Applied Statistical Regression
HS 2011 – Week 06

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 1, 2011

Applied Statistical Regression
HS 2011 – Week 06

Dummy Variables

So far, we only considered continuous predictors:

- temperature
- distance
- pressure
- …

It is perfectly valid to have categorical predictors, too:

- sex (male or female)
- status variables (employed or unemployed)
- working shift (day, evening, night)
- …

→ **Implementation in the regression with dummy variables**

Example: Binary Categorical Variable

The lathe dataset:

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

Dummy variable encoding:

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

Applied Statistical Regression
HS 2011 – Week 06

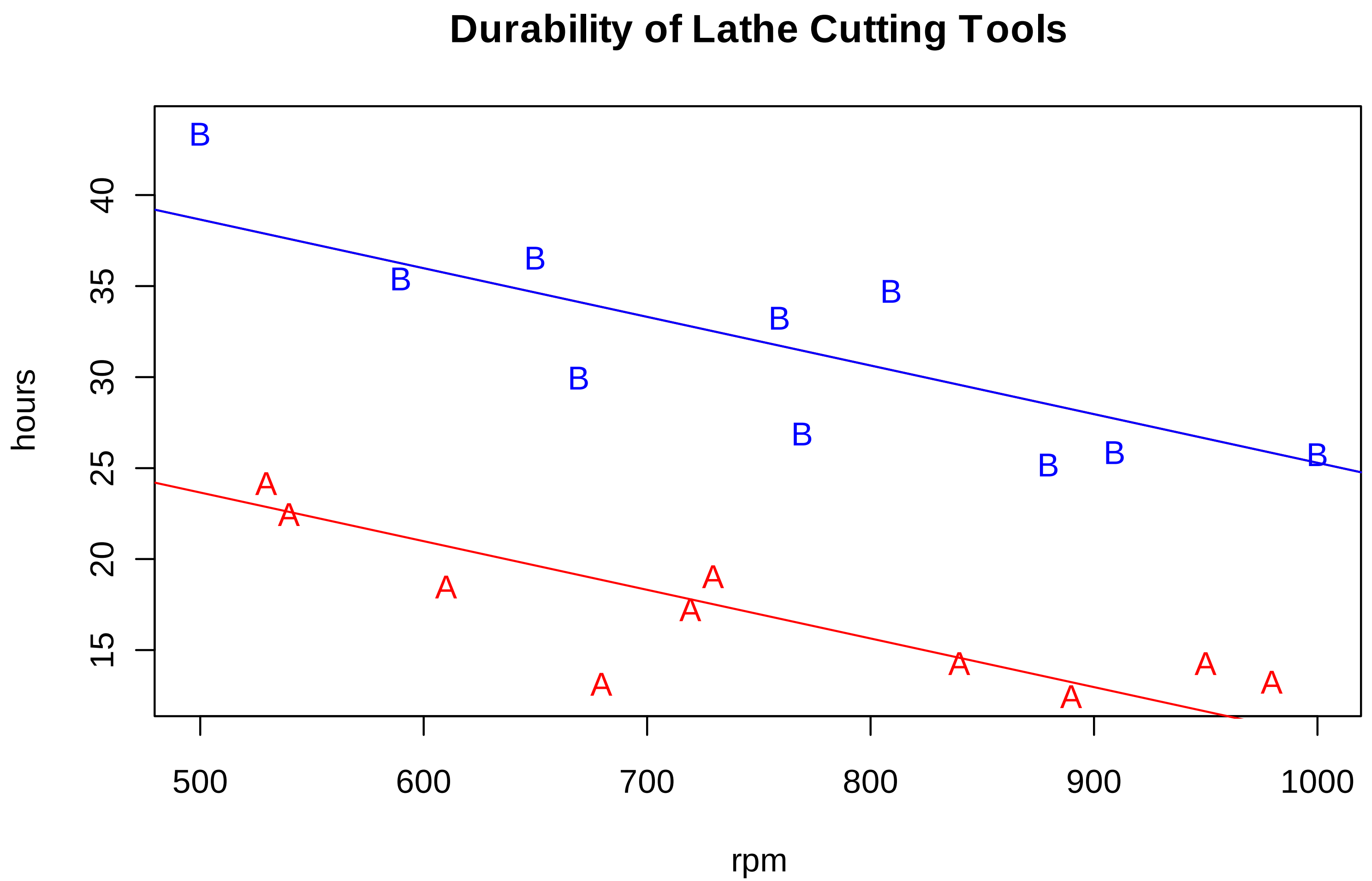
Interpretation of the Model

→ 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
```

Applied Statistical Regression
HS 2011 – Week 06

The Dummy Variable Fit



A Model with Interactions

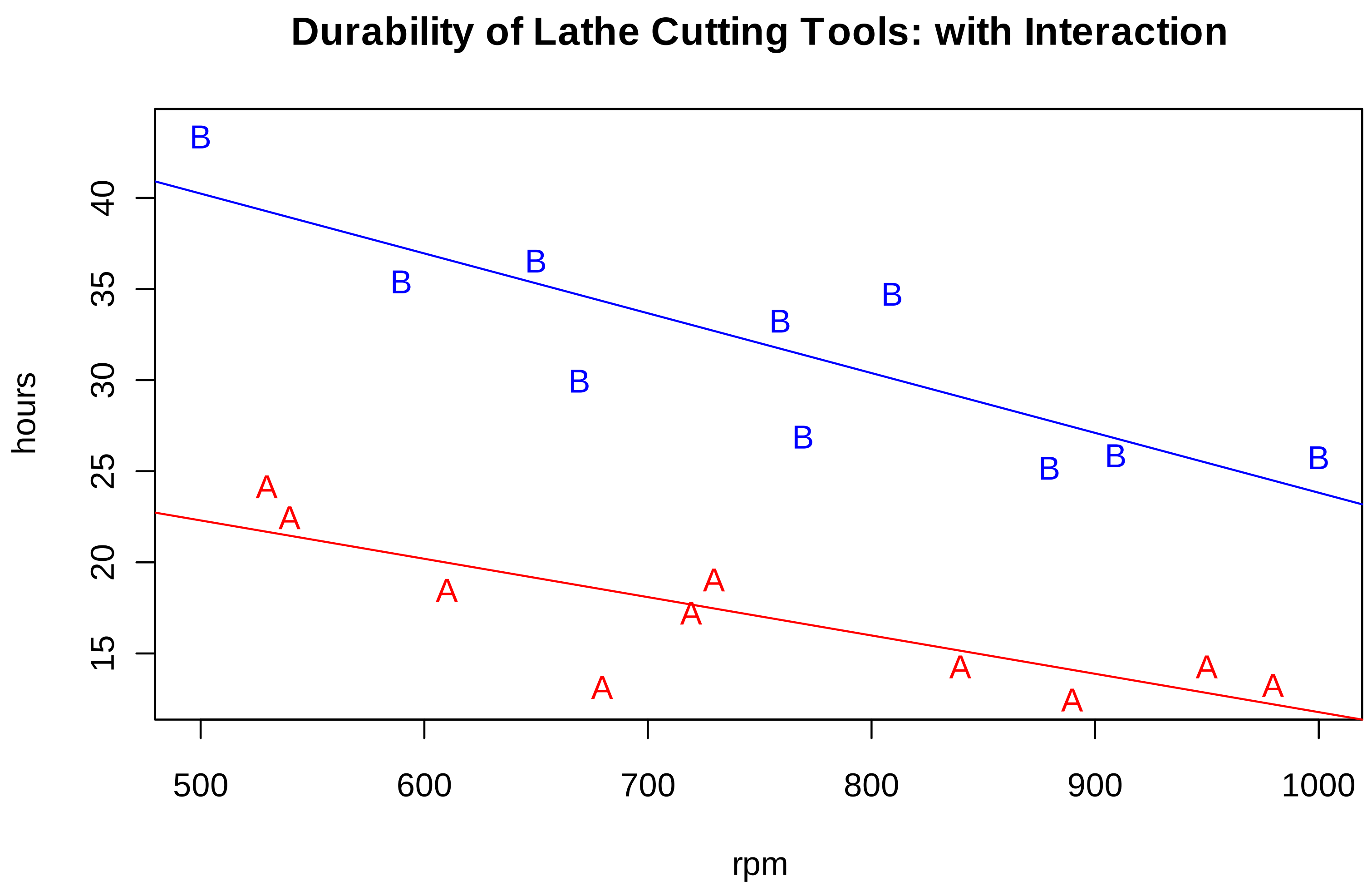
Question: do the slopes need to be identical?

→ 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 Slope 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?

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

→ individual parameter test for the interaction term!

Question 2: is there any difference altogether?

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

→ this is a partial F-test
→ we try to exclude interaction and dummy variable together

R offers convenient functionality for these tests!

Anova Output

Summary output for the interaction model

```
> fit1 <- lm(hours ~ rpm, data=lathe)
> fit2 <- lm(hours ~ rpm * tool, data=lathe)
> anova(fit1, fit2)
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 ***
```

→ no different slopes, but different intercept!

Categorical Input with More than 2 Levels

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

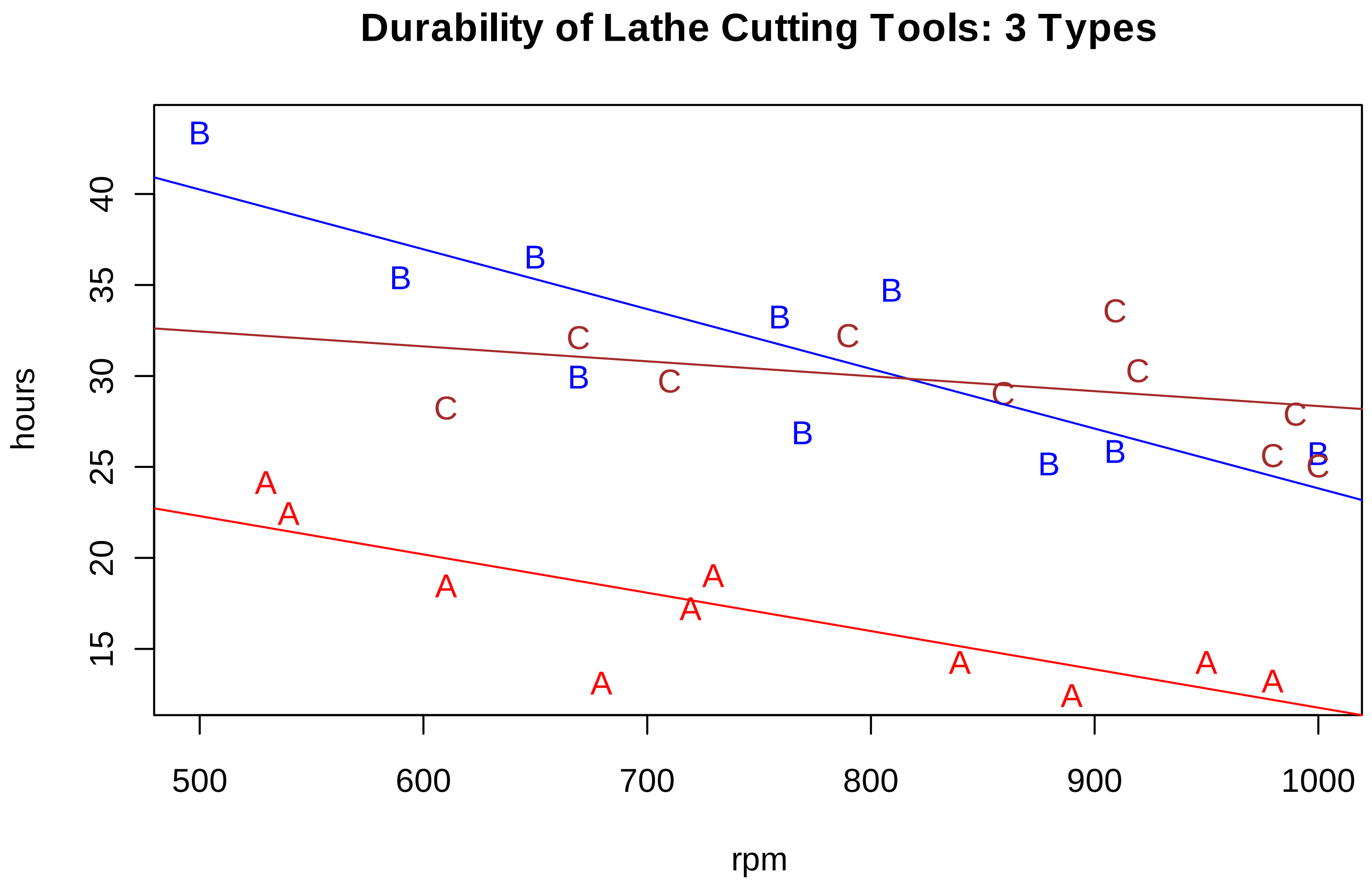
x_2	x_3	
0	0	*for observations of type A*
1	0	*for observations of type B*
0	1	*for observations of type C*

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

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

Applied Statistical Regression
HS 2011 – Week 06

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
```

Applied Statistical Regression
HS 2011 – Week 06

Inference with Categorical Predictors

Do not perform individual hypothesis tests on factors!

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 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
```

→ strong evidence that we need to distinguish the tools!
→ weak evidence for the necessity of different slopes

Applied Statistical Regression
HS 2011 – Week 06

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.

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.

  $\tilde{r}_i = \dfrac{r_i}{\hat{\sigma}_\varepsilon \cdot \sqrt{1-h_{ii}}}$, 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:

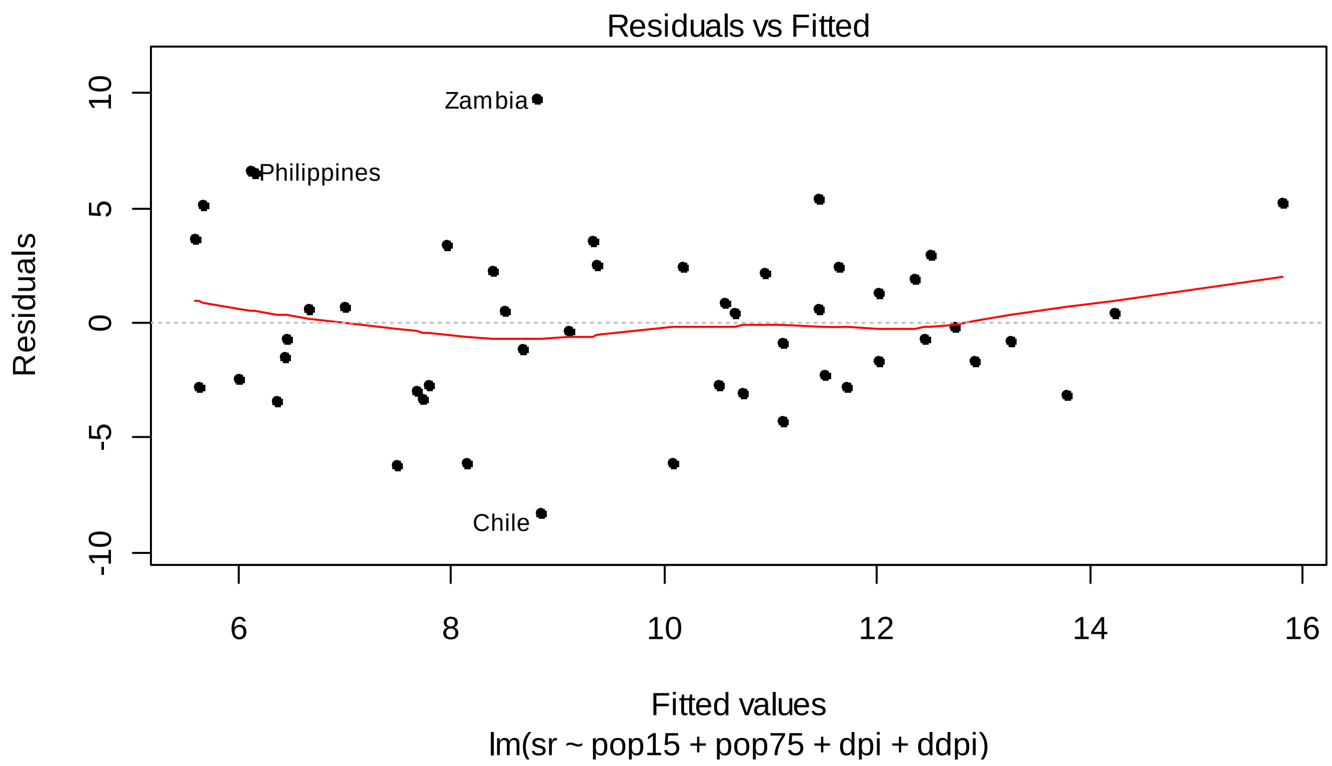
```
lm(sr ~ pop15 + pop75 + dpi + ddpi, data=LifeCycleSavings)
```

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:

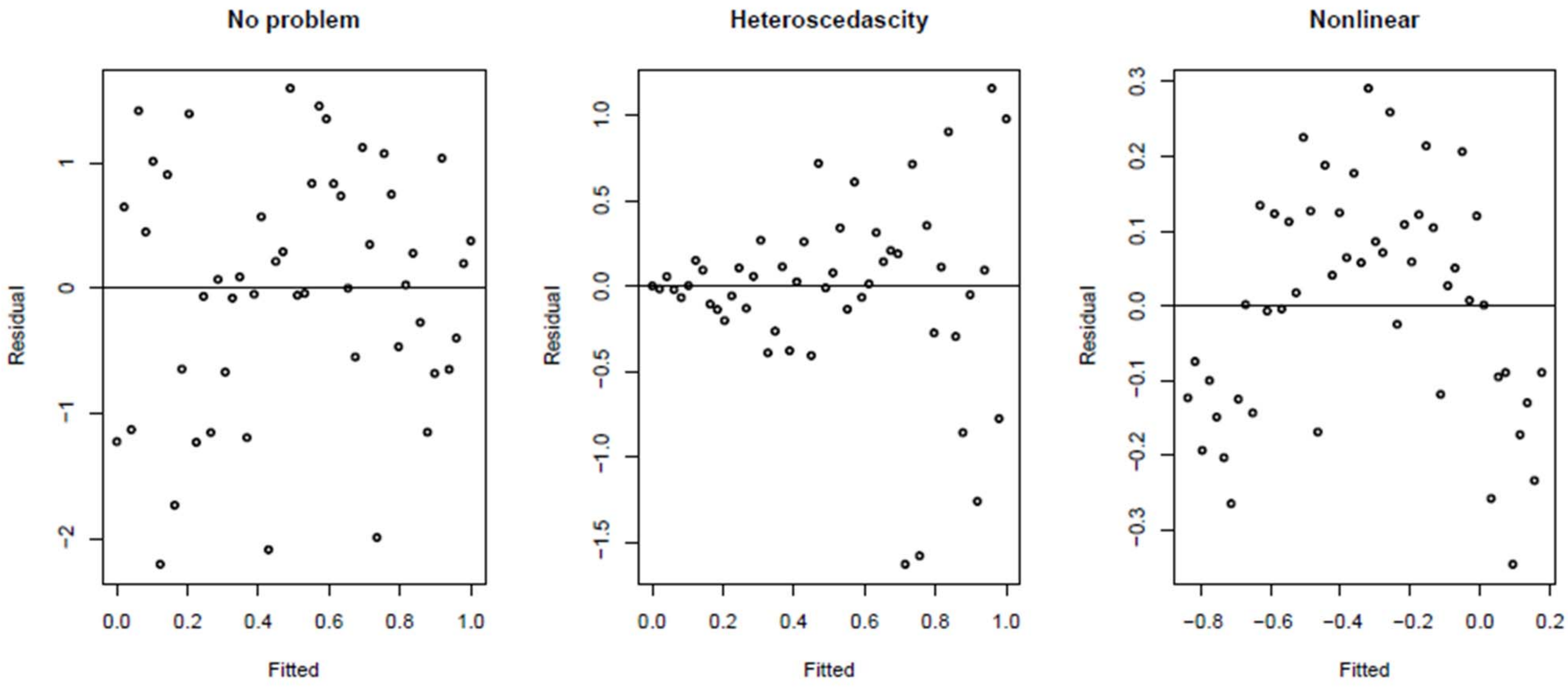
- finding structural model deficiencies, i.e. $E[E_i] \neq 0$
- if that is the case, the response/predictor relation could be nonlinear, or some predictors could be missing
- it is also possible to detect non-constant variance (→ then, the smoother does not deviate from 0)

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
HS 2011 – Week 06

Tukey-Anscombe-Plot



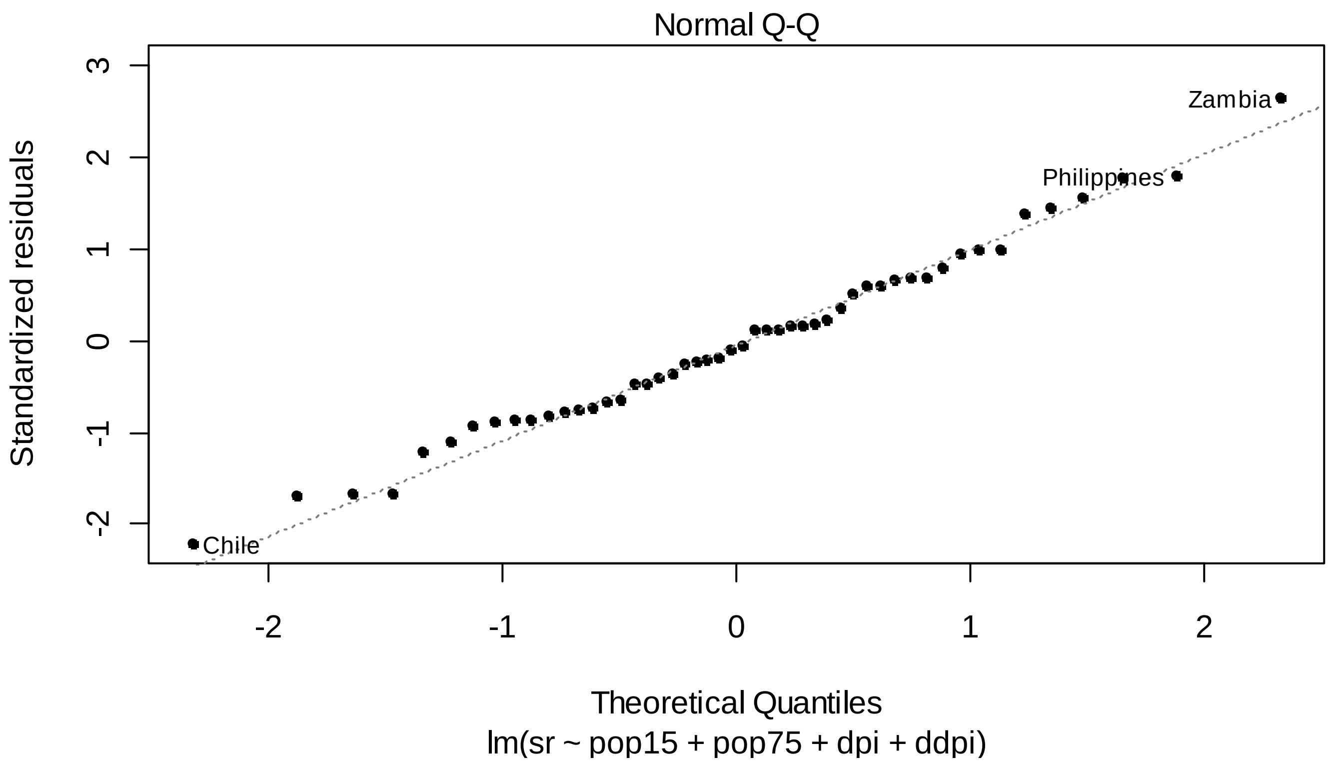
Tukey-Anscombe-Plot

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!

Normal Plot

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



Normal Plot

Is useful for:

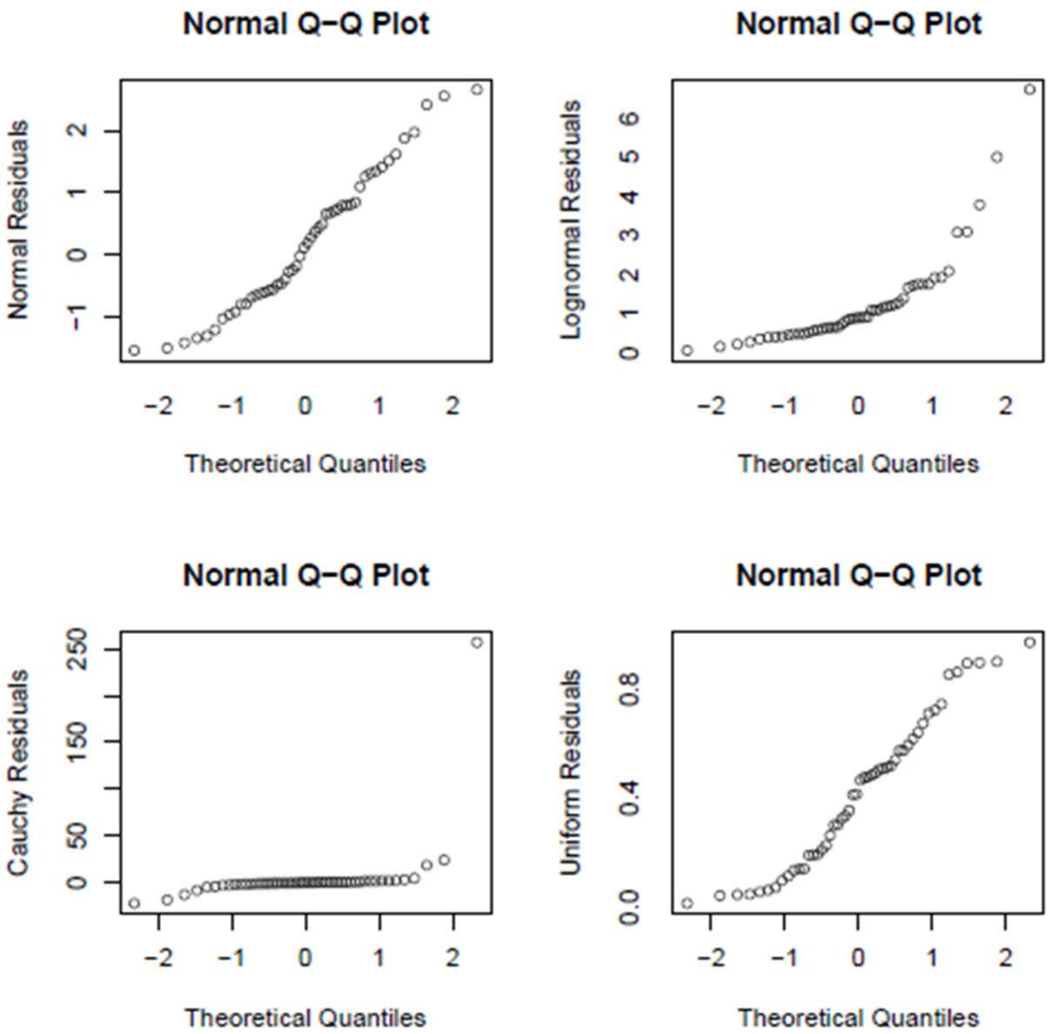
- for identifying non-Gaussian errors: $E_i \overset{!}{\sim} N(0, \sigma_E^2 I)$

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!

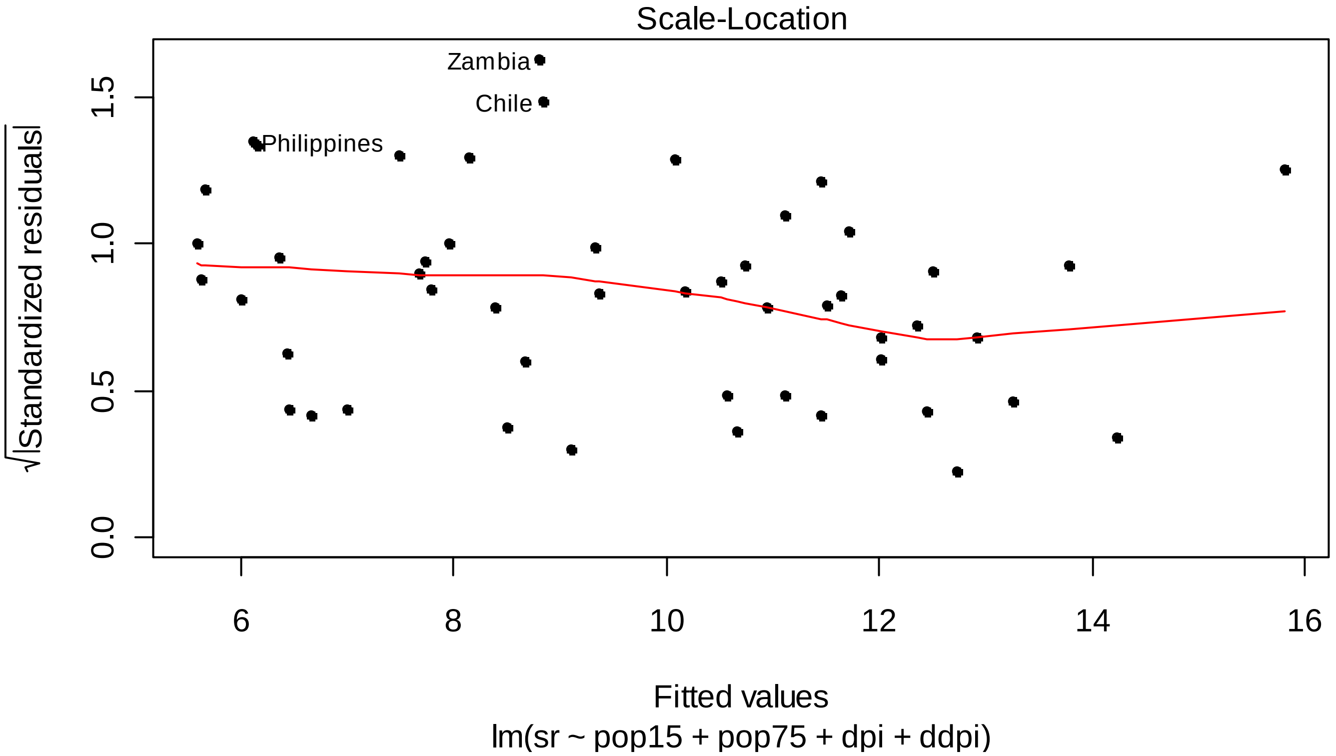
Applied Statistical Regression
HS 2011 – Week 06

Normal Plot



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.