Applied Statistical Regression
AS 2013 – Week 05

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, October 14, 2013

Model Extensions

So far, simple linear regression was considered as fitting a straight line into a xy-scatterplot. While this is correct, it does not reflect the full potential of linear regression. With creative use of variable transformations, many more possibilites open.

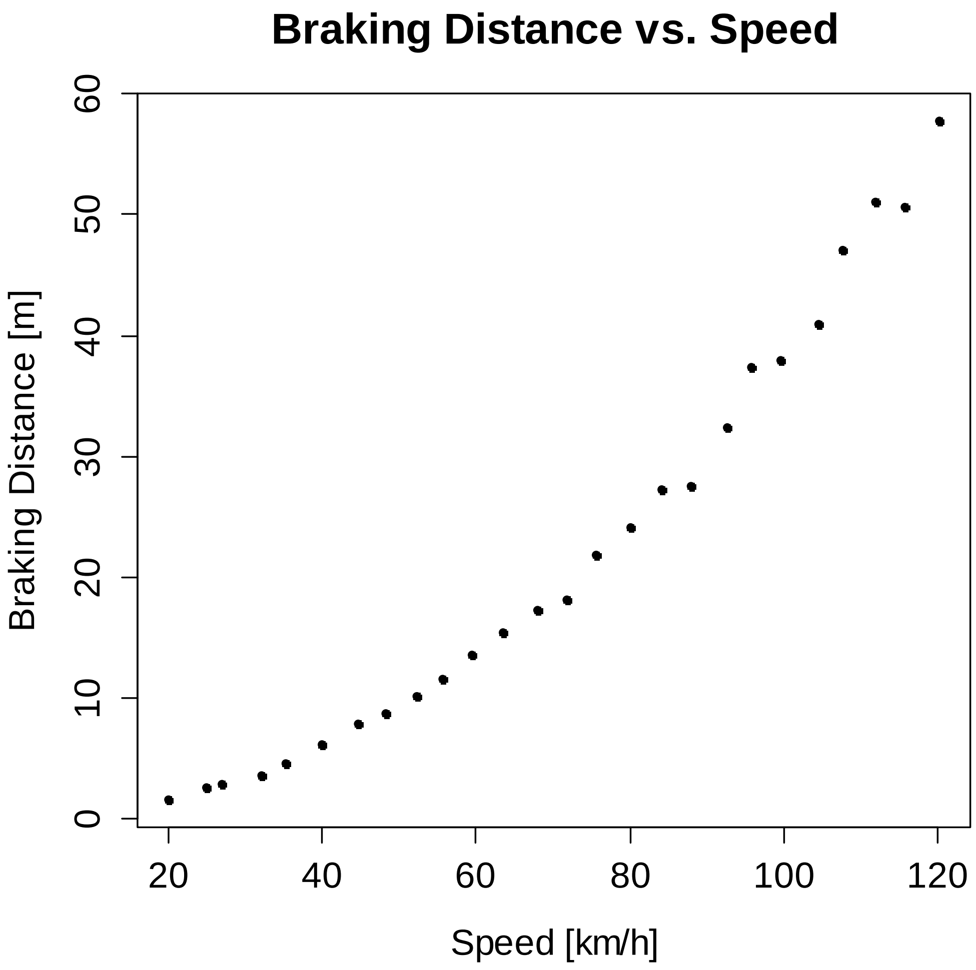
Example: Automobile Braking Distance



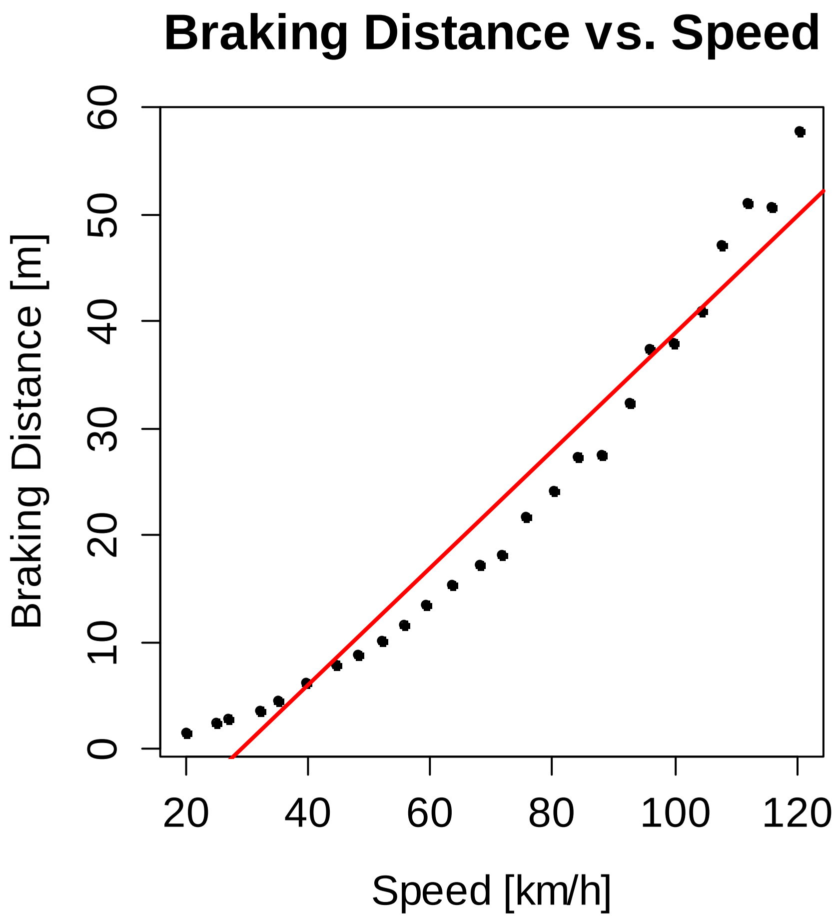
We have data from 26 test drives with differing speed. The goal was to estimate the braking behavior of a certain type of tires. The data are displayed on the next slide…

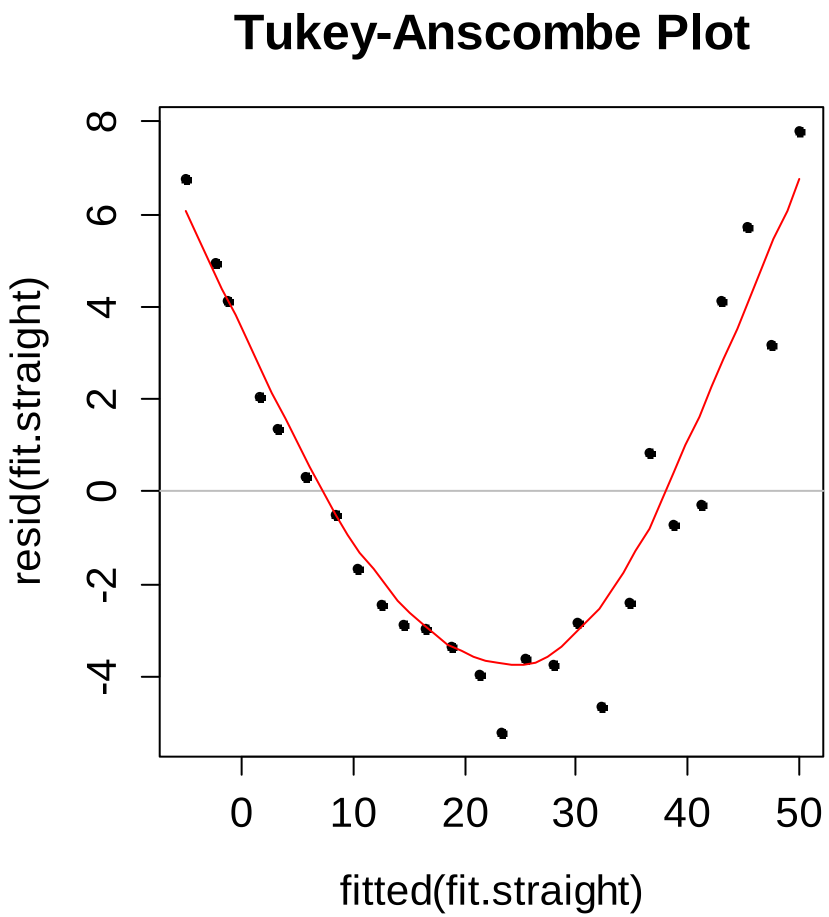
Braking Distance: Data

obs	speed	brdist
1	19.96	1.60
2	24.97	2.54
3	26.97	2.81
4	32.14	3.58
5	35.24	4.59
6	39.87	6.11
7	44.62	7.91
8	48.32	8.76
9	52.18	10.12
10	55.72	11.62
11	59.44	13.57
12	63.56	15.45
...	...	...
24	111.97	51.09
25	115.88	50.69
26	120.35	57.77



Braking Distance: Fitting a Straight Line





Braking Distance: Facts

Conclusions from the residual plots:

- The straight line has a systematic error and does not reflect the true relation between speed and braking distance. From physics, we know that a parabola is more appropriate.

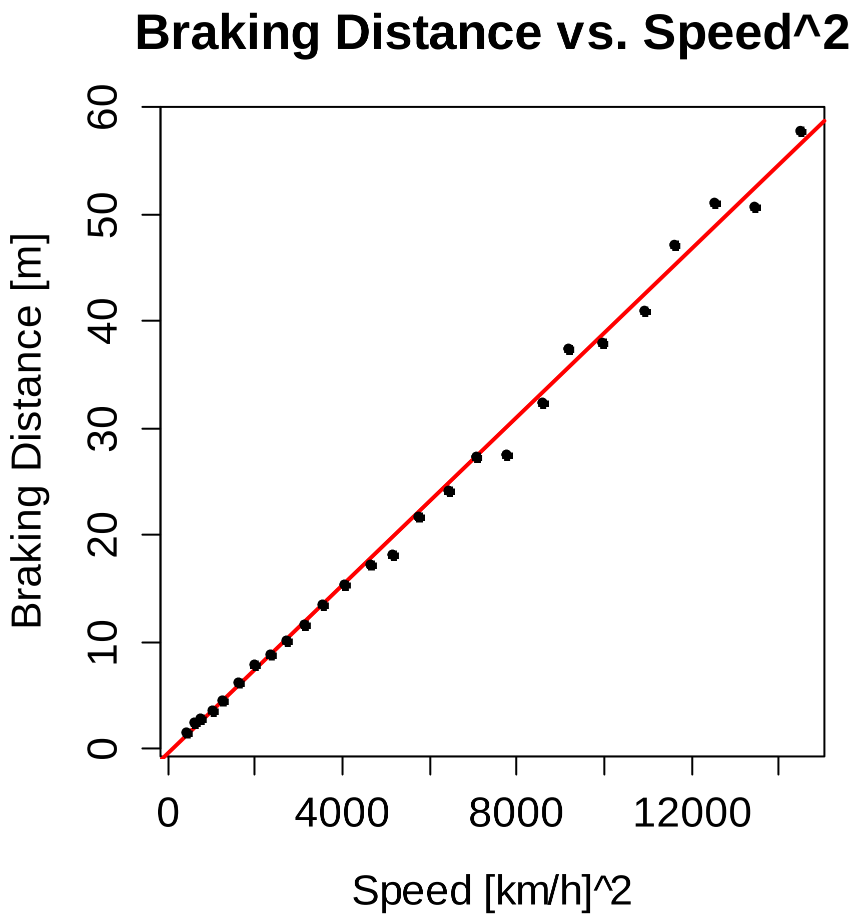
$$Distance_i = \beta_0 + \beta_1 \cdot Speed_i^2 + E_i$$

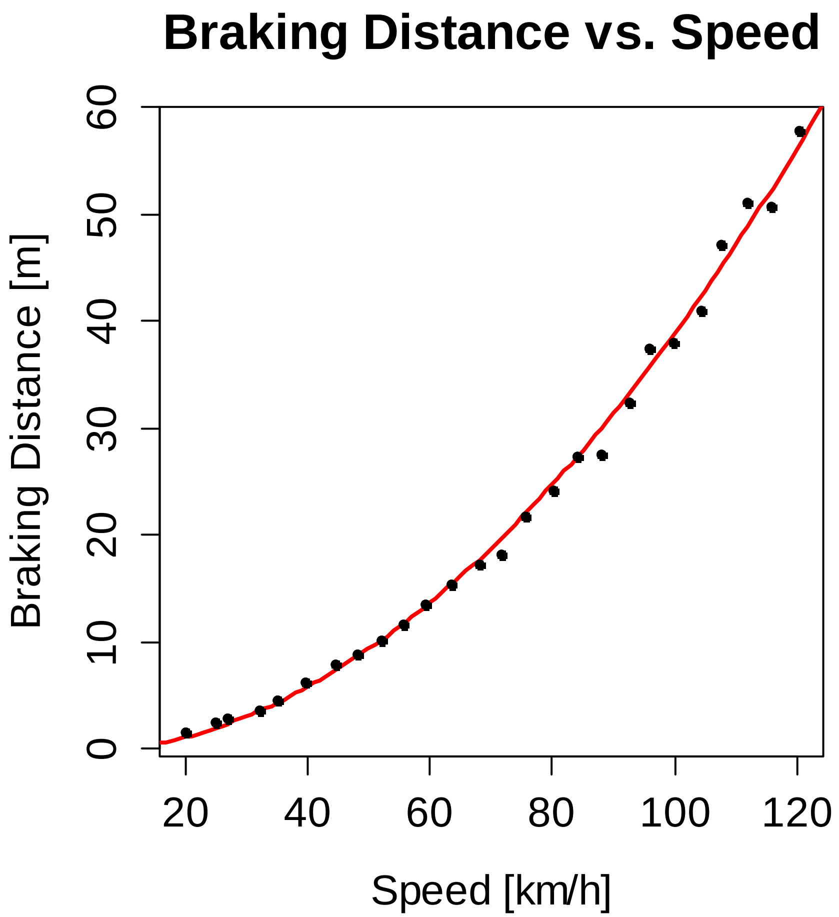
resp. $y_i = \beta_0 + \beta_1 \cdot x_i' + E_i$, where $x_i' = x_i^2 = Speed_i^2$

- Please note that this is a simple linear regression problem. There is only one single predictor and the coefficients $\hat{\beta}_0, \hat{\beta}_1$ can and need to be estimated with the LS algorithm by taking partial derivatives and setting them to zero.

Braking Distance: Distance vs. Speed^2

```
> fit <- lm(weg ~ I(speed^2))
```





Curvilinear Regression

Simple linear regression offers more than fitting straight lines! We can fit any curvilinear relation with the LS algorithm. Some examples include:

- $y_i = \beta_0 + \beta_1 \cdot \ln(x_i) + E_i$
- $y_i = \beta_0 + \beta_1 \cdot \sqrt{x} + E_i$
- $y_i = \beta_0 + \beta_1 \cdot x^{-1} + E_i$

We are using $x_i' = \ln(x_i)$, $x_i' = \sqrt{x_i}$, bzw. $x_i' = (x_i)^{-1}$. In this form, it is obvious that all these are simple linear regression problems that can be solved via LS.

→ **BUT...** see next slide

Braking Distance: Remarks

Curvilinear Models are often inadequate in practice:

- In our braking distance example, we should also consider the reaction time. This is a multiple regression model:

$$Distance_i = \beta_0 + \beta_1 \cdot Speed_i + \beta_2 \cdot Speed_i^2 + E_i$$

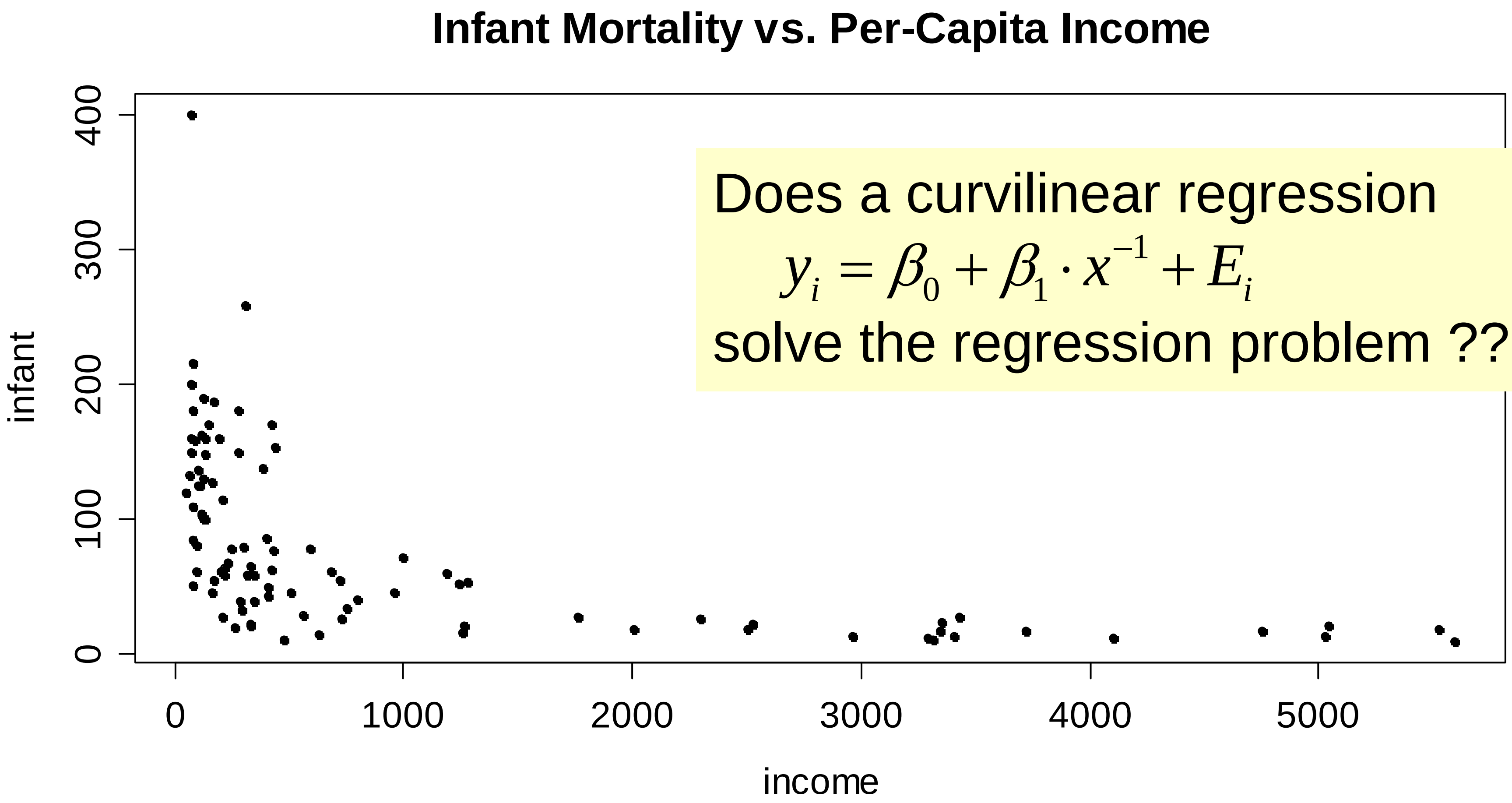
- Often, the variance/scatter of the errors is non-constant. In many examples, it increases with increasing.

- In many applications, the polynomial degree is not dictated by theorie, but needs to be estimated, too:

$$y_i = \beta_0 + \beta_1 \cdot x^{\beta_2} + E_i$$

Applied Statistical Regression
AS 2013 – Week 05

Infant Mortality vs. Per-Capita Income

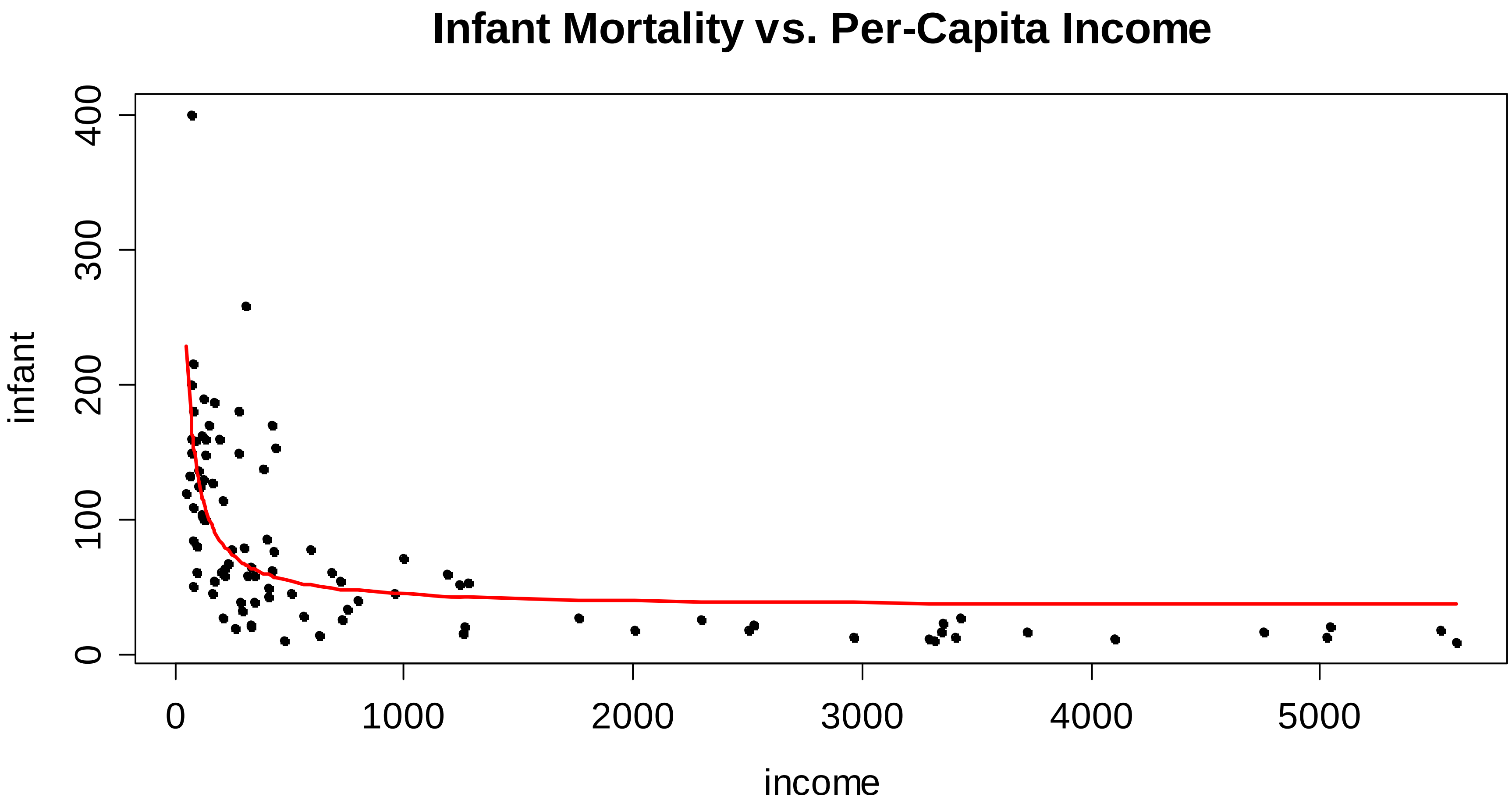


Does a curvilinear regression

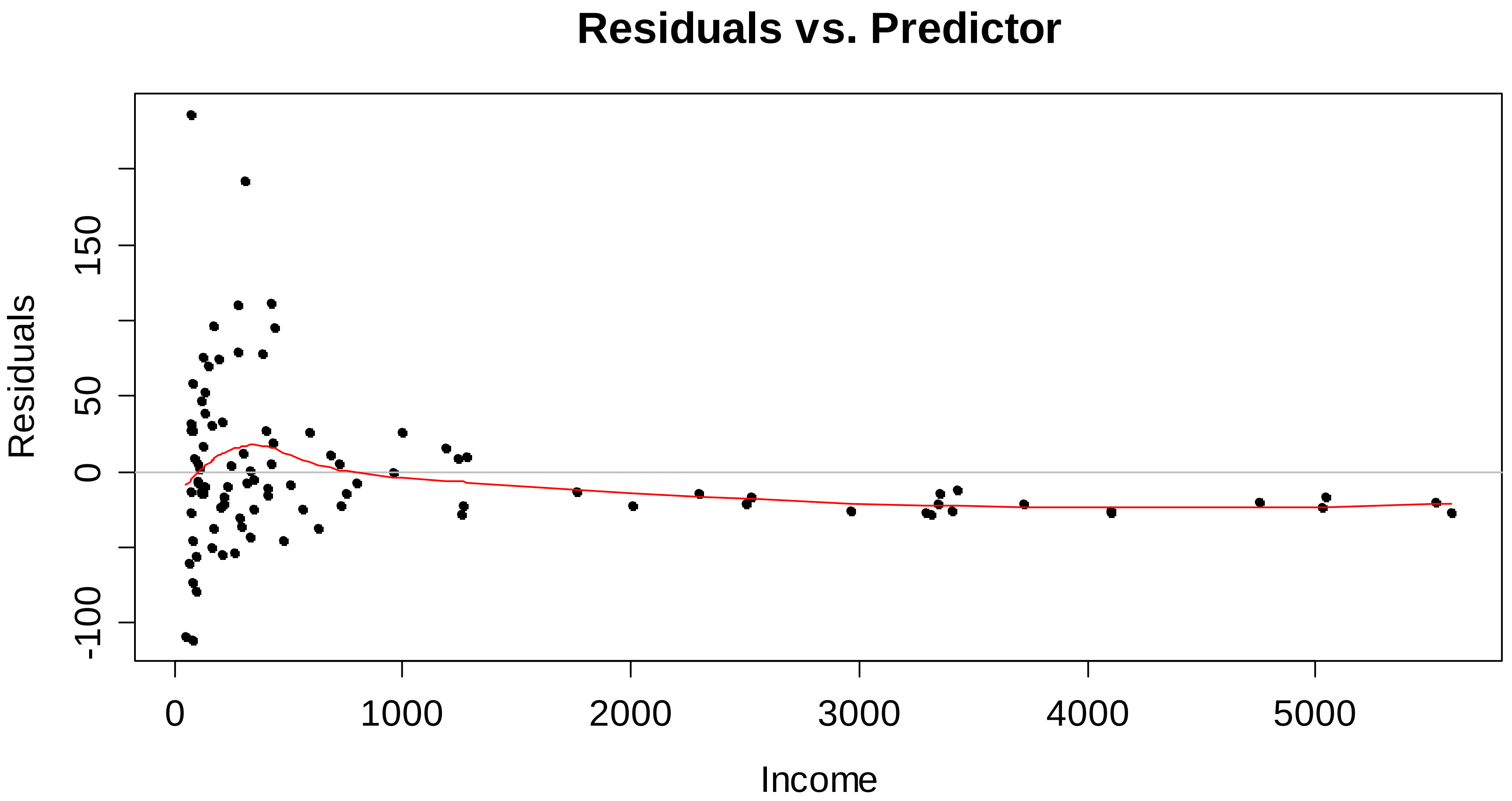
$$y_i = \beta_0 + \beta_1 \cdot x^{-1} + E_i$$

solve the regression problem ???

The Fitted Hyperbolic Regression Line



Residuals from Hyperbolic Fit



The Problem and the Solution

The hyperbolic fit shows some systematic error and is **not** the correct relation between mortality and income. We could try to estimate a power law such as:

$$y_i = \beta_0 + \beta_1 \cdot x_i^{\beta_2} + E_i$$

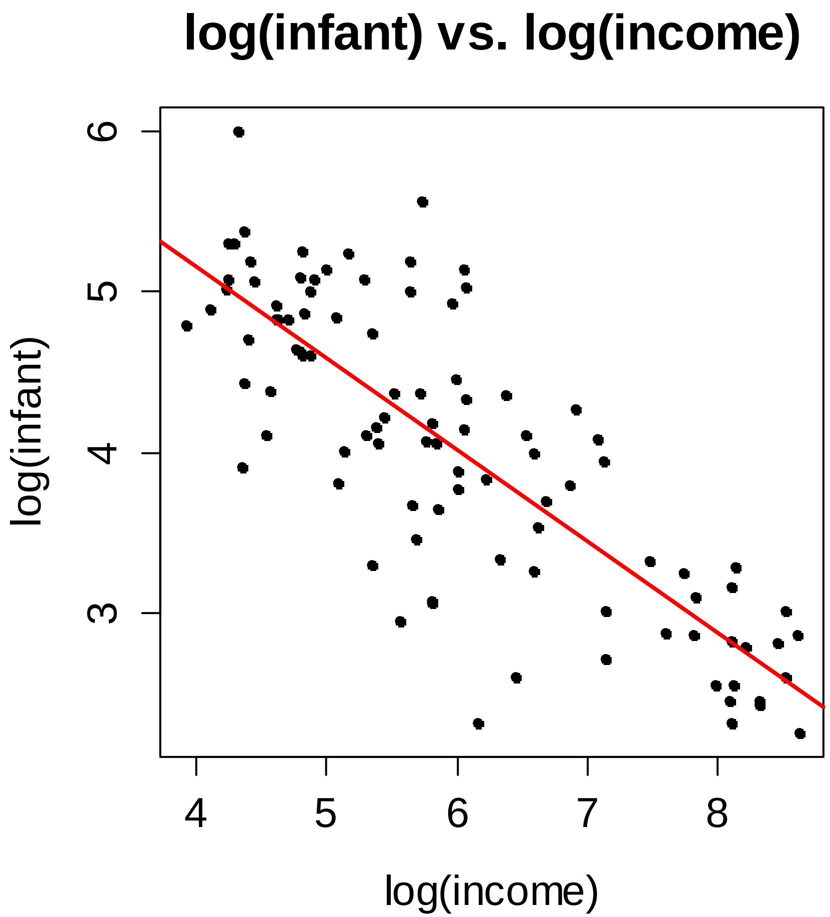
However, this problem is **non-linear** in the parameter β_2 and cannot be solved with the LS algorithm. Moreover, the error **variance** is **non-constant**.

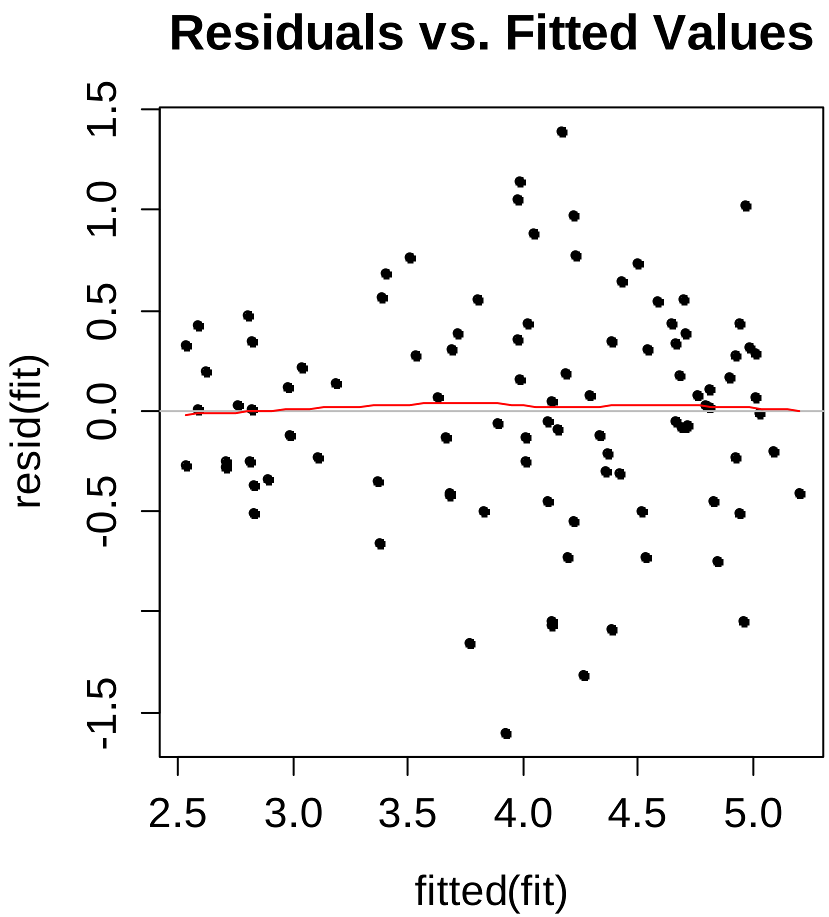
A simple yet very useful trick solves the problem:

$$y_i' = \log(y_i),\ \ x_i' = \log(x_i)$$

For details, **see the next slide and the blackboard...**

The Log-Transformation Helps!





Model and Coefficients

If a straight line is fitted on the log-log-scale,

$$y' = \beta_0' + \beta_1 \cdot x' + E', \text{ where } y' = \log(y),\ x' = \log(x) \quad ,$$

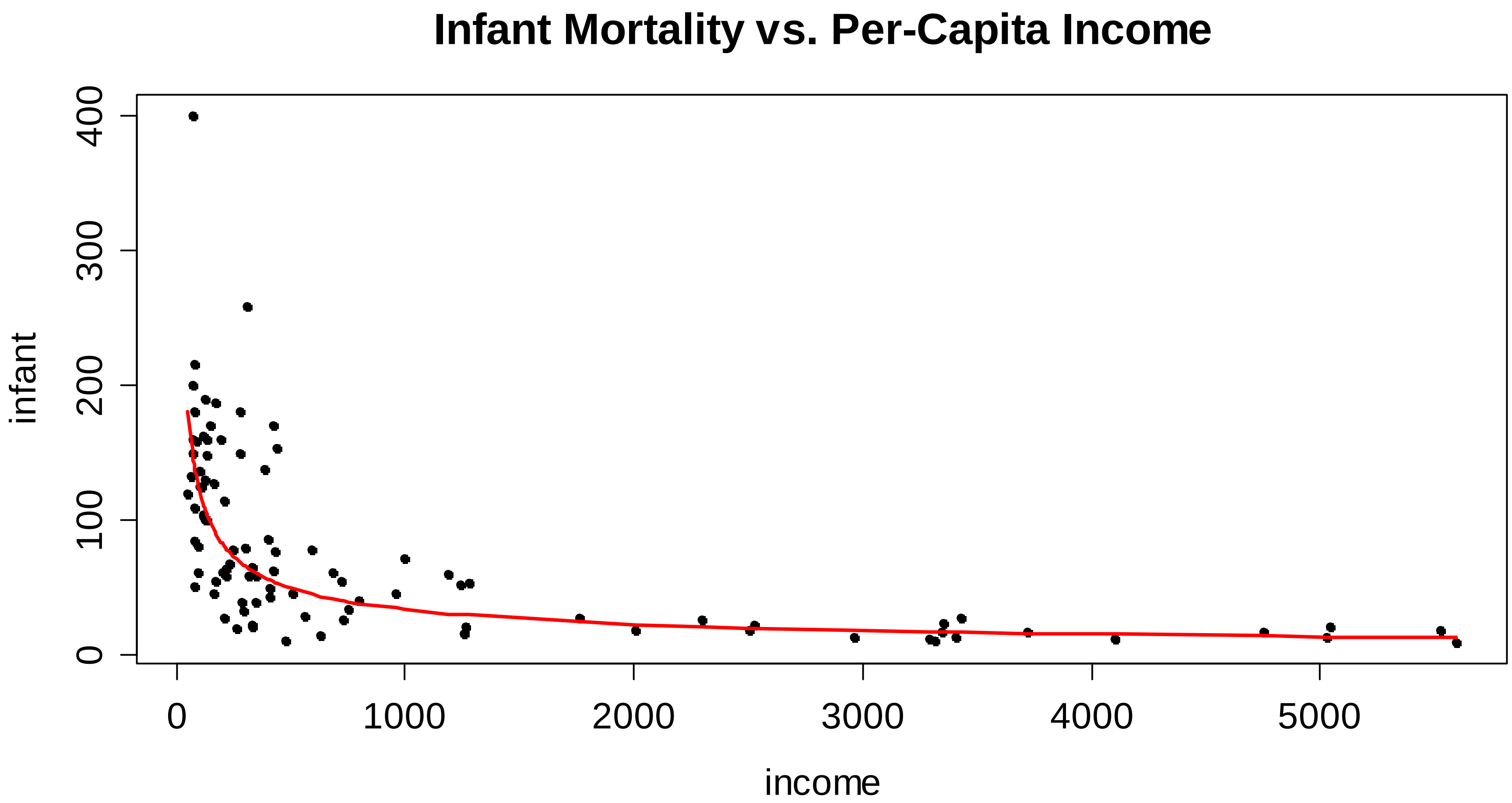
this means fitting the following relation on the original scale:

$$y = \beta_0 \cdot x^{\beta_1} \cdot E$$

The meaning of the parameter β_1 is as follows:

If x, i.e. the income increases by 1%, then y, i.e. the mortality decreases by $\hat{\beta}_1 = 0.56\%$. In other words: β_1 characterizes the relative change in y per unit of relative change in x .

The Fitted Relation



Fitted Values and Intervals

- For predicting the y-value on the original scale, we can just re-exponentiate to invert the log-transformation and hence:

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

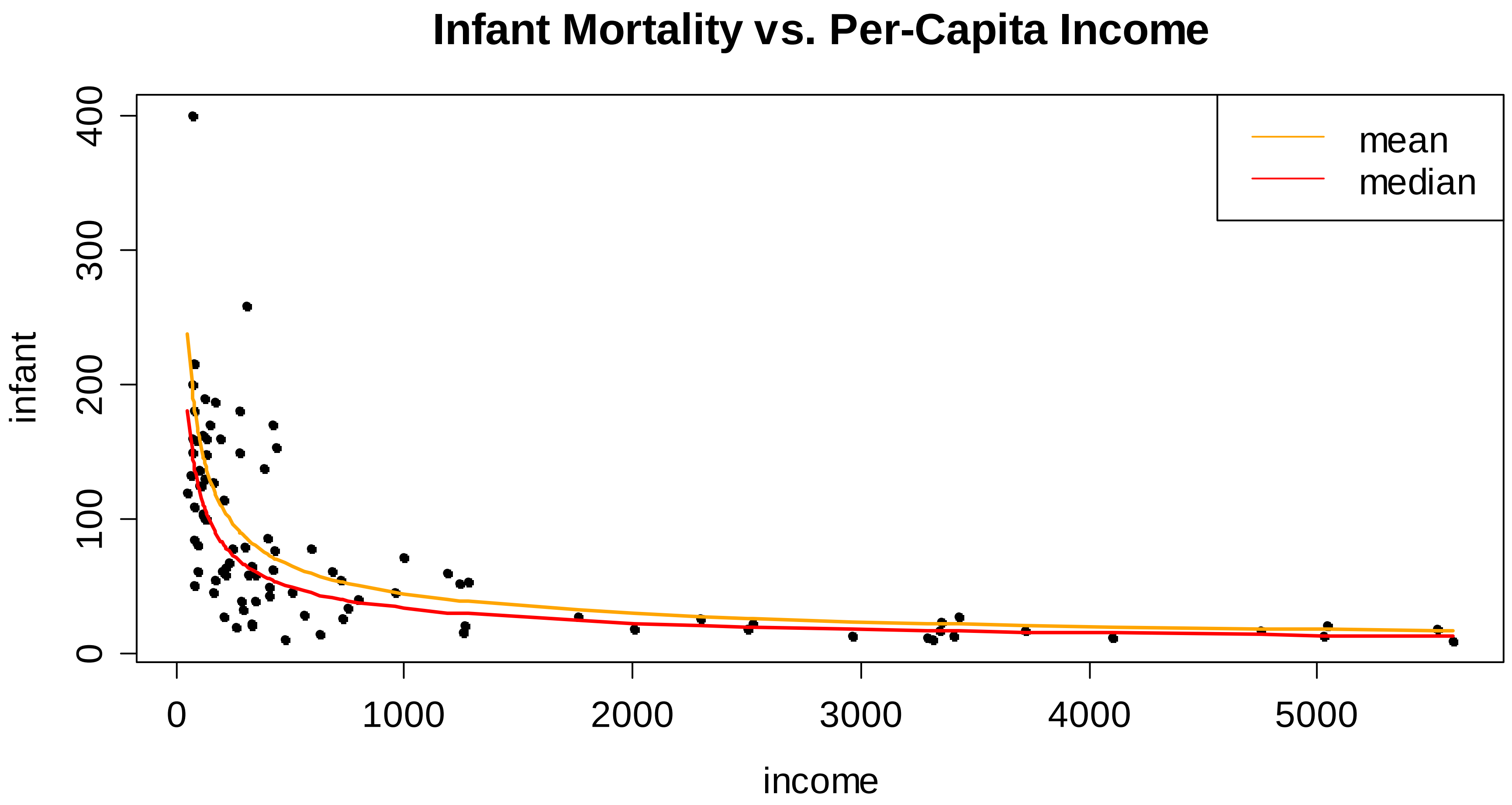
- **Beware:** this is an estimate of the conditional median, but not the conditional mean $E[y \mid x]$. If we require *unbiased estimation*, we need to use a correction factor :

$$\hat{y} = \exp(\hat{y}' + \hat{\sigma}_E^2 / 2)$$

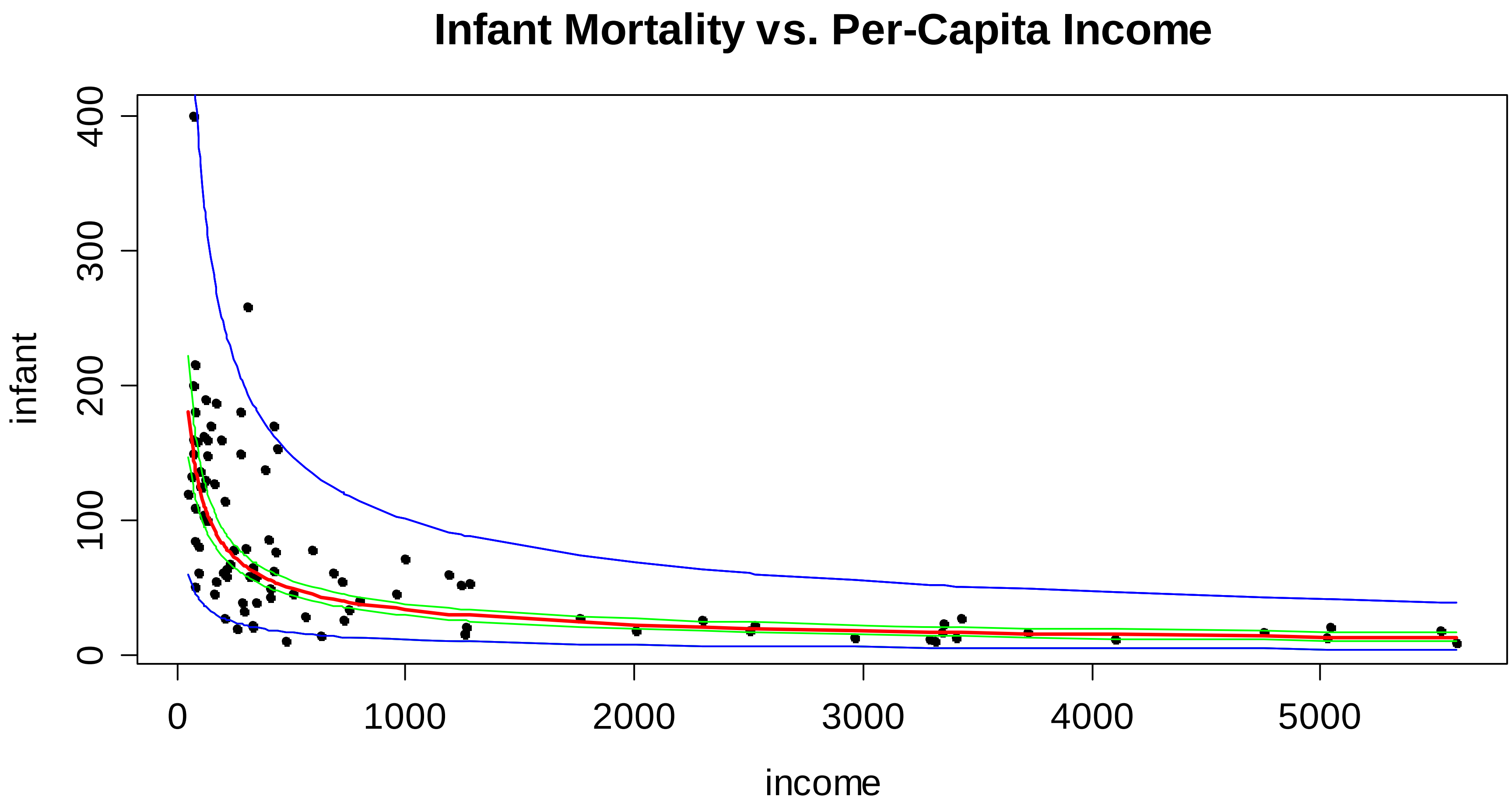
- The confidence and prediction intervals are easy:

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

Conditional Mean and Median



Confidence and Prediction Interval



What to do if y=0 and/or x=0?

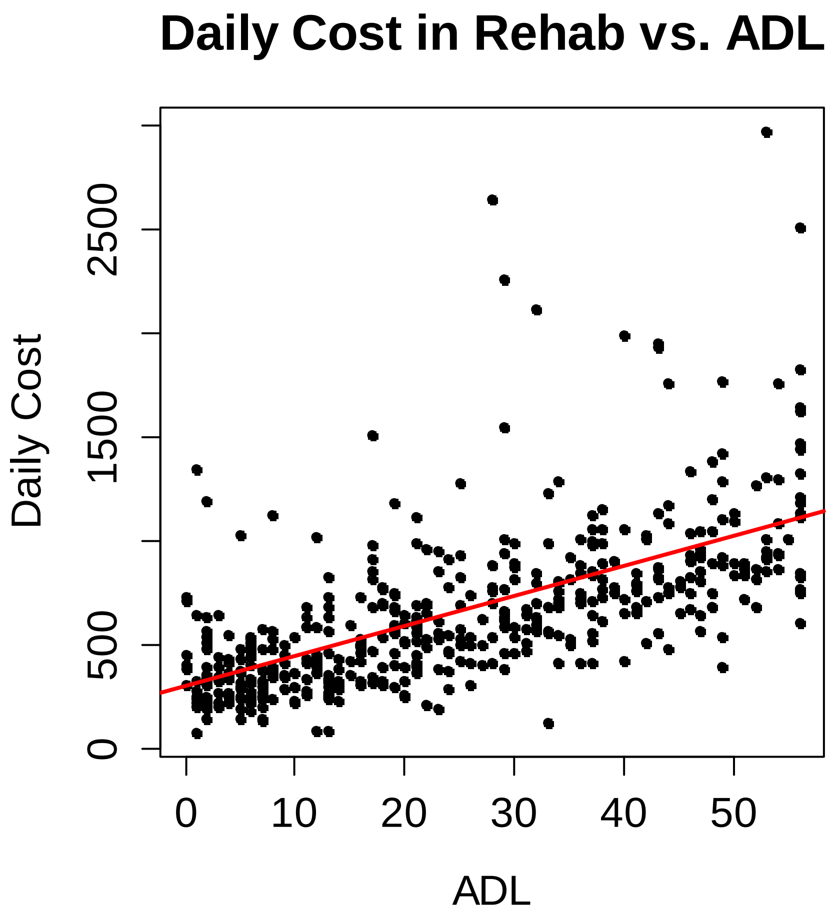
- We can only take logarithms if $x, y > 0$. In cases where the response and/or predictor takes negative values, we should not log-transform. If zero's occur, they need treatment.

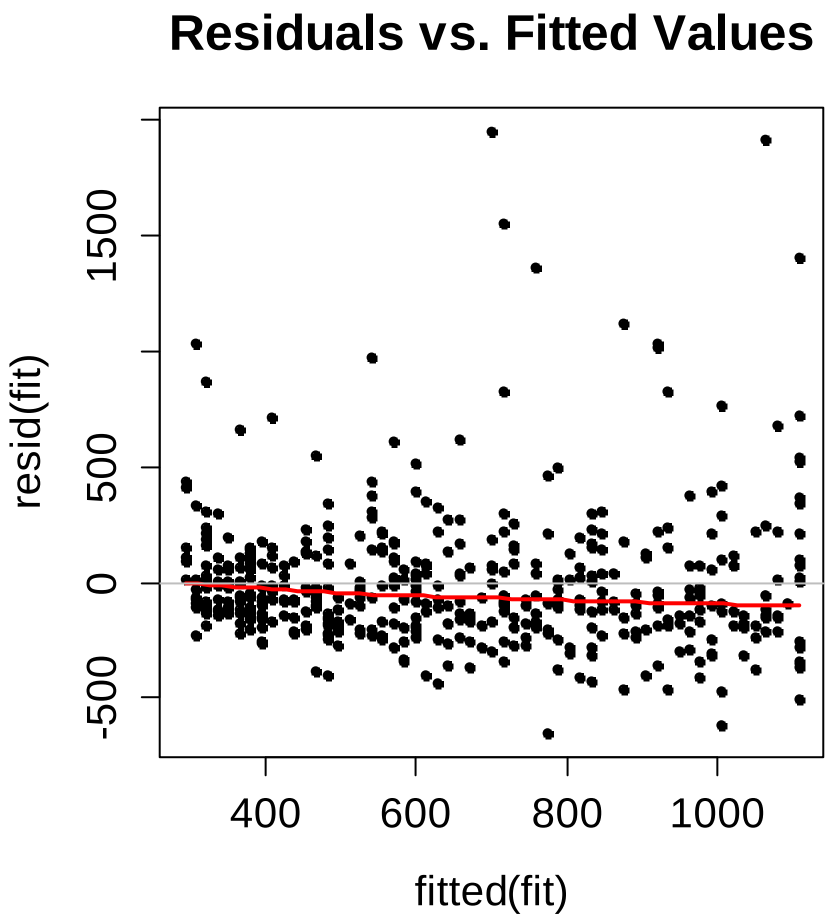
- What do we do with either $x = 0$ or $y = 0$?
  - do never exclude such data points!
  - adding a constant value is allowed!

- What about the choice of the constant?
  - standard choice: $c = 1$
  - scale dependent, thus not recommended!

→ **Set $c =$ smallest value > 0!**

Another Example: Daily Cost in Rehab





Logged Response Model

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

$$y' = \log(y) = \beta_0 + \beta_1 x + E$$

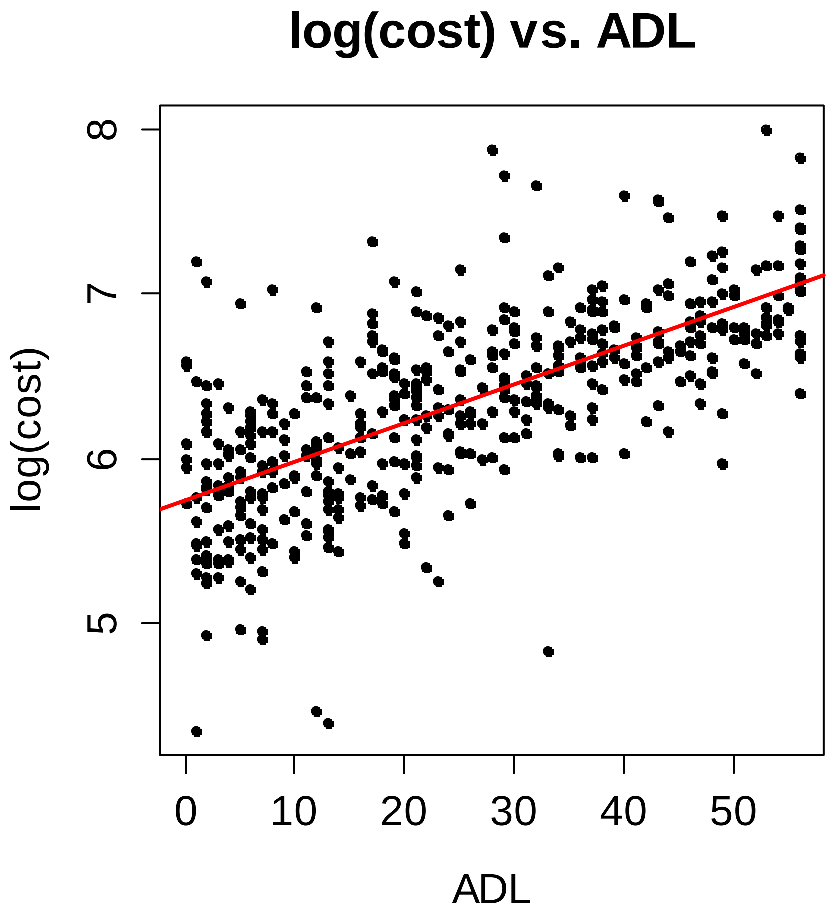
In the *original scale*, we can write the logged response model using the same predictors:

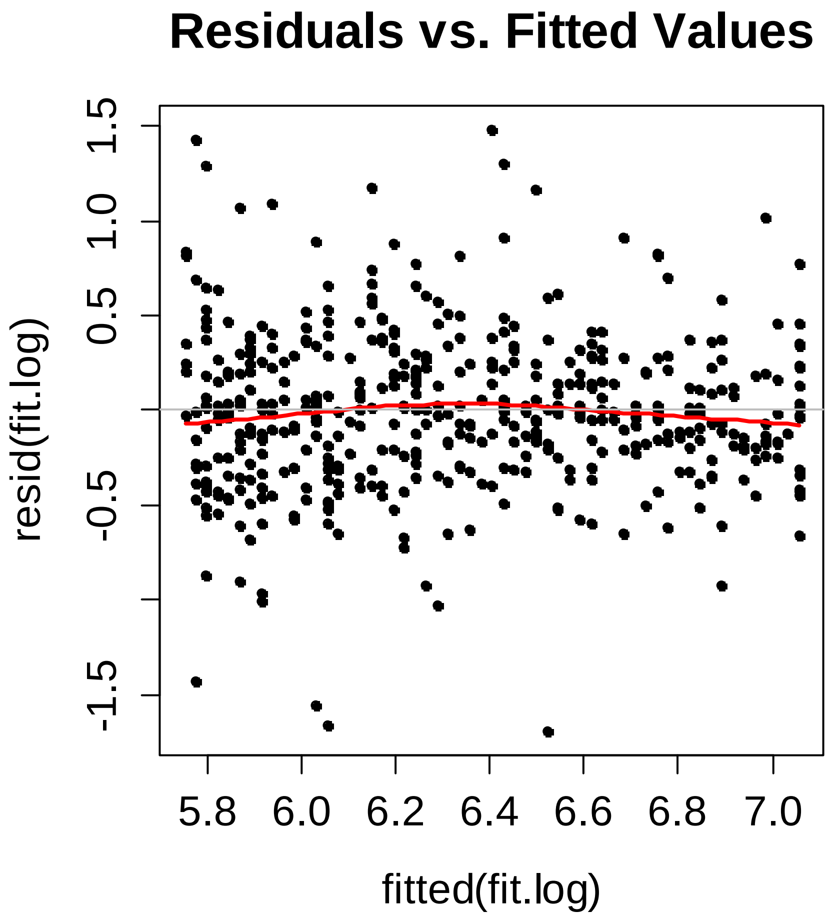
$$y = \exp(\beta_0) \cdot \exp(\beta_1 x) \cdot \exp(E)$$

→ **Multiplicative model**

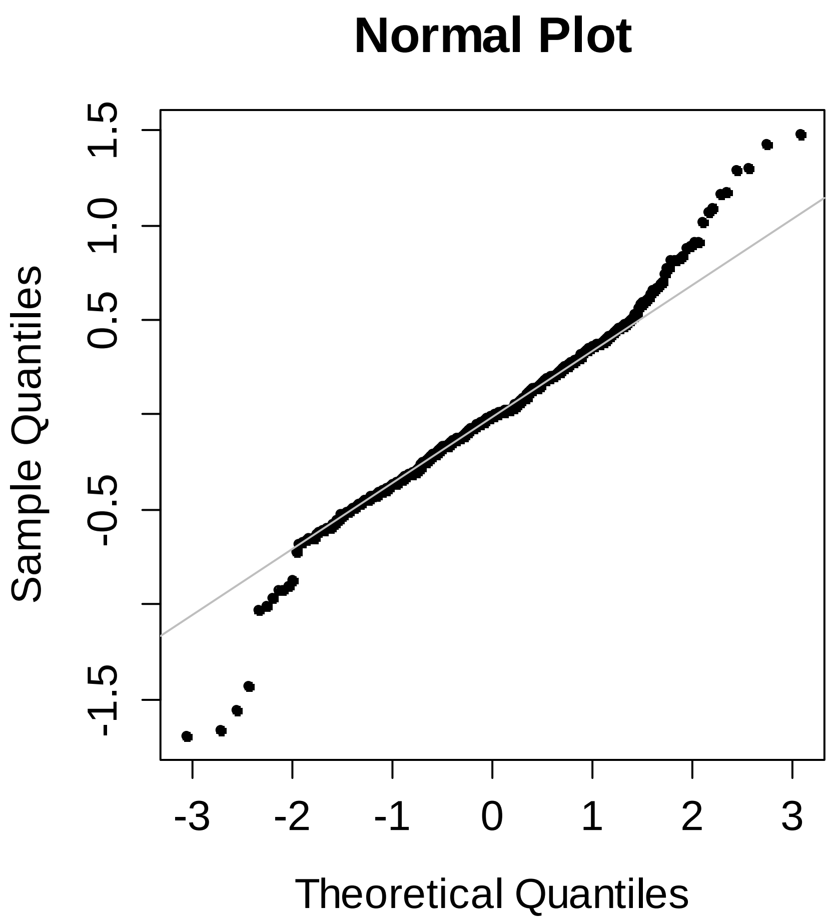
→ $E \sim N(0, \sigma_E^2)$, and thus, $\exp(E)$ has a *lognormal distribution*

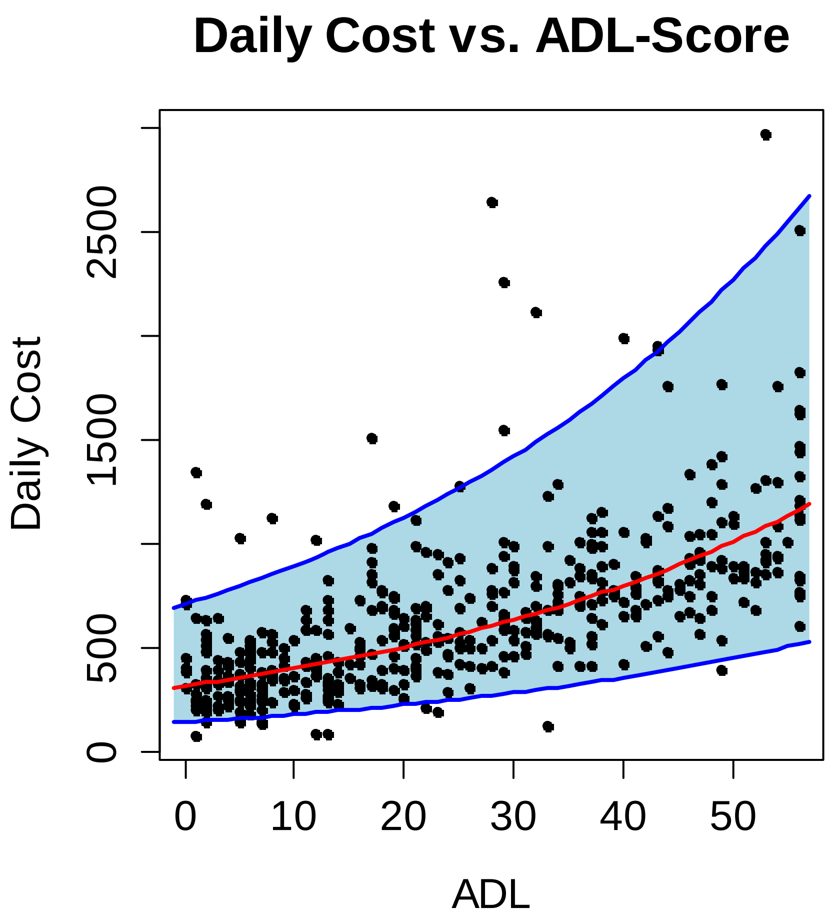
Fit and Residuals after the Transformation





Original Scale: Fit and Prediction Interval





Interpretation of the Coefficients

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

$$\begin{aligned} \log(y) &= \beta_0 + \beta_1 x + E \\ y &= \exp(\beta_0)\exp(\beta_1 x)\exp(E) \end{aligned}$$

An increase by one unit in x would multiply the fitted value in the original scale with $\exp(\beta_1)$.

→ **Coefficients are interpreted multiplicatively!**

When to Transform?

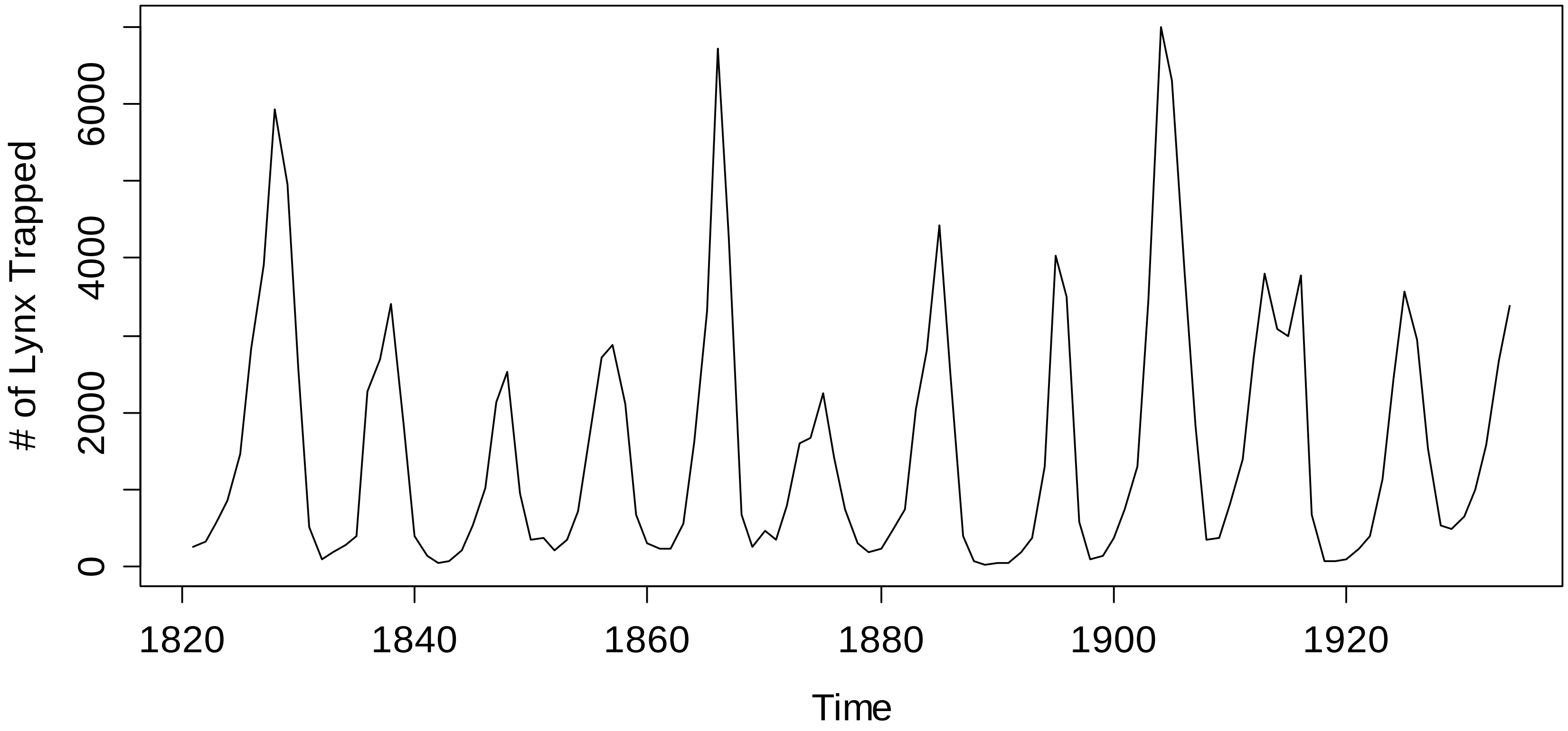
We have seen a few examples where a log-transformation of the response and/or the predictor yields a better fit. Some general rules about when to apply it:

- If the values are on a scale, that is **left-closed** (with 0 as the smallest possible value), but is **open** on the **right**.
- If the marginal distribution of the variable (as we can observe in a histogram) is clearly **right-skewed**.
- If the **scatter**, i.e. the magnitude of the uncertainty, **increases** with increasing value – be this due to theoretical considerations, or due to evidence in the data.

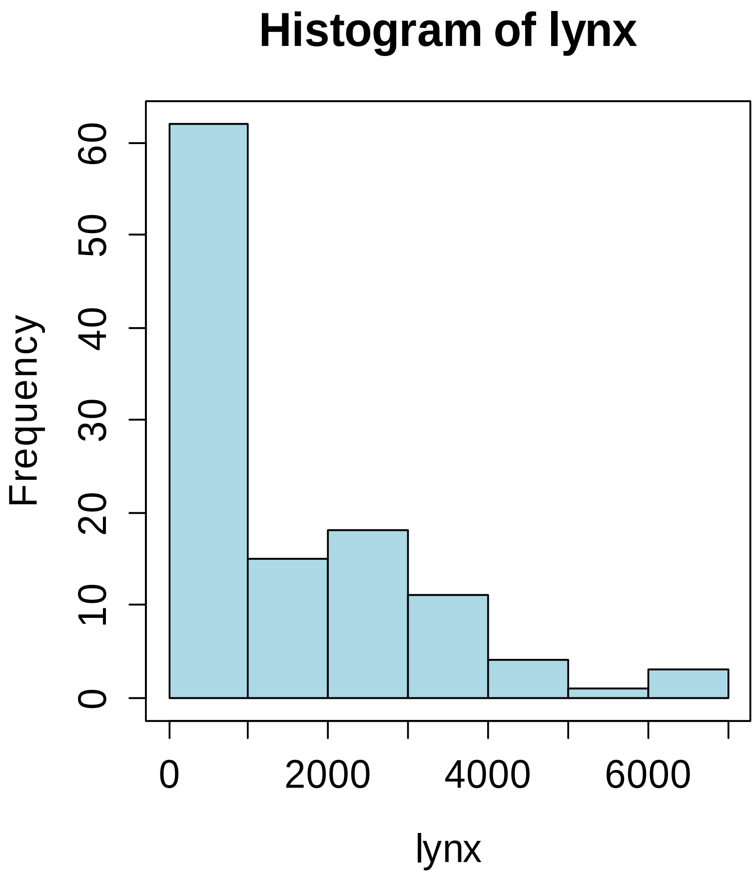
Applied Statistical Regression
AS 2013 – Week 05

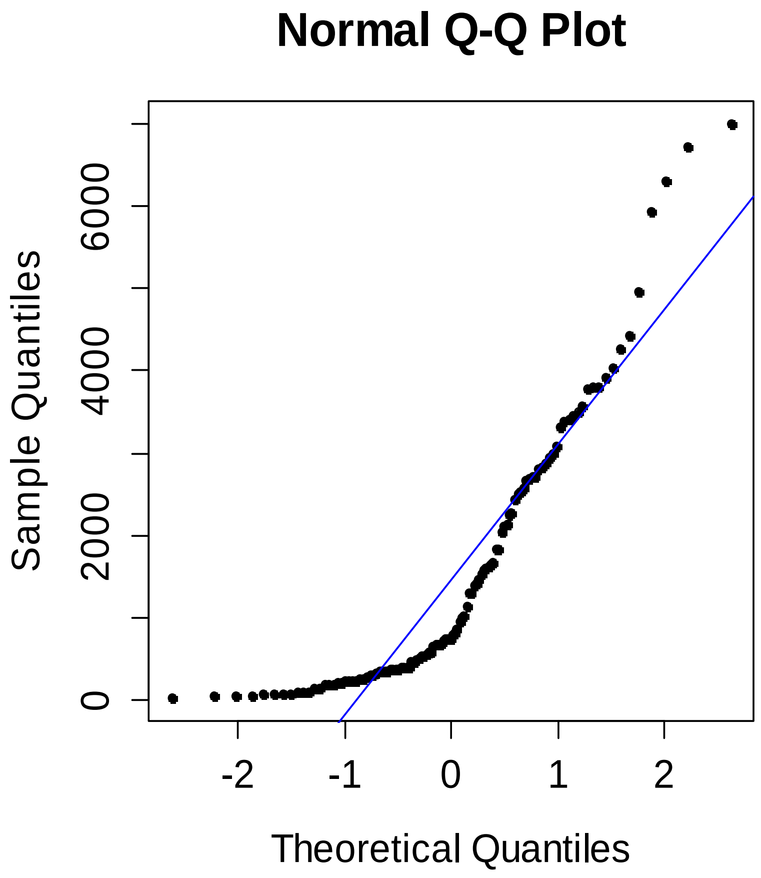
Transformations: Lynx Data

Lynx Trappings



Transformations: Lynx Data





Zurich Airport Data: Re-Evaluation

Both Pax and ATM are variables that only take values ≥ 0.
In our example, we do not observe any right-skewness, but we still try to apply the log-transformation:

$$ATM' = \log(ATM),\ Pax' = \log(Pax)$$

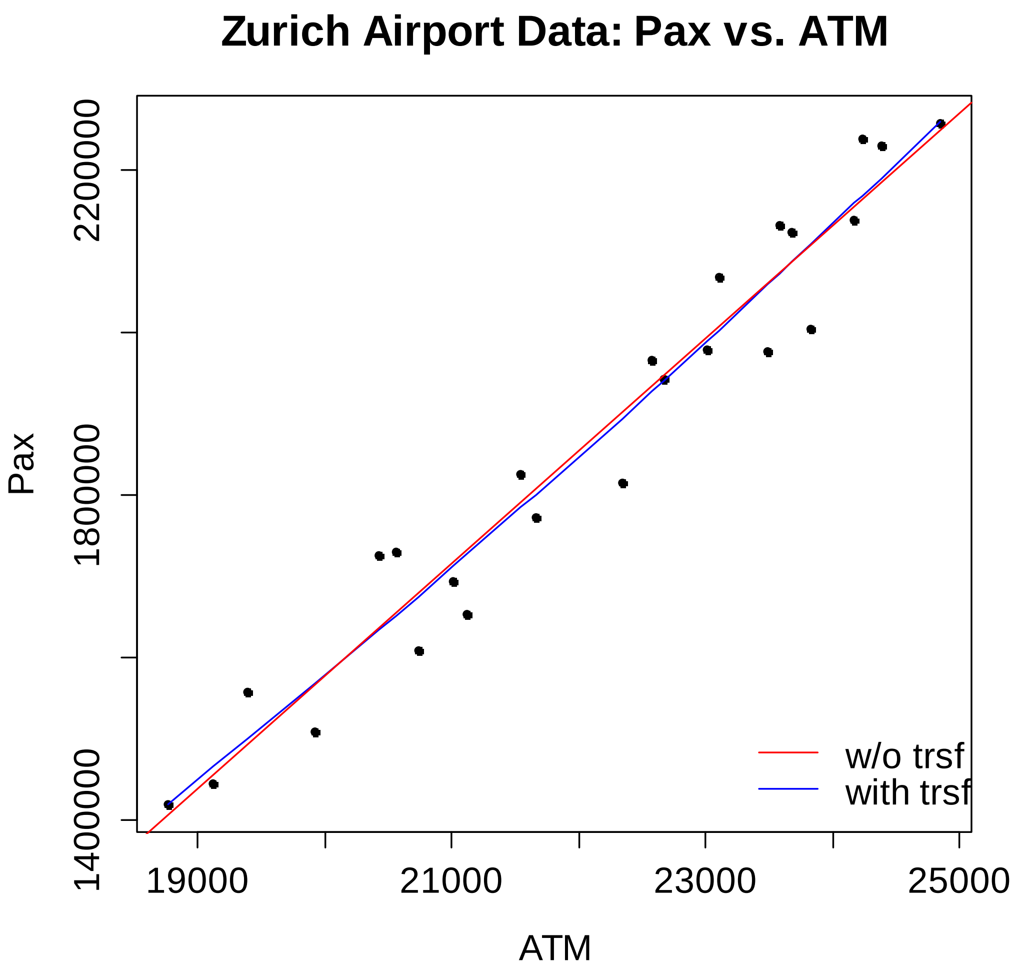
It also has the advantage that the fit goes through (0/0).

```
> fit        <- lm(Pax ~ ATM, data=unique2010)
> fit.log    <- lm(log(Pax) ~ log(ATM), data=unique2010)
> fit.y.orig <- exp(fitted(fit.log)[order(unique2010$ATM)])
> plot(Pax ~ ATM, data=unique2010, pch=20)
> lines(sort(unique2010$ATM), fit.y.orig, col="blue")
> abline(fit, col="red")
```

The difference in the fitted line is only small, but important!

Applied Statistical Regression
AS 2013 – Week 05

Zurich Airport Data: Re-Evaluation



We estimate $\hat{\beta}_1 = 1.655$. If ATM increases by 1% then Pax will increase by 1.655%.

This reflects that during high season, bigger air-planes are used, and the seat load factor is better.

Comparing the Residual Plots