Applied Statistical Regression
AS 2015 – Simple Regression

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, September 21, 2015

Applied Statistical Regression
AS 2015 – Simple Regression

Course Organization

Your Lecturer:

Marcel Dettling

Dr. Math. ETH, i.e. Statistician

Lecturer @ ZHAW and @ ETH Zürich

Researcher in Applied Statistics @ ZHAW

About this Course:

Topics: simple & multiple linear regression, GLM

Materials: script, slides, exercises, sample solutions

Teaching: 2 lectures weekly, exercises every 2nd week

Important: all details and materials are on the course webpage!

What is Regression?

The answer to an everyday question:

How does a target variable of special interest depend on several other (explanatory) factors or causes.

Examples:

- growth of plants, depends on fertilizer, soil quality, …
- apartment rents, depends on size, location, furnishment, …
- car insurance premium, depends on age, sex, nationality, …

Regression:

- quantitatively describes relation between predictors and target
- high importance, most widely used statistical methodology

Applied Statistical Regression
AS 2015 – Simple Regression

Regression Mathematics

→ **See blackboard...**

Applied Statistical Regression
AS 2015 – Simple Regression

What is Regression?

Example: *Fresh Water Tank on* edelweiss air *Planes*

- **Earlier**: it was impossible to predict the amount of fresh water needed, the tank was always filled to 100% at Zurich airport.
- **Goal**: Minimizing the amount of fresh water that is carried. This lowers the weight, and thus fuel consumption and cost.
- **Task***:* Modelling the relation between fresh water consumption and # *of passengers*, *flight duration*, *daytime*, *destination*, … Furthermore, quantifying what is needed as a reserve.
- **Method:** *Multiple linear regression model*

Regression: Goals

1) Understanding the relation between y and $x_1,...,x_p$

The aim is to pin down which of the predictors have influence on the response variable, as well as to quantify the strength of this relation. There is a battery of statistics and tests that address these questions.

2) Prediction

The regression equation can be used for predicting the expected response value $\hat{y}$ for an arbitrary predictor configuration $x_1,...,x_p$. We will not only generate point predictions, but can also attribute a prediction interval that quantifies the involved uncertainty.

Simple Regression

In this course, we first discuss *simple regression*, where there is only one single predictor variable. Later, we will extend this to *multiple regression*, where many predictors can be present.

Advantages of discussing simple regression:

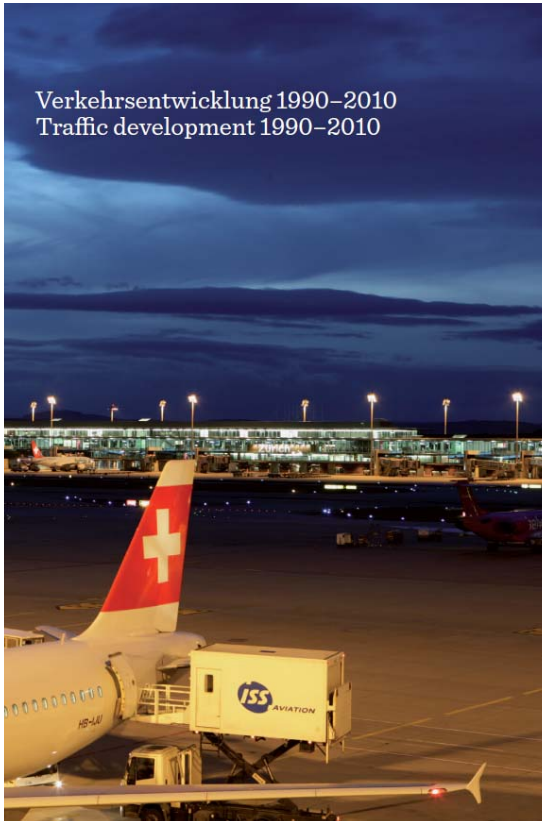
- **Visualization of data and fit is possible**
- **Corresponds to estimating a straight line or curve**
- **Is also mathematically simpler and more intuitive**

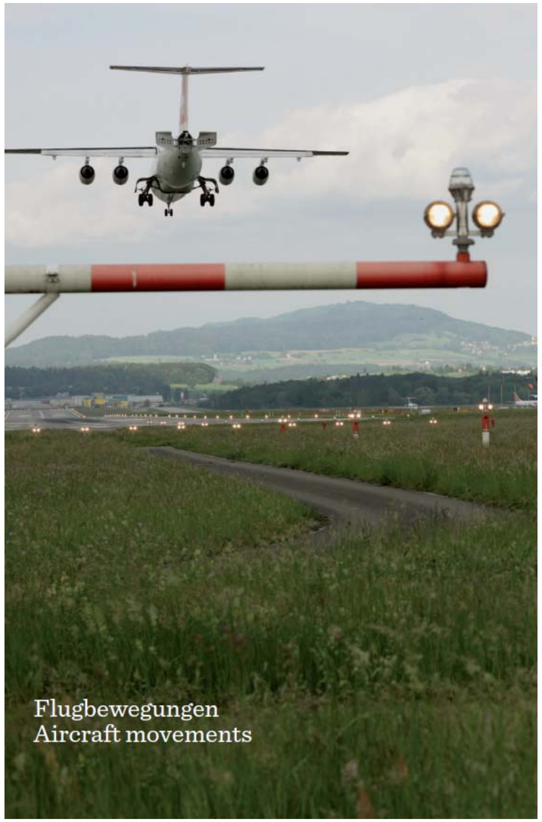
We start out with *smoothing*, i.e. fitting non-parametric curves. Then, we will proceed with discussing linear models, i.e. the classical parametric regression approach.

Applied Statistical Regression
AS 2015 – Simple Regression

Example: Airline Passengers

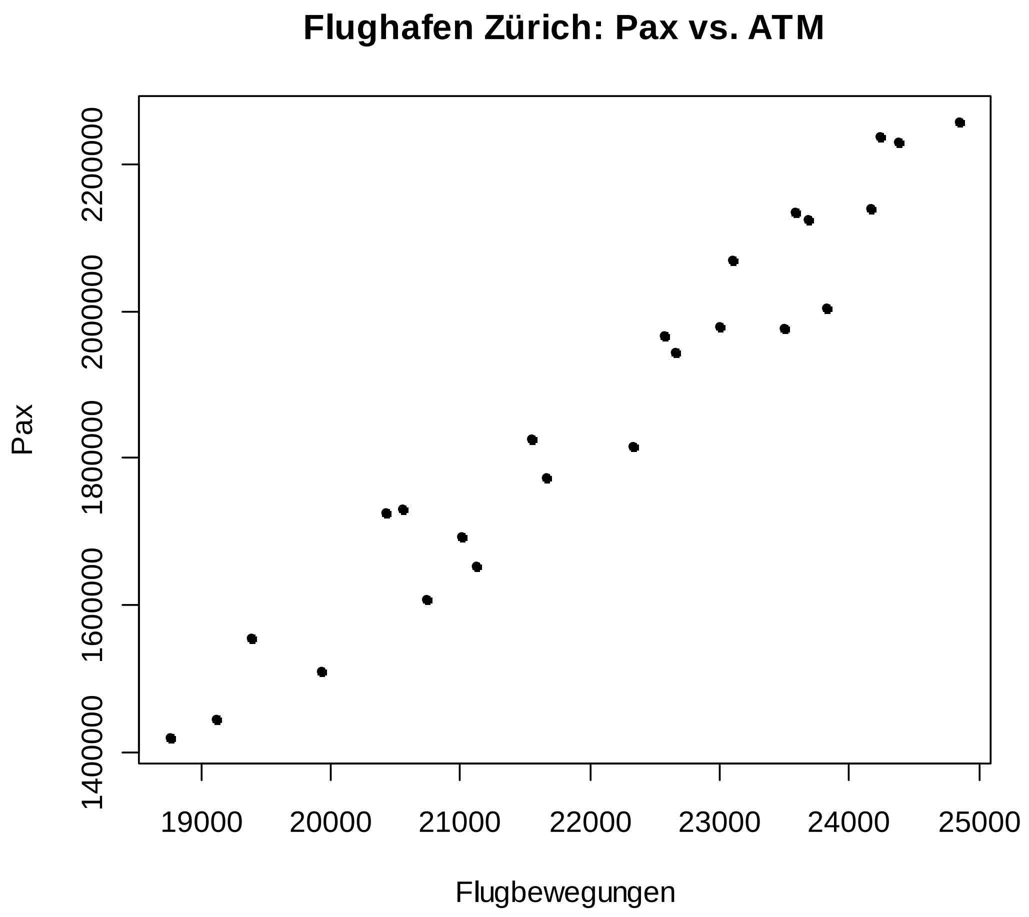
Each month, Zurich Airport publishes the number of air traffic movements and airline passengers. We study their relation.





Example: Airline Passengers

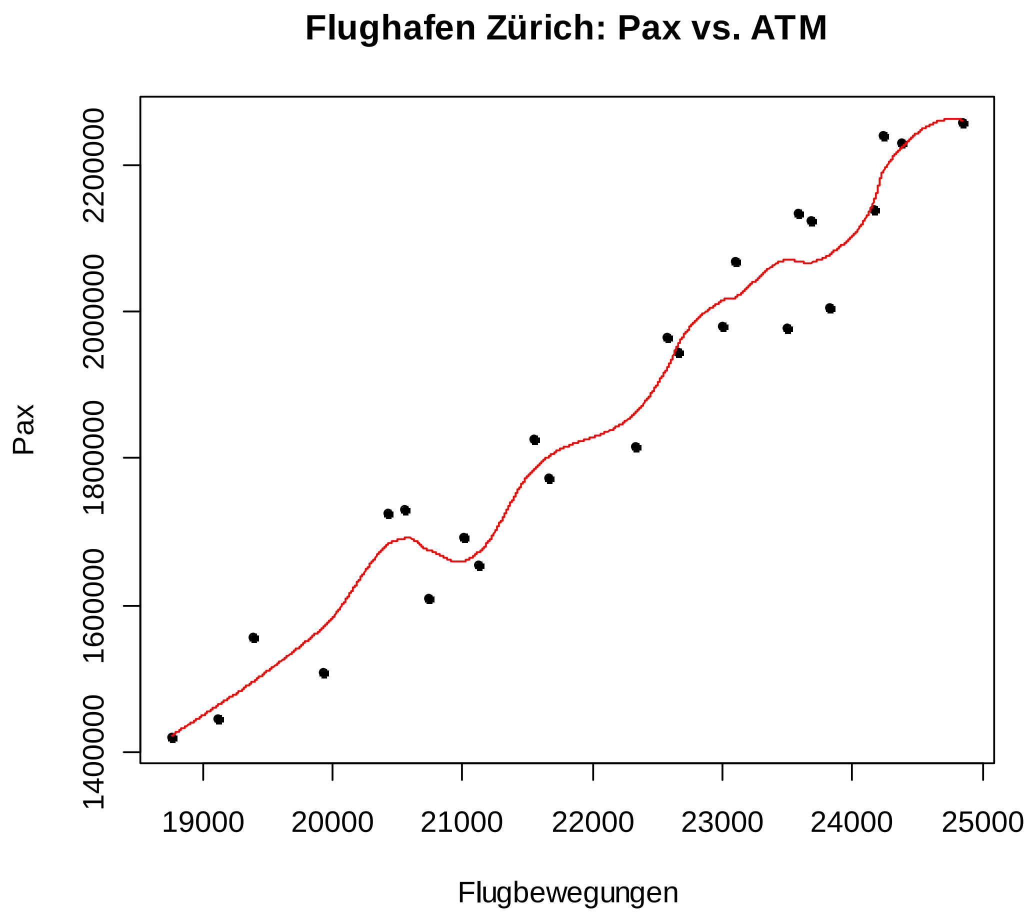
Month	Pax	ATM
2010-12	1‘730‘629	22‘666
2010-11	1‘772‘821	22‘579
2010-10	2‘238‘314	24‘234
2010-09	2‘139‘404	24‘172
2010-08	2‘230‘150	24‘377
...	...	...



Smoothing

We may use an arbitrary smooth function $f(\cdot)$ for capturing the relation between Pax and ATM.

- It should fit well, but not follow the data too closely.
- The question is how the line/function are obtained.

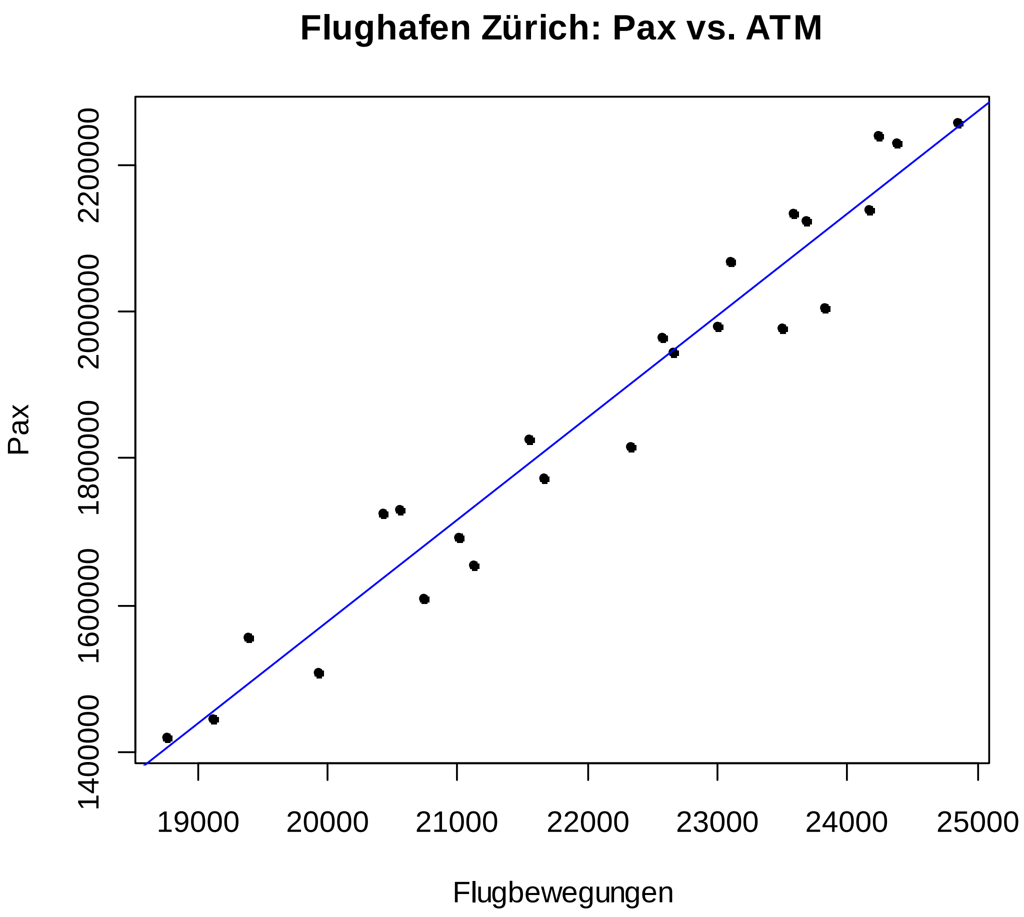


Applied Statistical Regression
AS 2015 – Simple Regression

Linear Modeling

A straight line represents the systematic relation between Pax and ATM.

- Only appropriate if the true relation is indeed a straight line
- The question is how the line/function are obtained.



Smoothing vs. Linear Modeling

Advantages and disadvantages of *smoothing*:

+ Flexibility
+ No assumptions are made
- Functional form remains unknown
- Danger of overfitting

Advantages and disadvantages of *linear modelling*:

+ Formal inference on the relation is possible
+ Better efficiency, i.e. less data required
- Only reasonable if the relation is linear
- Might falsely imply causality

Smoothing

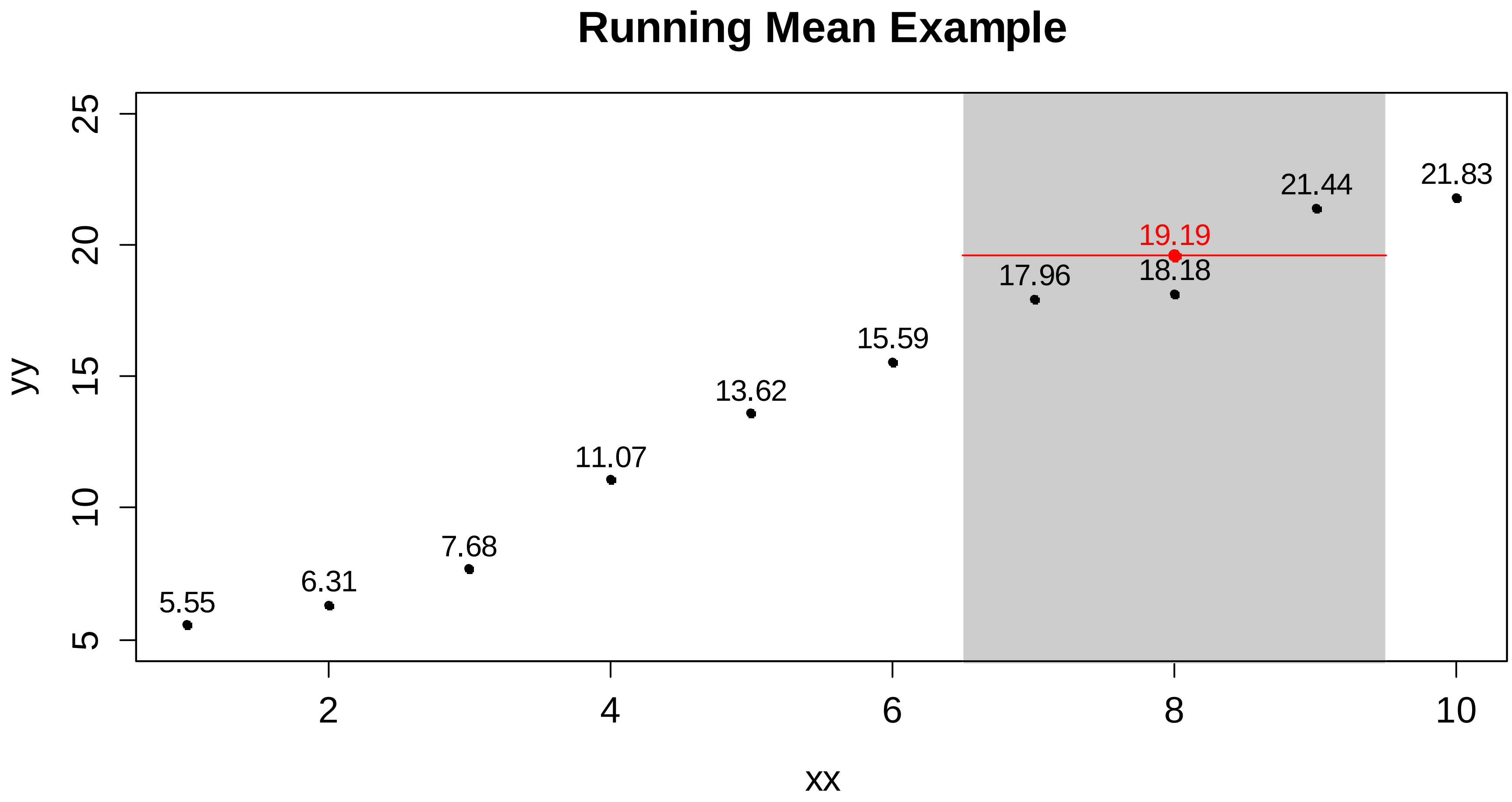
Our goal is *visualizing* the relation between the y / response variable Pax and the x / predictor variable ATM.

→ we are not after a functional description of $f(\cdot)$

Since there is no parametric function that describes the response vs. predictor relation, smoothing is also termed **non-parametric regression analysis**.

Method/Idea: *"Running Mean"*
- take a window of x-values
- compute the mean of the y-values within the window
- this is an estimate for the function value at the window center

Running Mean: Example



Running Mean: Mathematics

RunningMean(x) = Mean of y-values over a window with width $\pm\lambda/2$ around x.

The *estimate* for $f(\cdot)$, denoted as $\hat{f}_\lambda(\cdot)$, is defined as follows:

$$\hat{f}_\lambda(x) = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{j=1}^{n} w_i},$$

The *weights* are defined as $w_i = \begin{cases} 1 & if\ |x - x_j| \leq \lambda/2 \\ 0 & else \end{cases}$,
and λ is the *window width*.

Running Mean: R-Implementation

- As an introductory exercise, it is instructive to code a function that computes and visualizes the running mean.

  Arguments: `xx=` x values
  `yy=` y values
  `width=` window width
  `steps=` # of points computed

- Alternatively, one can simply use function `ksmooth()`. The window size can be adjusted by argument `bandwidth=...` Some other settings can be made, especially with respect to evaluation.

→ **We will now study the running mean fit...**

Running Mean: R-Implementation

Kernel Regression Smoother

Description

The Nadaraya–Watson kernel regression estimate.

Usage

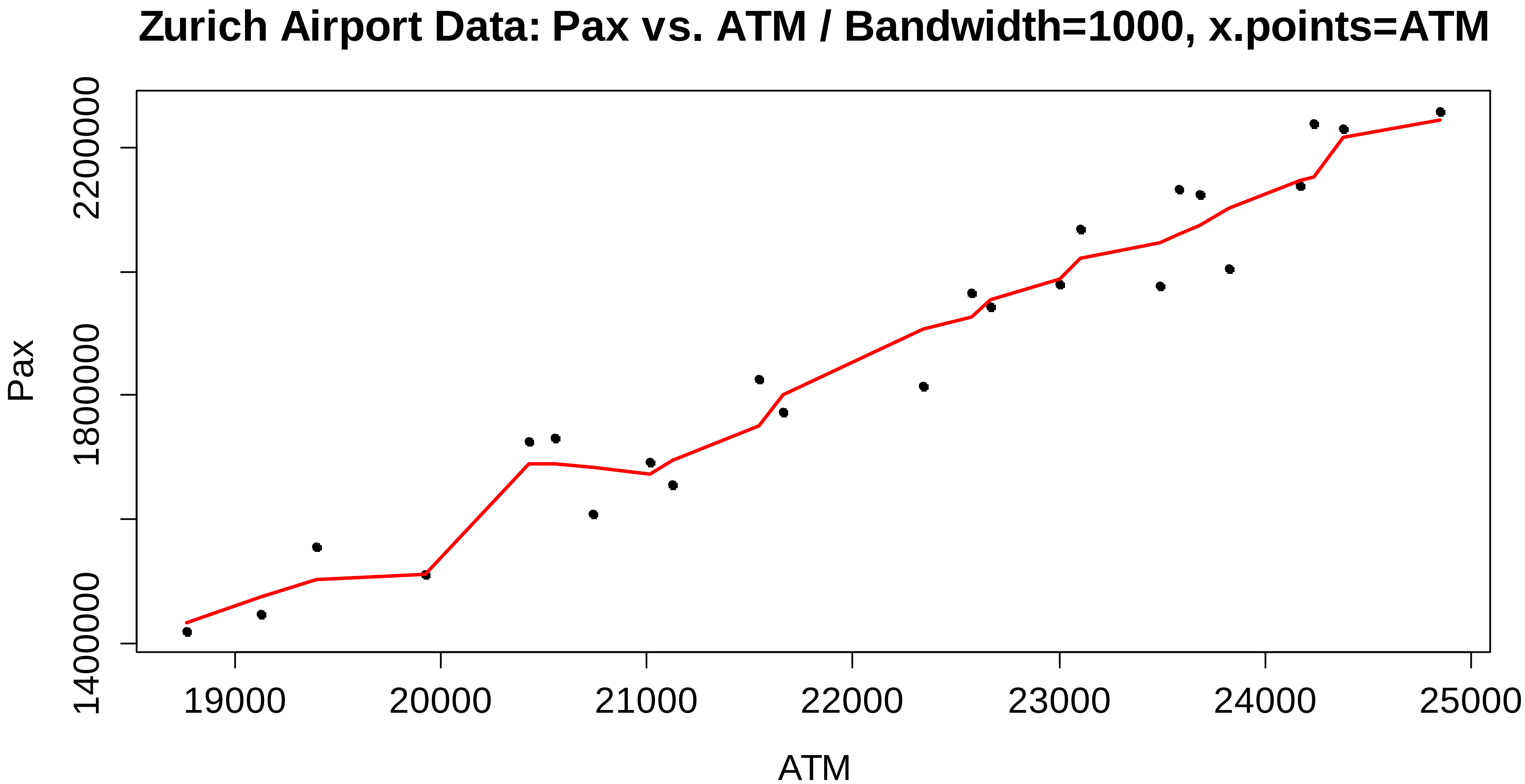
```
ksmooth(x, y, kernel = c("box", "normal"), bandwidth = 0.5,
        range.x = range(x),
        n.points = max(100, length(x)), x.points)
```

Arguments

`x`	input x values
`y`	input y values
`kernel`	the kernel to be used.
`bandwidth`	the bandwidth. The kernels are scaled so that their quartiles (viewed as probability densities) are at +/- `0.25*bandwidth`.
`range.x`	the range of points to be covered in the output.
`n.points`	the number of points at which to evaluate the fit.
`x.points`	points at which to evaluate the smoothed fit. If missing, `n.points` are chosen uniformly to cover `range.x`.

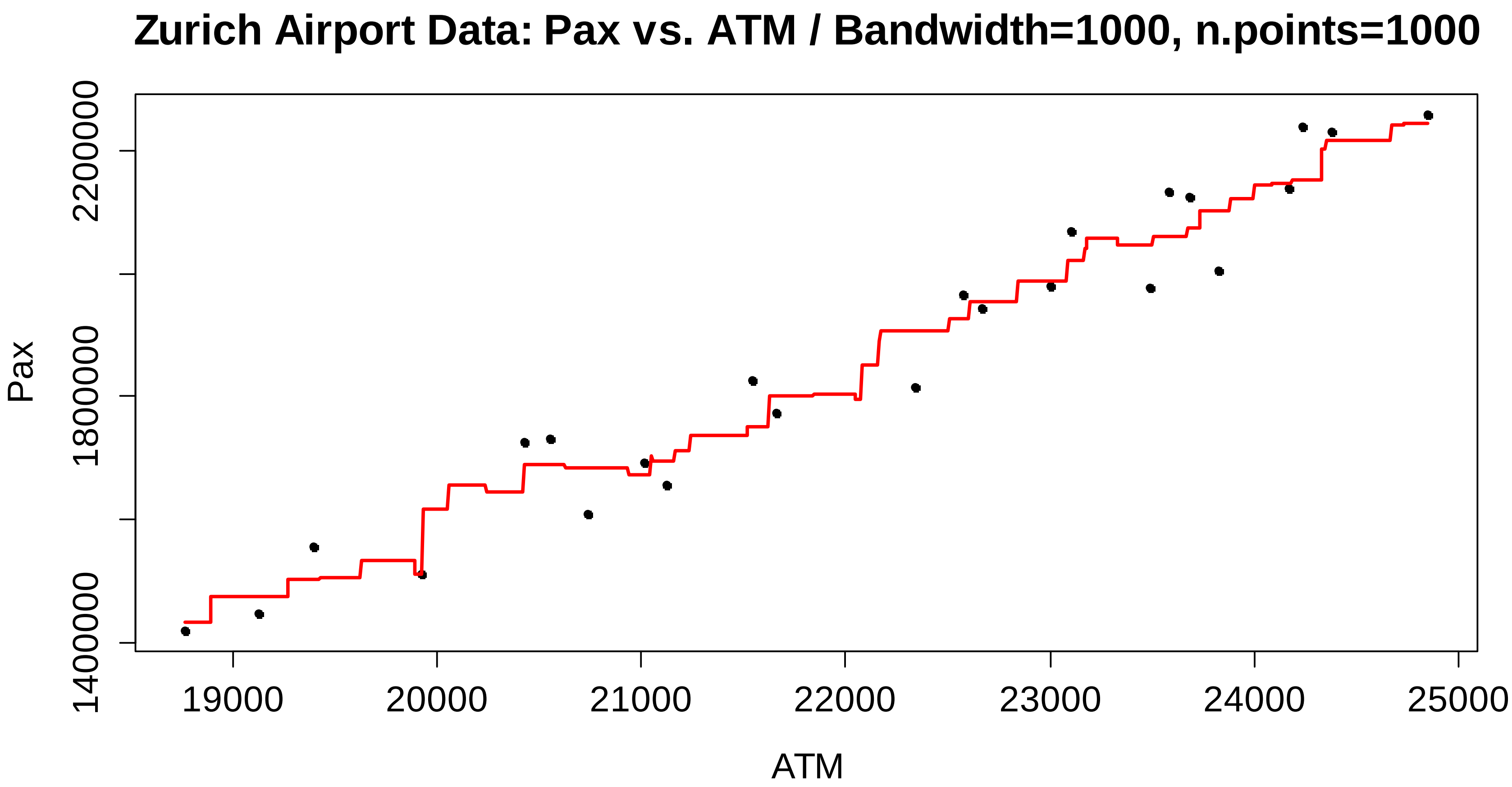
Running Mean: Unique Data

```
> fit <- ksmooth(ATM, Pax, kernel="box", bandwidth=1000,…)
> lines(fit, col="red", lwd=2)
```



Running Mean: Unique Data

```
> fit <- ksmooth(ATM, Pax, kernel="box", n.points=1000,…)
> lines(fit, col="red", lwd=2)
```



Running Mean: Drawbacks

- The finer grained the evaluation points are, the less smooth the fitted function turns out to be. This is unwanted.
  Reason: *data points are "lost" abruptly.*

- For large window width, we loose a lot of information on the boundaries. For small windows however, we may have too few points withing the window, and thus instability.

→ ***There are much better smoothing algorithms!***

We will introduce:

a) a *Gaussian Kernel Smoother*, and
b) the robust *LOESS-Smoother*

Gaussian Kernel Smoother

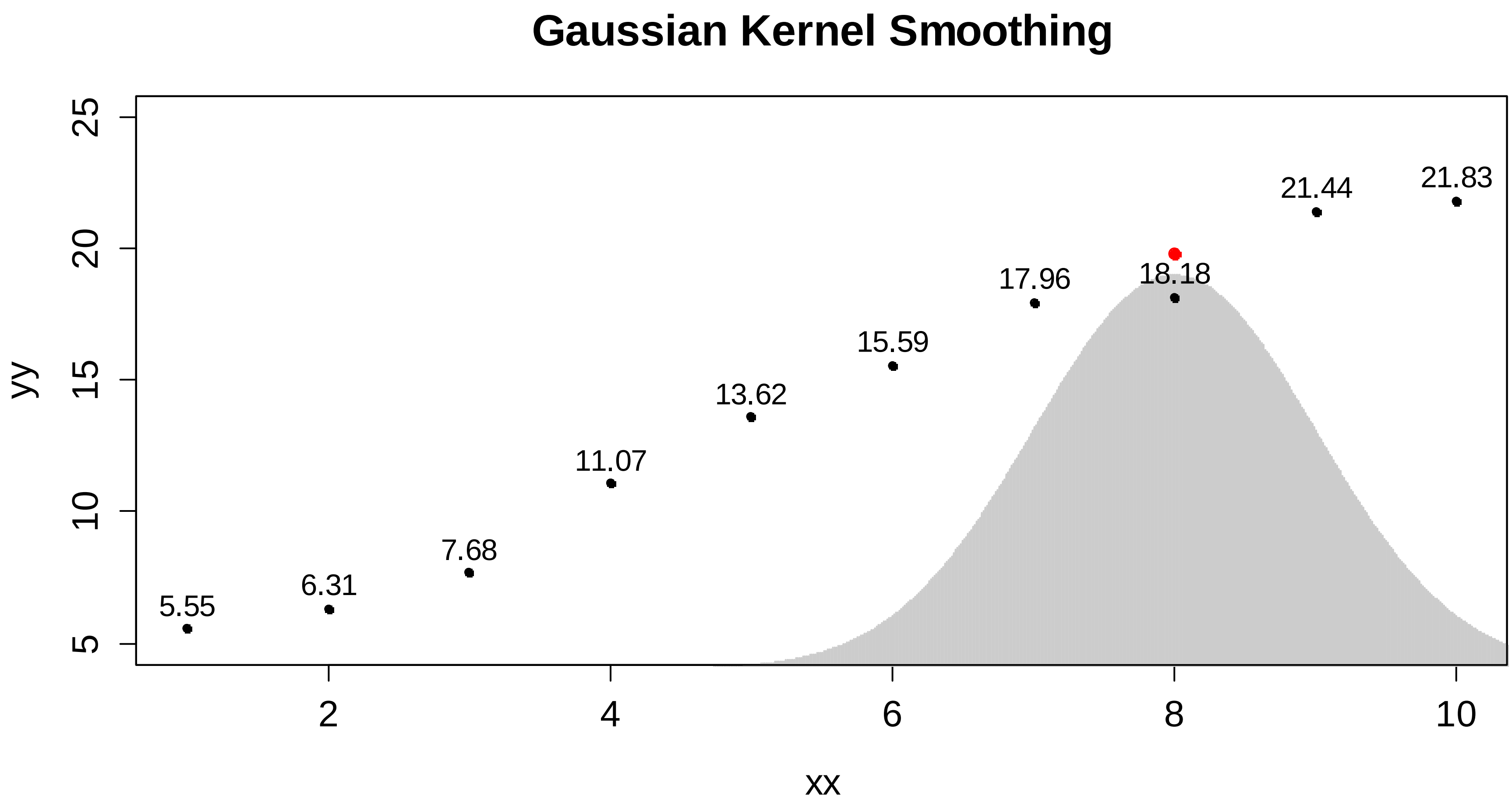
KernelSmoother(x) = Gaussian bell curve weighted average of y-values around x.

The estimate for $f(\cdot)$, denoted as $\hat{f}_\lambda(\cdot)$, is defined as:

$$\hat{f}_\lambda(x) = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \quad ,$$

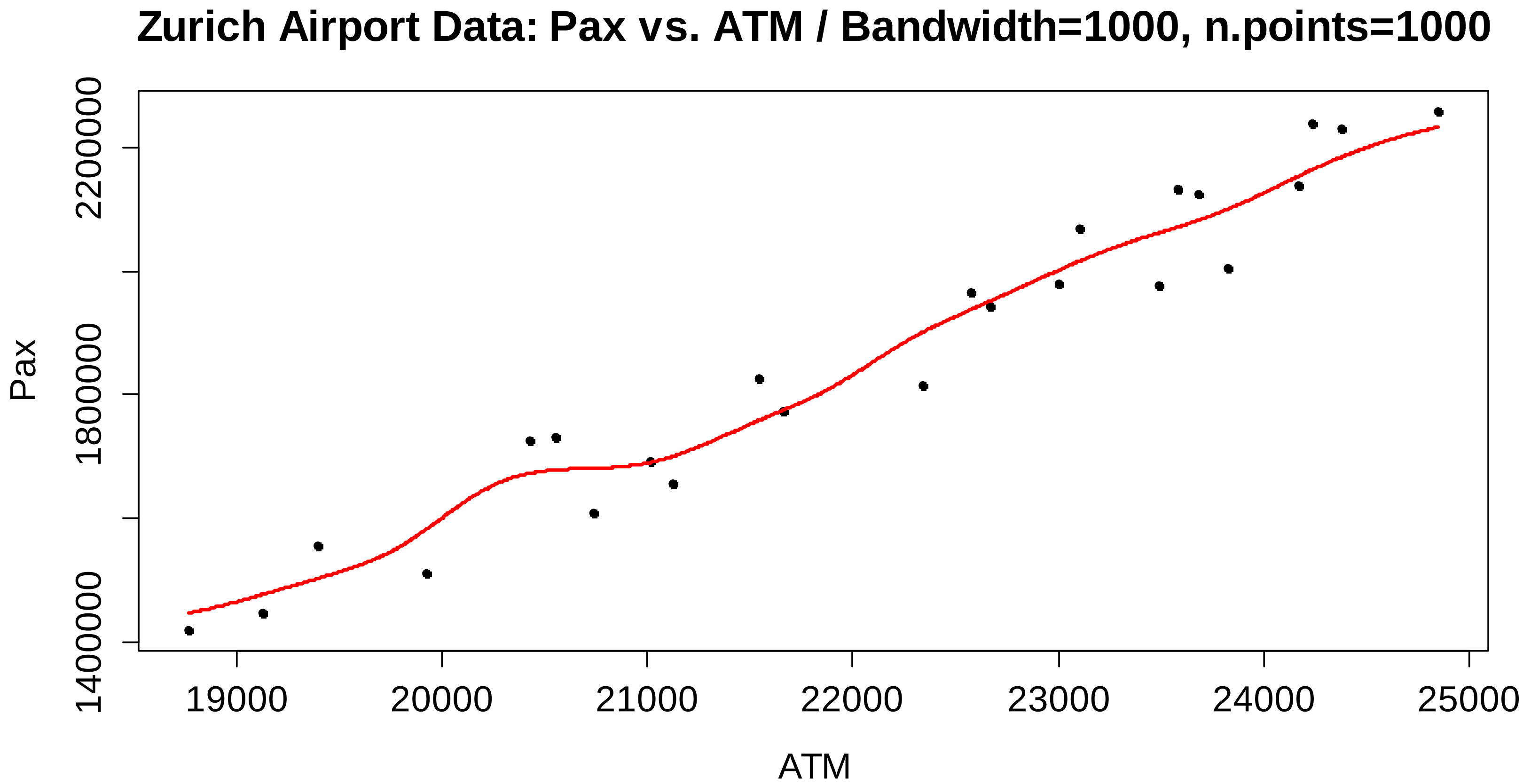
The weights are defined as: $w_i = \exp\left(-\frac{(x - x_i)^2}{\lambda} \right)$, i.e.
the window is infinitely wide,
but distant observation obtain little weight.

Gaussian Kernel Smoother: Idea



Gaussian Kernel Smoother: Unique Data

```
> ks.gauss <- ksmooth(ATM, Pax, kernel="normal", band=1000)
> plot(ATM, Pax, xlab="ATM", ylab="Pax", pch=20)
> lines(ks.gauss, col="darkgreen", lwd=1.5)
```



LOESS-Smoother

The LOESS-Smoother is better, more flexible and more robust than the Gaussian Kernel Smoother. It should be prefered!

It works as follows:

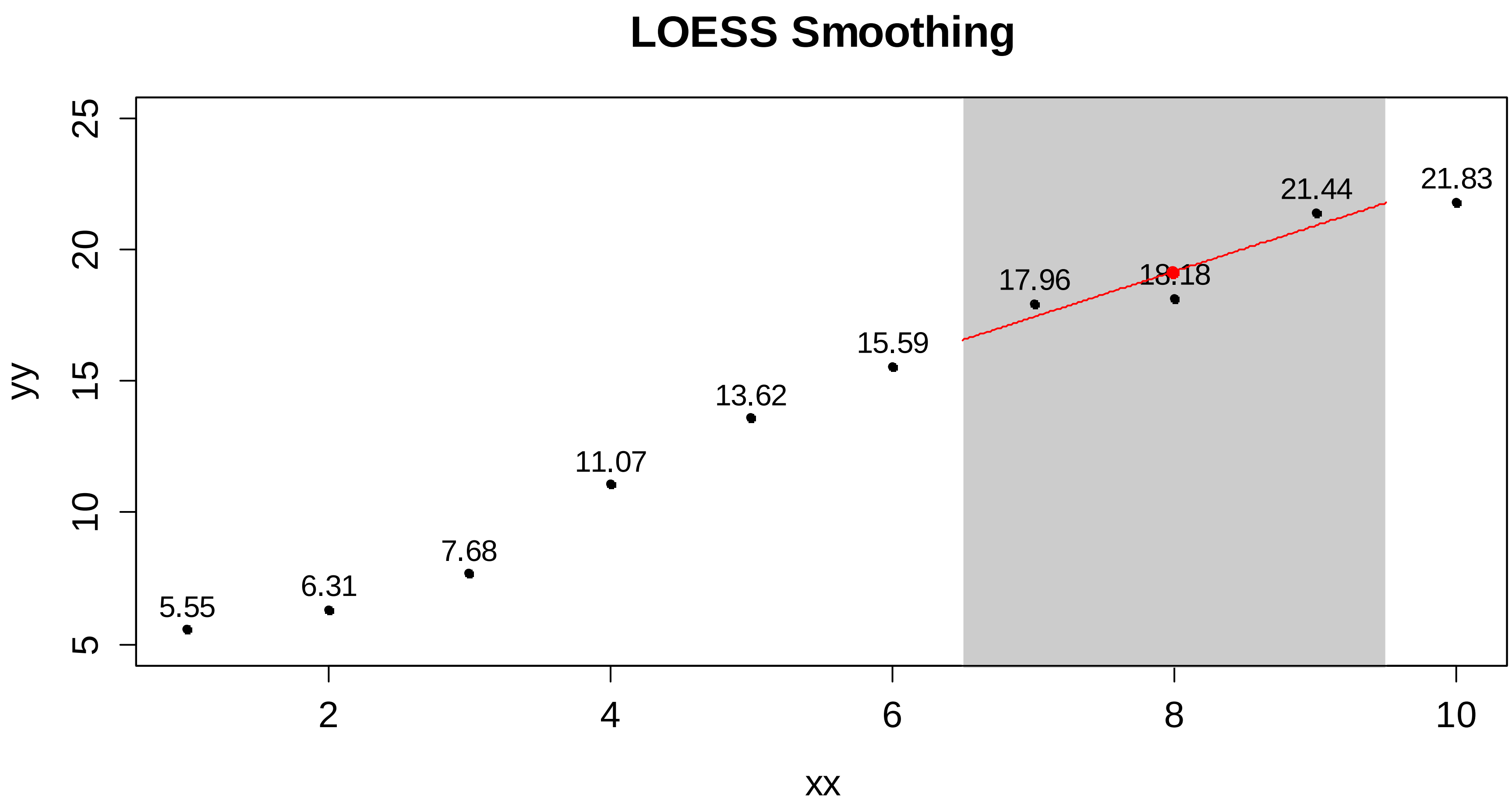
1) *Choose a window of fixed width*

2) *For this window, a straight line (or a parabola) is fitted to the datapoints within, using a robust fitting method.*

3) *Predicted value at window center := fitted value*

4) *Slide the window over the entire x-range*

LOESS-Smoother: Idea



LOESS-Smoother: R-Implementation

Scatter Plot with Smooth Curve Fitted by Loess

Description

Plot and add a smooth curve computed by `loess` to a scatter plot.

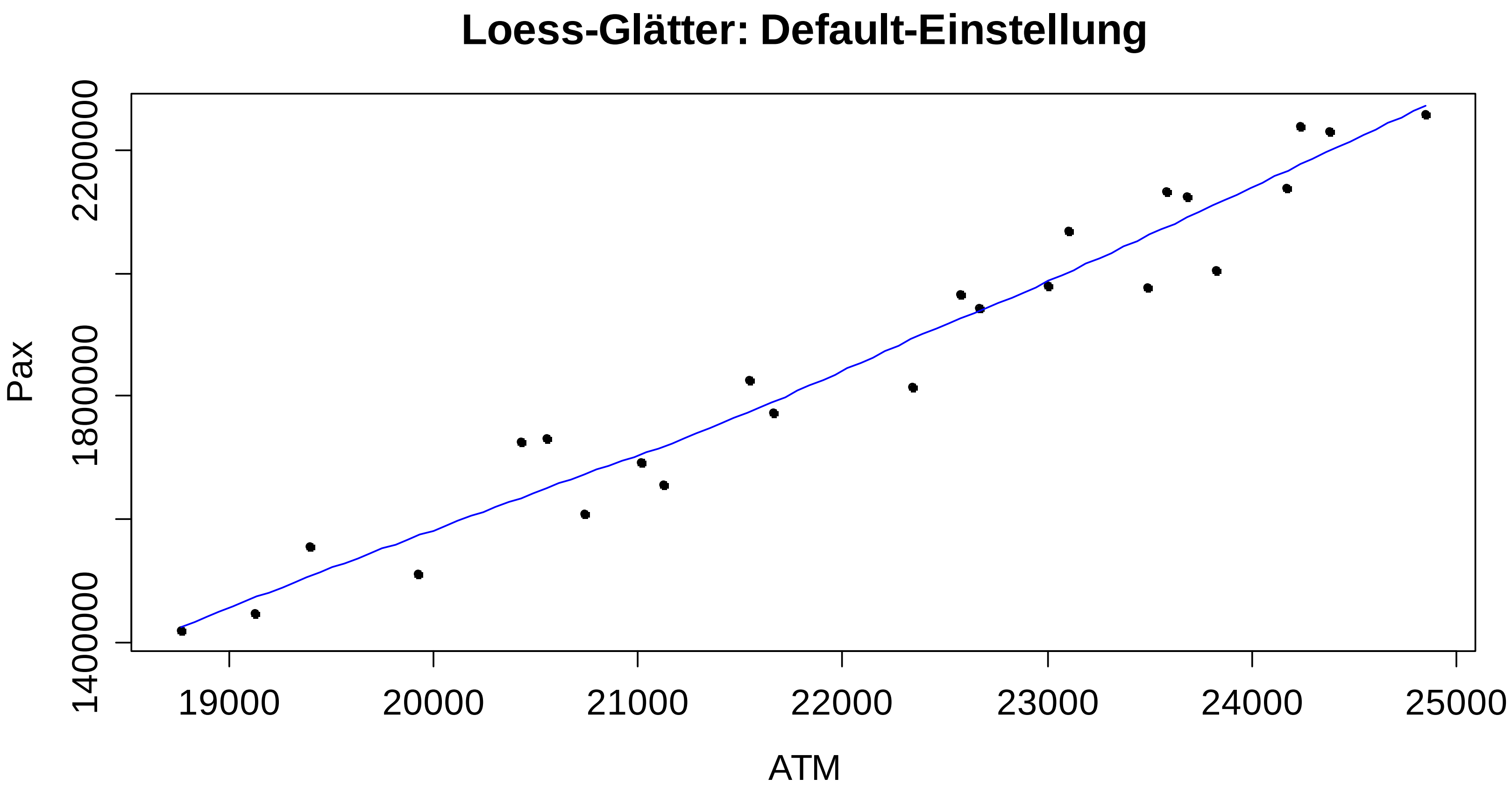
```
loess.smooth(x, y, span = 2/3, degree = 1,
    family = c("symmetric", "gaussian"), evaluation = 50, ...)
```

Arguments

`x, y`	the `x` and `y` arguments provide the x and y coordinates for the plot. Any reasonable way of defining the coordinates is acceptable. See the function `xy.coords` for details.
`span`	smoothness parameter for `loess`.
`degree`	degree of local polynomial used.
`family`	if `"gaussian"` fitting is by least-squares, and if `family="symmetric"` a re-descending M estimator is used.
`xlab`	label for x axis.
`ylab`	label for y axis.
`ylim`	the y limits of the plot.
`evaluation`	number of points at which to evaluate the smooth curve.

LOESS-Smoother: Unique Data

```
> smoo  <- loess.smooth(unique2010$ATM, unique2010$Pax)
> plot(Pax ~ ATM, data=unique2010, main=...)
> lines(smoo, col="blue")
```



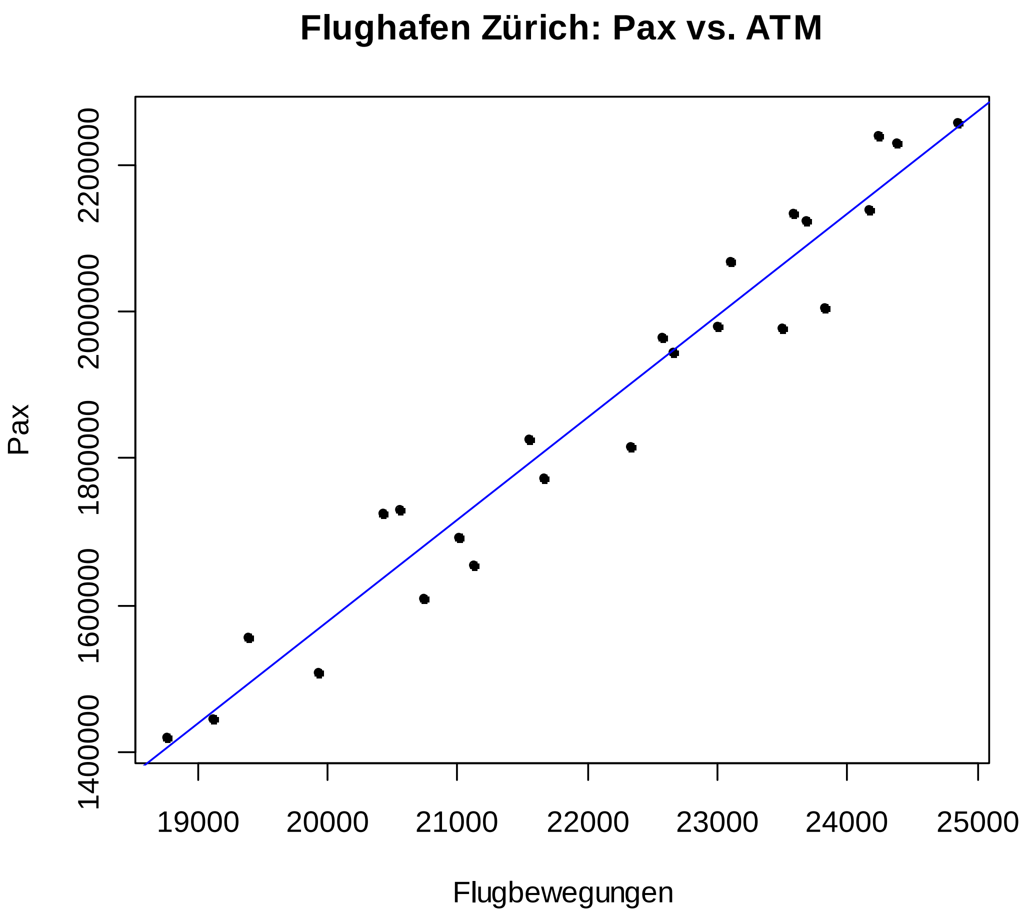
Applied Statistical Regression
AS 2015 – Simple Regression

Linear Modeling

A straight line represents the systematic relation between Pax and ATM.

- Only appropriate if the true relation is indeed a straight line

- The question is how the line/function are obtained.



Simple Linear Regression

The more air traffic movements, the more passengers there are. The relation seems to be linear, which is of course also the mathematically most simple way of describing the relation.

$$f(x) = \beta_o + \beta_1 x, \text{ resp. } Pax = \beta_0 + \beta_1 \cdot ATM$$

Name/meaning of the two parameters in the equation: $\beta_0 =$ "Intercept" $\beta_1 =$ "Slope"

Fitting a straight line into a 2-dimensional scatter plot is known as **simple linear regression**. This is because:

- there is just one single predictor variable ("*simple*").
- the relation is linear in the parameters ("*linear*").

Model, Data & Random Errors

No we are bringing the data into play. The regression line will not run through all the data points. Thus, there are random errors:

$$y_i = \beta_0 + \beta_1 x_i + E_i, \text{ for all } i = 1, \ldots, n$$

Meaning of variables/parameters:

y_i is the response variable (Pax) of observation i.

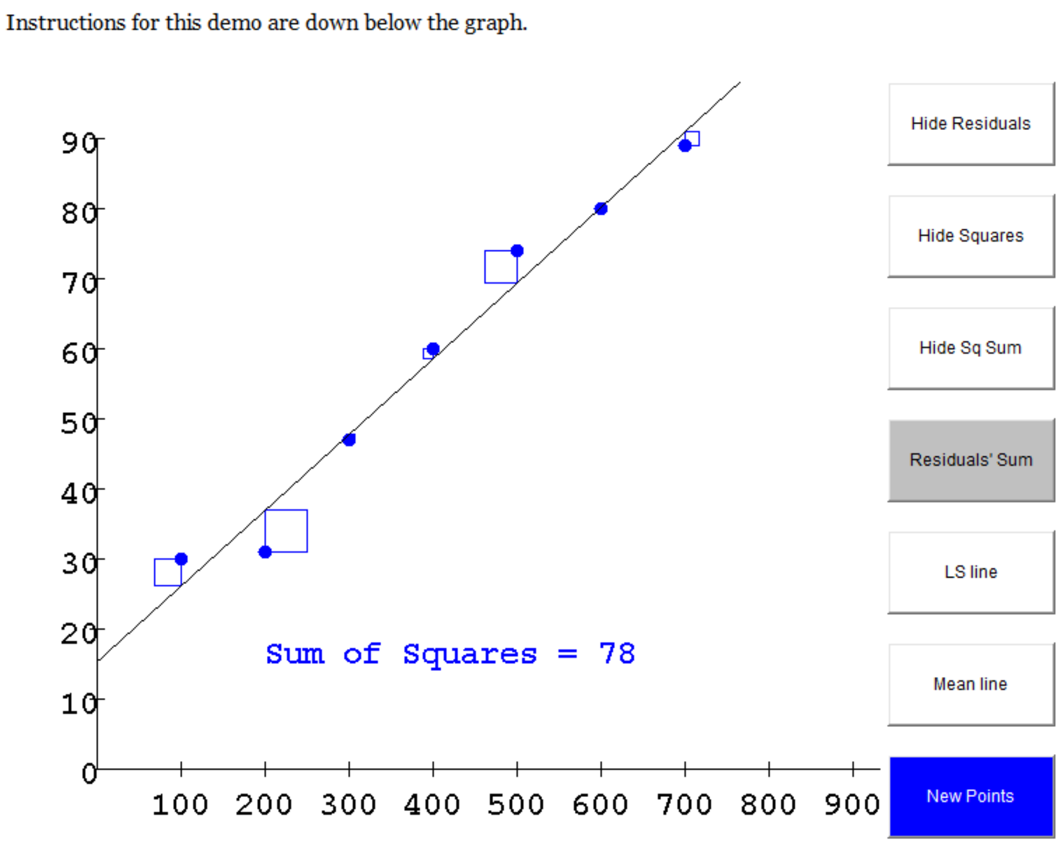
x_i is the predictor variable (ATM) of observation i.

β_0, β_1 are the regression coefficients. They are unknown previously, and need to be estimated from the data.

E_i is the residual or error, i.e. the random difference between observation and regression line.

Least Squares Fitting

→ http://demonstrations.wolfram.com/LeastSquaresCriteriaForTheLeastSquaresRegressionLine/



We need to fit a straight line that fits the data well.

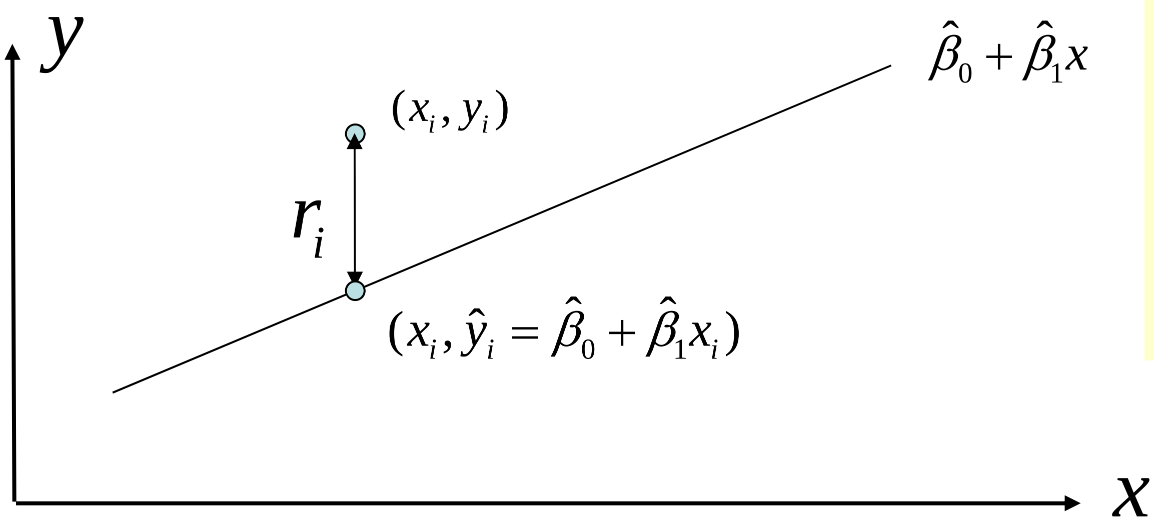
Many possible solutions exist, some are good, some are worse.

Our paradigm is to fit the line such that the squared errors are minimal.

Residuals vs. Errors

The residual $r_i = y_i - \hat{y}_i$ is the difference between the observed and the fitted y-value for the ith observation. While the error is a concept and random variable, the residuals are numerical values

Illustration of the residuals



The paradigm remains to fit a straight line such that the sum of squared residuals is minimized: $\sum_{i=1}^{n} r_i^2 = \min$

Least Squares: Mathematics

The paradigm in verbatim...

Given a set of data points $(x_i, y_i)_{i=1,\dots,n}$, the goal is to fit the regression line such that the sum of squared differences between observed value y_i and regression line is minimal. The function

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n} r_i^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-(\beta_0+\beta x_i))^2=\min!$$

measures, how well the regression line, defined by β_0, β_1, fits the data. The goal is to minimize this "quality function".

Solution: → **see next slide...**

Solution Idea: Partial Derivatives

- We are taking partial derivatives on the function $Q(\beta_0, \beta_1)$ with respect to both arguments β_0 and β_1. As we are after the minimum of the function, we set them to zero:

$$\frac{\partial Q}{\partial \beta_0} = 0 \text{ and } \frac{\partial Q}{\partial \beta_1} = 0$$

- This results in a linear equation system, which (here) has two unknowns β_0, β_1, but also two equations. These are also known under the name *normal equations*.

- The solution for β_0, β_1 can be written explicitly as a function of the data pairs $(x_i, y_i)_{i=1,\ldots,n}$, **see next slide...**

Least Squares: Solution

According to the least squares paradigm, the best fitting regression line is, i.e. the optimal coefficients are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

- For a given set of data points $(x_i, y_i)_{i=1,...,n}$ we can determine the solution with a pocket calculator (...or better, with R).

- **The solution for our example Pax vs. ATM:**
  $\hat{\beta}_1 = 138.8,\ \hat{\beta}_0 = -1'197'682$ obtained from

```
> lm(Pax ~ ATM, data=unique2010)
```

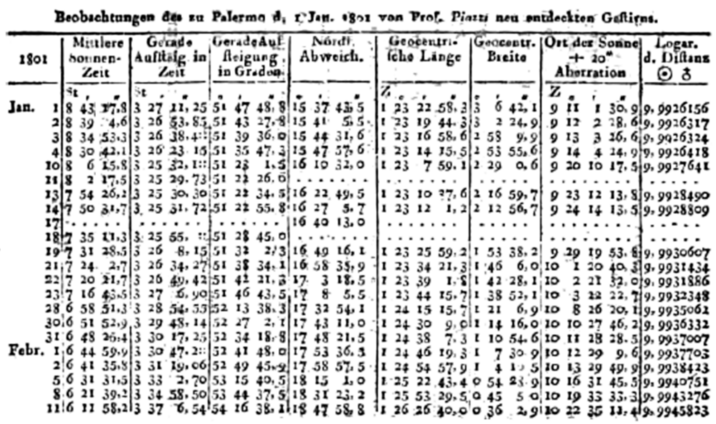
Applied Statistical Regression
AS 2015 – Simple Regression

Why Least Squares?

History...

Within a few years (1801, 1805), the method was developed independently by Gauss and Legendre. Both were after solving applied problems in astronomy...

Source: → http://de.wikipedia.org/wiki/Methode_der_kleinsten_Quadrate



Beobachtungen des zu Palermo d. 1. Jan. 1801 von Prof. Piazzi neu entdeckten Gestirns.

1801		Mittlere Sonnen-Zeit	Gerade Aufsteig. in Zeit	Gerade Aufsteigung in Graden	Nördl. Abweich.	Geocentrische Länge	Geocentr. Breite	Ort der Sonne + 20" Aberration	Logar. d. Distanz ☉ ♂
		St ′ ″	St ′ ″	° ′ ″	° ′ ″	Z ° ′ ″	° ′ ″	Z ° ′ ″	
Jan.	1	8 43 27,8	3 27 11,25	51 47 48,8	15 37 43,5	1 23 22 58,3	3 6 42,1	9 11 1 30,9	9,9926156
	2	8 39 4,6	3 26 53,85	51 43 27,8	15 41 5,5	1 23 19 44,3	3 2 24,9	9 12 2 28,6	9,9926317
	3	8 34 53,3	3 26 38,4::	51 39 36,0	15 44 31,6	1 23 16 58,6	2 58 9,9	9 13 3 26,6	9,9926324
	4	8 30 42,1	3 26 23,15	51 35 47,3	15 47 57,6	1 23 14 15,5	2 53 55,6	9 14 4 24,9	9,9926418
	10	8 6 15,8	3 25 32,1::	51 23 1,5	16 10 32,0	1 23 7 59,1	2 29 0,6	9 20 10 17,5	9,9927641
	11	8 2 17,5	3 25 29,73	51 22 26,0	. . . . . .	. . . . . .	. . . . .	. . . .	. . . . .
	13	7 54 26,2	3 25 30,30	51 22 34,5	16 22 49,5	1 23 10 27,6	2 16 59,7	9 23 12 13,8	9,9928490
	14	7 50 31,7	3 25 31,72	51 22 55,8	16 27 5,7	1 23 12 1,2	2 12 56,7	9 24 14 13,5	9,9928809
	17	. . . . . .	. . . . . .	. . . . . .	16 40 13,0	. . . . . .	. . . . .	. . . .	. . . . .
	18	7 35 11,3	3 25 55,::	51 28 45,0	. . . . . .	. . . . . .	. . . . .	. . . .	. . . . .
	19	7 31 28,5	3 26 8,15	51 32 2,3	16 49 16,1	1 23 25 59,2	1 53 38,2	9 29 19 53,8	9,9930607
	21	7 24 2,7	3 26 34,27	51 38 34,1	16 58 35,9	1 23 34 21,3	1 46 6,0	10 1 20 40,3	9,9931434
	22	7 20 21,7	3 26 49,42	51 42 21,3	17 3 18,5	1 23 39 1,8	1 42 28,1	10 2 21 32,0	9,9931886
	23	7 16 43,5	3 27 6,90	51 46 43,5	17 8 5,5	1 23 44 15,7	1 38 52,1	10 3 22 22,7	9,9932348
	28	6 58 51,3	3 28 54,55	52 13 38,3	17 32 54,1	1 24 15 15,7	1 21 6,9	10 8 26 20,1	9,9935062
	30	6 51 52,9	3 29 48,14	52 27 2,1	17 43 11,0	1 24 30 9,0	1 14 16,0	10 10 27 46,2	9,9936332
	31	6 48 26,4	3 30 17,25	52 34 18,8	17 48 21,5	1 24 38 7,3	1 10 54,6	10 11 28 28,5	9,9937007
Febr.	1	6 44 59,9	3 30 47,2::	52 41 48,0	17 53 36,3	1 24 46 19,3	1 7 30,9	10 12 29 9,6	9,9937703
	2	6 41 35,8	3 31 19,06	52 49 45,9	17 58 57,5	1 24 54 57,9	1 4 1[illegible],5	10 13 29 49,9	9,9938423
	5	6 31 31,5	3 33 2,70	53 15 40,5	18 15 1,0	1 25 22 43,4	0 54 23,9	10 16 31 45,5	9,9940751
	8	6 21 39,2	3 34 58,50	53 44 37,5	18 31 23,2	1 25 53 29,5	0 45 5,0	10 19 33 33,3	9,9943276
	11	6 11 58,2	3 37 6,54	54 16 38,1	18 47 58,8	1 26 26 40,0	0 36 2,9	10 22 35 11,4	9,9945823

Carl Friedrich Gauss

Adrien-Marie Legendre

Why Least Squares?

Mathematics...

- Least Squares is simple in the sense that the solution is known in closed form as a function of $(x_i, y_i)_{i=1,...,n}$.
- The line runs through the center of gravity $(\overline{x}, \overline{y})$
- The sum of residuals adds up to zero: $\sum_{i=1}^{n} r_i = 0$
- Some deeper mathematical optimality can be shown when analyzing the large sample properties of the estimates $\hat{\beta}_0, \hat{\beta}_1$ This is especially true under the assumption of normally distributed errors E_i.

Fitted Values

The estimated parameters $\hat{\beta}_0, \hat{\beta}_1$ can be used for determining the fitted values $\hat{y}$. Please note that mathematically, this is a conditional expected value::

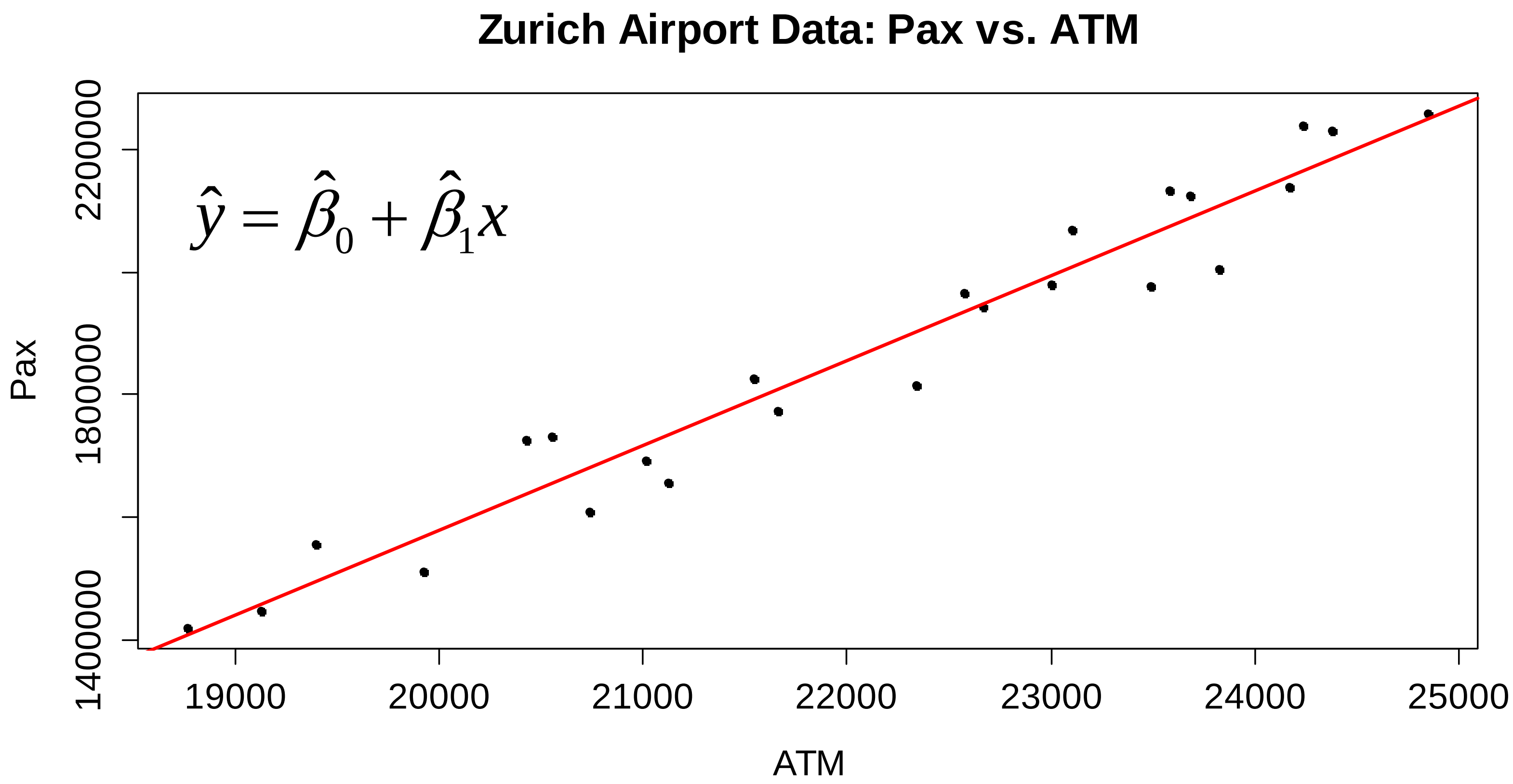
$$\hat{y} = E[y \mid x] = \hat{\beta}_0 + \hat{\beta}_1 x$$

In R, the fitted values are obtained as follows:

```
> fit <- lm(Pax ~ ATM, data=unique2010)
> fitted(fit)
      1       2       3       4       5
1654841 1808312 2165068 2156465 2184911
      6       7       8       9     ...
2250545 2108731 2062107 1493184     ...
```

Drawing the Regression Line

```
> plot(Pax ~ ATM, data=unique2010, pch=20)
> title("Zurich Airport Data: Pax vs. ATM")
> abline(fit, col="red", lwd=2)
```



Is This a Good Model for Predicting the Pax Number from the ATM?

a) Beyond the range of observed data

Unknown, but most likely not...

b) Within the range of observed data

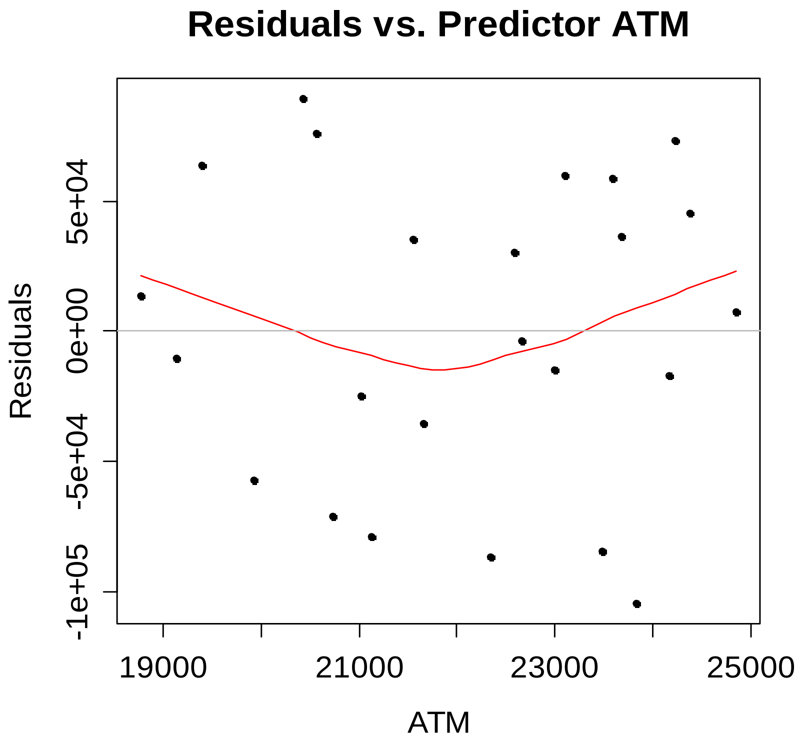
Yes, under the following conditions:

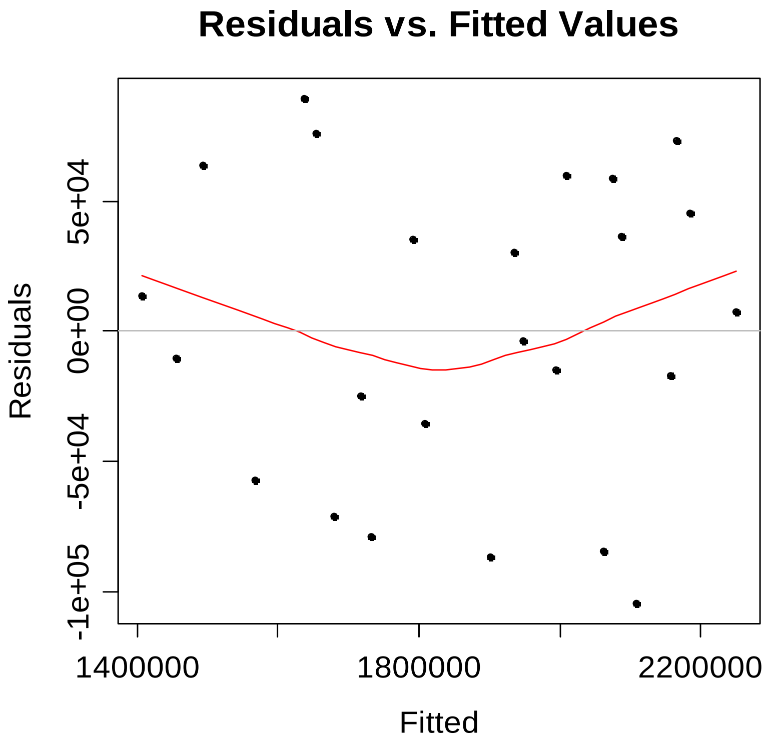
- the relation is in truth a straight line, i.e. $E[E_i] = 0$
- the scatter of the errors is constant, i.e. $Var(E_i) = \sigma^2$
- the errors are uncorrelated (from a representative sample)
- the errors are approximately normally distributed

→ **Fodder for thougt: 9/11, SARS, Eyjafjallajökull...?**

Model Diagnostics

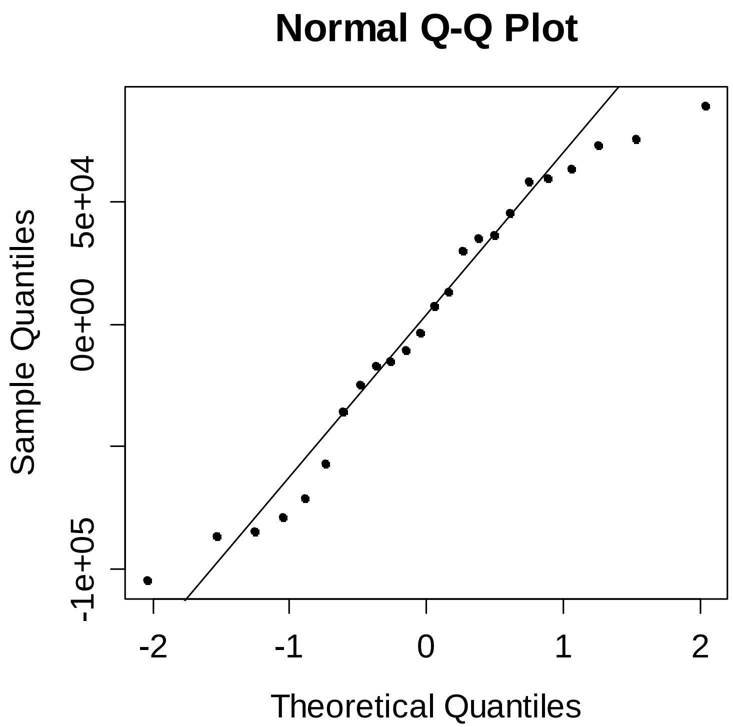
For assessing the quality of the regression line, we need to (at least roughly) check whether the assumptions are met:
$E[E_i]=0$ and $Var(E_i)=\sigma^2$ can be reviewed by:





Model Diagnostics

For assessing the quality of the regression line, we need to (at least roughly) check whether the assumptions are met: Gaussian distribution can be reviewed by:



We will revisit model diagnostics again later in this course, where it will be discussed more deeply.

"Residuals vs. Fitted" and the "Normal Plot" will always stay at the heart of model diagnostics.

Gauss-Markov-Theorem

A mathematical optimality result for the Least Squares line

It only holds if the following conditions are met:

- the relation is in truth a straight line, i.e. $E[E_i] = 0$
- the scatter of the errors is constant, i.e. $Var(E_i) = \sigma^2$
- the errors are uncorrelated, i.e. $Cov(E_i, E_j) = 0, \; if \; i \neq j$

Not explicitly (but implicitly) required:

- the errors are normally distributed: $E_i \sim N(0, \sigma_E^2)$

Gauss-Markov-Theorem:

- Least Squares yields the *best linear unbiased estimates*

Properties of the Least Square Estimates

Under the conditions above, the estimates are unbiased:

$E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$

The variances of the estimates are as follows:

$Var(\hat{\beta}_0) = \sigma_E^2 \cdot \left(\frac{1}{n} + \frac{\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \right)$ and $Var(\hat{\beta}_1) = \frac{\sigma_E^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$

Precise estimates are obtained with:

- a large number of observations n
- a good scatter in the predictor x_i
- an informative/useful predictor, making σ_E^2 small
- (an error distribution which is approximately Gaussian)

Estimating the Error Variance

Besides the regression coefficients, we also need to estimate the *error variance*. It is a necessary ingredient for all tests and confidence intervals that will be discussed shortly.

The estimate is based on the *residual sum of squares* (RSS):

$$\hat{\sigma}_E^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} r_i^2$$

In R, the regression summary provides the estimate for the error's standard deviation as `Residual standard error`:

```
> summary(fit)
...
Residual standard error: 59700 on 22 degrees of freedom
```

Benefits of Linear Regression

- **Inference on the relation between y and x**

The goal is to understand if and how strongly the response variable depends on the predictor. There are performance indicators as well as statistical tests adressing the issue.

- **Prediction of (future) observations**

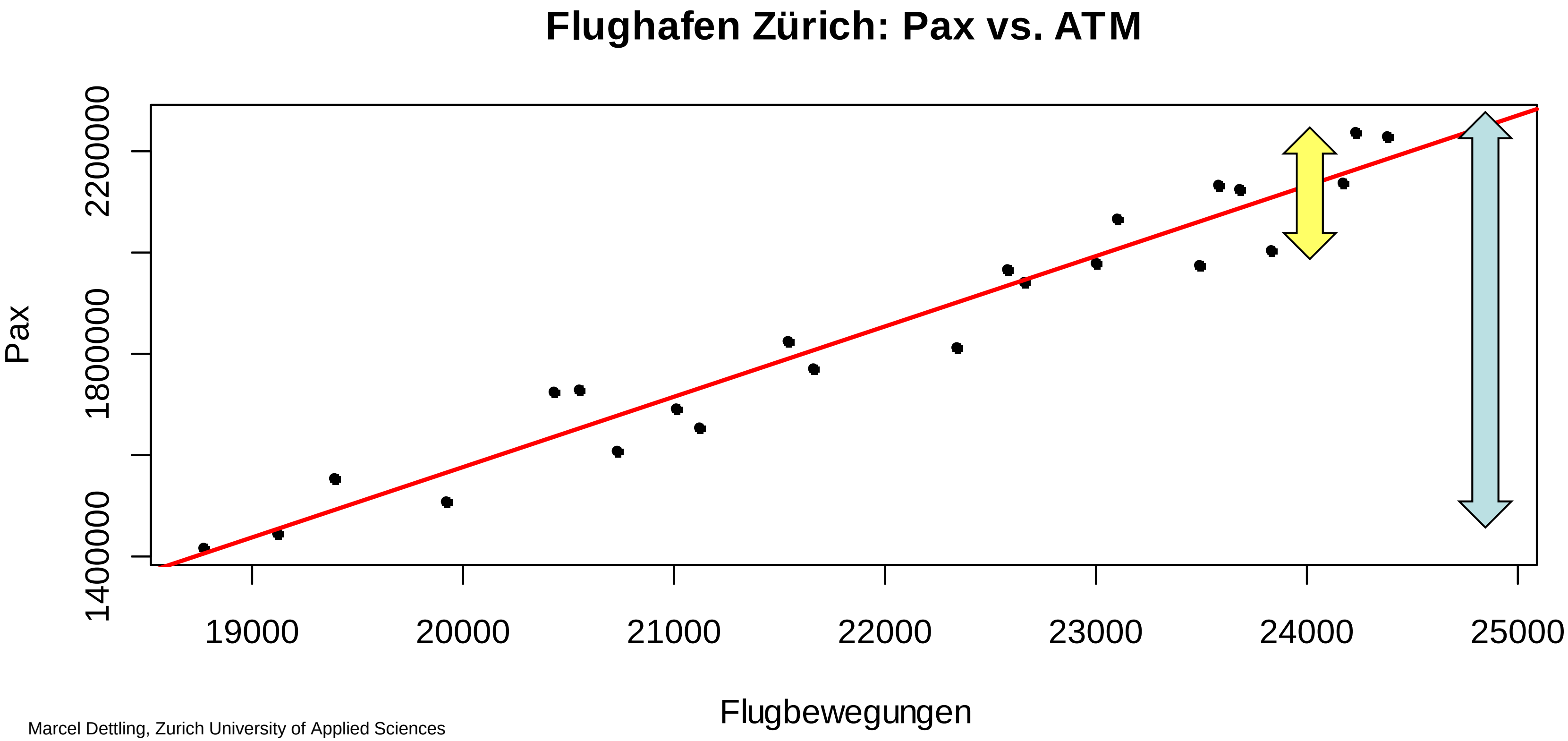
The regression line/equation can be employed to predict the PAX number for any given ATM value.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

However, this mostly will not work well for extrapolation!

R^2: *The Coefficient of Determination*

The coefficient of determination R^2 is also known as *multiple R-squared*. It tells which portion of the total variation is accounted for by the regression line.



Computation of R^2

R^2 is the portion of the total variation that is explained through regression. It is determined as one minus the quotient of the yellow arrow divided by the blue arrow.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \in [0,1]$$

The closer to 1 the value is, the tighter the datapoints are packed around the regression line. However, there are no formal criteria which R^2 value needs to be met such that the regression can be said to be useful/valid.

Confidence Interval for the Slope β_1

A 95%-CI for the slope β_1 tells which values (besides the point estimate $\hat{\beta}_1$) are plausible, too. The uncertainty is due to estimation/sampling effects.

95%-CI for β_1 : $\hat{\beta}_1 \pm qt_{0.975;n-2} \cdot \hat{\sigma}_{\hat{\beta}_1}$, resp.

$$\hat{\beta}_1 \pm qt_{0.975;n-2} \cdot \sqrt{\hat{\sigma}_E^2 \Big/ \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

In R:

```
> fit <- lm(Pax ~ ATM, data=unique2010)
> confint(fit, "ATM")
        2.5 %  97.5 %
ATM 124.4983 153.025
```

Testing the Slope β_1

There is a statistical hypothesis test which can be used to check whether the slope is significantly different from zero, or any other arbitrary value b. The null hypothesis is:

$$H_0 : \beta_1 = 0, \text{ resp. } H_0 : \beta_1 = b$$

One usually tests two-sided on the 95%-level. The alternative is:

$$H_A : \beta_1 \neq 0, \text{ resp. } H_A : \beta_1 \neq b$$

As a test statistic, we use:

$$T_{H_0:\beta_1=0} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}, \text{ resp. } T_{H_0:\beta_1=b} = \frac{\hat{\beta}_1 - b}{\hat{\sigma}_{\hat{\beta}_1}}, \text{ both have a } t_{n-2} \text{ distribution.}$$

Reading R-Output

```
> summary(lm(Pax ~ ATM, data=dat))

Call: lm(formula = Pax ~ ATM, data = dat)

Residuals:     Min       1Q  Median       3Q      Max
           -104188   -40885    2099    48588    89154

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.198e+06  1.524e+05  -7.858 7.94e-08 ***
ATM          1.388e+02  6.878e+00  20.176 1.11e-15 ***

---

Residual standard error: 59700 on 22 degrees of freedom
Multiple R-squared: 0.9487, Adjusted R-squared: 0.9464
F-statistic: 407.1 on 1 and 22 DF,  p-value: 1.110e-15
```

→ **Will be explained in detail on the blackboard!**

Testing the Slope β_1

Practical Example:

Use the Pax vs. ATM data and perform a statistical test for the null hypothesis $H_0 : \beta_1 = 150$. The information from the summary on slide 51 can be used as a basis. Then, also answer:

a) *Explain in colloquial language what was just tested. What is the benefit of this test? What claims could motivate the test?*

b) *How does the testing result relate with the 95%-CI that we computed on slide 49? Would we be able to tell the test results from the CI alone?*

→ **See blackboard for the answers**

Testing the Intercept β_0

An analogous test can be done for the intercept.

- No matter what the test result will be, the intercept should generally not be omitted from the regression model.

- The presence of the intercept protects against possible non-linearities and calibration errors of measurement devices. If it is kicked out of the model, the results are generally worse.

- If theory dictates that there should not be an intercept but it is still significant, take this as evidence that the linear relation does not hold when extrapolating to $x = 0$.

Prediction

Using the regression line, we can predict the y-value for any desired x-value. The result is the expectation for y given x.

$$E[y \mid x] = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$ a.k.a. ***"fitted value"***

Example: With 24'000 air traffic movements, we expect

$$-1'197'682 + 24'000 \cdot 138.8 = 2'133'518$$ Passengers

Be careful:

At best, interpolation within the range of observed x-values is trustworthy. Extrapolation with ATM values such as 50'000, 5'000 or even 0 usually produces completely useless results.

Prediction with R

We can use the regression fit object for prediction. The syntax for obtaining the fitted value(s) is as follows:

```
> fit <- lm(Pax ~ ATM, data=unique2010)
> dat <- data.frame(ATM=c(24000))
> predict(fit, newdata=dat)
1 2132598
```

The x-values need to be provided in a data frame, where the variable/column name is identical to the predictor name.

Then, the `predict()` procedure is invoked with the regression fit and the new x-values as arguments.

Confidence Interval for $E[y \mid x]$

We just computed the fitted value $\hat{\beta}_0 + \hat{\beta}_1 x$, i.e. the expected number of passengers for 24'000 ATMs. This is not a deterministic value, but an estimate that is subject to variability.

A 95%-CI for the fitted value at position x is given by:

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm qt_{0.975;n-2} \cdot \hat{\sigma}_E \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

In R:
```
> predict(fit, newdata=dat, interval="confidence")
      fit     lwr     upr
1 2132598 2095450 2169746
```

Prediction Interval for y

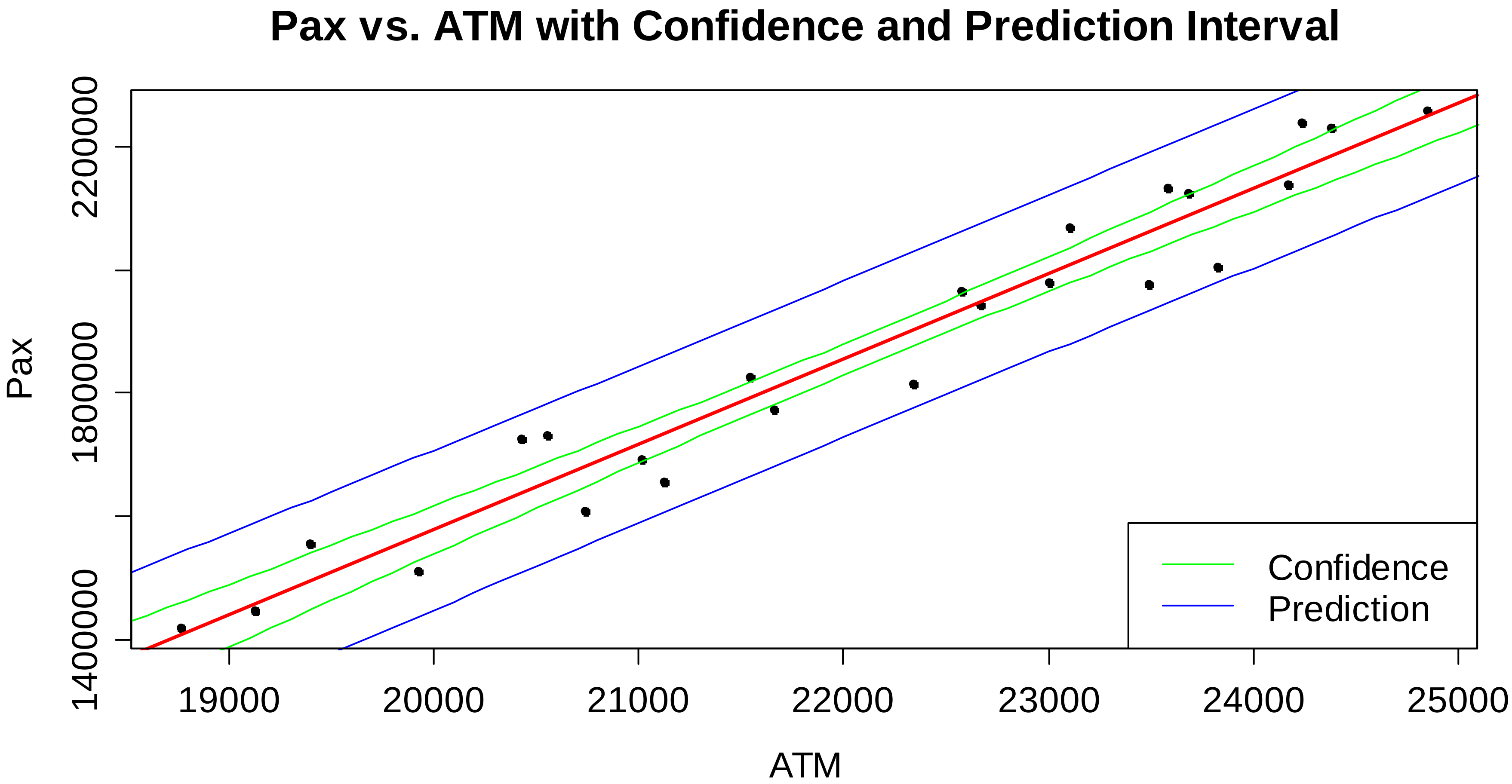
The confidence interval for $E[y \mid x]$ tells about the variabilty of the fitted value. It does not account for the scatter of the data points around the regression line and thus does not define a region where we have to expect the observed value. A 95% prediction interval at position x is given by:

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm qt_{0.975;n-2} \cdot \hat{\sigma}_E \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

In R: `> predict(fit,newdata=dat,interval="prediction")`

```
      fit     lwr     upr
1 2132598 2003343 2261853
```

Confidence and Prediction Interval



Confidence and Prediction Interval

Note:

Visualizing the confidence and prediction intervals in R is not straightforward, but requires some tedious handwork.

R-Hints:

```
dat  <- data.frame(ATM=seq(...,..., length=200))
pred <- predict(fit, newdata=dat, interval=...)
plot(..., ..., main="...")
lines(dat$ATM, pred[,2], col=...)
lines(dat$ATM, pred[,3], col=...)
```

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.

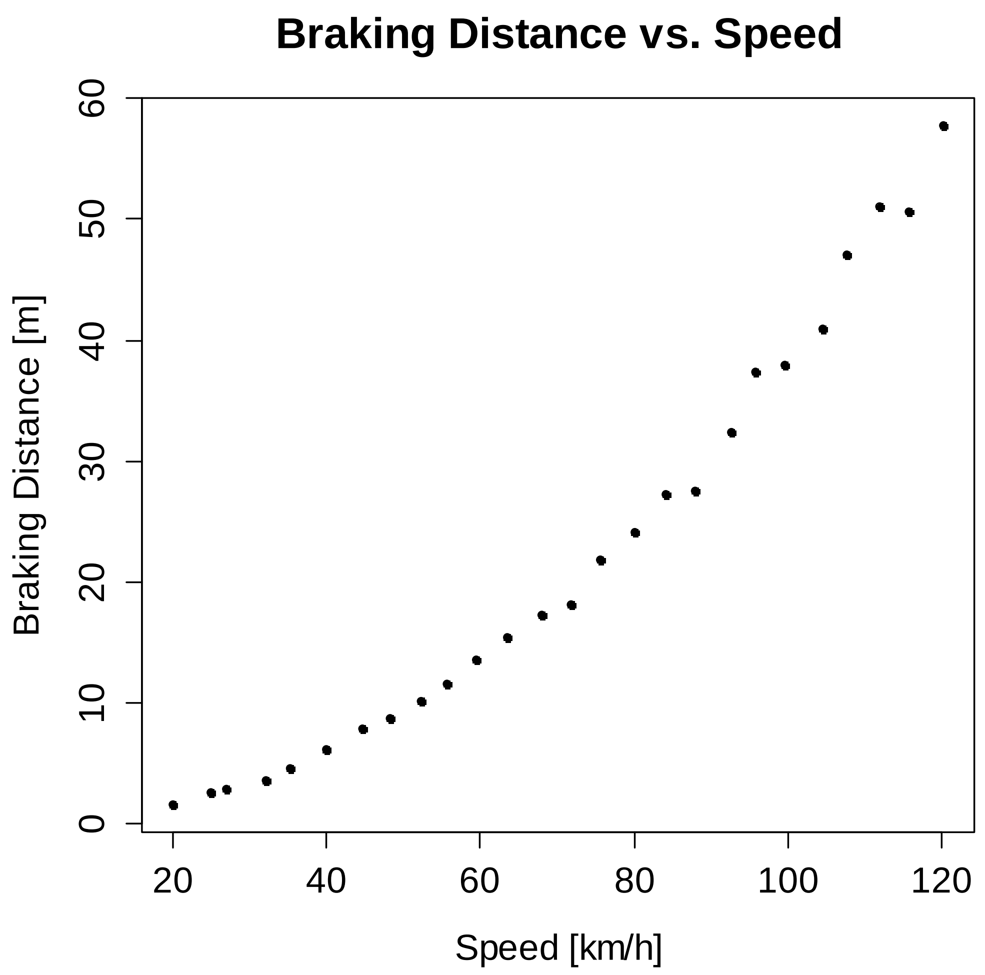
Example: Automobile Braking Distance



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…

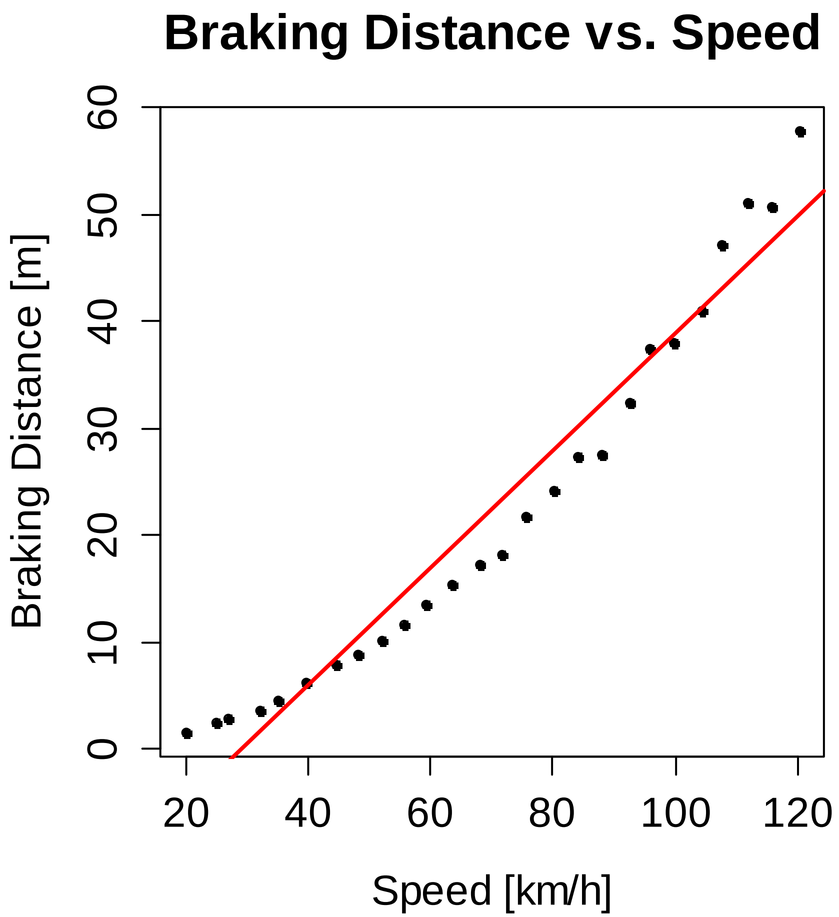
Braking Distance: Data

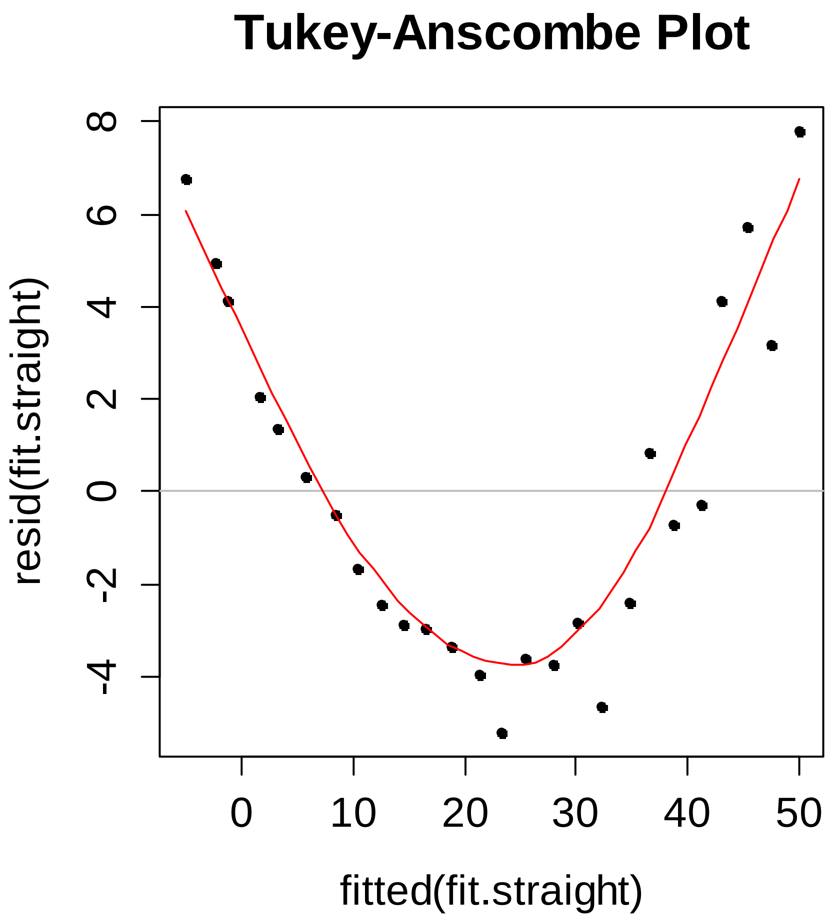
obs	speed	brdist
1	19.96	1.60
2	24.97	2.54
3	26.97	2.81
4	32.14	3.58
5	35.24	4.59
6	39.87	6.11
7	44.62	7.91
8	48.32	8.76
9	52.18	10.12
10	55.72	11.62
11	59.44	13.57
12	63.56	15.45
...	...	...
24	111.97	51.09
25	115.88	50.69
26	120.35	57.77



Applied Statistical Regression
AS 2015 – Simple Regression

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.

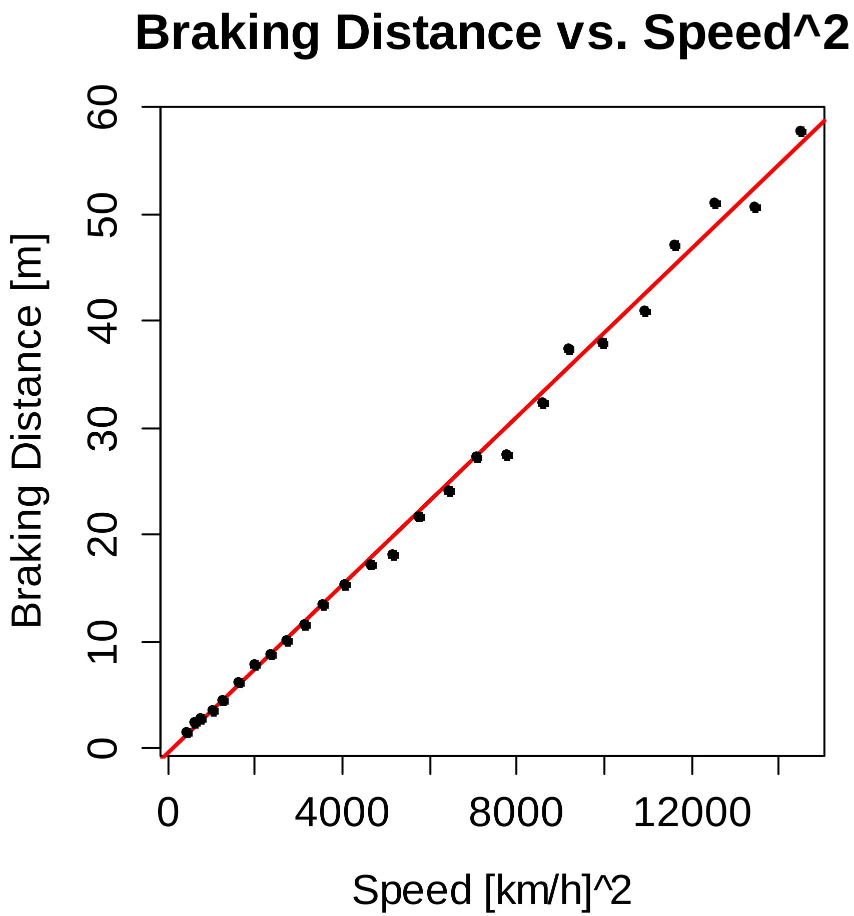
$$Distance_i = \beta_0 + \beta_1 \cdot Speed_i^2 + E_i$$

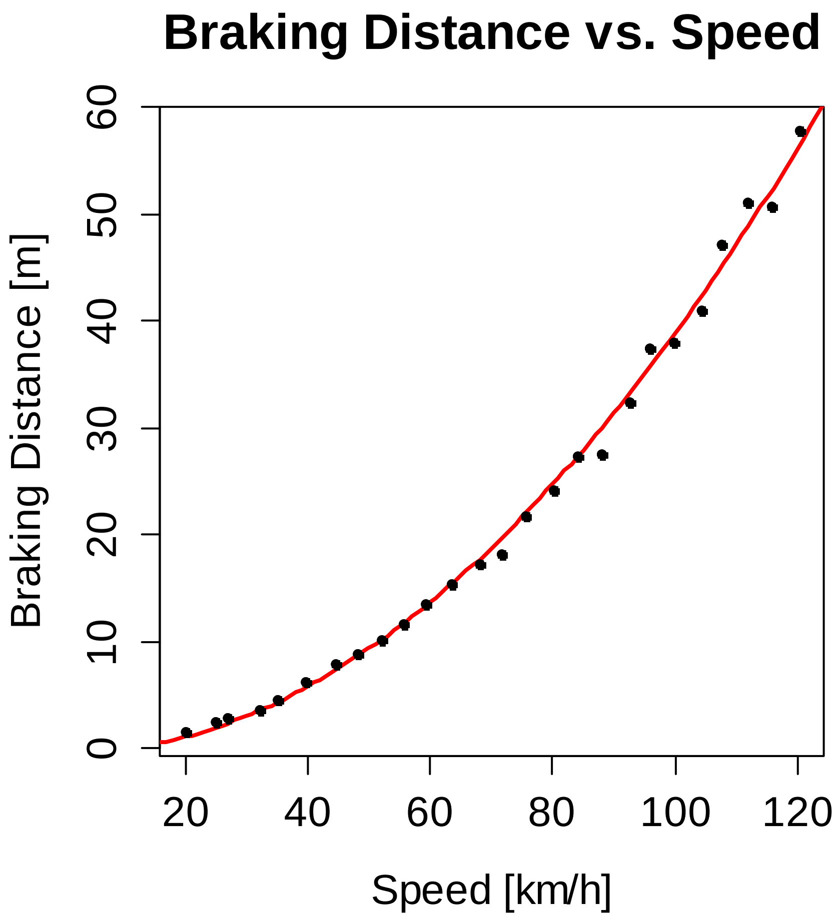
resp. $y_i = \beta_0 + \beta_1 \cdot x_i' + E_i$, where $x_i' = x_i^2 = Speed_i^2$

- 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:

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

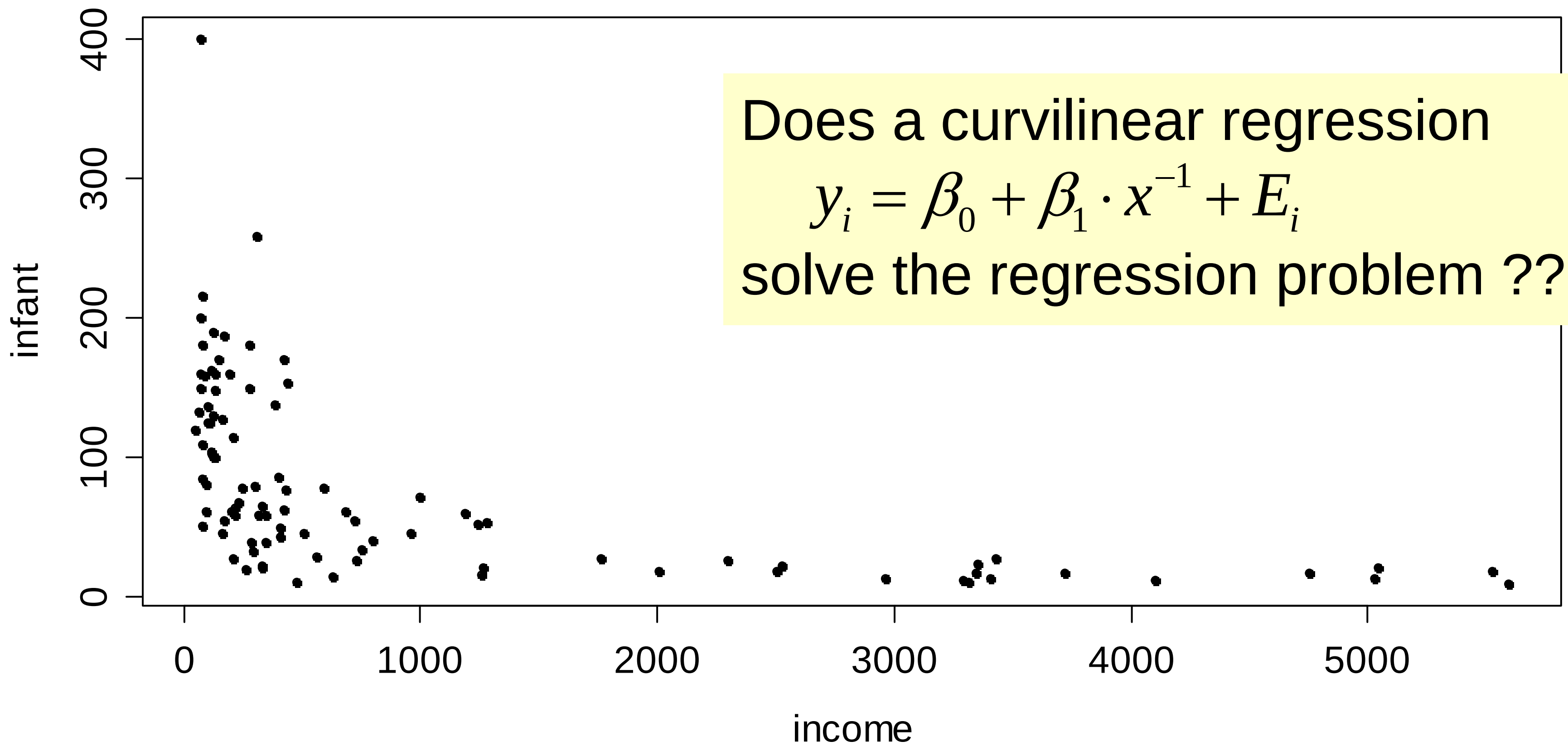
- Often, the variance/scatter of the errors is non-constant. In many examples, it increases with increasing.

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

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

Infant Mortality vs. Per-Capita Income

Infant Mortality vs. Per-Capita Income

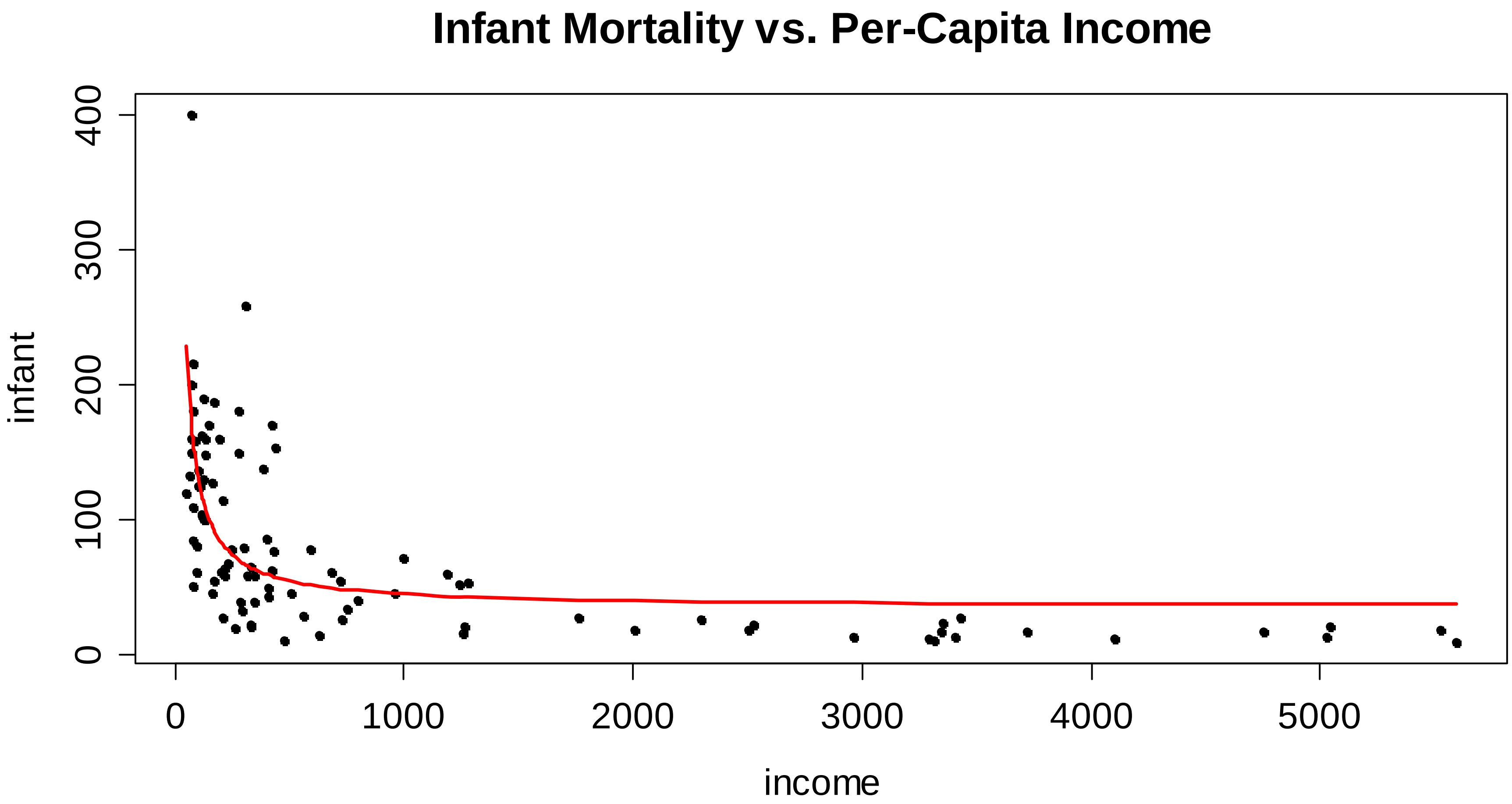


Does a curvilinear regression

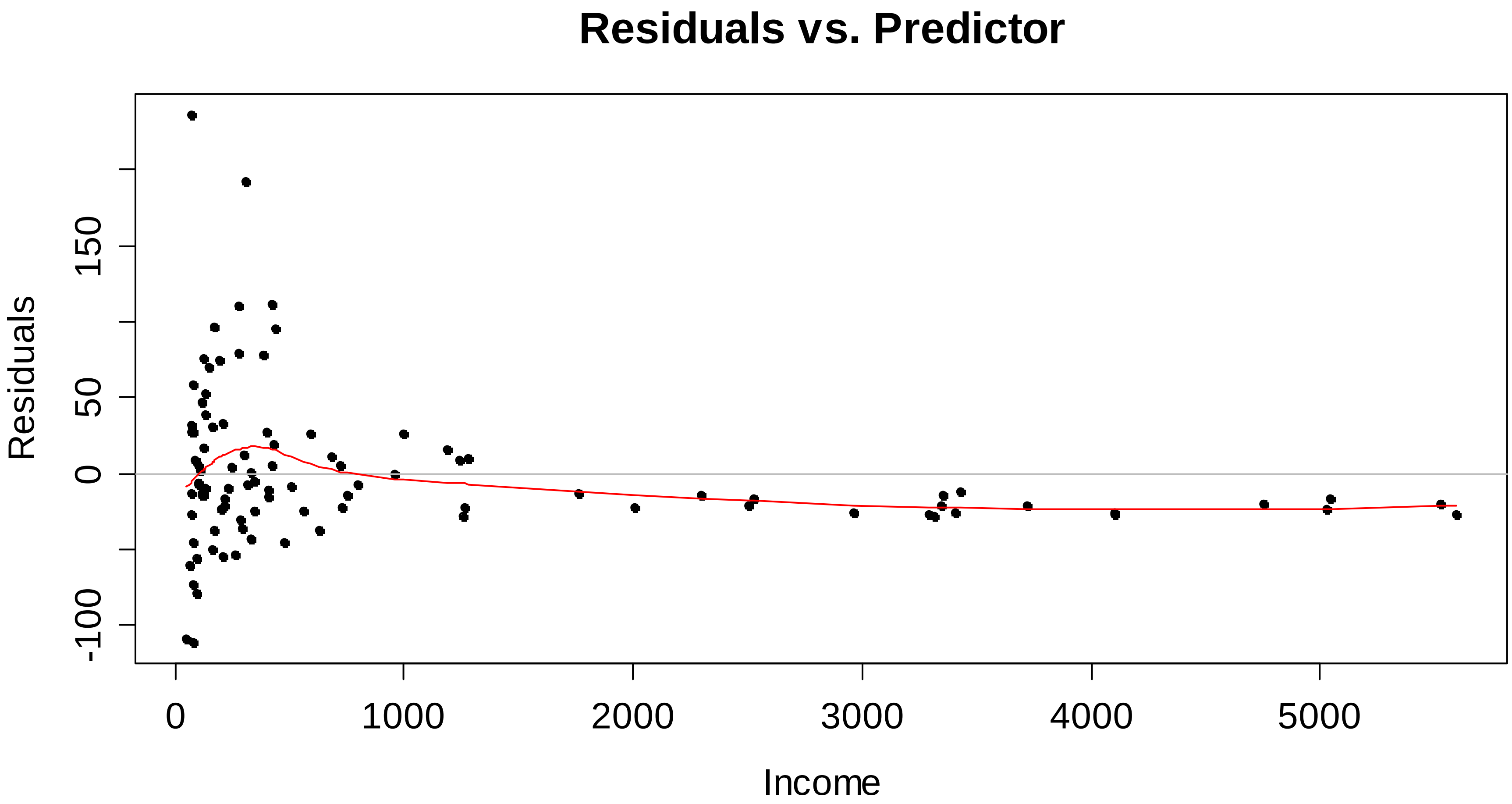
$$y_i = \beta_0 + \beta_1 \cdot x^{-1} + E_i$$

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**.

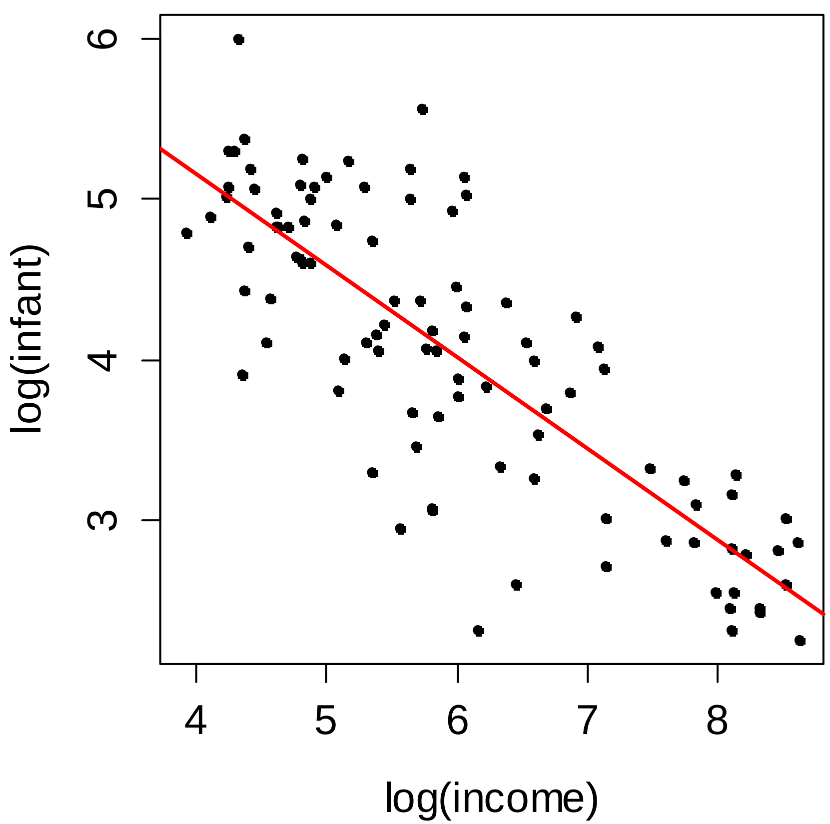
A simple yet very useful trick solves the problem:

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

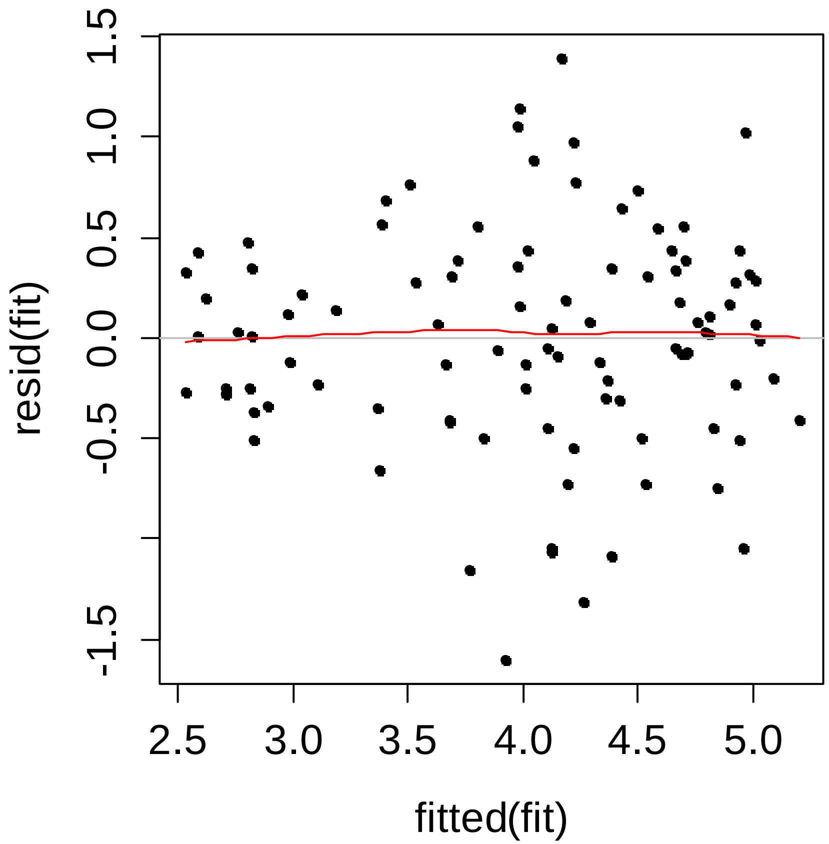
For details, **see the next slide and the blackboard...**

The Log-Transformation Helps!

log(infant) vs. log(income)



Residuals vs. Fitted Values



Model and Coefficients

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

$$y' = \beta_0' + \beta_1 \cdot x' + E'\text{, with } y' = \log(y),\ x' = \log(x),\ E' \sim N(0, \sigma_E^2)$$

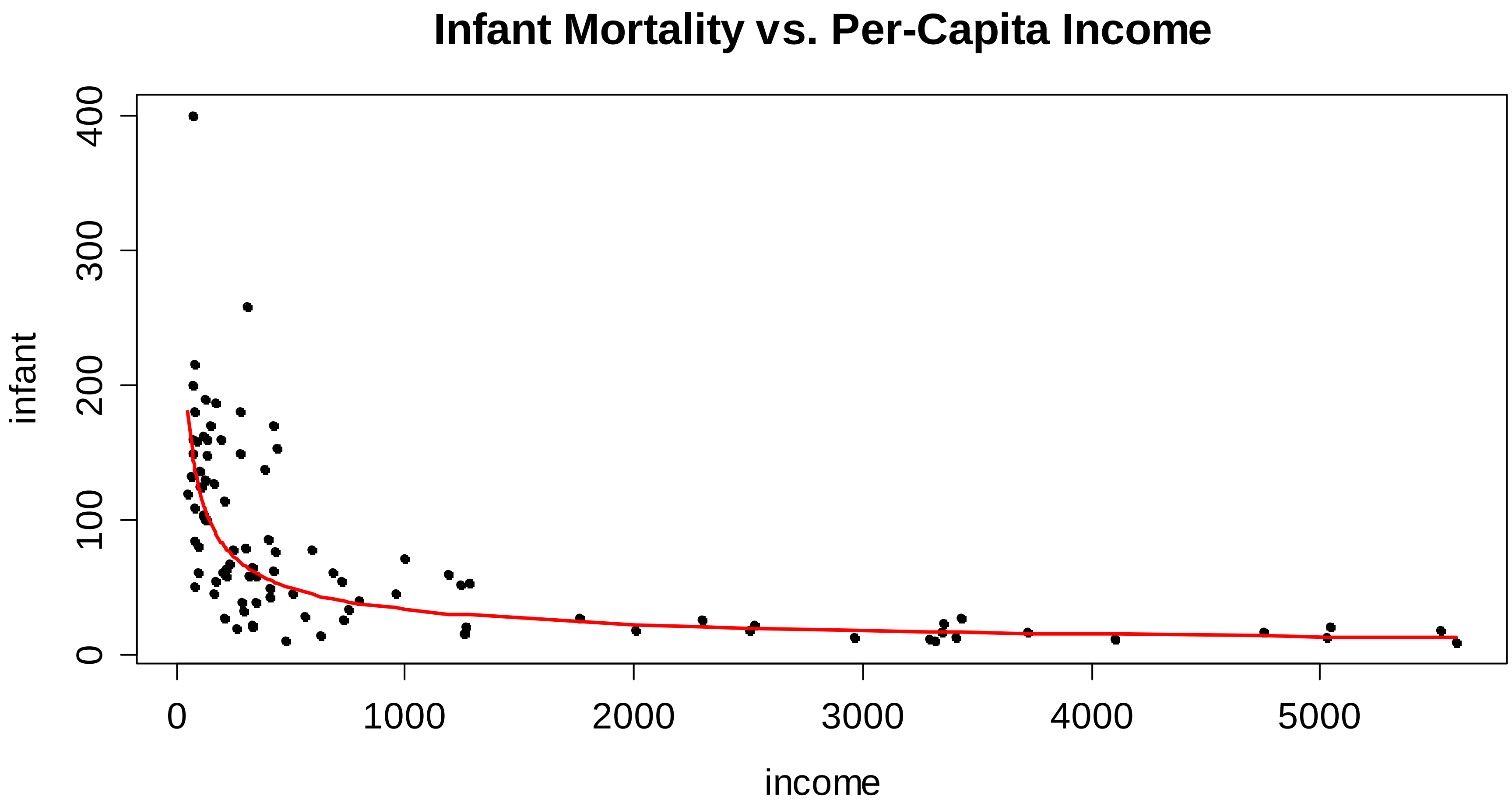
this amounts to fitting a **power law with multiplicative, lognormal error** on the original scale, i.e.:

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

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 on the Original Scale

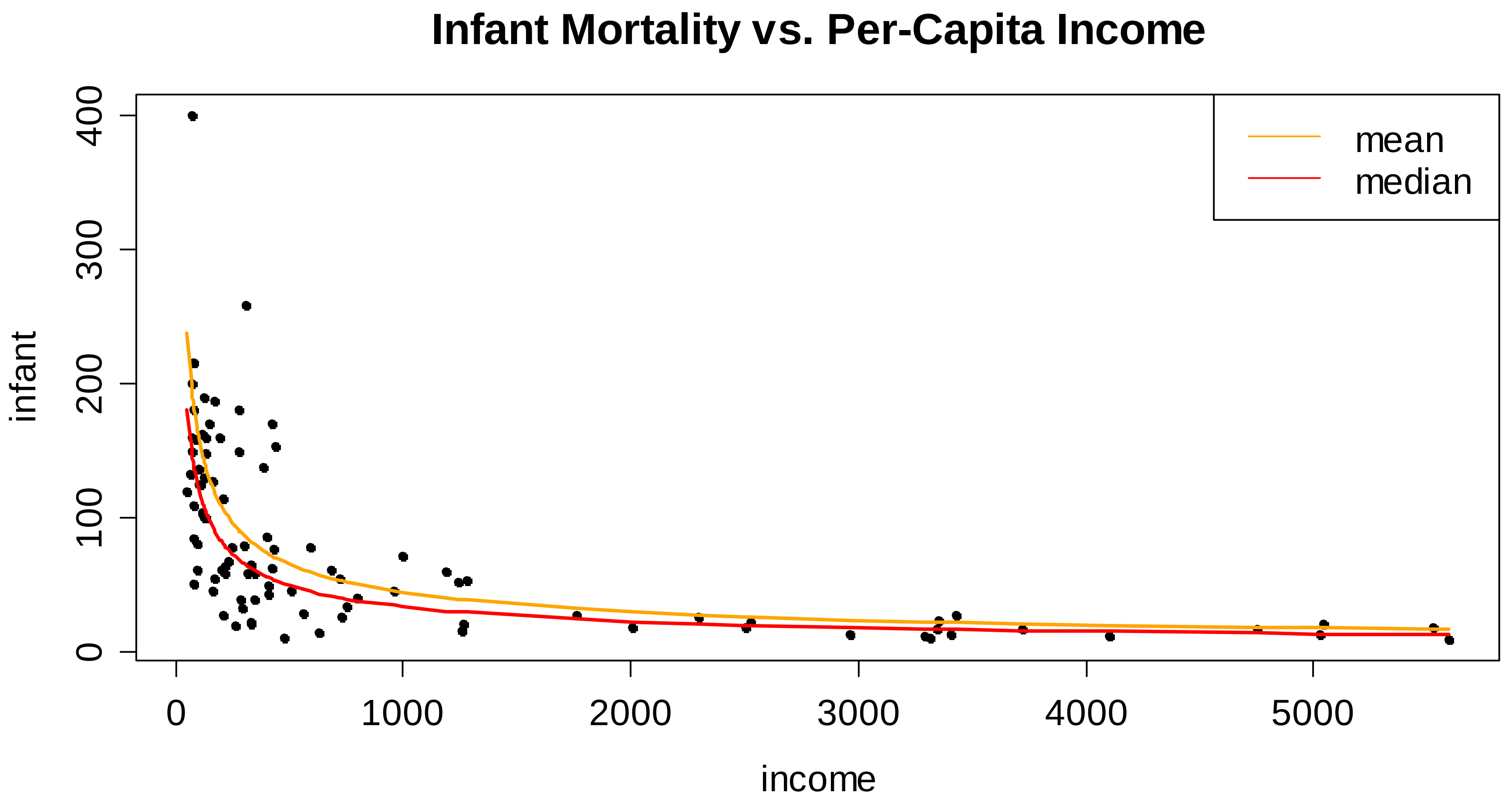
- For a simple prediction of the y-value on the original scale, we can exponentiate to invert the log-transformation: $\hat{y} = \exp(\hat{y}')$

- **Caution:** this is an estimate for the median of the conditional distribution $y \mid x$, but not the conditional mean $E[y \mid x]$. If we require unbiased fitted values on the original scale, applying a correction factor is required!

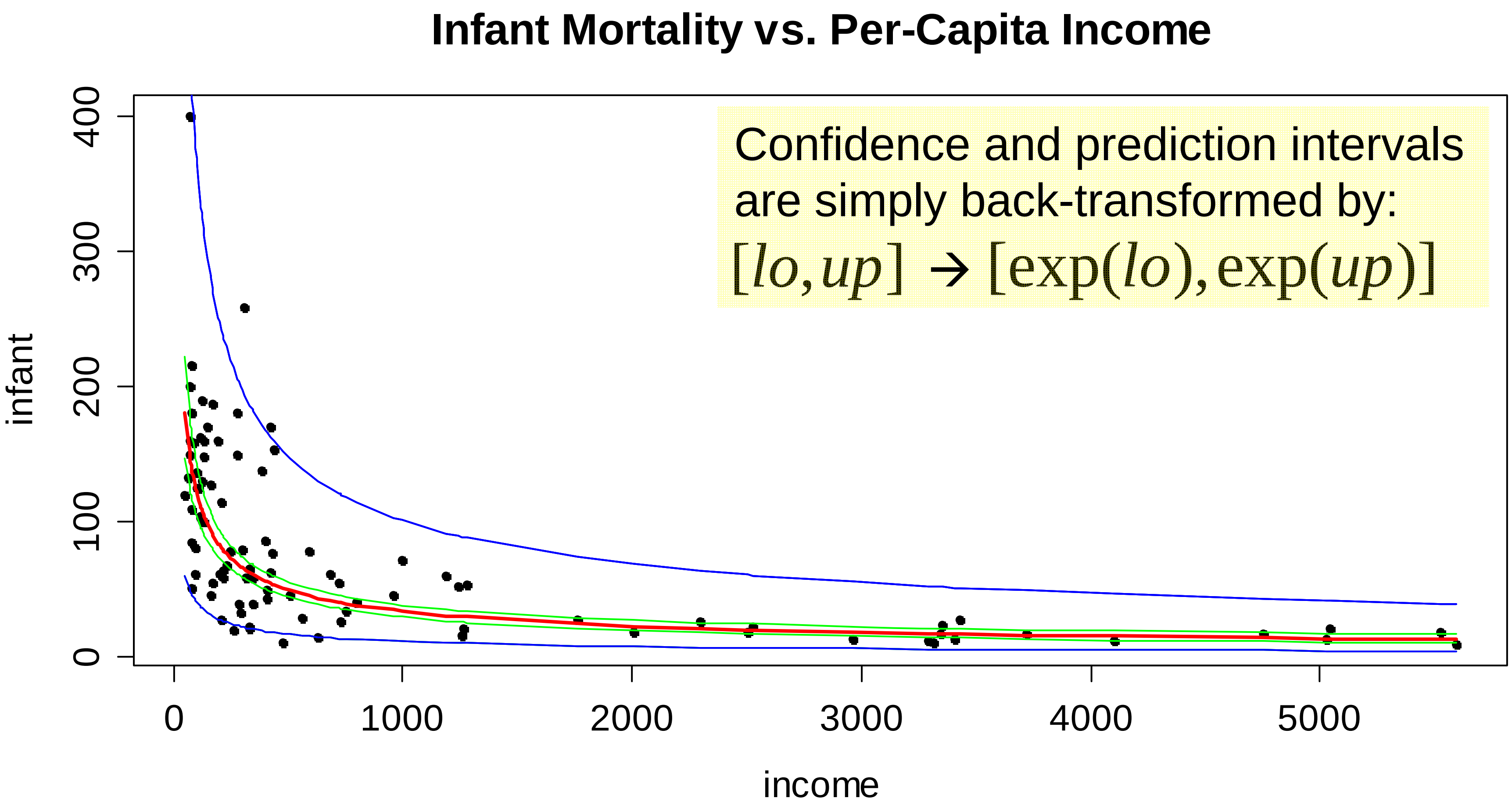
- We can either use $\hat{y} = \exp(\hat{y}' + \hat{\sigma}_E^2 / 2)$ which is motivated by the link between Gaussian and lognormal distribution, or the smearing estimator proposed by Duan (1983):

$$\hat{y} = \exp(\hat{y}') \cdot \frac{1}{n} \sum_{i=1}^{n} \exp(r_i')$$

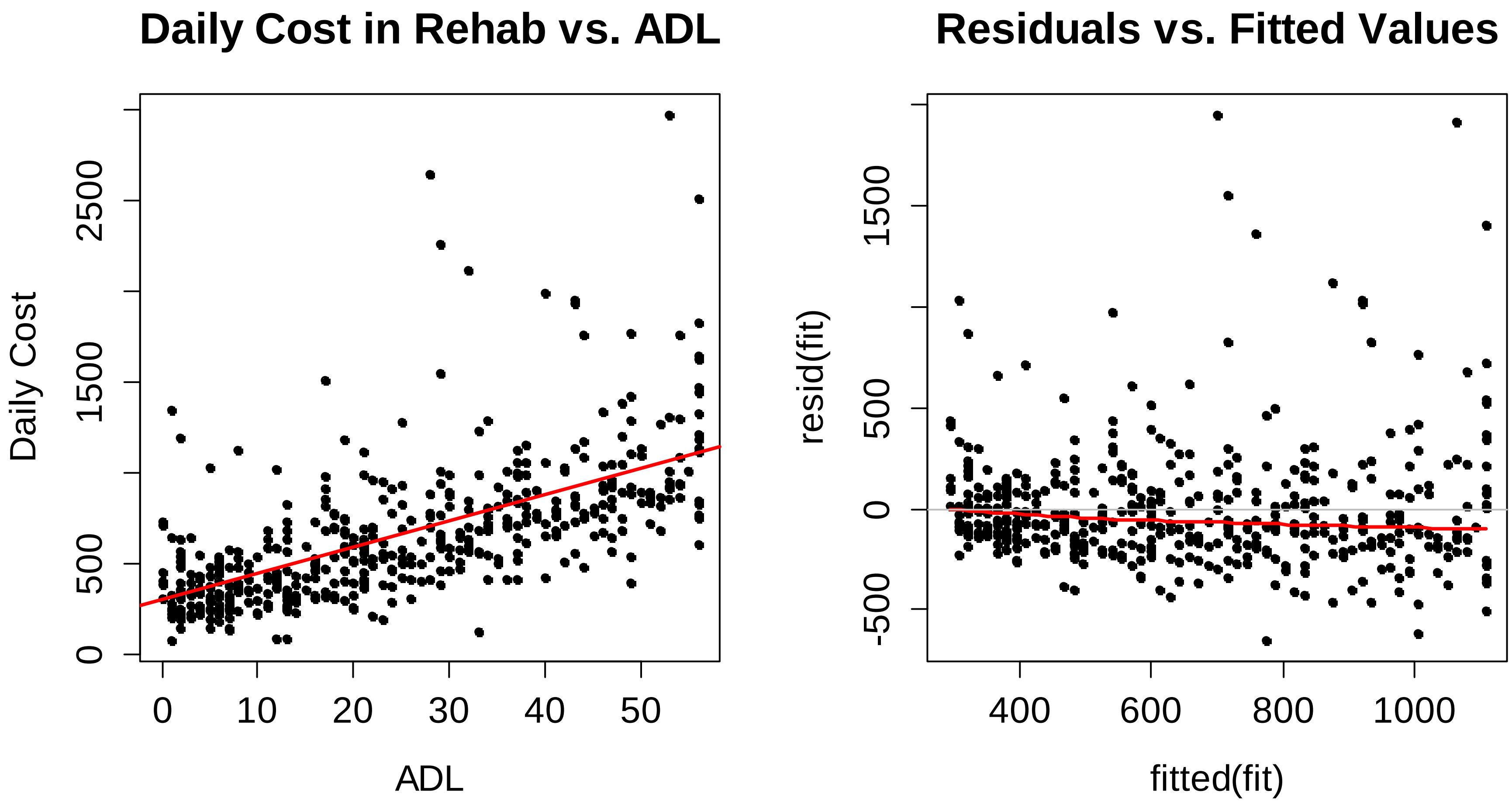
Conditional Mean and Median



Confidence and Prediction Interval



Another Example: Daily Cost in Rehab



Logged Response Model

We *transform* only the *response* variable: $y_i' = \log(y_i)$. Then, we use a linear model with predictor x_i and hence:

$$y_i' = \log(y_i) = \beta_0' + \beta_1' x_i + E_i'$$

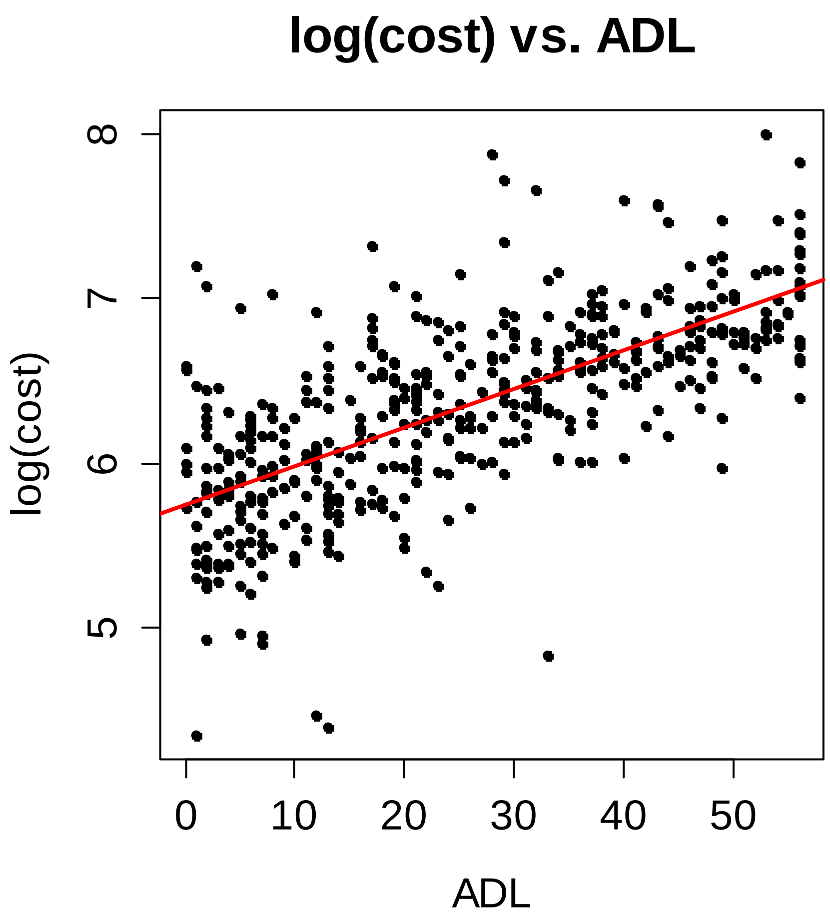
On the *original scale*, we can write the logged response model using the same predictors an obtain an **exponential function**:

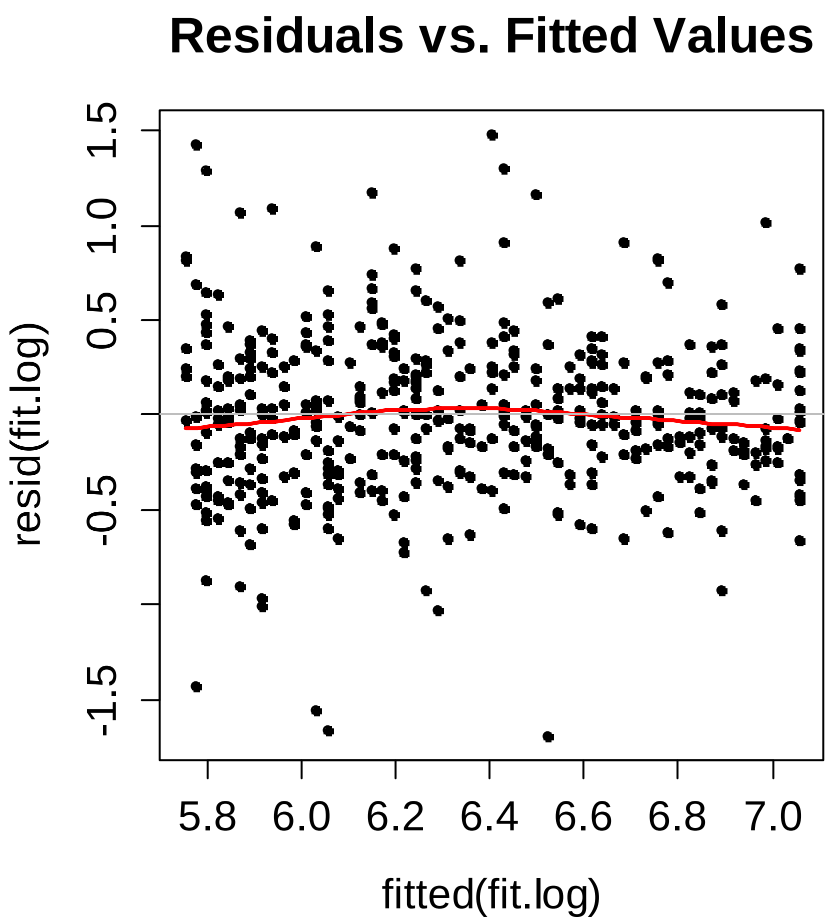
$$y_i = \exp(\beta_0') \cdot \exp(\beta_1' x_i) \cdot \exp(E_i') = \beta_0 \cdot \beta_1^{x_i} \cdot E_i$$

→ Predictor and error effects are **multiplicative !!!**

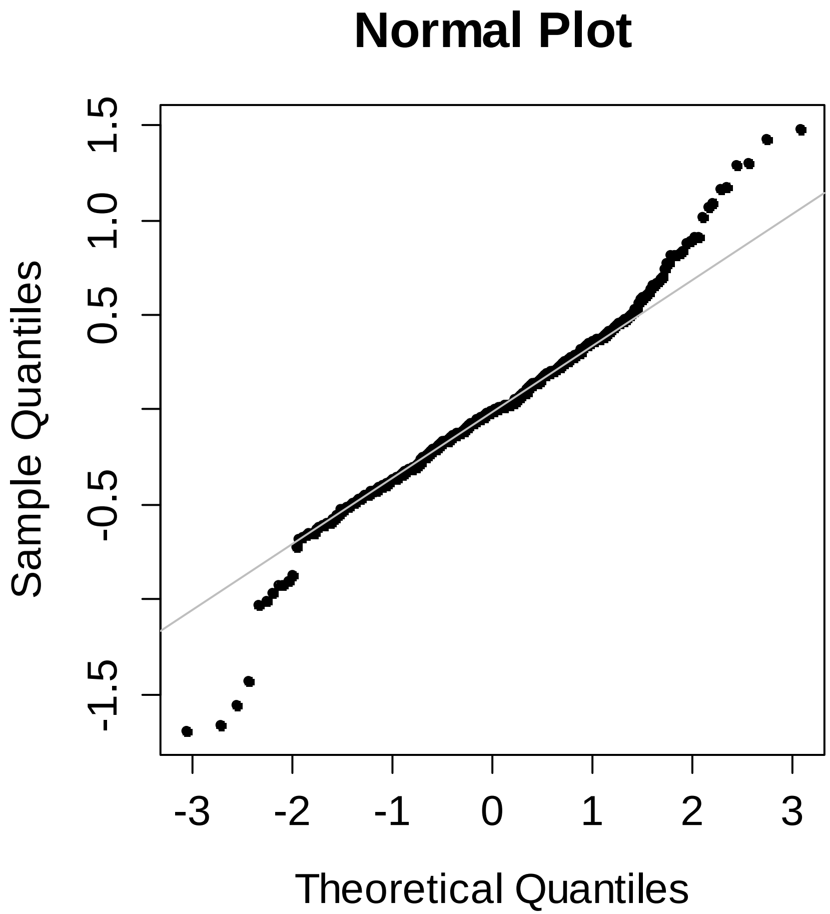
→ $E_i' \sim N(0, \sigma_E^2)$, thus, $E_i = \exp(E_i')$ has *lognormal distribution*

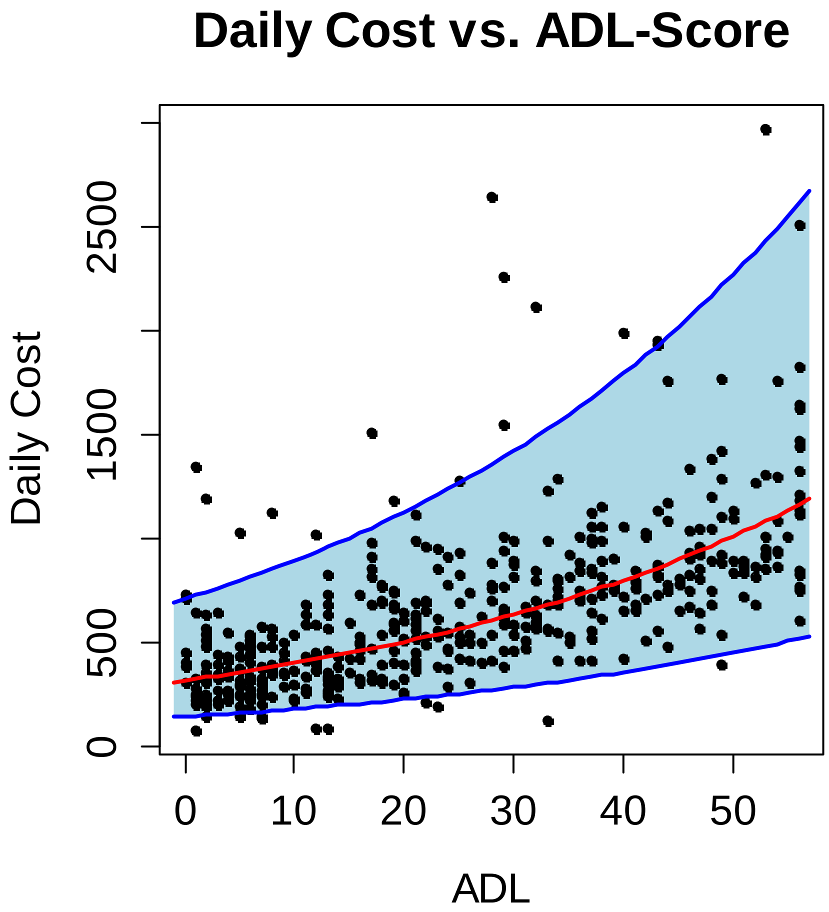
Fit and Residuals after the Transformation





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

$$\begin{aligned} \log(y_i) &= \beta_0' + \beta_1' x_i + E' \\ y_i &= \exp(\beta_0')\exp(\beta_1' x_i)\exp(E_i') \\ y_i &= \beta_0 \cdot \beta_1^{x_i} \cdot E_i \end{aligned}$$

An increase by one unit in x multiplies the fitted value on the original scale with the $\beta_1 = \exp(\beta_1')$. Furthermore, also the error term is multiplicative and has a lognormal distribution! Unbiased predictions require correction, as explained on slide 74 !!!

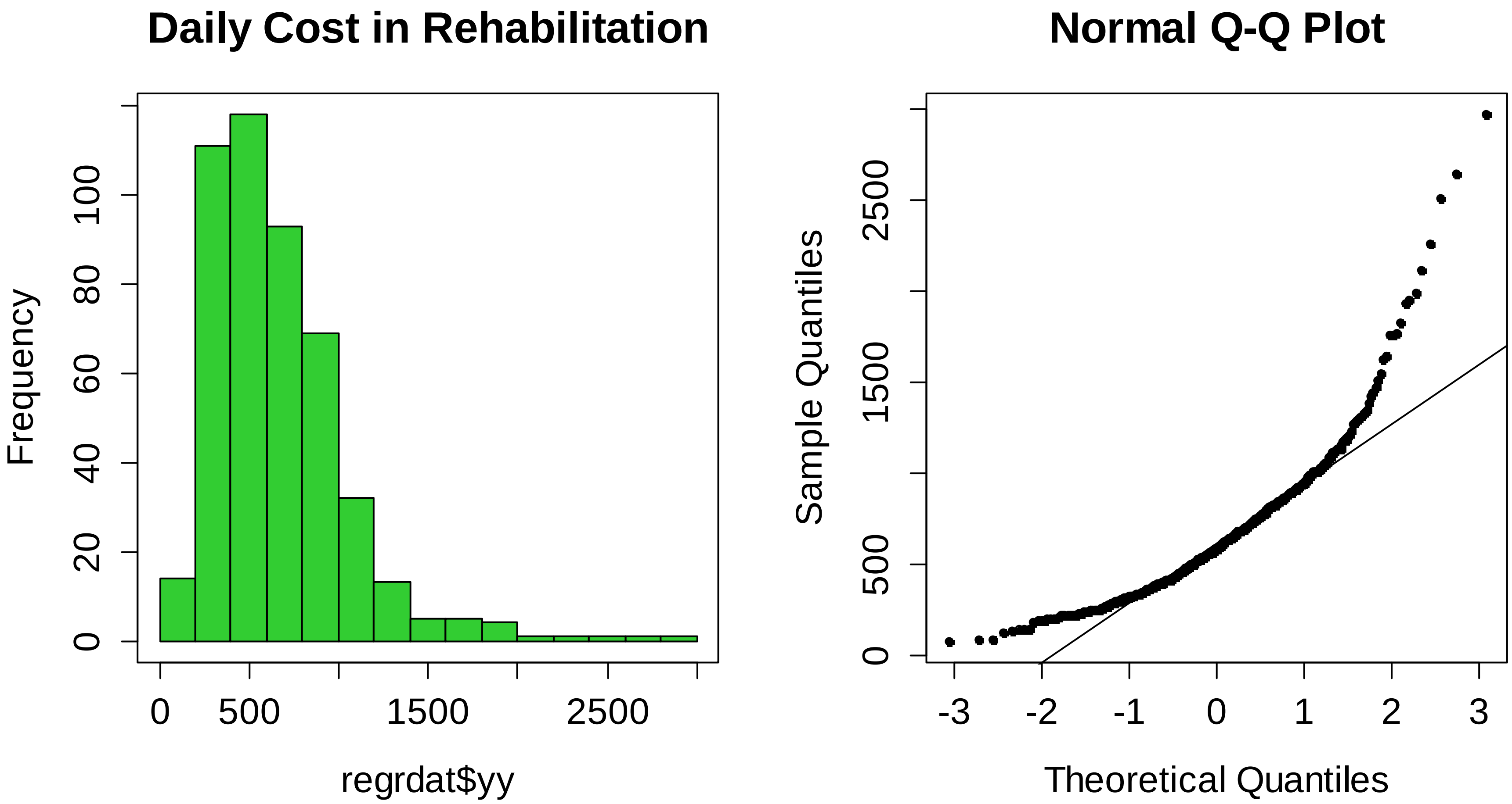
When to Transform?

Often, a log-transformation of response and/or predictor improves the fit. Some general guidelines for when to transform a variable:

- If on a **relative scale**, meaning that an increase from 10 → 11 is non-identical to 100 → 101, i.e. percent changes are required.
- **Left-closed** (with 0 as the smallest possible value), and **right-open** variables are often relative and require transformation.
- If the **scatter**, i.e. the magnitude of the uncertainty, **increases** with increasing value, as is often the case for relative scales.
- If the marginal distribution of the variable (as observed in a histogram) is clearly **right-skewed**.

Applied Statistical Regression
AS 2015 – Simple Regression

When to Transform?



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!**

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:

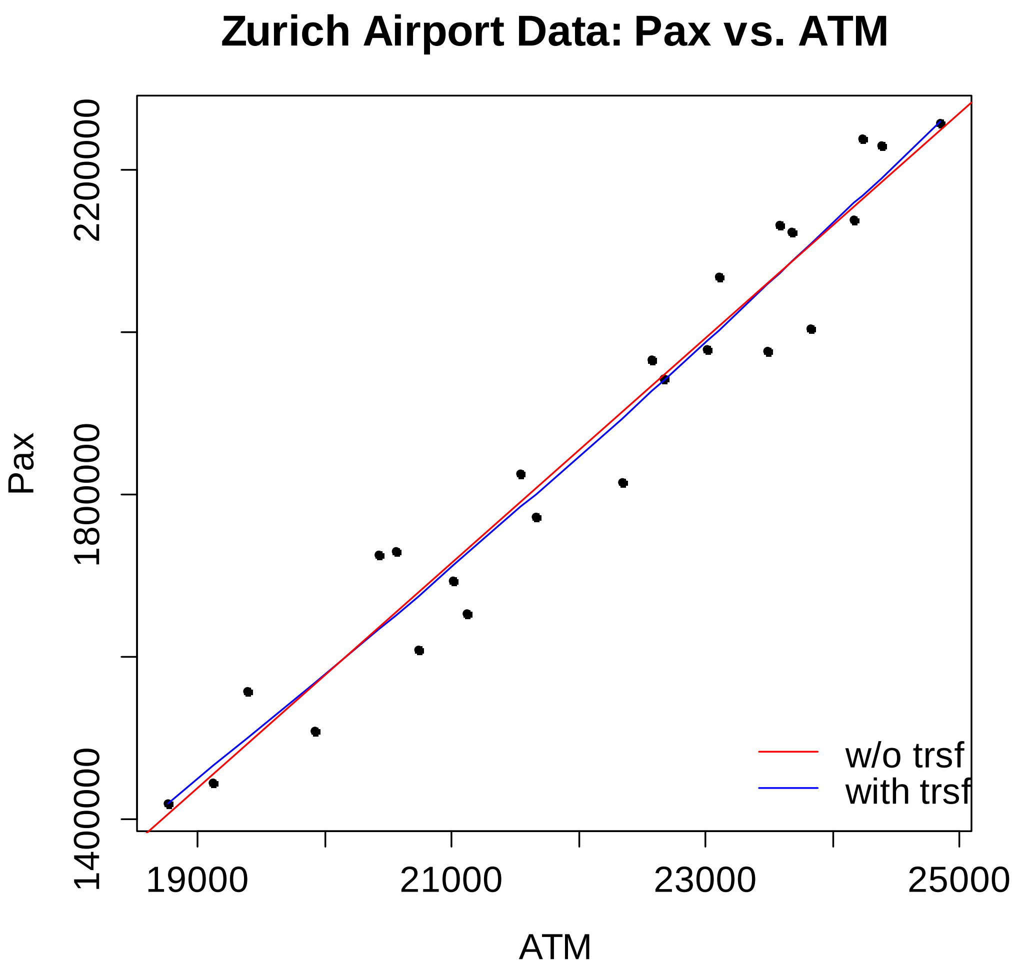
$$ATM' = \log(ATM),\; Pax' = \log(Pax)$$

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!

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