Applied Statistical Regression
HS 2010 – 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, 2010

Applied Statistical Regression
HS 2010 – 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**

Applied Statistical Regression
HS 2010 – Week 06

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 2010 – Week 06

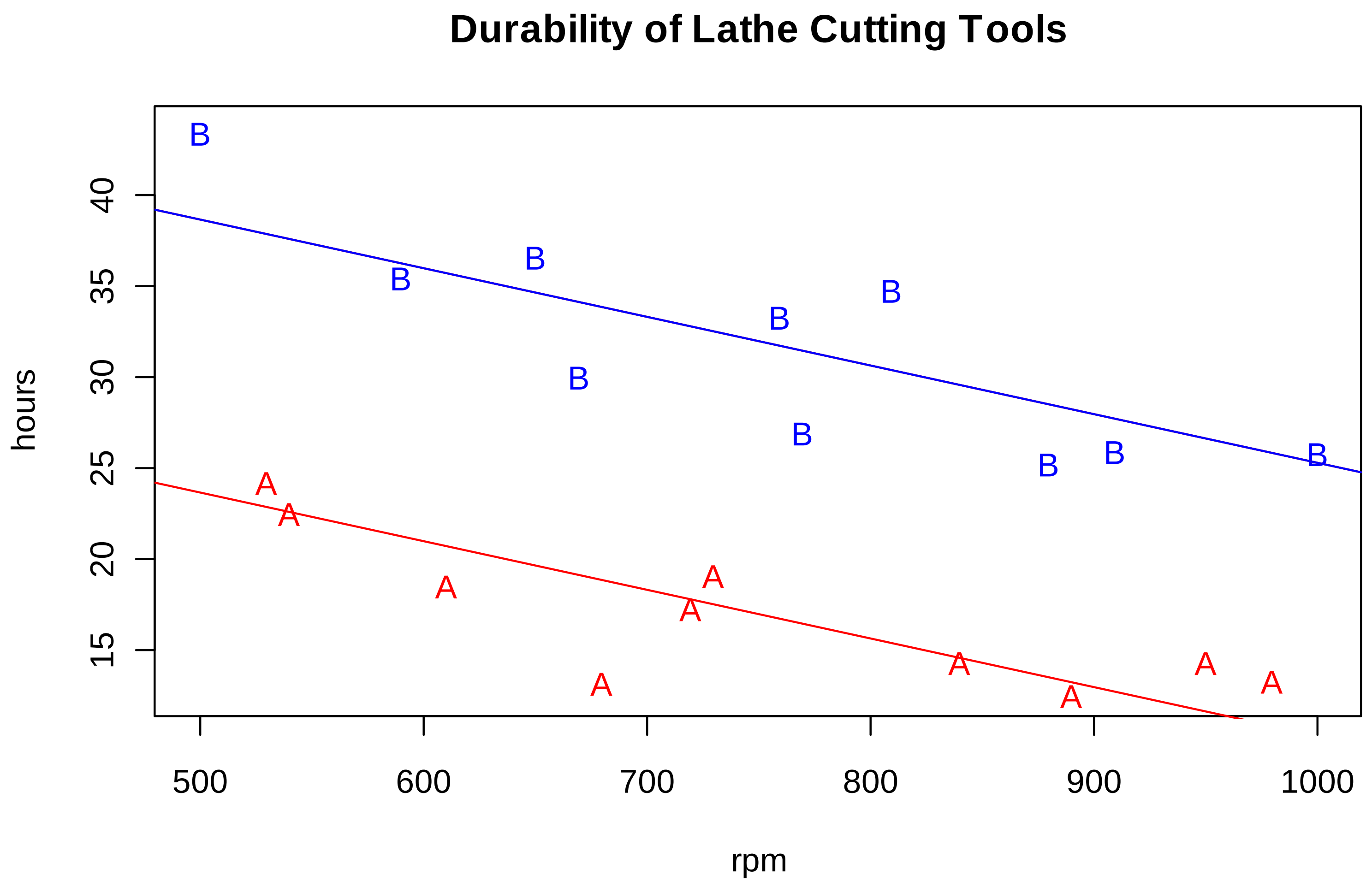
Interpretation of the Model

→ 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 2010 – Week 06

The Dummy Variable Fit



Applied Statistical Regression
HS 2010 – Week 06

A Model with Interactions

Question: do the slopes need to be identical?

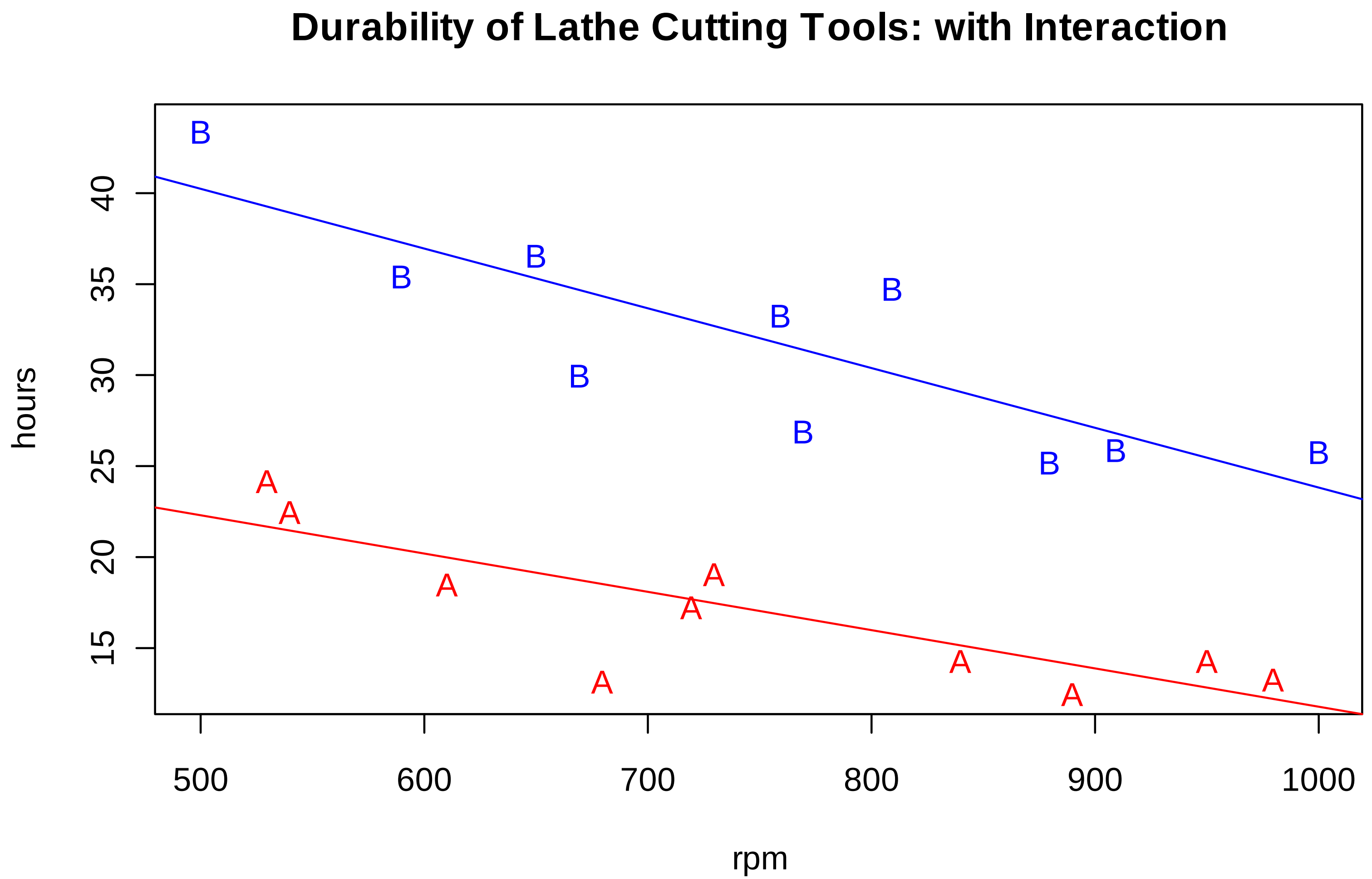
→ with the appropriate model, the answer is no!

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

→ see blackboard for model interpretation…

Zurich University of Applied Sciences

Different Slope for the Regression Lines



Applied Statistical Regression
HS 2010 – Week 06

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!

Applied Statistical Regression
HS 2010 – Week 06

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:

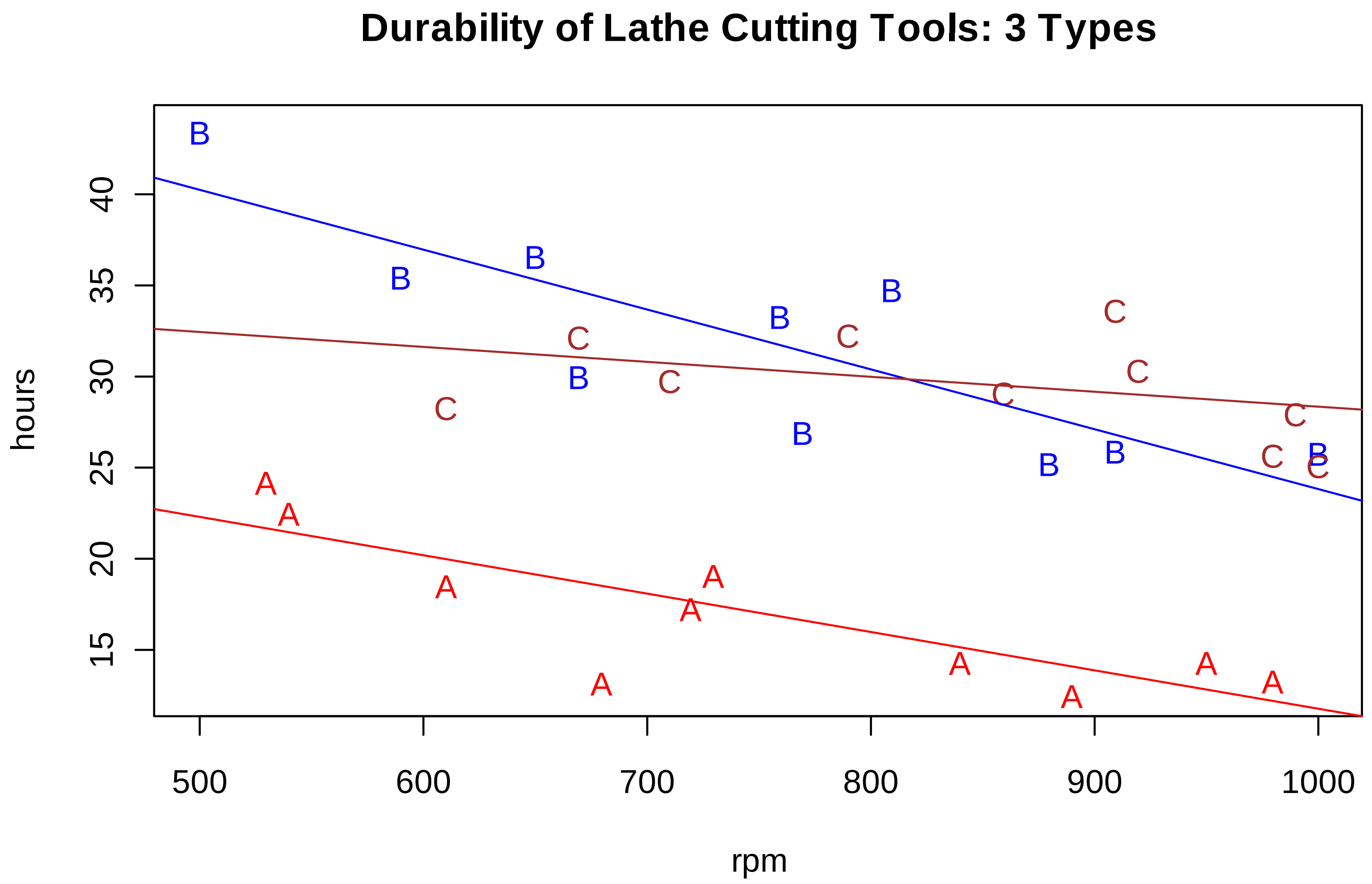
x_2	x_3	
0	0	*for observations of type A*
1	0	*for observations of type B*
0	1	*for observations of type C*

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

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

Applied Statistical Regression
HS 2010 – Week 06

Three Types of Cutting Tools



Applied Statistical Regression
HS 2010 – Week 06
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 2010 – 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

Applied Statistical Regression
HS 2010 – Week 06

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

Transformations

Scope:

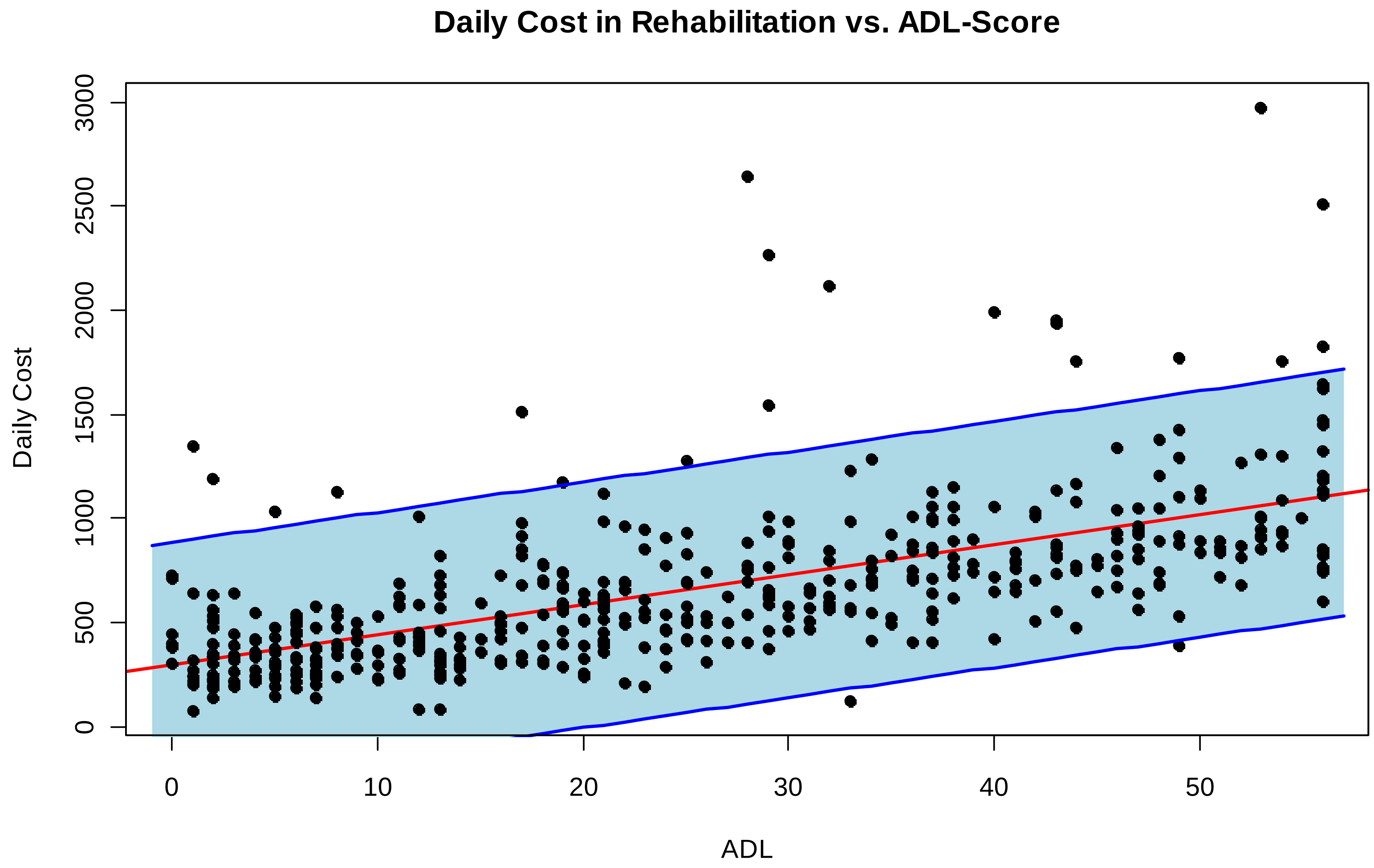
- For both response and the predictors

Goals:

- Dealing with violated model assumptions
- Extension to linear modeling
- More versatility

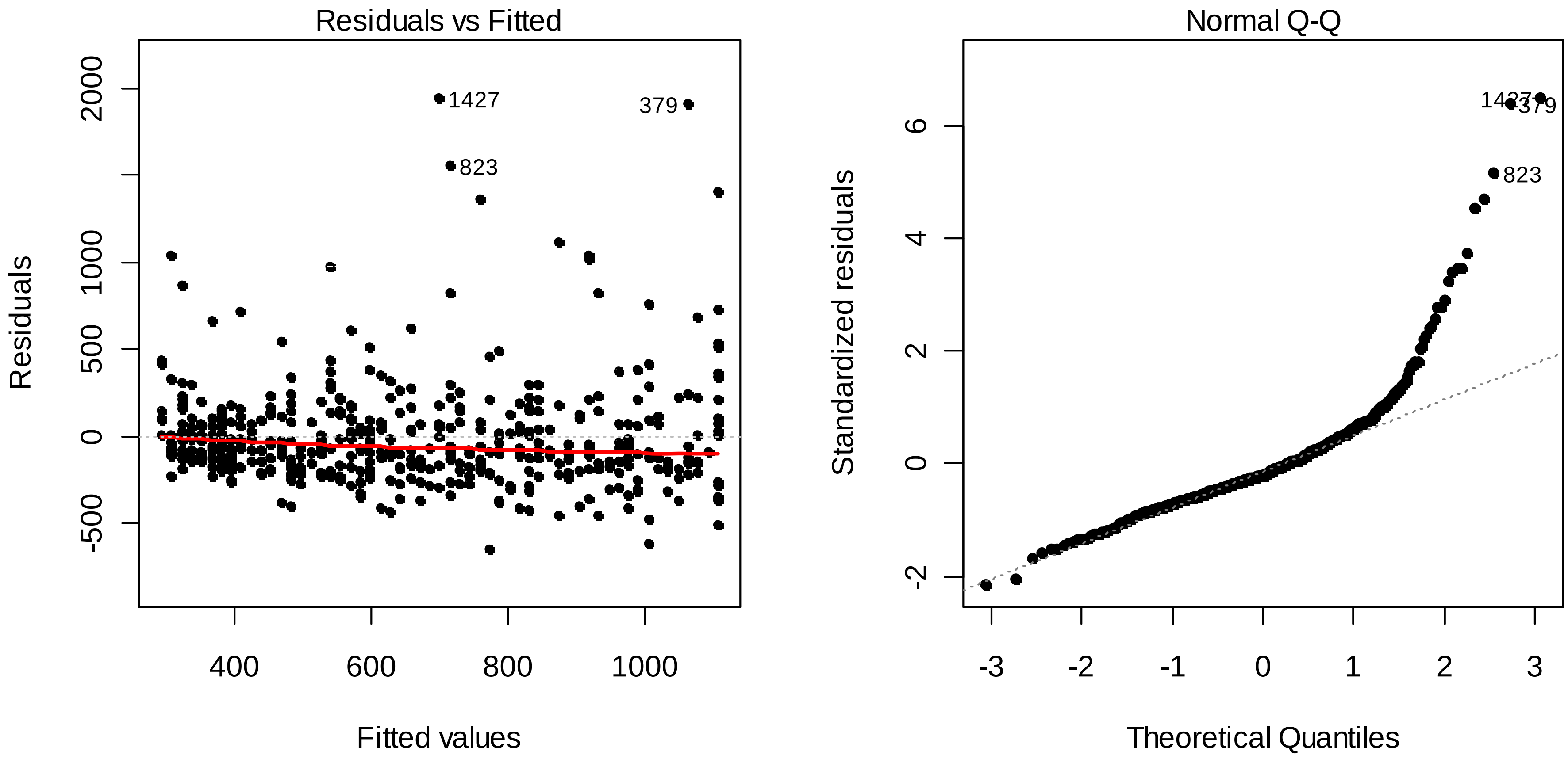
Applied Statistical Regression
HS 2010 – Week 06

Transformations: Example



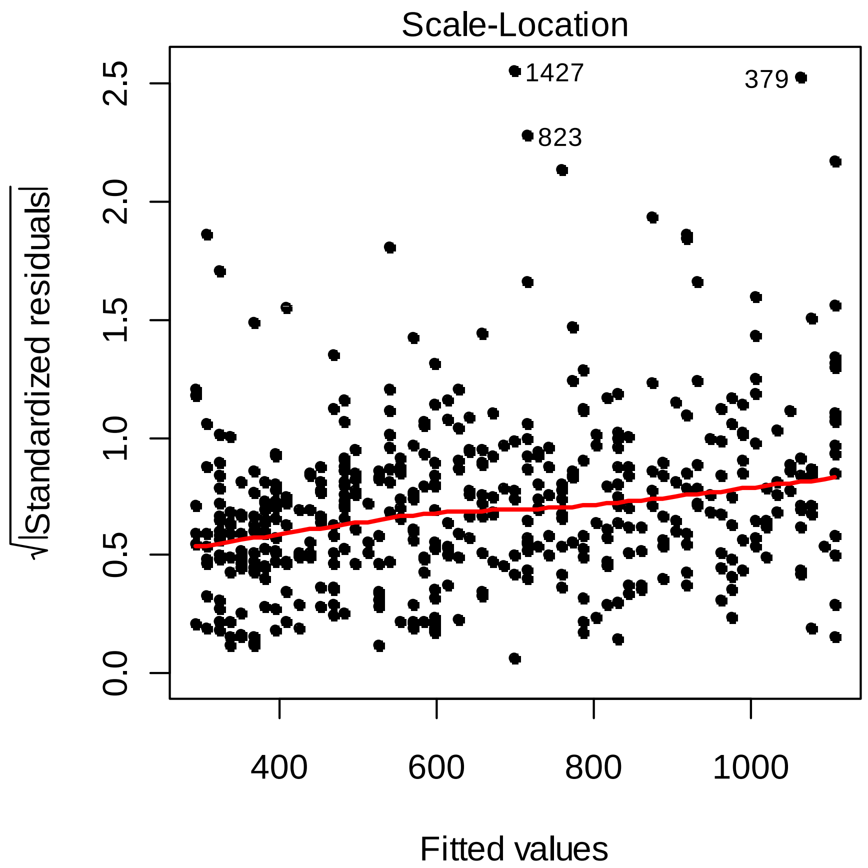
Applied Statistical Regression
HS 2010 – Week 06

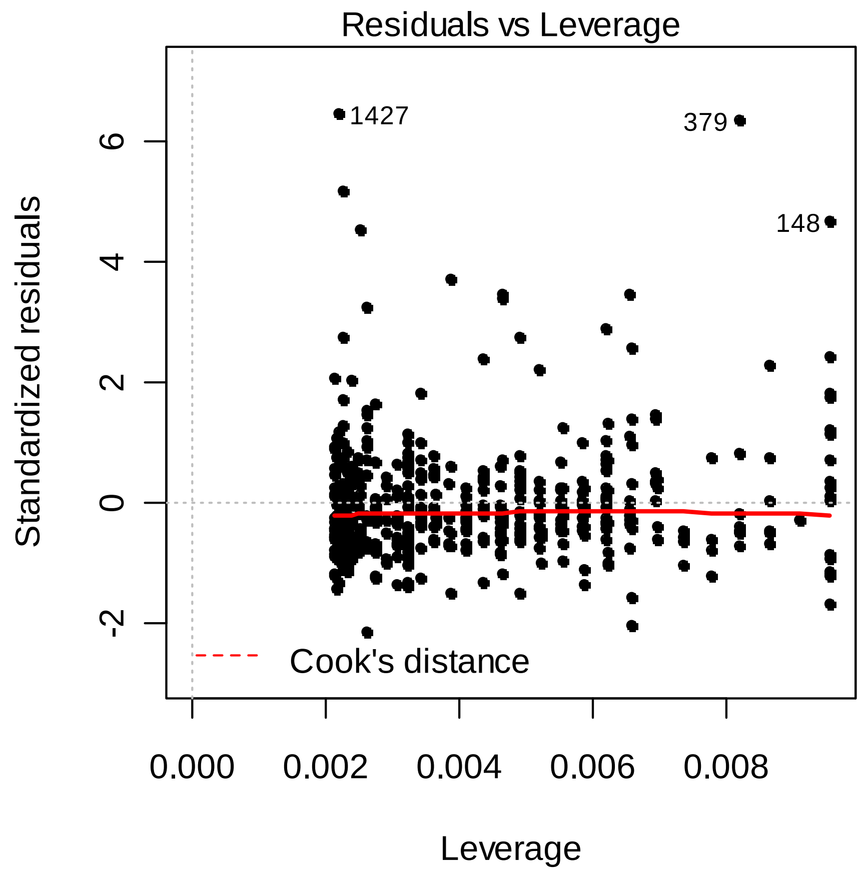
Transformations: Example



Applied Statistical Regression
HS 2010 – Week 06

Transformations: Example





Problems with this Example

Non-zero expectation:

- visible with the Tukey-Anscombe plot

Non-constant variance:

- apparent in the Scale-Location plot

Skewed residuals:

- very prominent in the Normal plot

Unwanted negative values:

- prediction interval is partly below zero

Positive Skewness:

- Unwanted

- Often present…

Logged Response Model

We transform the response variable and try to explain it using a linear model with our previous predictors:

$$Y' = \log(Y) = \beta_0 + \beta_1 x + \varepsilon$$

In the original scale, we can write the logged response model using the same predictors:

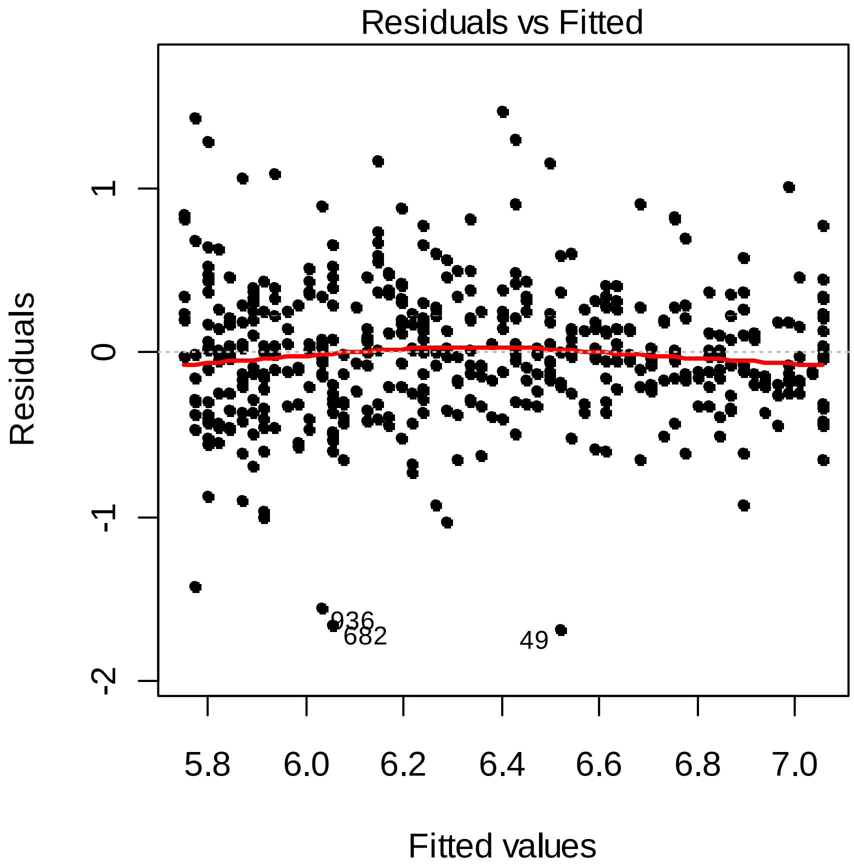
$$Y = \exp(\beta_0 + \beta_1 x) \cdot \exp(\varepsilon)$$

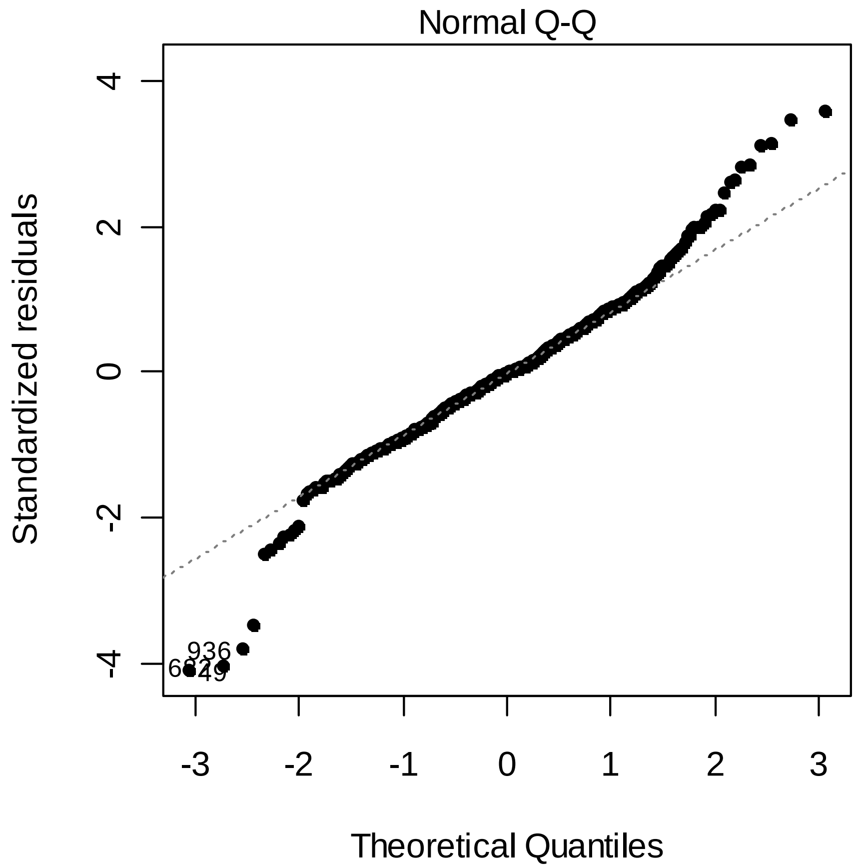
→ Multiplicative model

→ $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, and thus, $\exp(\varepsilon)$ has a lognormal distribution

Applied Statistical Regression
HS 2010 – Week 06

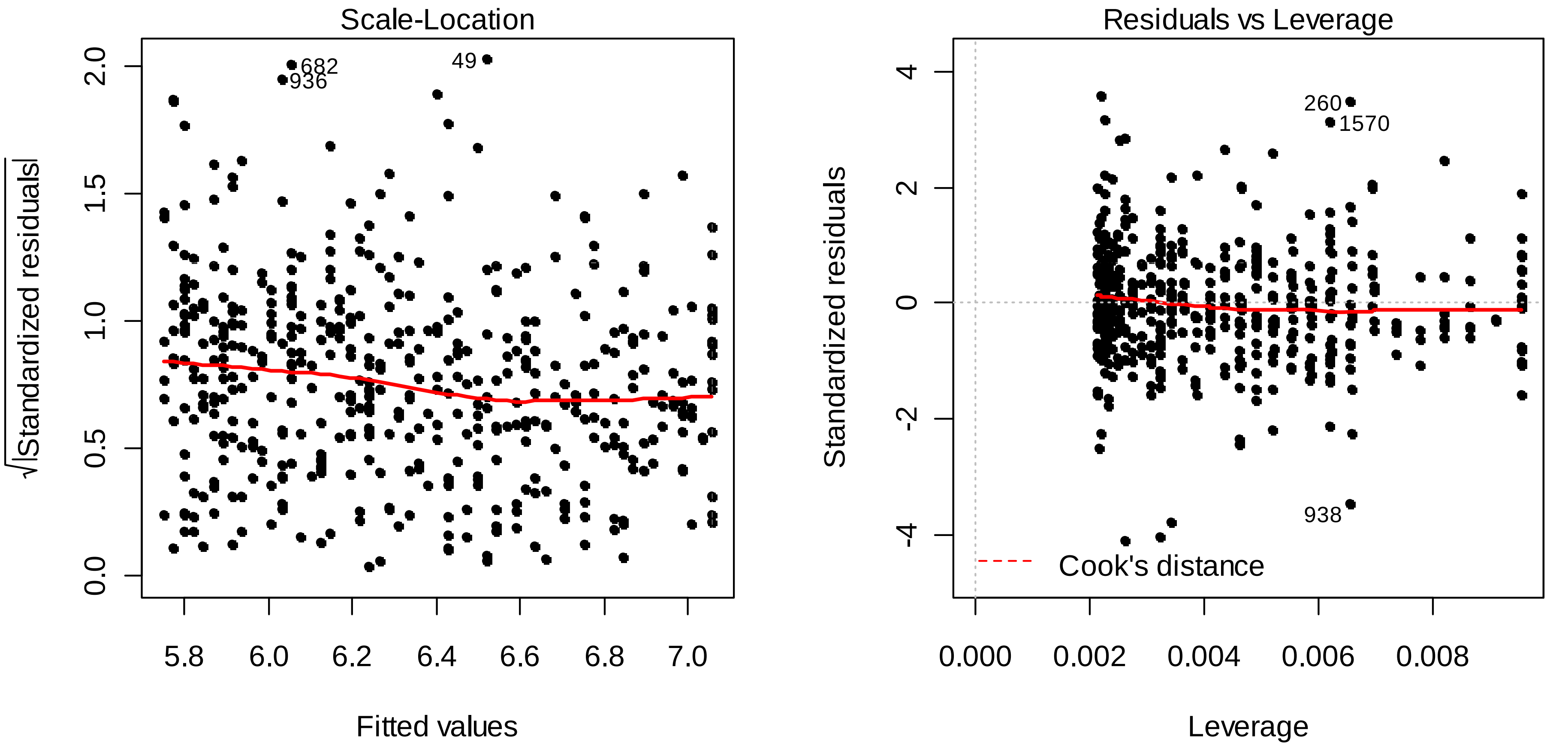
Does It Work?





Applied Statistical Regression
HS 2010 – Week 06

Does It Work?



Improvements with the Logged Response

Non-zero expectation:

- much better now!

Non-constant variance:

- better now, some over-correction?

Skewed residuals:

- now perfectly symmetric, and a bit long-tailed

Unwanted negative values:

- issue solved!

Applied Statistical Regression
HS 2010 – Week 06

Dealing with Zero Response

- Logged response model is only applicable when the response is strictly positive…

- What if there are some cases with $Y = 0$?
  - never omit these
  - additive shifting is possible

- How to additively shift?
  - usual choice: c=1
  - not good, because effect is scale-dependent

→ **Shift with the value of the smallest positive observation!**

Back Transforming the Fitted Values

- In principle, we can „simply back transform"

$$\hat{y} = \exp(\hat{y}')$$

- This is an estimate for the median, but not the mean!

- If unbiased estimation is required, then use:

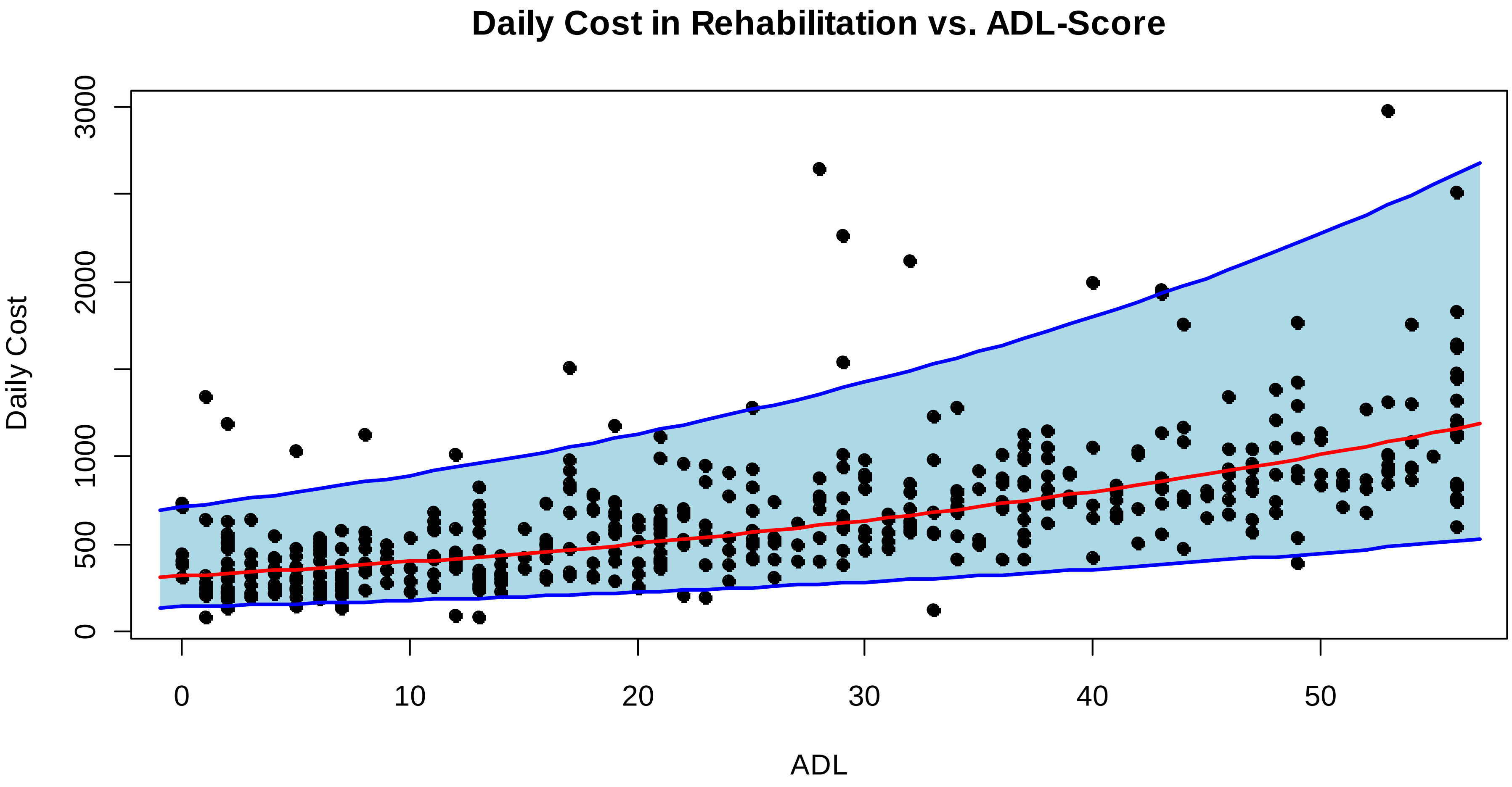
$$\hat{y} = \exp\left(\hat{y}' + \frac{\hat{\sigma}_\varepsilon^2}{2}\right)$$

- Confidence/prediction intervals are not problematic

$$[l, u] \;\rightarrow\; [\exp(l), \exp(u)]$$

Applied Statistical Regression
HS 2010 – Week 06

Back Transforming: Example



Interpretation of the Coefficients

Important: there is no back transformation for the coefficients to the original scale, but still a good interpretation

$$\log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$$

$$\hat{y} = \exp(\hat{\beta}_0)\exp(\hat{\beta}_1 x_1)\ldots\exp(\hat{\beta}_p x_p)$$

An increase by one unit in x_1 would multiply the fitted value in the original scale with $\exp(\hat{\beta}_1)$.

→ **Coefficients are interpreted multiplicatively!**

First-Aid Transformations

There are more transformations than the logged response model!

First-Aid Transformations:

→ do always apply these (if no practical reasons against it)

→ to both response and predictors

Absolute values and concentrations:

log-transformation: $y' = \log(y)$

Count data:

square-root transformation: $y' = \sqrt{y}$

Proportions:

arcsine transformation: $y' = \sin^{-1}\left(\sqrt{y}\right)$