuals is proportional to a predictor.
   Then, we use: $w_i = 1/x_i$

   Example: **see Exercises...**

3) If non-constant variance is only "observed", but the cause is unknown (with respect to 1) and 2) above), the we can still try to first fit an ordinary least squares regression and use it for estimating weights, which will then be used in an weighted linear regression.

   Example: none...