Statistical Computing
Week 7—20. November 2017
Basic Statistics: Regression

Statistical Models: linear Regression

Statistical Models: linear Regression

- Do variables W and X affect the value of variable Y?
- Can you express Y as a function of W und X, even though Y is observed with variability?

Examples

- Rent depends on floor area
- Fuel consumption of a car depends on speed
- Income depends on education level

In each of the examples there is an **independent variable**[1] (X), which has an effect on the **dependent variable**[2] (Y).

Regression function
Each observation of the independent variable y_i is a realisation from the regression function $y = f(x)$, which is a function of the dependant variable x_i. The value of y_i is observed “with error” Possible reasons for the “error” term are:

- natural variation
- imprecise measurement
- other unobserved variables
- ...

[1] German: Einflussgröße, “influence variable”

[2] German: Zielgröße, “target variable”

Regression function

Regression function examples

Simple linear	:	$f(x) = b_0 + b_1 x$
Quadratic	:	$f(x) = b_0 + b_1 x + b_2 x^2$
Multiple linear	:	$f(x) = b_0 + b_1 x + b_2 w$

Simple linear Regression

True function	$f(x)$	$= b_0 + b_1 x$
Observed data	y_i	$= b_0 + b_1 x_i + \epsilon_i$
i-th fitted value	$\widehat{y}_i$	$= \widehat{b}_0 + \widehat{b}_1 x_i$
i-th residual	$\widehat{\epsilon}_i$	$= y_i - \widehat{y}_0$

ϵ_i is the i-th error term, which is unknown.

Regression parameter:

b_0: true regression **intercept** coefficient (unknown)
b_1: true regression **gradient** coefficient (unknown)
$\widehat{b}_0$: estimated intercept coefficient
$\widehat{b}_1$: estimated gradient coefficient

Estimated values are written with a “hat”.

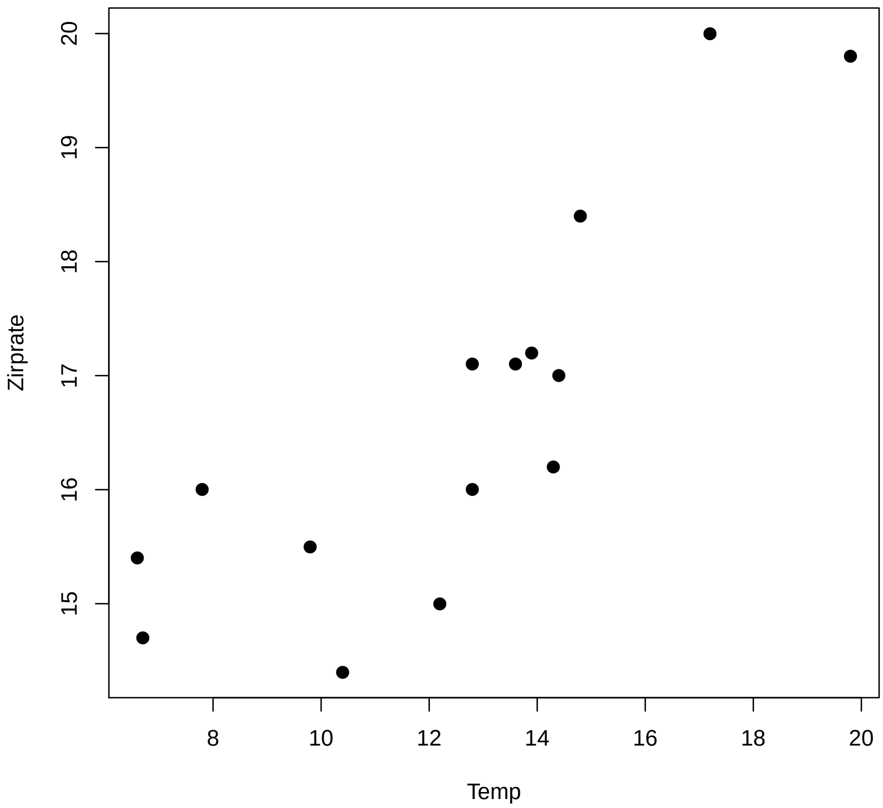
Data Example

Grasshoppers chirp at a rate which depends on the temperature.

	1	2	3	4	5	6	7	8	9
Temperature	17.2	7.8	19.8	14.8	12.8	9.8	6.7	13.6	6.6
Chirp rate	20.0	16.0	19.8	18.4	17.1	15.5	14.7	17.1	15.4

	10	11	12	13	14	15
Temperature	14.3	12.2	13.9	12.8	14.4	10.4
Chirp rate	16.2	15.0	17.2	16.0	17.0	14.4

Grasshopper Data: Chirp rate against temperature



Dependent variable: X is temperature
Independent variable: Y is chirp rate

The regression function is $y_i = b_0 + b_1 x_i + \epsilon_i$

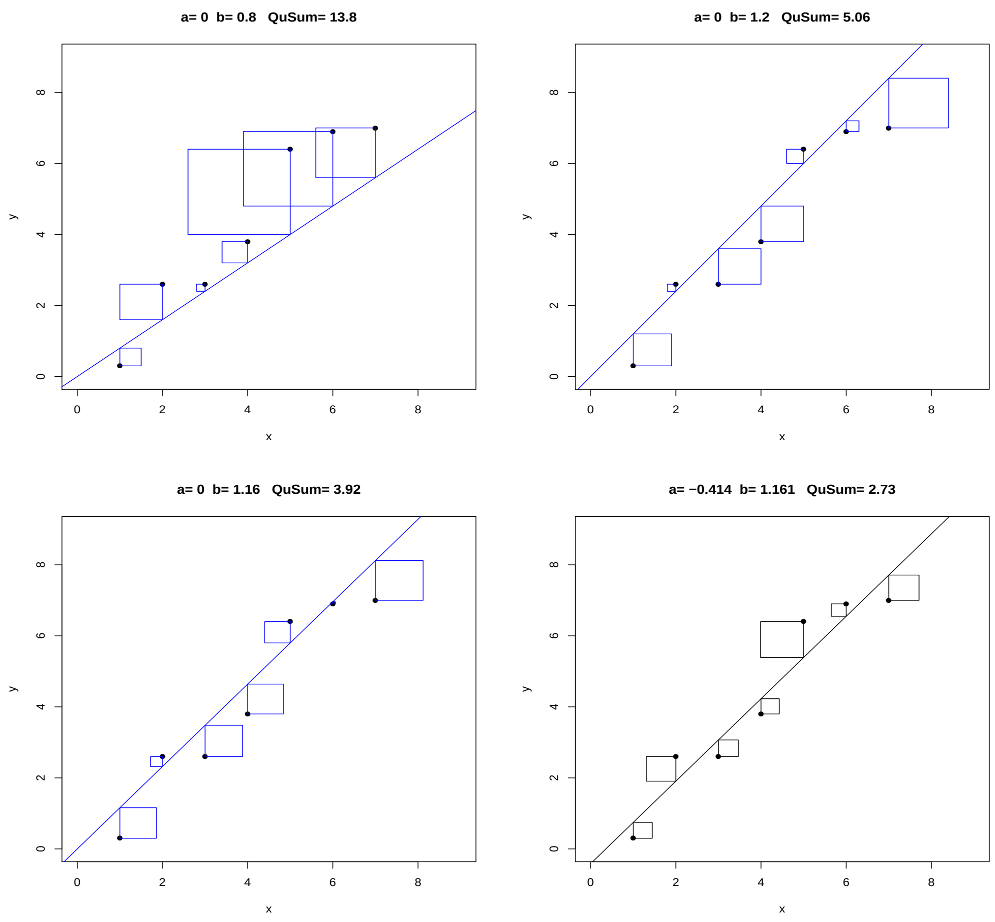
The chirp rate values don't all lie on a straight line.
ϵ_i is the error term corresponding to the natural variability in the chirp rates.

The true values of b_0 and b_1 are unknown. We have to *estimate* them.

The regression line is *„the line which best fits the data“*

When the total distance between the data points and the regression line is small, then the line fits the data well.

Our approach is to find the best line (the values of $\widehat{b}_0$ and $\widehat{b}_1$) using the method of least squares minimisation ...



QuSum = Sum of Squares

The squares in the diagrams are the values $\widehat{\epsilon}_i^2$ for a given $\widehat{b}_0$ and $\widehat{b}_1$. We want to minimise the total area of the squares, i.e. minimise

$$\widehat{\epsilon}_1^2 + \widehat{\epsilon}_2^2 + \cdots + \widehat{\epsilon}_n^2 = \sum_{i=1}^{n} \widehat{\epsilon}_i^2.$$

The values of $\widehat{b}_0$ and $\widehat{b}_1$ which minimise the sum of squares are our *best* intercept and gradient.

In this example:

the *best* intercept is $\widehat{b}_0 = -0.414$ and
the *best* gradient is $\widehat{b}_1 = 1.161$.

Method

$$\text{Minimise} \quad RSS = \sum_{i=1}^{n} (\widehat{\epsilon}_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

(RSS = Residual Sum of Squares)

Computing the Coeffizients

We don't need to explicitly minimise the residual sum of squares.
The best estimates can be calculated directly using two formulae:

Linear coefficient (gradient)

$$\widehat{b}_1 = \frac{s_{xy}}{s_x^2} = \frac{\text{Covariance}}{\text{Variance X}}$$

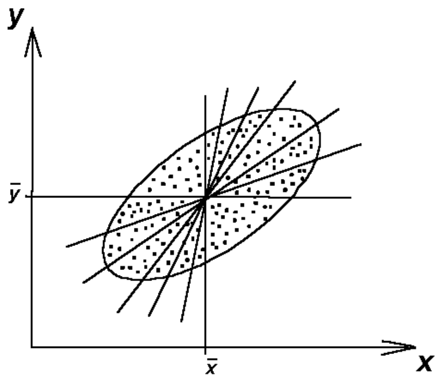
Intercept coefficient

$$\widehat{b}_0 = \overline{y} - \widehat{b}_1 \overline{x}$$

A consequence of the two formulae above is

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i) = \sum_{i=1}^{n}\widehat{\epsilon}_i = 0 \quad \text{or} \quad \overline{\widehat{\epsilon}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}_i = 0$$

Graphical interpretation: the regression line passes through $(\overline{x}, \overline{y})$.

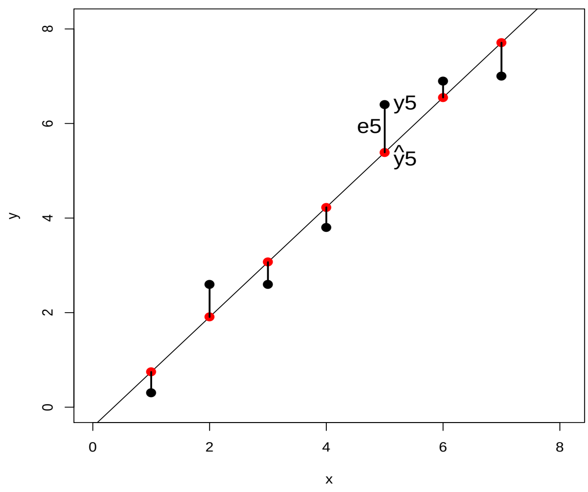


For every x_i, there is a point on the regression line. The y-coordinate of this point is called the **fitted value** $\widehat{y}_i$.

$$\widehat{y}_i = f(x_i) = \widehat{b}_0 + \widehat{b}_1 x_i$$

The difference between observed value y_i and the fitted value $\widehat{y}_i$ is called **residual**.

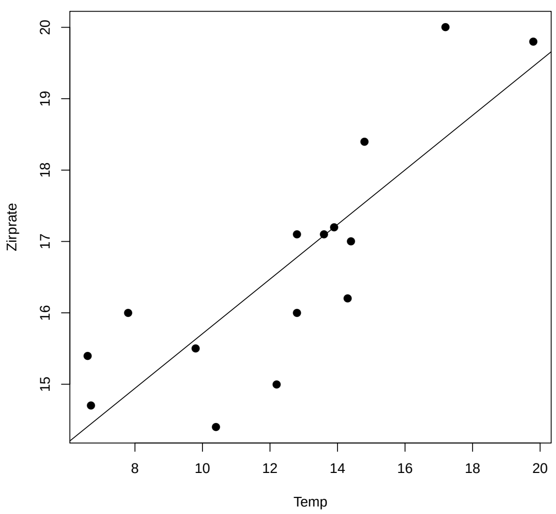
$$\widehat{\epsilon}_i = y_i - \widehat{y}_i$$



A **predicted value** is similar to a fitted value. In theory any real number x can be used in the regression formula, even if it is not a value in the data set.

For $x \in \mathbb{R}$ the predicted value is $f(x) = \widehat{b}_0 + \widehat{b}_1 x$.

Grasshopper Beispiel



Mean temp	$\overline{x} = 12.47$
Mean chirp	$\overline{y} = 16.65$
Variance temp	$s_x^2 = 13.80$
Covarianz	$S_{xy} = 5.278$
Gradient $\widehat{b}_1$	S_{xy}/S_x^2
	$= 13.80/5.278 = 0.3825$
Intercept $\widehat{b}_0$	$\bar{y} - b\bar{x}$
	$= 16.65 - 0.3825 \cdot 12.47$
	$= 11.88$

The regression line has the form: $f(x) = 11.88 + 0.3825x$

The fitted value and the residual for $x_{11} = 12.2$ are:

The predicted values for $10^\circ C$ and $20^\circ C$ are:

Regression in R

Output from the grasshopper regression model

```
> lm.obj<-lm(formula = chirp ~ temp, data = Grasshoppers)
> summary(lm.obj) Call: lm(formula = chirp ~ temp, data =
Grasshoppers)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54879 -0.58426  0.01574  0.60056  1.53880

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.882   0.908      13.084 7.36e-09 ***
Temp         0.382   0.069       5.466 0.000108 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Residual standard error: 0.9725 on 13 degrees of freedom
Multiple R-squared: 0.6968, Adjusted R-squared: 0.6735
F-statistic: 29.88 on 1 and 13 DF, p-value: 0.0001081
```


Coefficient of Determination, Goodness of Fit, R^2

How good does the regression line fit the data?

The R^2 statistic is the ratio $\dfrac{\text{Variance of the fitted values)}}{\text{Variance of (Y)}}$ is a measure of the "model fit".

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{s^2_{\widehat{Y}}}{s^2_Y}$$

Properties:

$$R^2 = r^2_{XY} \quad \Rightarrow \quad 0 \leqslant R^2 \leqslant 1$$

When all the data points lie on a straight line then $R^2 = 1$

When X has no influence on Y then R^2 is near to zero.

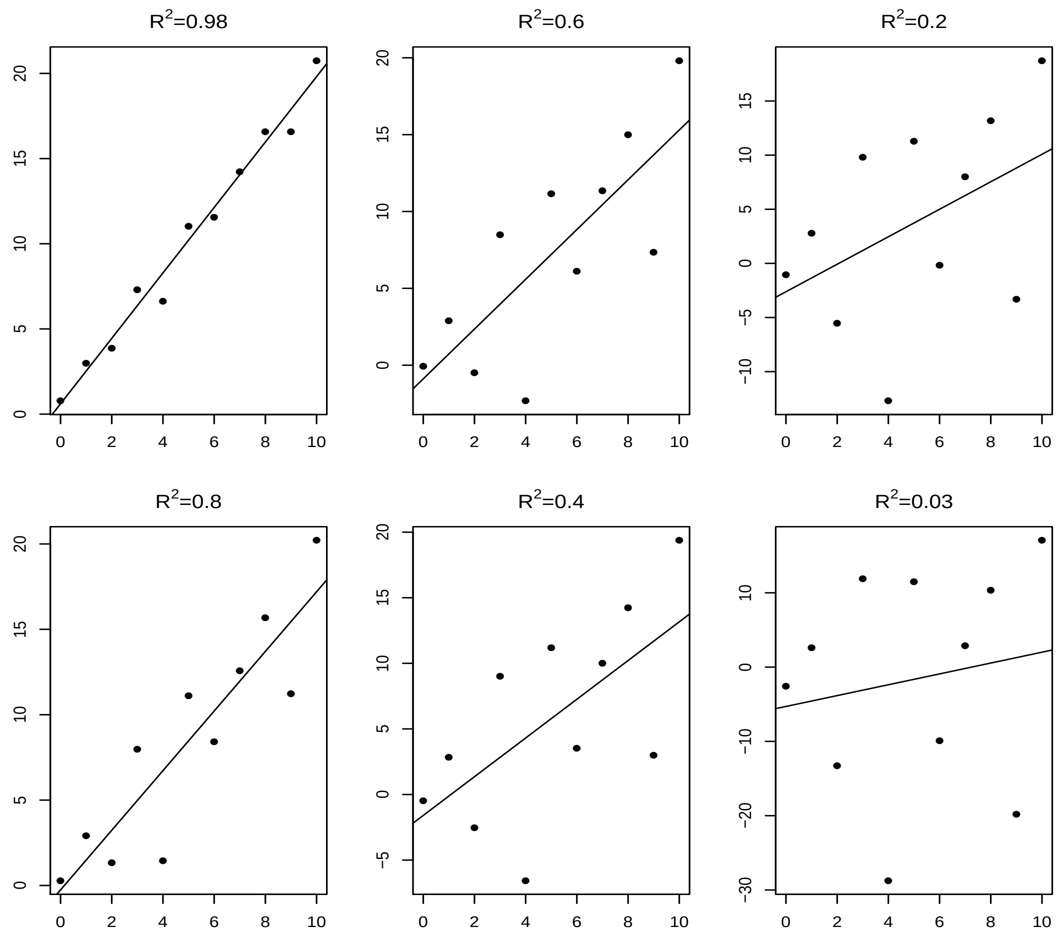
Grasshopper example:
(in the R Output: `Multiple R-squared: 0.6968`).

$s^2_{\widehat{Y}} = 2.018729$ $\quad$ $s^2_Y = 2.896952$

$$R^2 = \frac{2.018729}{2.896952} = 0.6968 \approx 0.7.$$

On a scale from 0 to 1 the model fit is *good*.

Graphical Examples of R^2



Warning!
R^2 is a popular but dangerous coefficient to use.
In a simple linear regression it can be useful to know how good the model fit is, but when choosing which variables should be included in a multiple regression maximising R^2 can quickly lead to *over fitting*, which is a common problem when analysing large datasets.

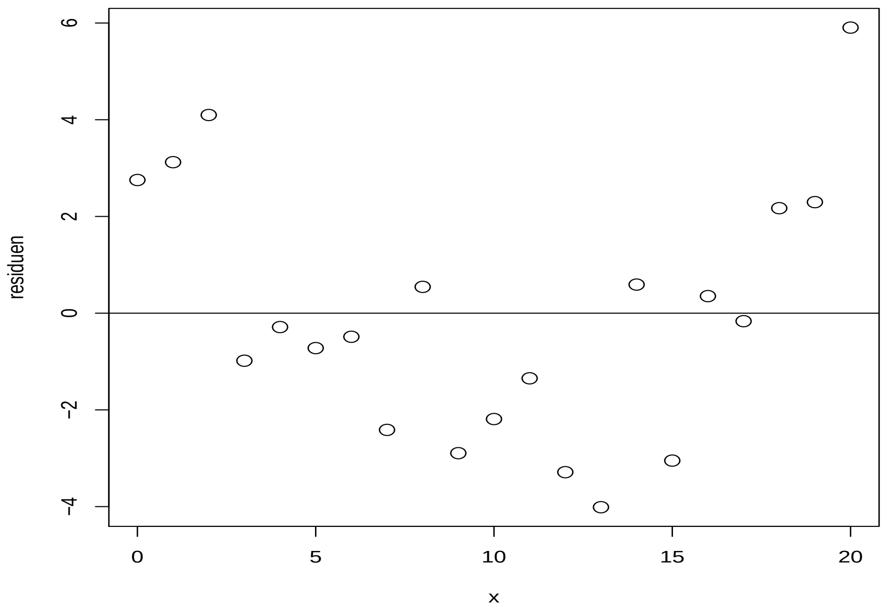
Checking the residuals

We have assumed that the dependent variable has a linear effect on the dependent variable.
We should check this assumption using a "residual plot"

The dependent variable x or the fitted values $\widehat{y}$ are plotted on the x axis, and the residuals $\widehat{\epsilon}$ are plotted on the y axis.

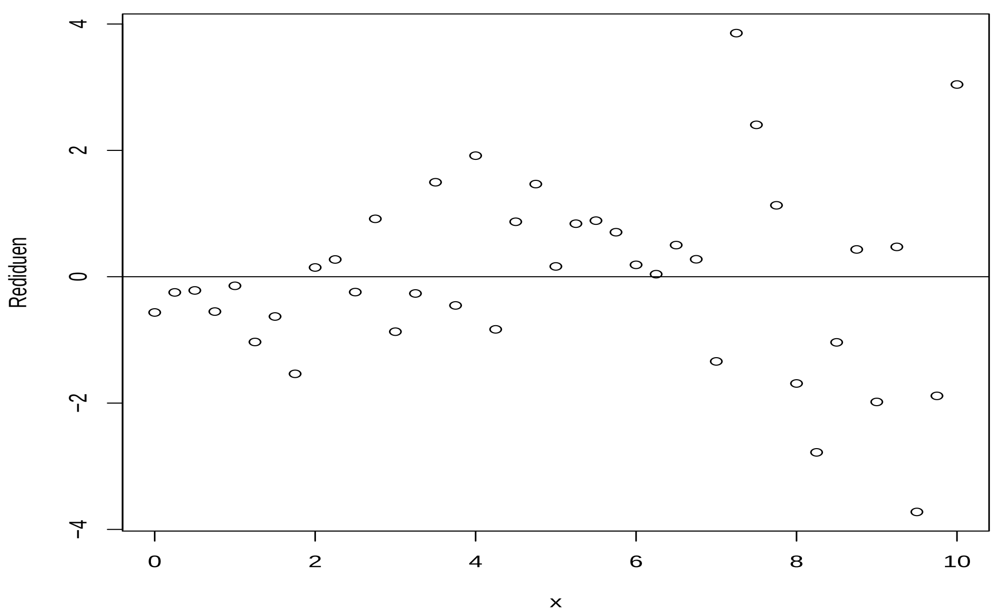
The most common problems are:

a) y is not linear in x:



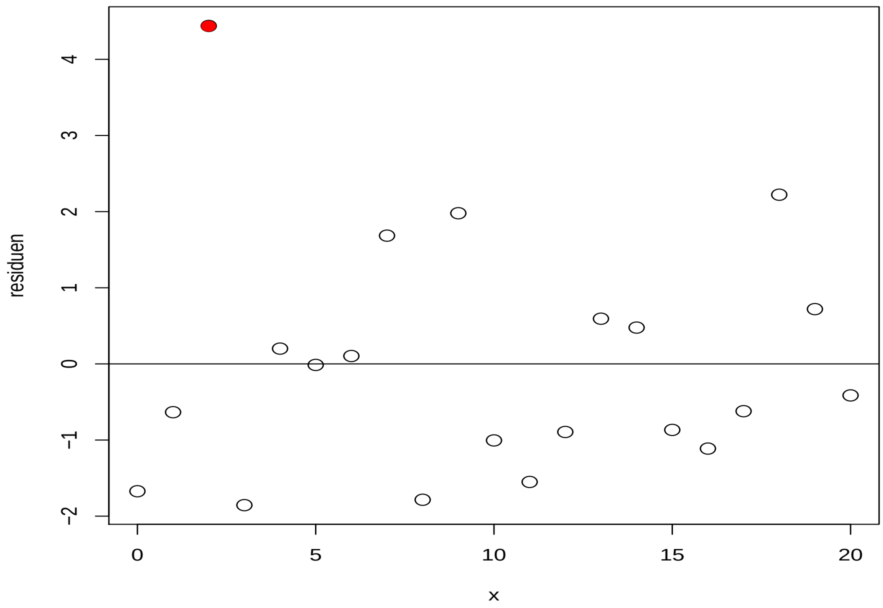
Possible solution:
try fitting a quadratic regression model $f(x) = b_0 + b_1 x + b_2 x^2$.

b) The variance of y varies with the x values (Heteroscedasticity):



Possible solution: Try transforming the variables e.g. use $\log(y)$ and/or $\log(x)$

c) There are outliers present.



Possible solution: Use a weighted regression model, or drop the outlier-observation from the model.

Quadratic Regression

In quadratic regression a parabola is fitted to the data: $f(x) = b_0 + b_1 x + b_2 x^2$

The best parabola is again defined by minimising the residual sum of squares, as with linear regression.

The model residuals are:

$$\widehat{\epsilon}_i = y_i - f(x_i) = y_i - (\widehat{b}_0 + \widehat{b}_1 x_i + \widehat{b}_2 x_i^2).$$

The model estimates $\widehat{b}_0$, $\widehat{b}_1$, $\widehat{b}_2$ minimise $\sum_{i=1}^{n} \widehat{\epsilon}_i^2$.

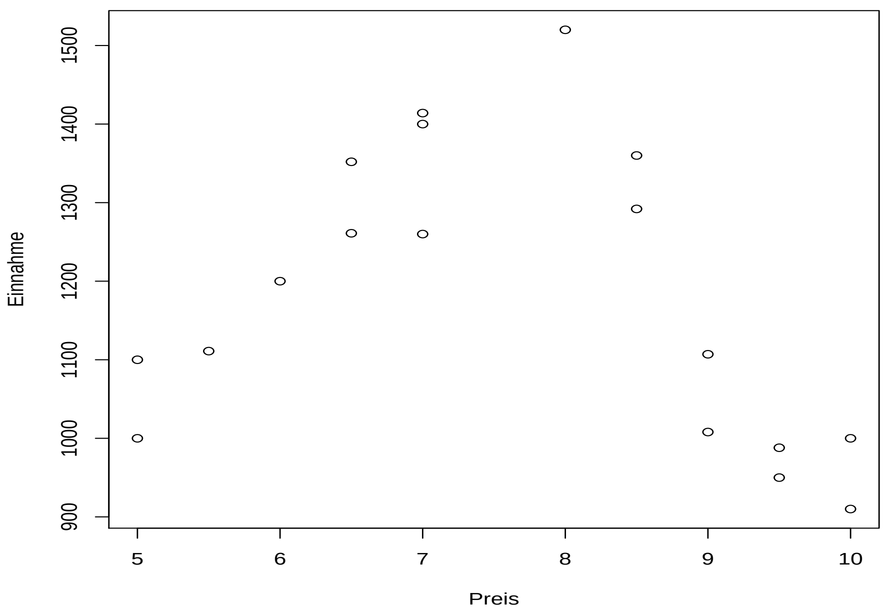
The fitted values are:

$$\widehat{y}_i = \widehat{b}_0 + \widehat{b}_1 x_i + \widehat{b}_2 x_i^2.$$

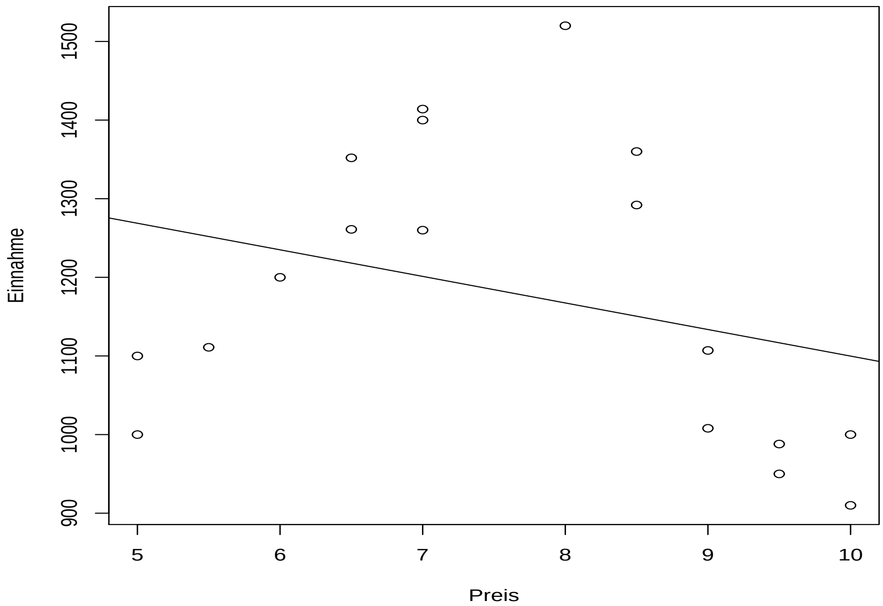
Eaxample

The management of a museum varied the entrance price of an exhibition in order to asses the influence on the daily taking. The entry price x_i in Euros is the independent variable and the daily taking y_i in Euros is the dependent variable. The data are:

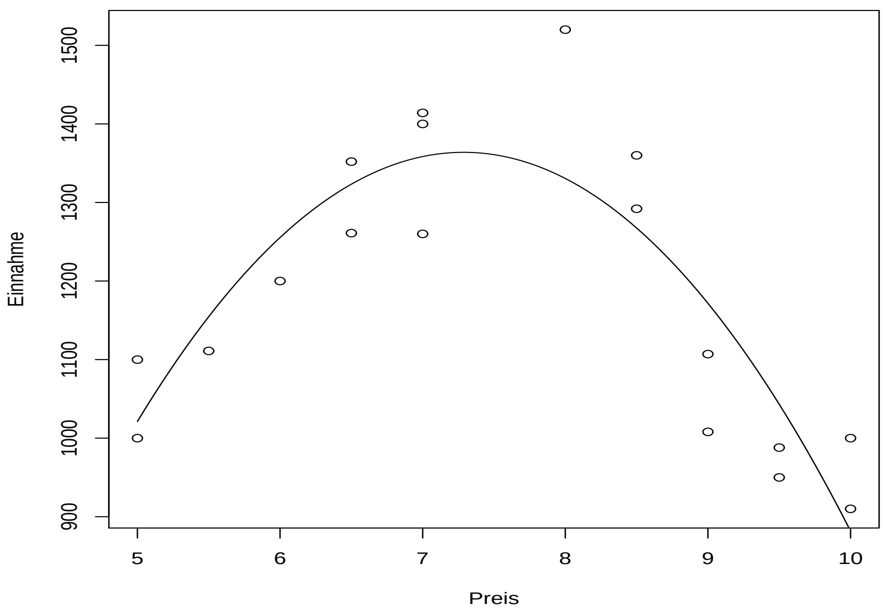
(5 ;1000)	(5 ;1100)	(5.5;1111)
(6 ;1200)	(6.5;1352)	(6.5;1261)
(7 ;1260)	(7 ;1400)	(7 ;1414)
(8 ;1520)	(8.5;1360)	(8.5;1292)
(9 ;1107)	(9 ;1008)	(9.5; 950)
(9.5; 988)	(10 ;1000)	(10 ; 910)



A straight line does not fit the data at all well



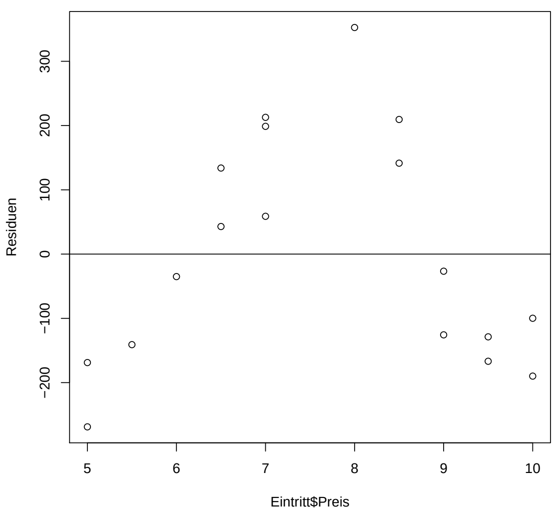
A quadratic regression fits much better.



$$F(x) = -2114 + 954.7x - 65.50x^2$$

Residual plots for a) linear and b) quadratic regression

a)



b)