Applied Statistical Regression
AS 2012 – 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 12, 2012

Applied Statistical Regression
AS 2012 – Week 08

Residual Analysis for Multiple Regression

Toolbox:

Model diagnostics for multiple linear regressions is based on a set of 4 different residual plots. These are routinely checked with every fitted model.

- **Tukey-Anscombe Plot**
- **Normal Plot**
- **Scale-Location Plot**
- **Leverage Plot with Cook's Distance**

In R: `> plot(fit)`

Applied Statistical Regression
AS 2012 – Week 08

More Residual Plots

General Remark:

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

- predictors that were used
- potential predictors which were not (yet) used
- other variables, e.g. time/sequence of the observations

The rule is:

No matter what the residuals are plotted against, there must not be any non-random structure. Else, the model has some deficiencies, and needs improvement!

Residuals vs. (Potential) Predictors

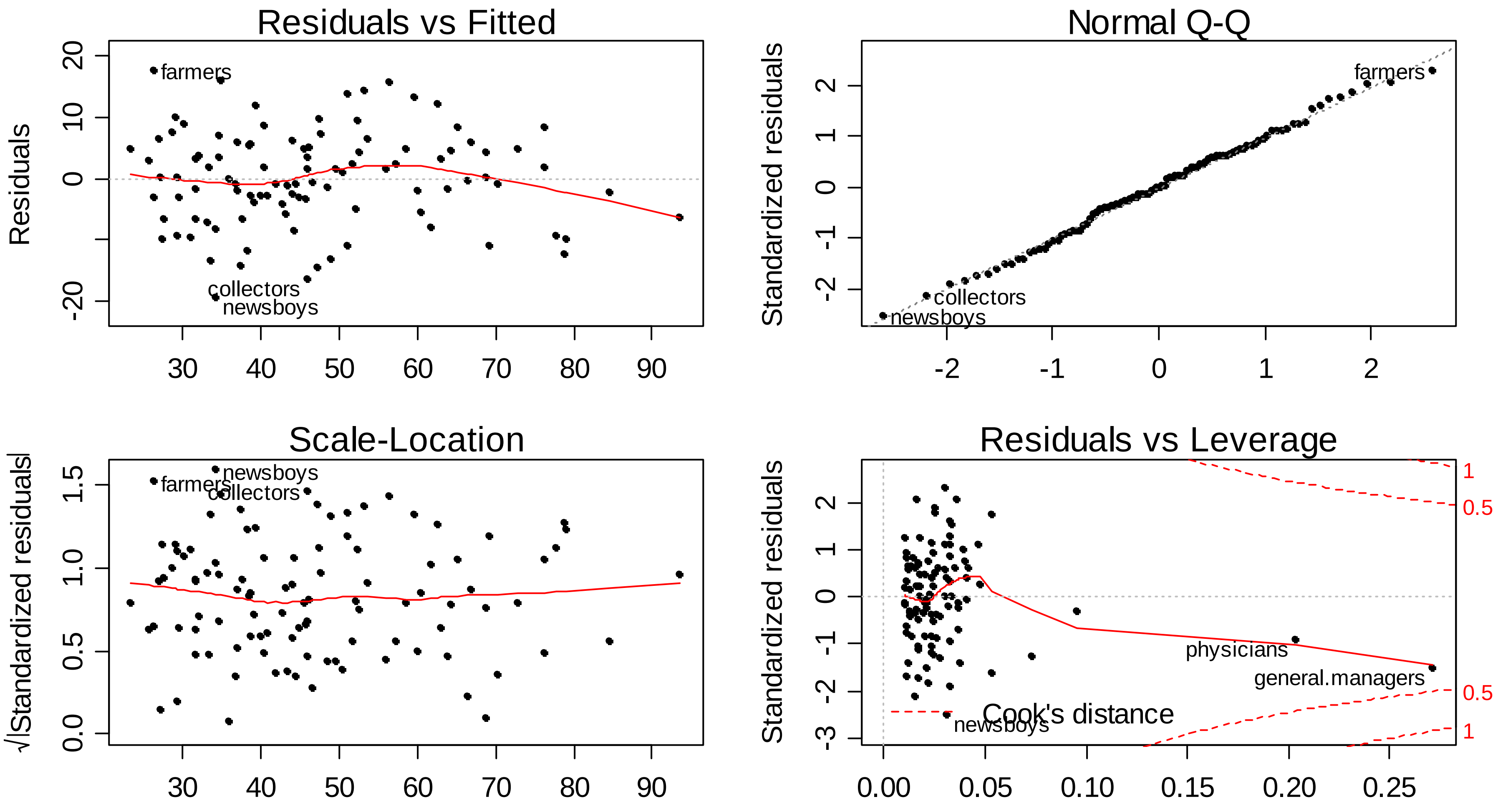
Example:

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

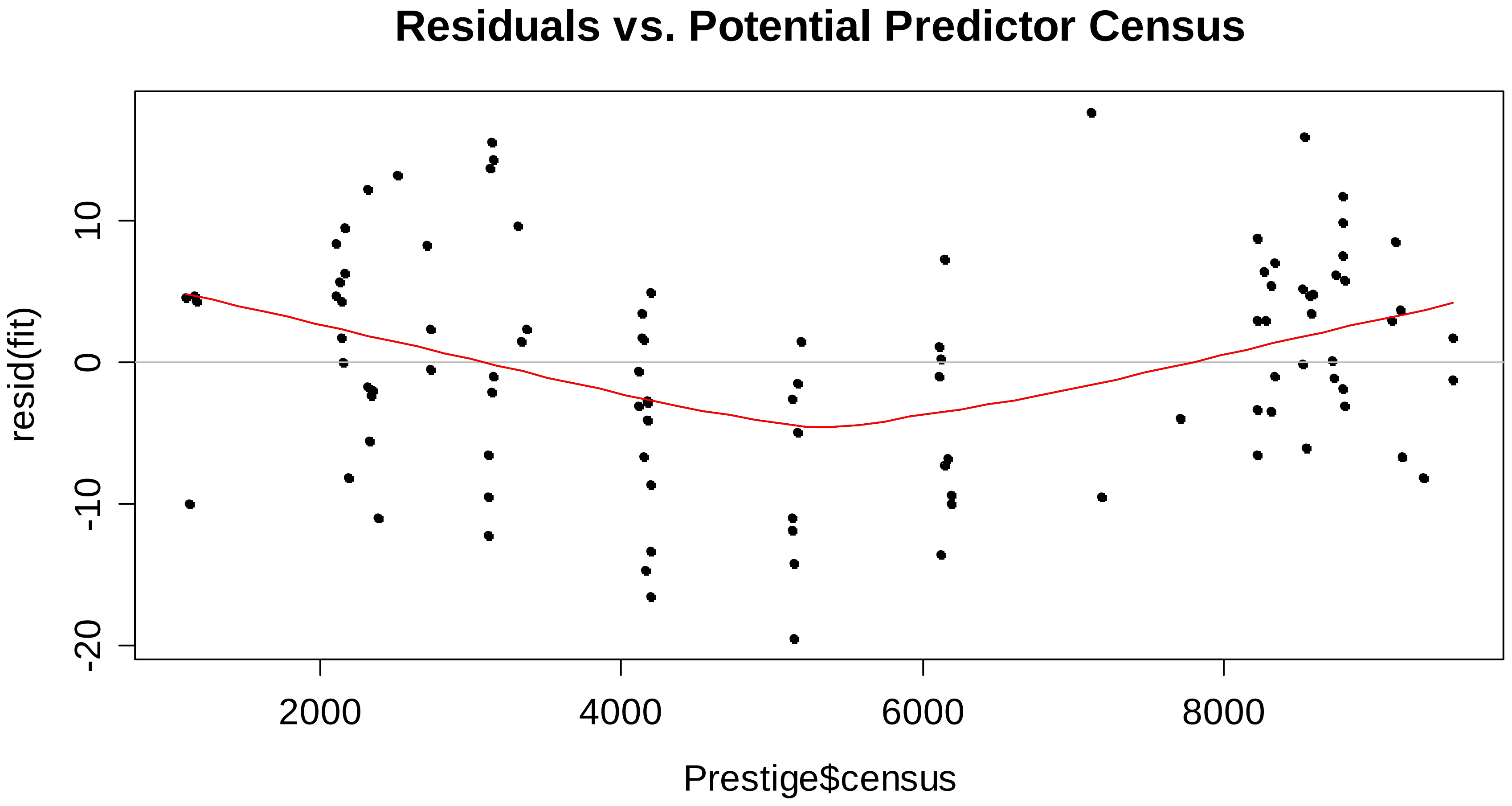
```
                     educ  income  women prest  cens  type
gov.administrators  13.11   12351  11.16  68.8  1113  prof
general.managers    12.26   25879   4.02  69.1  1130  prof
accountants         12.77    9271  15.70  63.4  1171  prof
```

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

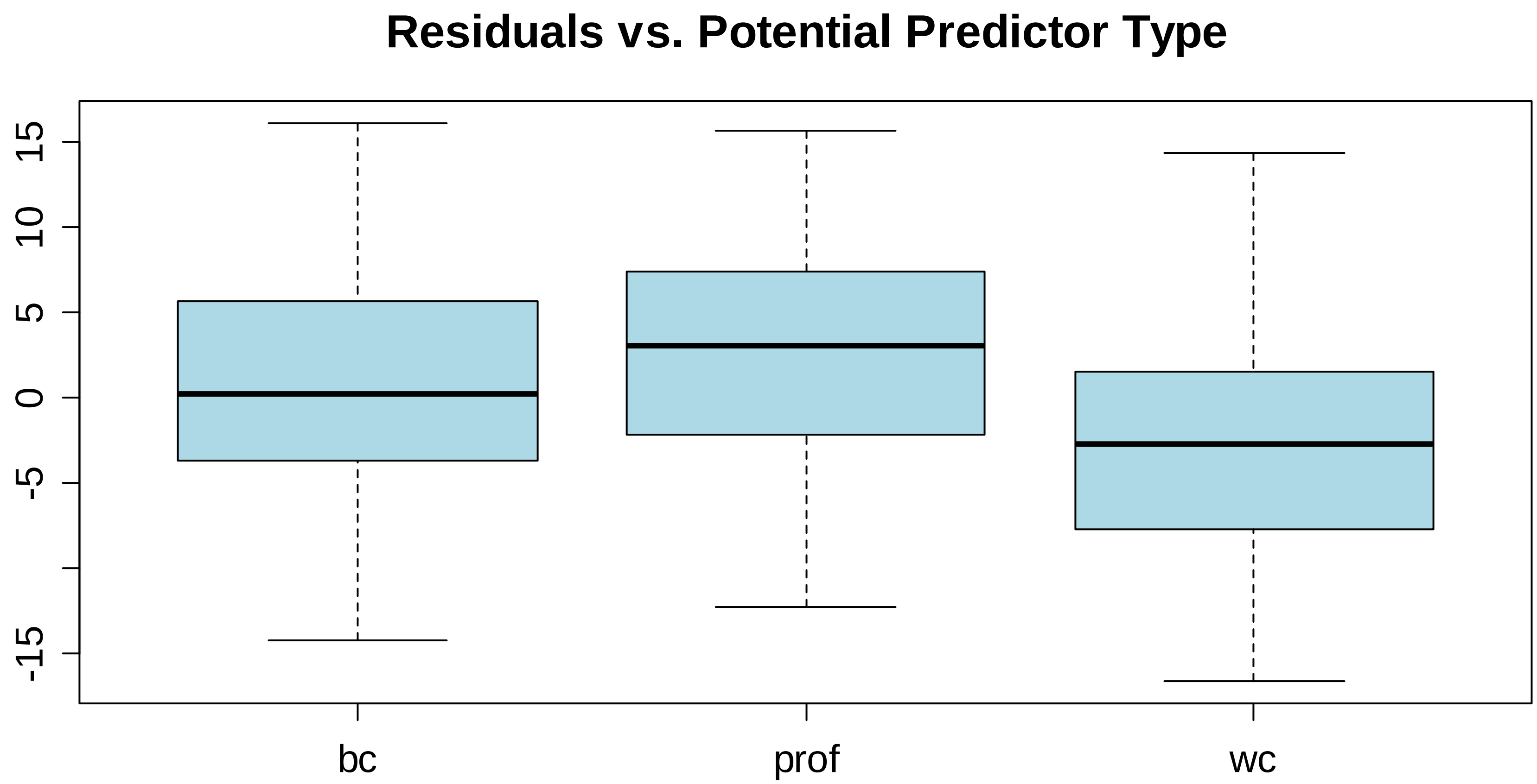
Residuals vs. (Potential) Predictors



Residuals vs. Potential Predictors



Residuals vs. Potential Predictors



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
AS 2012 – Week 08

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

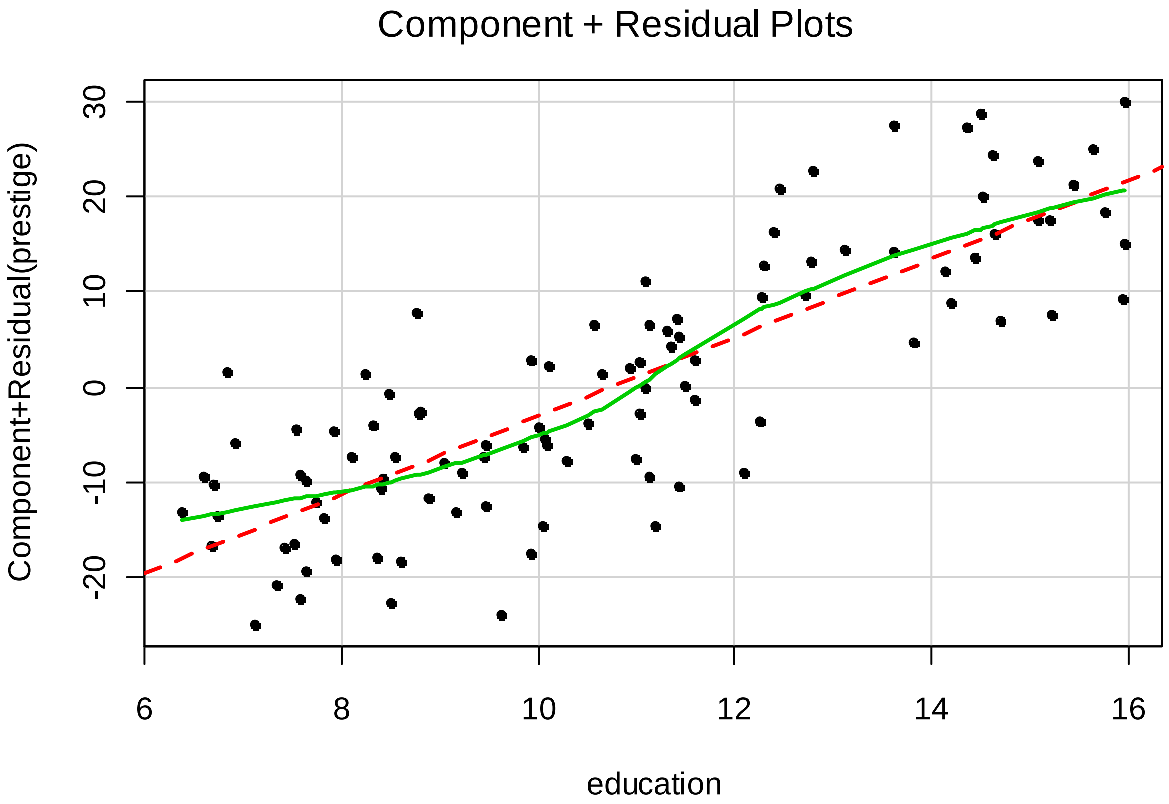
We try to predict the prestige of a number of 102 different profession with a set of 2 predictors:

```
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
...
```

Partial Residual Plots: Example

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

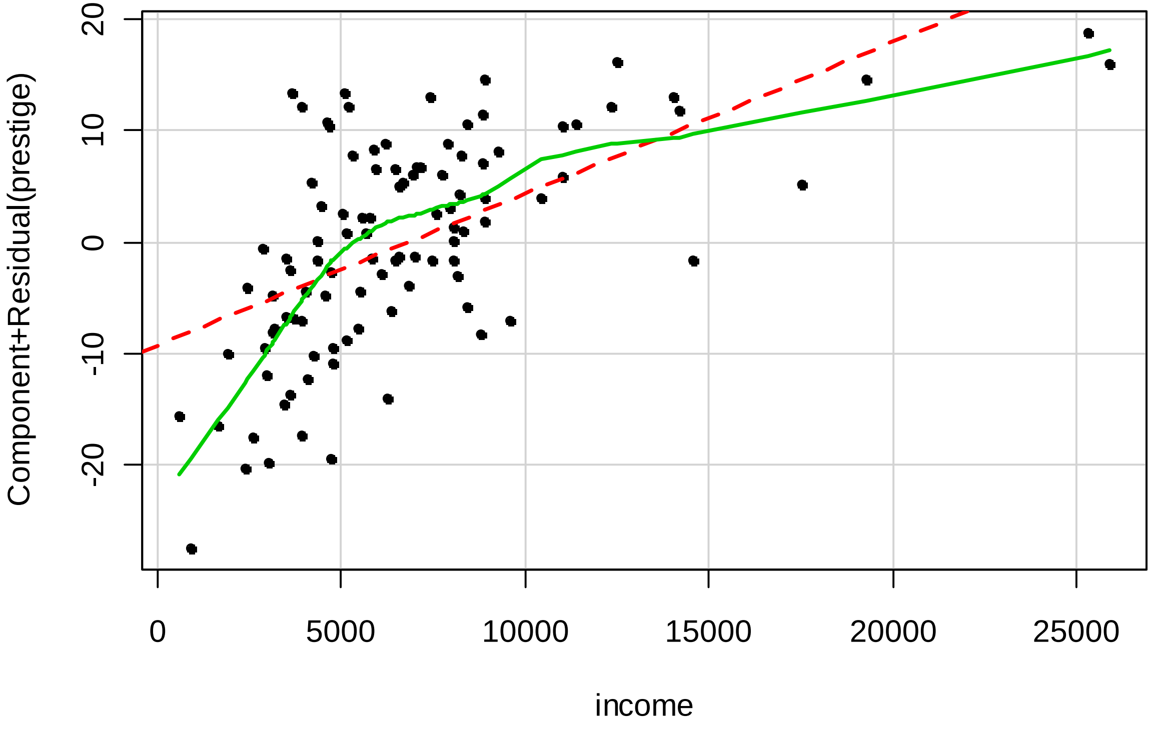


Applied Statistical Regression

AS 2012 – Week 08

Partial Residual Plots: Example

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



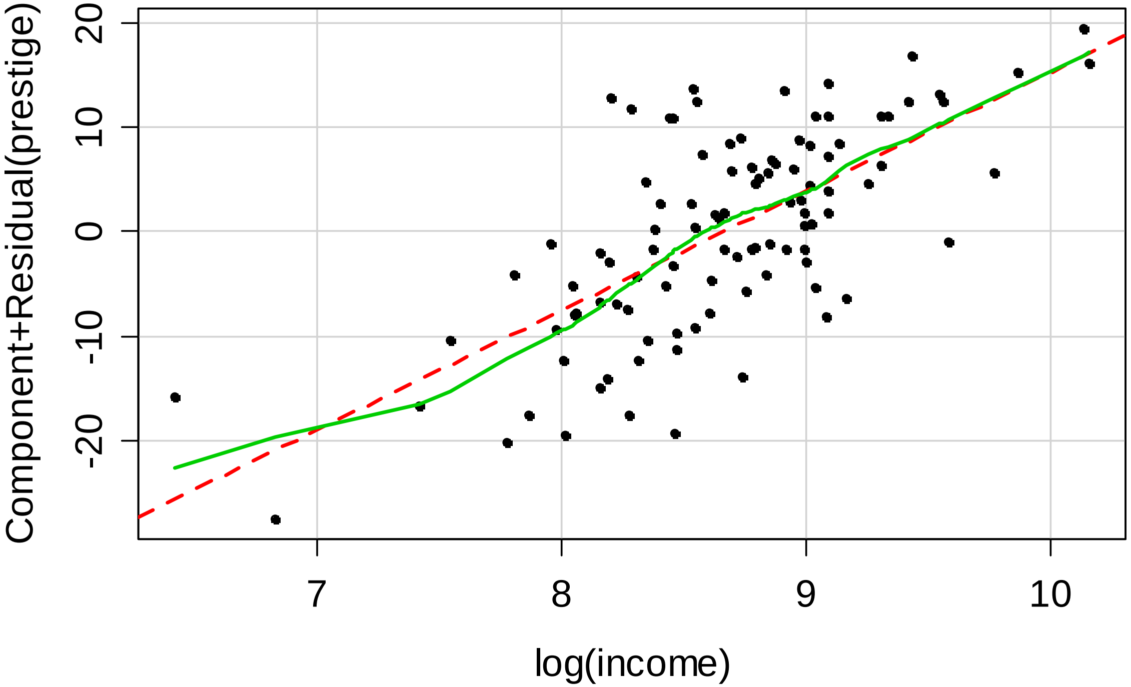
Evident non-linear influence of income on prestige.

→ not a good fit!
→ correction needed

Applied Statistical Regression
AS 2012 – Week 08

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

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)**

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:

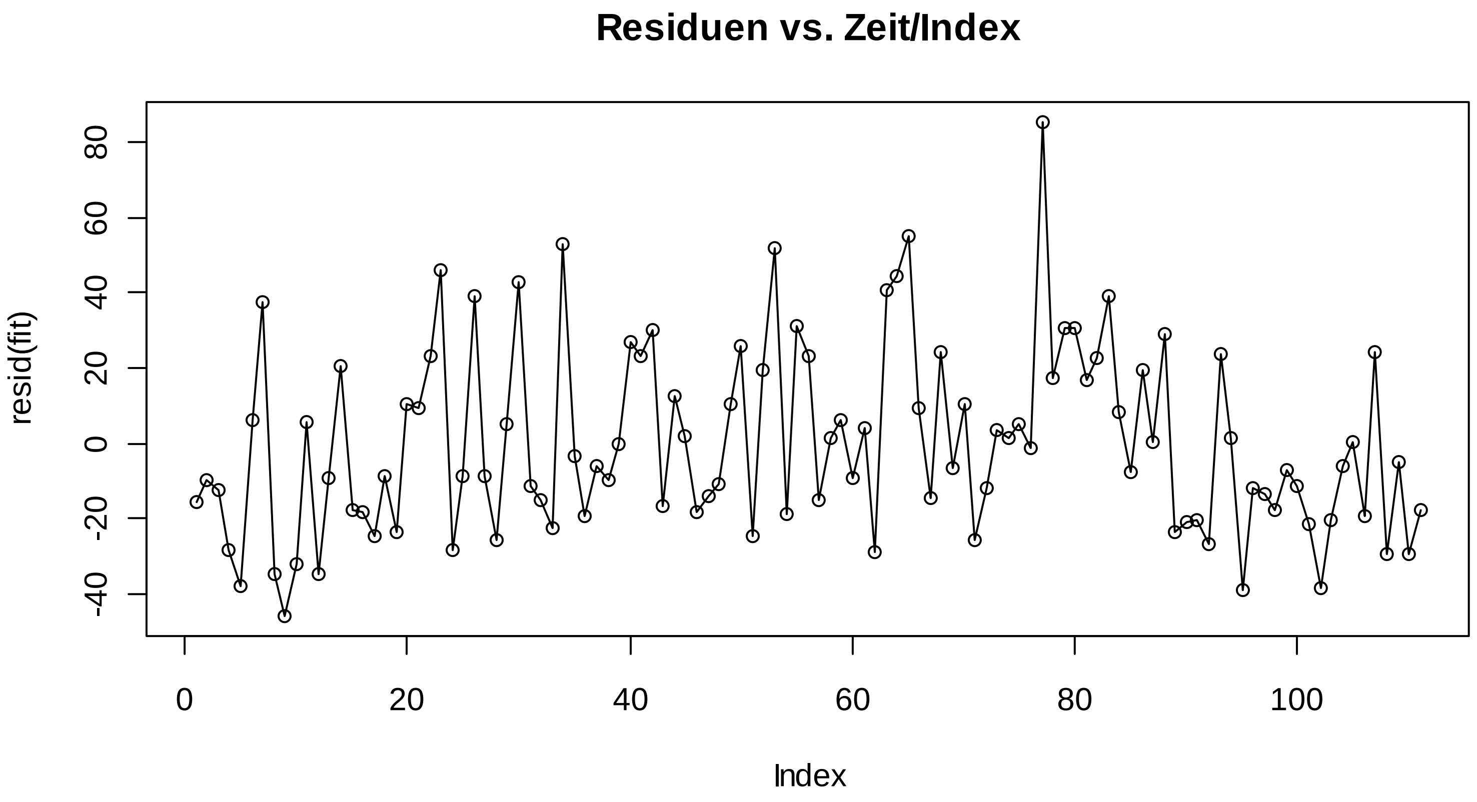
- `library(faraway); data(airquality)`
- `Ozone ~ Solar.R + Wind`
- Measurements from 153 consecutive days in New York
- data have a timely sequence

→ **to be handled with care!**

Applied Statistical Regression
AS 2012 – Week 08

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.

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!

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**.

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 skewed errors, then weighted regression is not the right tool. In that case, better search for a response/predictor transformation.

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 constant variance. The covariance matrix can thus be:

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

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.

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...

Robust Regression

When to use?

Robust regression is used if the residuals are symmetrically distributed and have expectation zero, but are more heavy-tailed than the Gaussian distribution suggests.

Be careful:

If long-tailed resdiuals appear in conjunction with a non-idle Tukey-Anscombe-Plot, and/or with non-constant variance, or if the residuals are skewed, then applying transformations is more appropriated than using robust regression.

Also if there are a few gross outliers, it's better to study these in detail, rather than just applying robust regression.

Robust Regression: Model

The model in robust regression is:

$$y = X\beta + E \text{, where } E \sim^{!} N(0, \sigma_E^2 \Sigma)$$

→ The errors are assumed to be symmetrically distributed, but more heavy-tailed than the Gaussian.

→ In this case, the LS-method is no longer optimal/efficient. There are better estimators for the regression coefficients.

→ Short-tailed errors do not need special attention. In such cases, it is fine to apply the ordinary LS method.

Robust Regression: Idea

In robust regression, observations with large residuals obtain a smaller weight. This is implemented by using a modified "loss function", i.e. no longer the LS-criterion, that measures the quality of the fit:

$$\sum_{i=1}^{n} \rho(r_i), \ where \ \rho(x) = \begin{cases} x^2/2 & if \ |x| \le c \\ c|x| - c^2/2 & if \ |x| > c \end{cases}$$

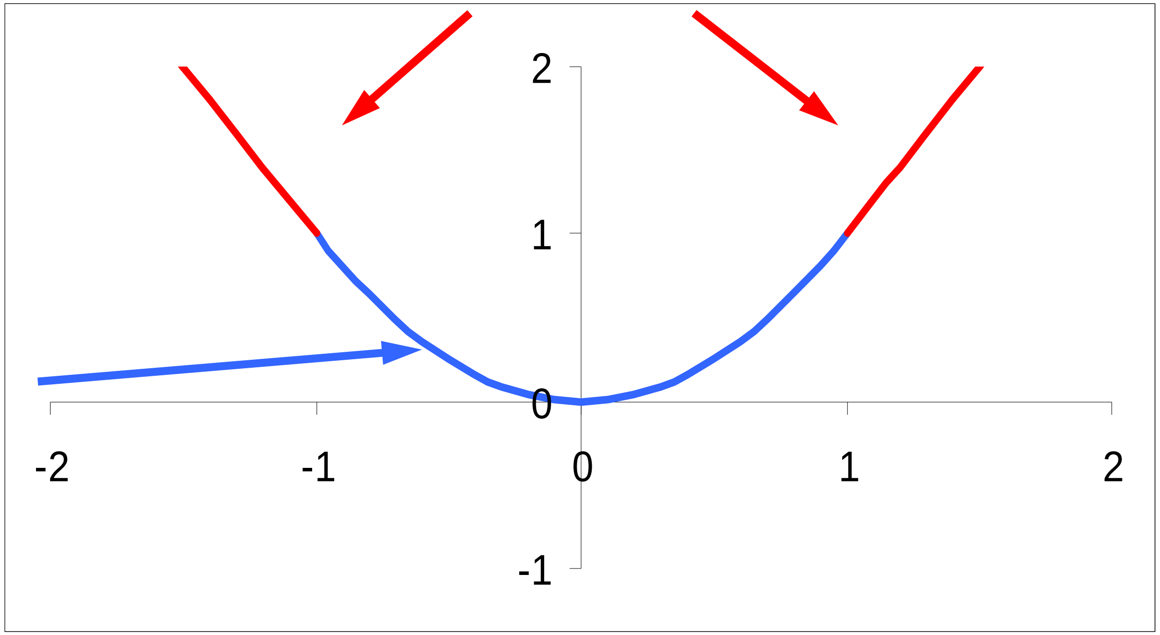
Visualization: see next slide!

There is no solution which can be written in closed form, and an optimization procedure needs to be employed. This is done by solving iteratively reweighted least squares regressions.

Huber Loss Function

This function is used as the default in R-function `rlm()` from `library(MASS)`. There are many other suggestions…

linear



quadratic

Robust Regression: R-Code

```
> library(MASS)

> fit.rlm <- rlm(Mortality ~ JanTemp + … + log(SO2), data=…)
```

→ This uses the Huber loss function

→ The summary is different!

```
summary(fit.rlm)

Coefficients: Value Std. Error    t value

(Intercept) 945.4414  251.6184     3.7574

JanTemp       -1.2313   0.6788     -1.8139

log(SO2)      13.0484   4.6444      2.8095

---

Residual standard error: 30.17 on 46 degrees of freedom
```