Applied Statistical Regression
AS 2012 – 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 5, 2012

Applied Statistical Regression
AS 2012 – Week 07

Multiple Linear Regression

We use linear modeling for a multiple predictor regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + E$$

- there are now p predictors
- the problem cannot be visualized in a scatterplot
- there will be n observations of response and predictors
- goal: estimating the coefficients $\beta_0, \beta_1, ..., \beta_p$ from the data

IMPORTANT: simple linear regression of the response on each of the predictors does not equal multiple regression, where *all predictors are used simultanously*.

Versatility of Multiple Linear Regression

Despite that we are using linear models only, we have a versatile and powerful tool. While the response is always a continuous variable, different predictor types are allowed:

- **Continuous Predictors**
  Default case, e.g. *temperature*, *distance*, *pH-value*, …
- **Transformed Predictors**
  For example: $log(x), sqrt(x), arcsin(\sqrt{x}),...$
- **Powers**
  We can also use: $x^{-1}, x^2, x^3, ...$
- **Categorical Predictors**
  Often used: *sex*, *day of week*, *political party*, …

Categorical Predictors

The canonical case in linear regression are *continuous predictor variables* such as for example:

→ *temperature, distance, pressure, velocity, ...*

While in linear regression, we cannot have categorical response, it is perfectly valid to have *categorical predictors*:

→ *yes/no, sex (m/f), type (a/b/c), shift (day/evening/night), ...*

Such categorical predictors are often also called **factor variables**. In a linear regression, each level of such a variable is encoded by a dummy variable, so that $(\ell - 1)$ degrees of freedom are spent.

Example: Binary Categorical Variable

The lathe (*in German: Drehbank*) **dataset:**

- y lifetime of a cutting tool in a turning machine
- x_1 speed of the machine in rpm
- x_2 tool type A or B

Dummy variable encoding:

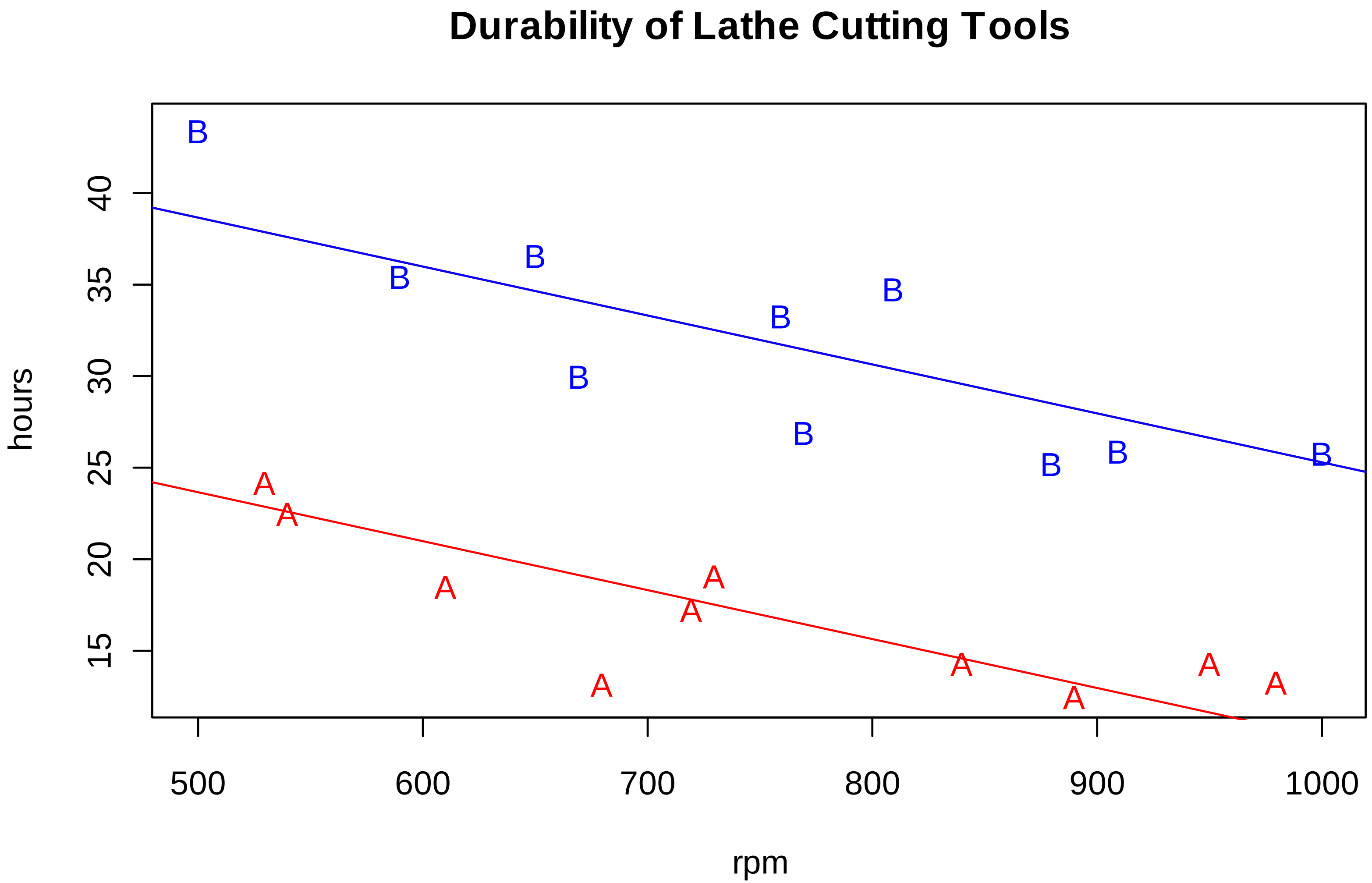
$$x_2 = \begin{cases} 0 & tool \ \ type \ \ A \\ 1 & tool \ \ type \ \ B \end{cases}$$

Interpretation of the Model

→ **see blackboard…**

```
> summary(lm(hours ~ rpm + tool, data = lathe))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.98560    3.51038  10.536 7.16e-09 ***
rpm         -0.02661    0.00452  -5.887 1.79e-05 ***
toolB       15.00425    1.35967  11.035 3.59e-09 ***
---
Residual standard error: 3.039 on 17 degrees of freedom
Multiple R-squared: 0.9003,  Adjusted R-squared: 0.8886
F-statistic: 76.75 on 2 and 17 DF,   p-value: 3.086e-09
```

The Dummy Variable Fit



A Model with Interactions

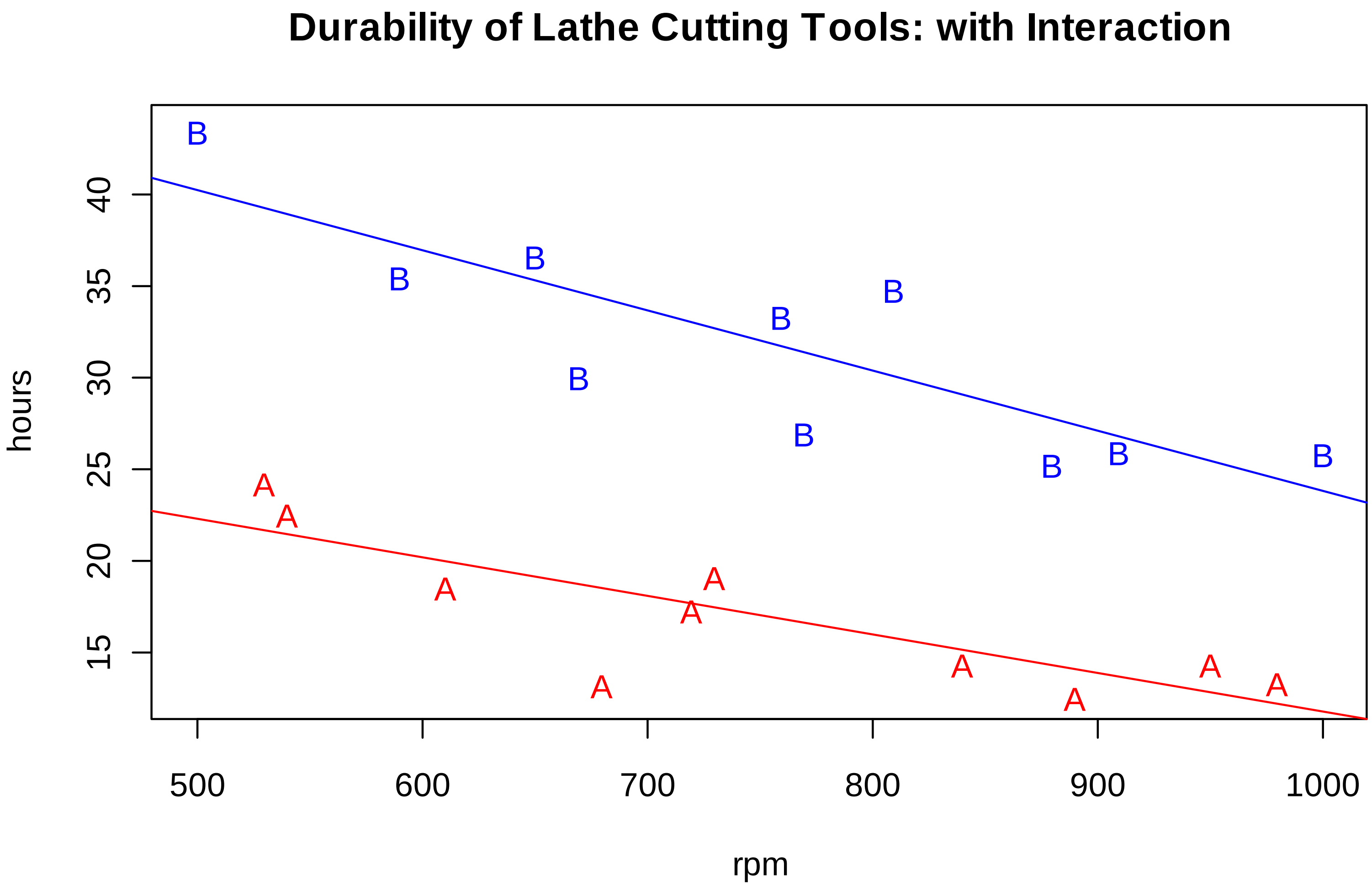
Question: do the slopes need to be identical?

→ with the appropriate model, the answer is no!

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + E$$

→ **see blackboard for model interpretation…**

Different Slopes for the Regression Lines



Summary Output

```
> summary(lm(hours ~ rpm * tool, data = lathe))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.774760   4.633472   7.073 2.63e-06 ***
rpm         -0.020970   0.006074  -3.452  0.00328 **
toolB       23.970593   6.768973   3.541  0.00272 **
rpm:toolB   -0.011944   0.008842  -1.351  0.19553
---
Residual standard error: 2.968 on 16 degrees of freedom
Multiple R-squared: 0.9105,  Adjusted R-squared: 0.8937
F-statistic: 54.25 on 3 and 16 DF,  p-value: 1.319e-08
```

How Complex the Model Needs to Be?

Question 1: do we need different slopes for the two lines?

$$H_0 : \beta_3 = 0 \quad \text{against} \quad H_A : \beta_3 \neq 0$$

→ no, see individual test for the interaction term on previous slide!

Question 2: is there any difference altogether?

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{against} \quad H_A : \beta_2 \neq 0 \;\; and\,/\,or \;\; \beta_3 \neq 0$$

→ this is a hierarchical model comparison
→ we try to exclude interaction and dummy variable together

R offers convenient functionality for this test, see next slide!

Testing the Tool Type Variable

Hierarchical model comparison with anova():

```
> fit.small <- lm(hours ~ rpm, data=lathe)
> fit.big <- lm(hours ~ rpm * tool, data=lathe)
> anova(fit.small, fit.big)
Model 1: hours ~ rpm
Model 2: hours ~ rpm * tool
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     18 1282.08
2     16  140.98  2    1141.1 64.755 2.137e-08 ***
```

→ The bigger model, i.e. making a distinction between the tools, is significantly better. The main effect is enough, though.

Categorical Input with More Than 2 Levels

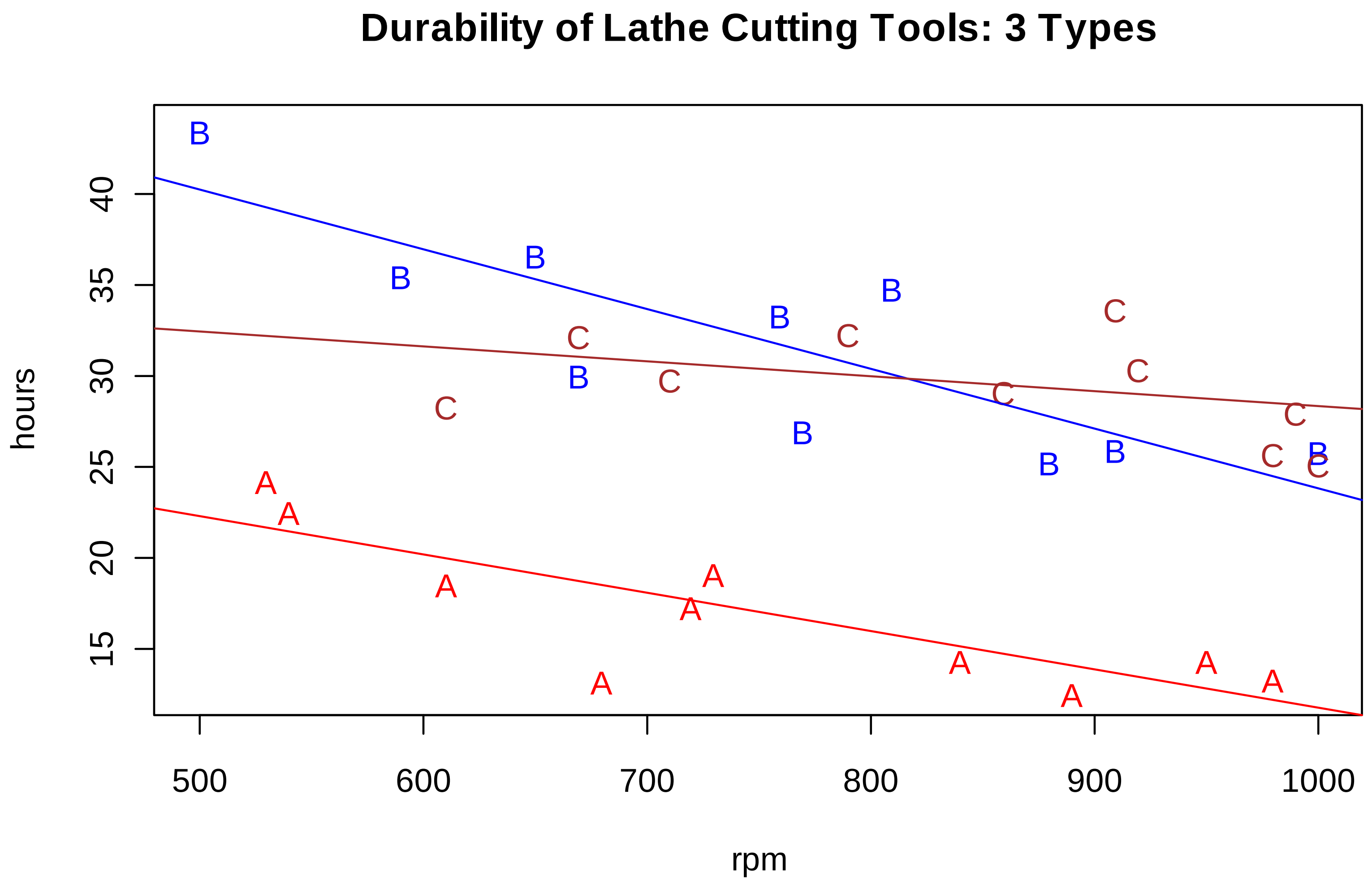
There are now 3 tool types A, B, C:

x_2	x_3	
0	0	*for observations of type A*
1	0	*for observations of type B*
0	1	*for observations of type C*

Main effect model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + E$

With interactions: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + E$

Three Types of Cutting Tools



Summary Output

```
> summary(lm(hours ~ rpm * tool, data = abc.lathe)

Coefficients:Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.774760   4.496024   7.290 1.57e-07 ***
rpm         -0.020970   0.005894  -3.558  0.00160 **
toolB       23.970593   6.568177   3.650  0.00127 **
toolC        3.803941   7.334477   0.519  0.60876
rpm:toolB   -0.011944   0.008579  -1.392  0.17664
rpm:toolC    0.012751   0.008984   1.419  0.16869
---
Residual standard error: 2.88 on 24 degrees of freedom
Multiple R-squared: 0.8906,    Adjusted R-squared: 0.8678
F-statistic: 39.08 on 5 and 24 DF,  p-value: 9.064e-11
```

This summary is of limited use for deciding about model complexity. We require hierarchical model comparisons!

Inference with Categorical Predictors

Do not perform individual hypothesis tests on factors that have more than 2 levels, they are meaningless!

Question 1: do we have different slopes?

$$H_0 : \beta_4 = 0 \ \ and \ \ \beta_5 = 0 \text{ against } H_A : \beta_4 \neq 0 \ \ and/or \ \ \beta_5 \neq 0$$

Question 2: is there any difference altogether?

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \text{ against } H_A : any\ of \ \ \beta_2, \beta_3, \beta_4, \beta_5 \neq 0$$

→ Again, R provides convenient functionality: `anova()`

Applied Statistical Regression
AS 2012 – Week 07

Anova Output

```
> anova(fit.abc)

Analysis of Variance Table
          Df  Sum Sq Mean Sq F value    Pr(>F)
rpm        1  139.08  139.08 16.7641  0.000415 ***
tool       2 1422.47  711.23 85.7321 1.174e-11 ***
rpm:tool   2   59.69   29.84  3.5974  0.043009 *
Residuals 24  199.10    8.30
```

→ The interaction term is weakly significant. Thus, there is some weak evidence for the necessity of different slopes.

→ The p-value for the tool variable includes omitting interaction and main effect. Being strongly significant, we have strong evidence that tool type distinction is needed.

Polynomial Regression

Polynomial Regression = Multiple Linear Regression !!!

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_d x^d + E$$

Goals:

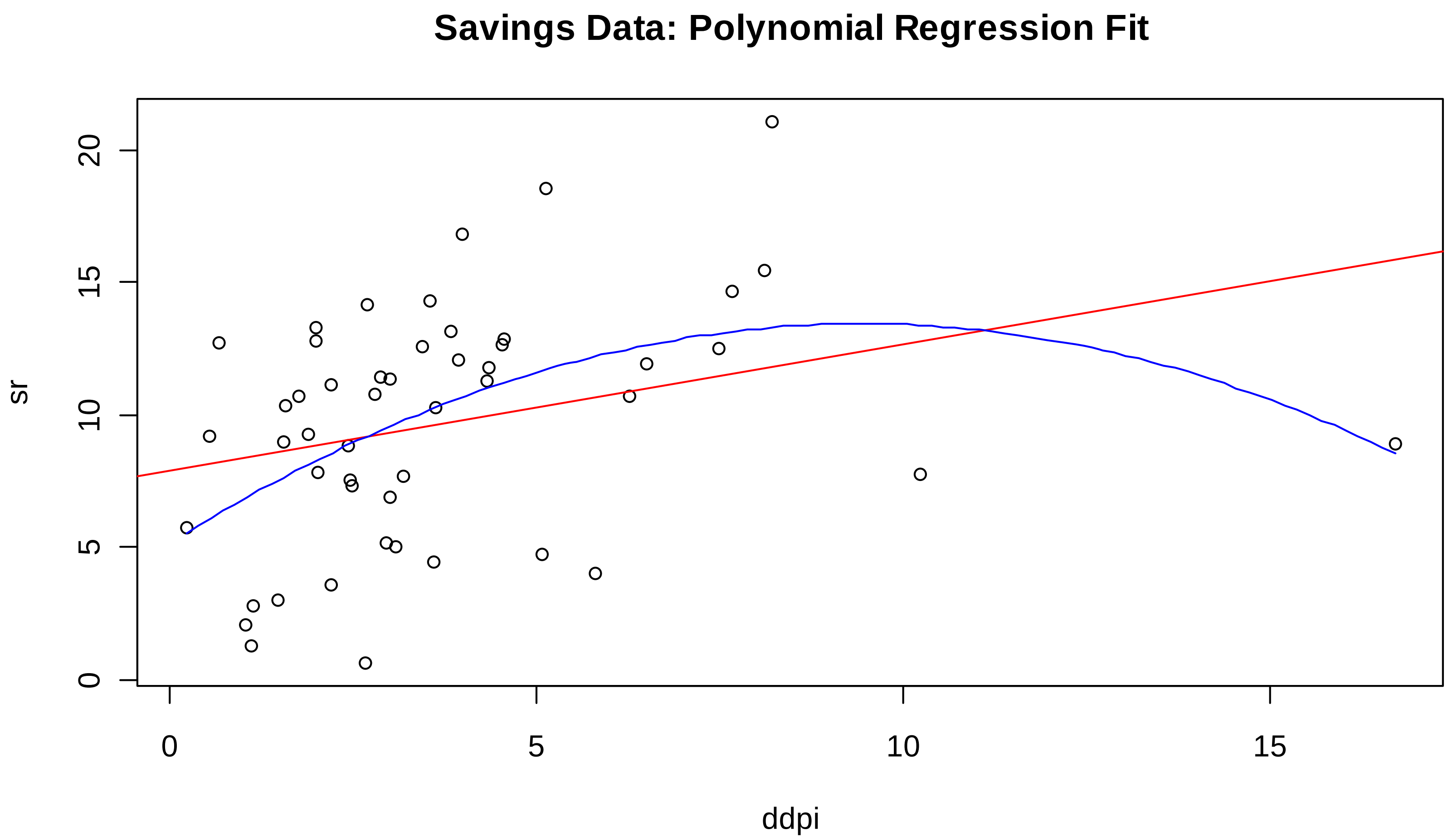
- fit a curvilinear relation
- improve the fit between x and y
- determine the polynomial order d

Example:

- Savings dataset: personal savings ~ income per capita

Applied Statistical Regression
AS 2012 – Week 07

Polynomial Regression Fit



Polynomial Regression

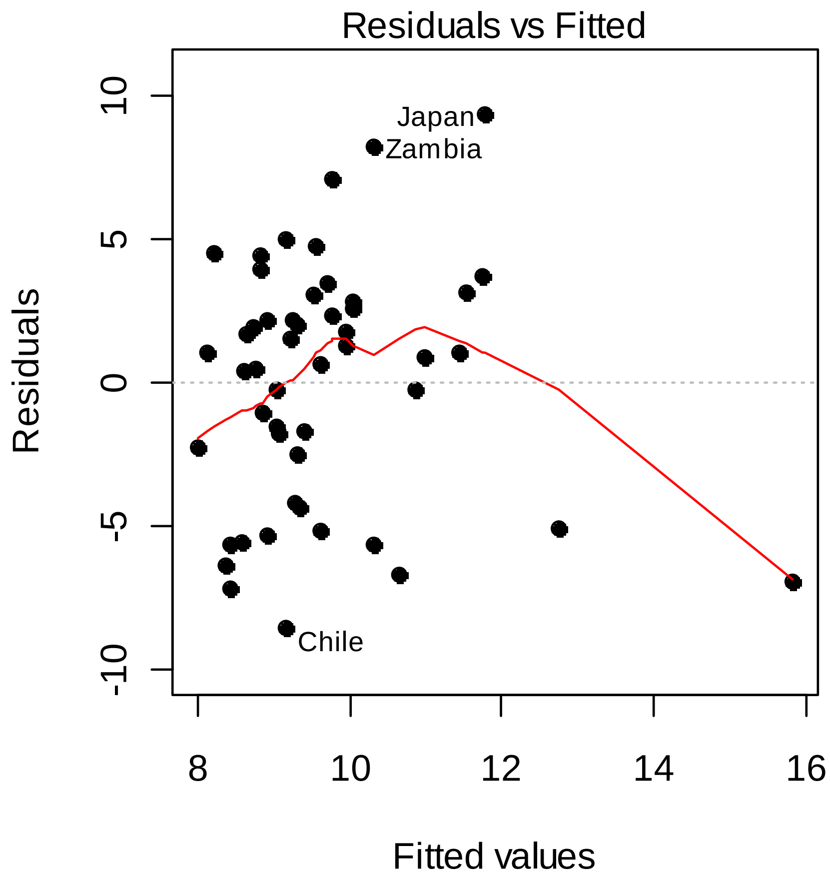
Output from the model with the linear term only:

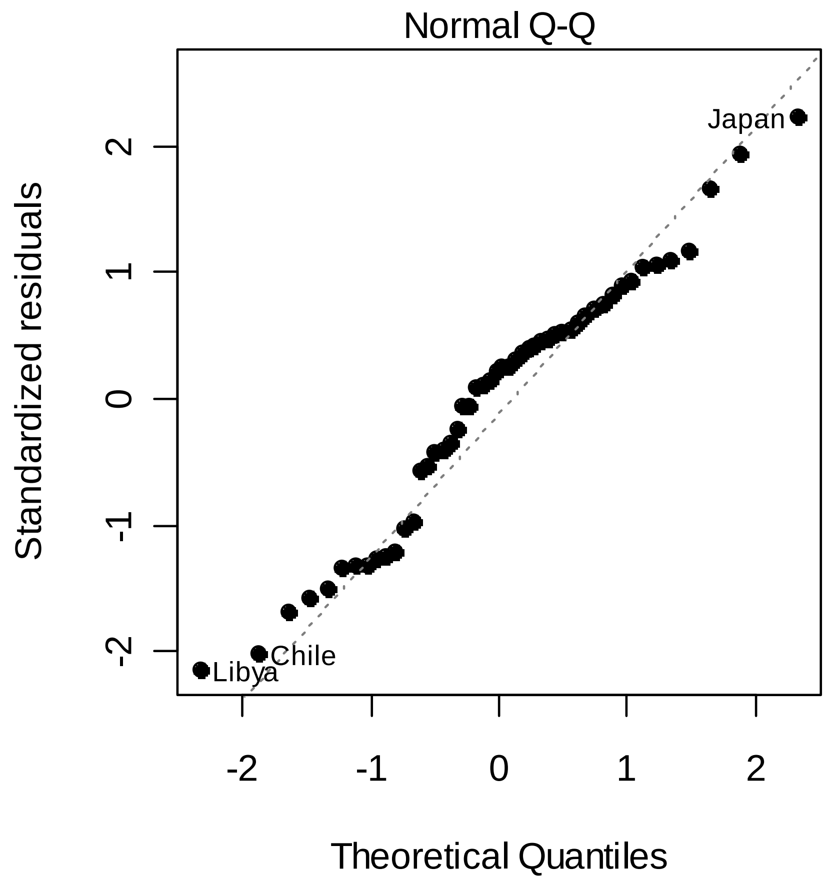
```
> summary(lm(sr ~ ddpi, data = savings))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.8830     1.0110   7.797 4.46e-10 ***
ddpi          0.4758     0.2146   2.217   0.0314 *
---
Residual standard error: 4.311 on 48 degrees of freedom
Multiple R-squared: 0.0929, Adjusted R-squared: 0.074
F-statistic: 4.916 on 1 and 48 DF,  p-value: 0.03139
```

Applied Statistical Regression
AS 2012 – Week 07

Diagnostic Plots





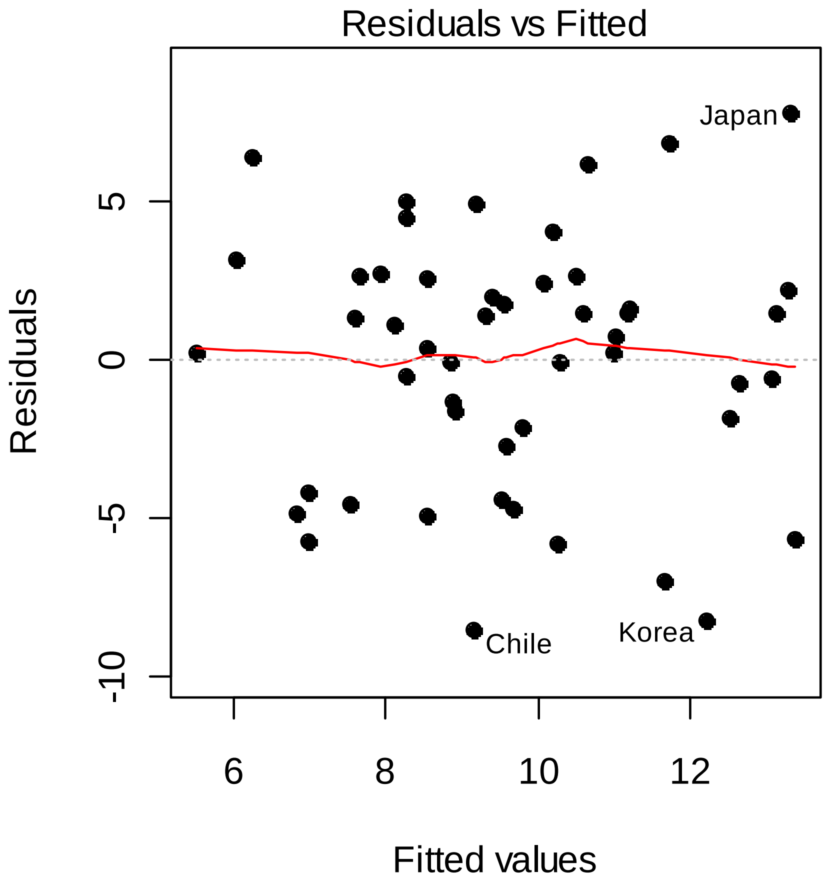
Quadratic Regression

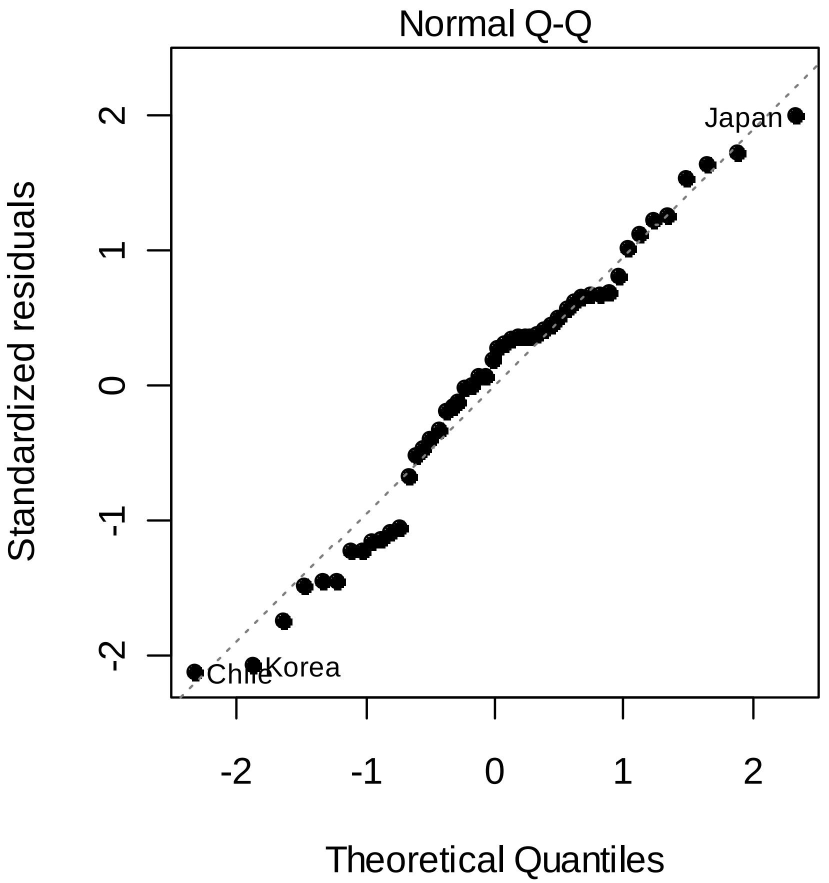
Add the quadratic term: $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + E$

```
> summary(lm(sr ~ ddpi + I(ddpi^2), data = savings))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.13038    1.43472   3.576 0.000821 ***
ddpi         1.75752    0.53772   3.268 0.002026 **
I(ddpi^2)   -0.09299    0.03612  -2.574 0.013262 *
---
Residual standard error: 4.079 on 47 degrees of freedom
Multiple R-squared: 0.205, Adjusted R-squared: 0.1711
F-statistic: 6.059 on 2 and 47 DF,  p-value: 0.004559
```

Diagnostic Plots: Quadratic Regression





Cubic Regression

Add the cubic term: $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + E$

```
> summary(lm(sr~ddpi + I(ddpi^2) + I(ddpi^3), data = savings)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.145e+00  2.199e+00   2.340   0.0237 *
ddpi         1.746e+00  1.380e+00   1.265   0.2123
I(ddpi^2)   -9.097e-02  2.256e-01  -0.403   0.6886
I(ddpi^3)   -8.497e-05  9.374e-03  -0.009   0.9928
---
Residual standard error: 4.123 on 46 degrees of freedom
Multiple R-squared: 0.205,     Adjusted R-squared: 0.1531
F-statistic: 3.953 on 3 and 46 DF,  p-value: 0.01369
```

Powers Are Strongly Correlated Predictors!

The smaller the x-range, the bigger the problem!

```
> cor(cbind(ddpi, ddpi2=ddpi^2, ddpi3=ddpi^3))
           ddpi     ddpi2     ddpi3
ddpi  1.0000000 0.9259671 0.8174527
ddpi2 0.9259671 1.0000000 0.9715650
ddpi3 0.8174527 0.9715650 1.0000000
```

Way out: use centered predictors!

$$z_i = (x_i - \overline{x})$$

$$z_i^2 = (x_i - \overline{x})^2$$

$$z_i^3 = (x_i - \overline{x})^3$$

Powers Are Strongly Correlated Predictors!

```
> summary(lm(sr~z.ddpi+I(z.ddpi^2)+I(z.ddpi^3),dat=z.savings)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.042e+01  8.047e-01  12.946  < 2e-16 ***
z.ddpi       1.059e+00  3.075e-01   3.443  0.00124 **
I(z.ddpi^2) -9.193e-02  1.225e-01  -0.750  0.45691
I(z.ddpi^3) -8.497e-05  9.374e-03  -0.009  0.99281
```

→ Coefficients, standard error and tests are different
→ Fitted values and global inference remain the same
→ Not overly beneficial on this dataset!

→ **Be careful: extrapolation with polynomials is dangerous!**

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.

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!

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?

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.

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

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:

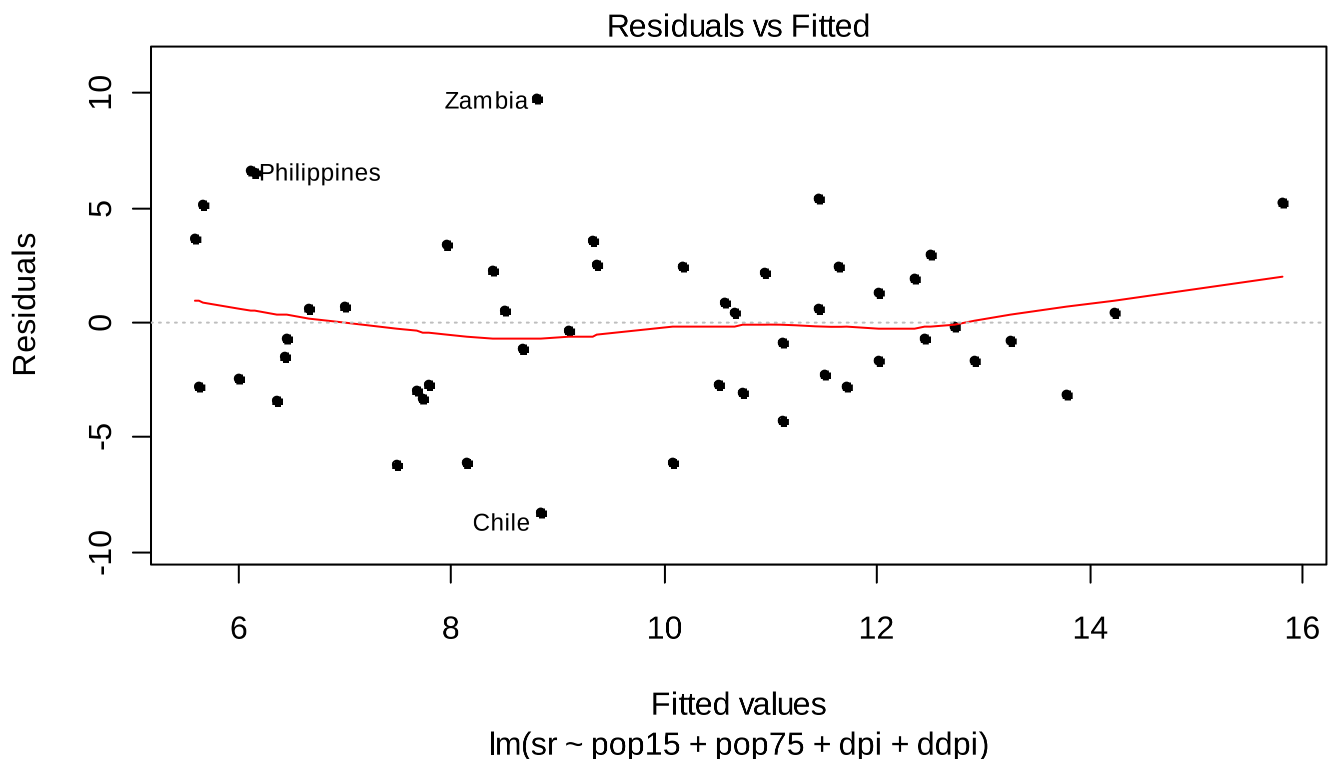
```
lm(sr ~ pop15 + pop75 + dpi + ddpi, data=LifeCycleSavings)
```

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.

Tukey-Anscombe-Plot

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



Tukey-Anscombe-Plot

Is useful for:

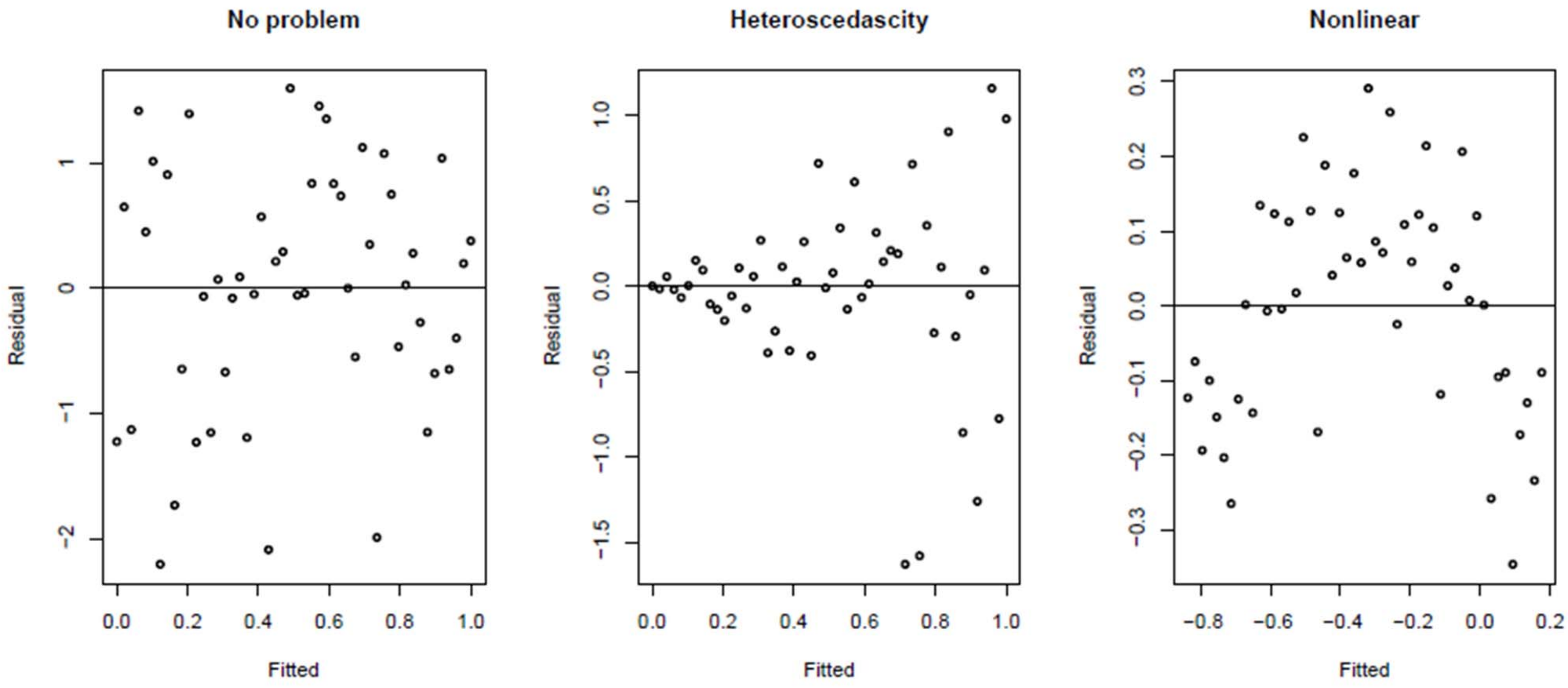
- finding structural model deficiencies, i.e. $E[E_i] \neq 0$
- if that is the case, the response/predictor relation could be nonlinear, or some predictors could be missing
- it is also possible to detect non-constant variance (→ then, the smoother does not deviate from 0)

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 2012 – Week 07

Tukey-Anscombe-Plot



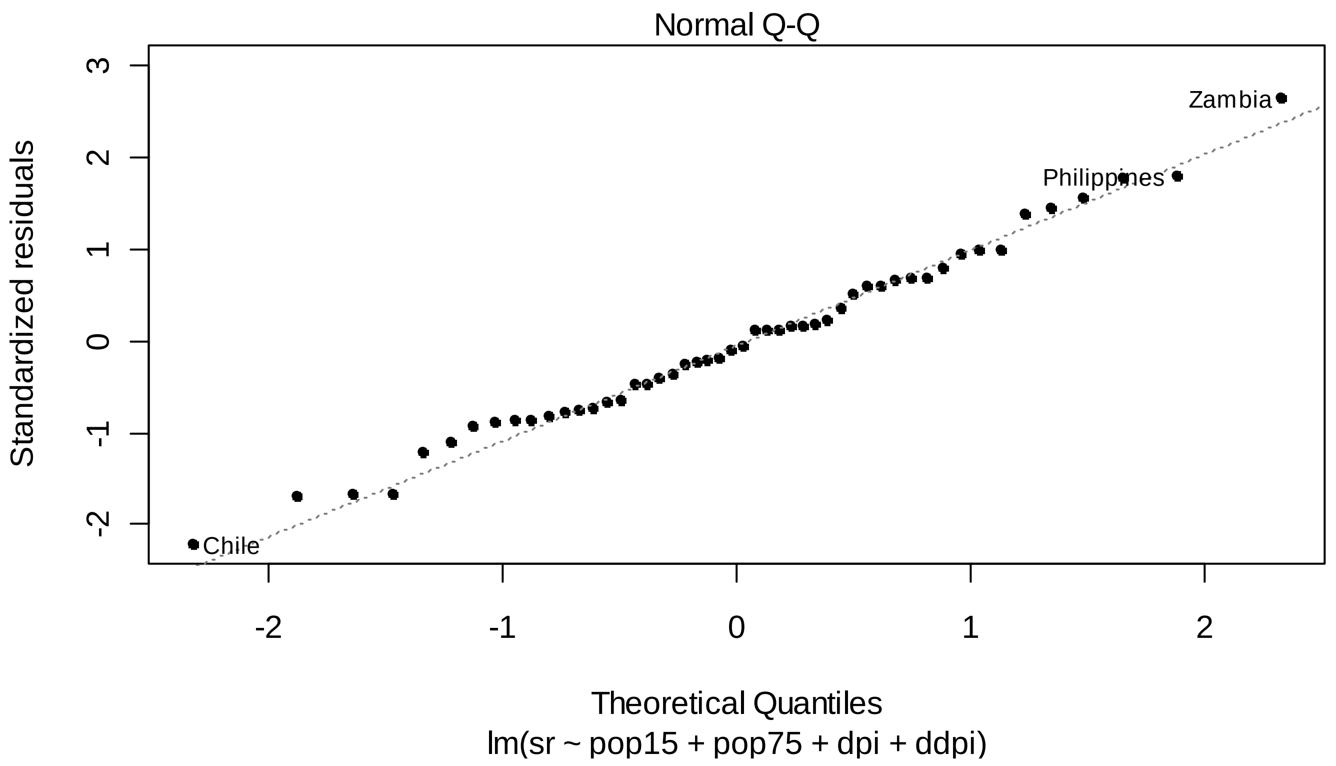
Tukey-Anscombe-Plot

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)`



Normal Plot

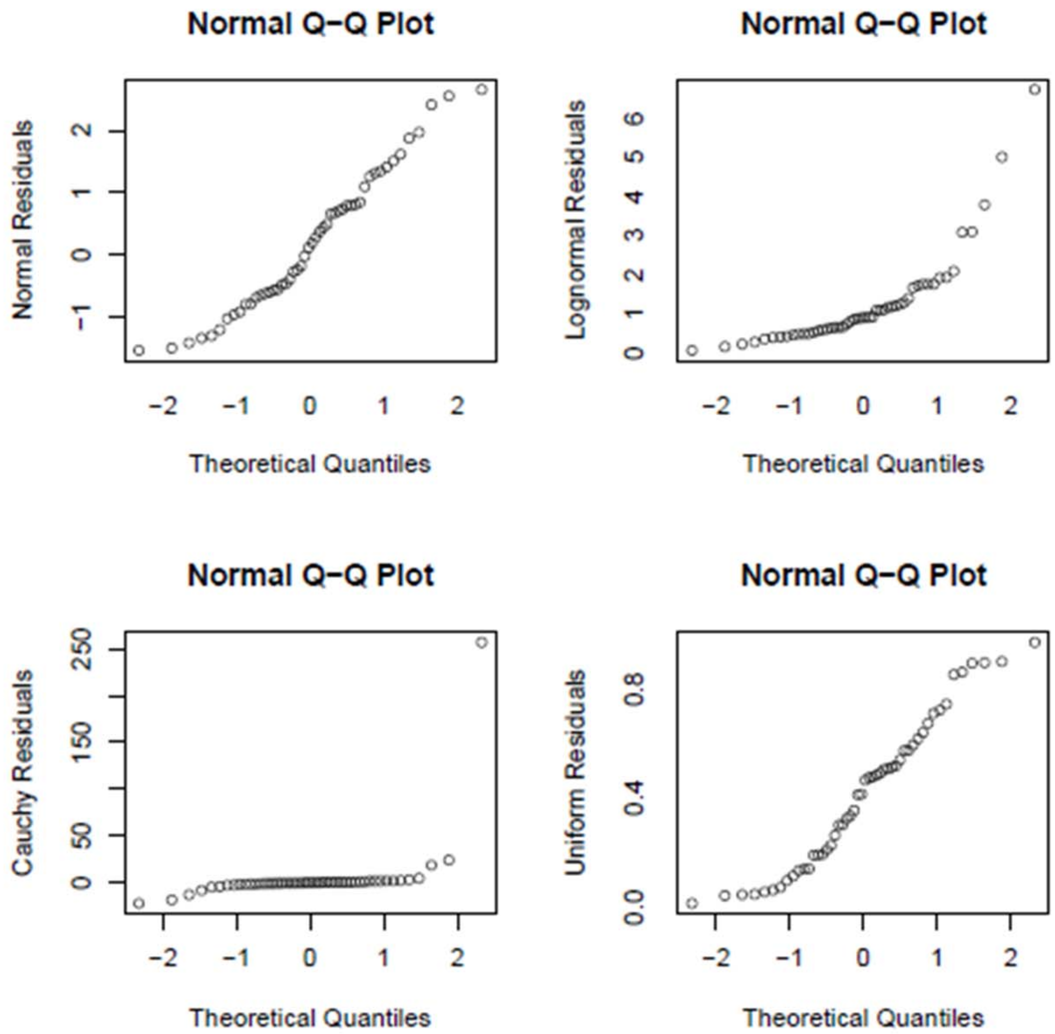
Is useful for:

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

When is the plot OK?

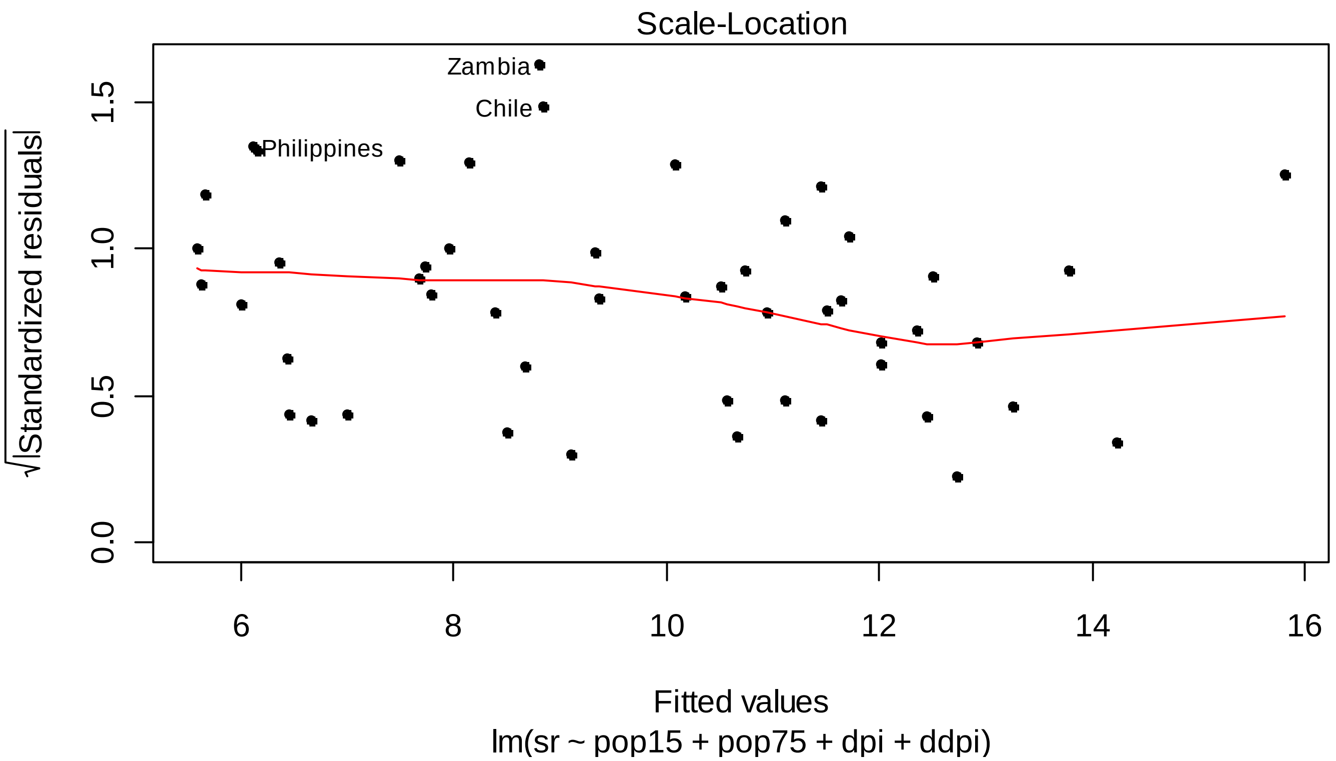
- the residuals $\tilde{r}_i$ must not show any systematic deviation from line which leads to the 1st and 3rd quartile.
- a few data points that are slightly "off the line" near the ends are always encountered and usually tolerable
- skewed residuals need correction: they usually tell that the model 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
Normal Q−Q Plot
Lognormal Residuals
Theoretical Quantiles
Normal Q−Q Plot
Cauchy Residuals
Theoretical Quantiles
Normal Q−Q Plot
Uniform Residuals
Theoretical Quantiles

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?

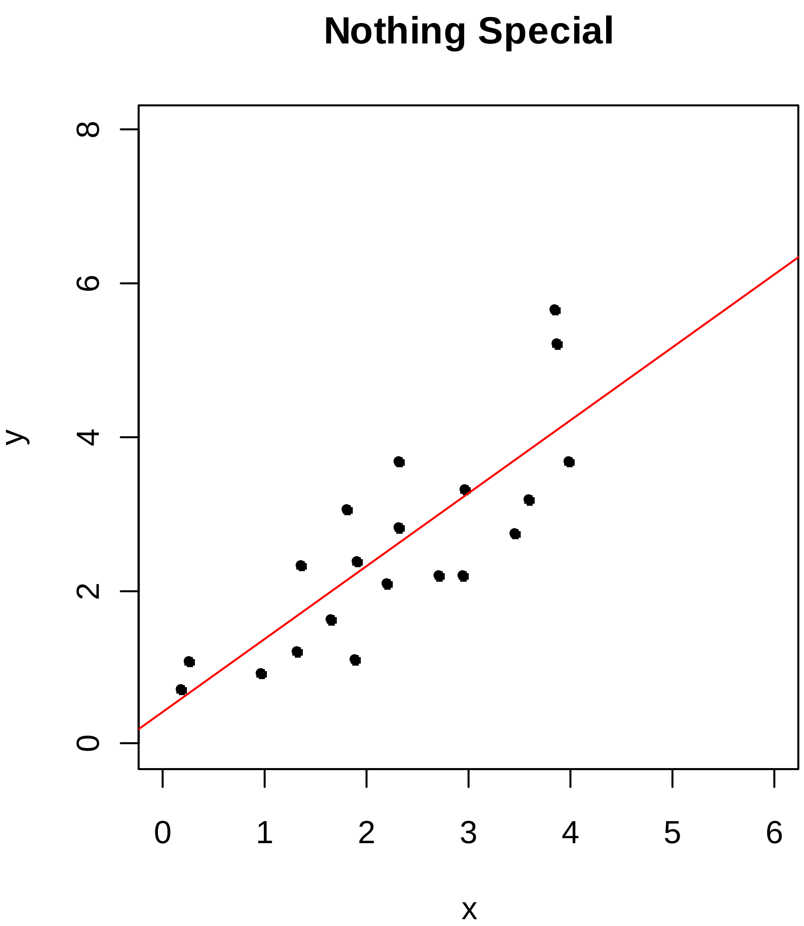
- the smoother line runs horizontally along the x-axis, without any systematic deviations.

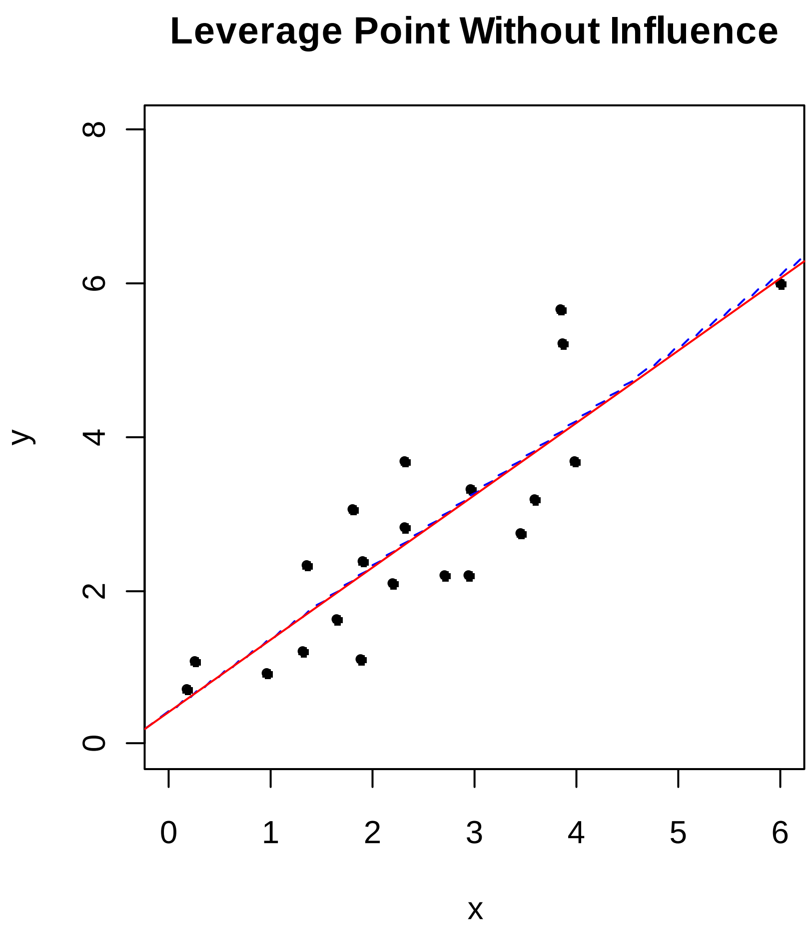
Unusual Observations

- 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
AS 2012 – Week 07

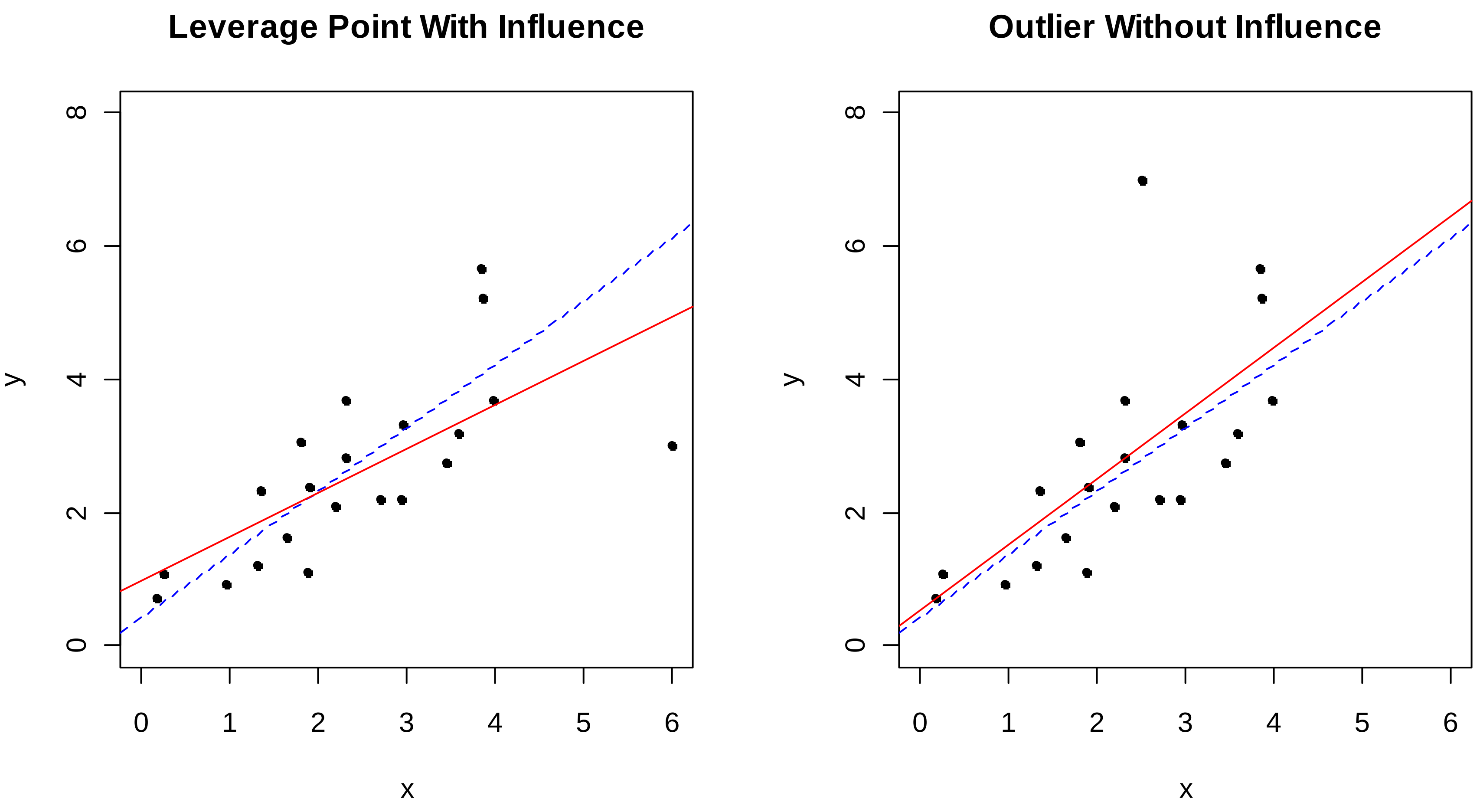
Unusual Observations





Applied Statistical Regression
AS 2012 – 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.

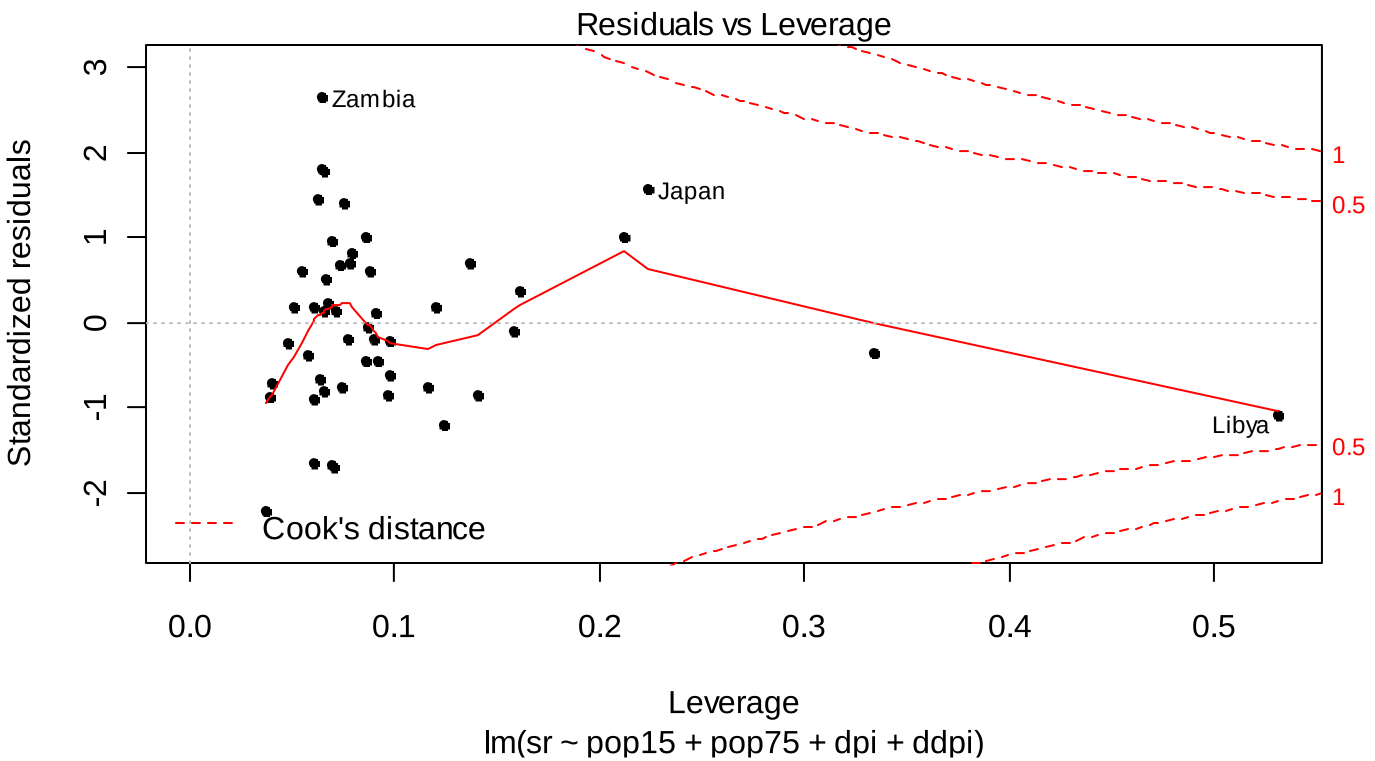
3) **Cook‘s Distance**

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

   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.