Regression Analysis

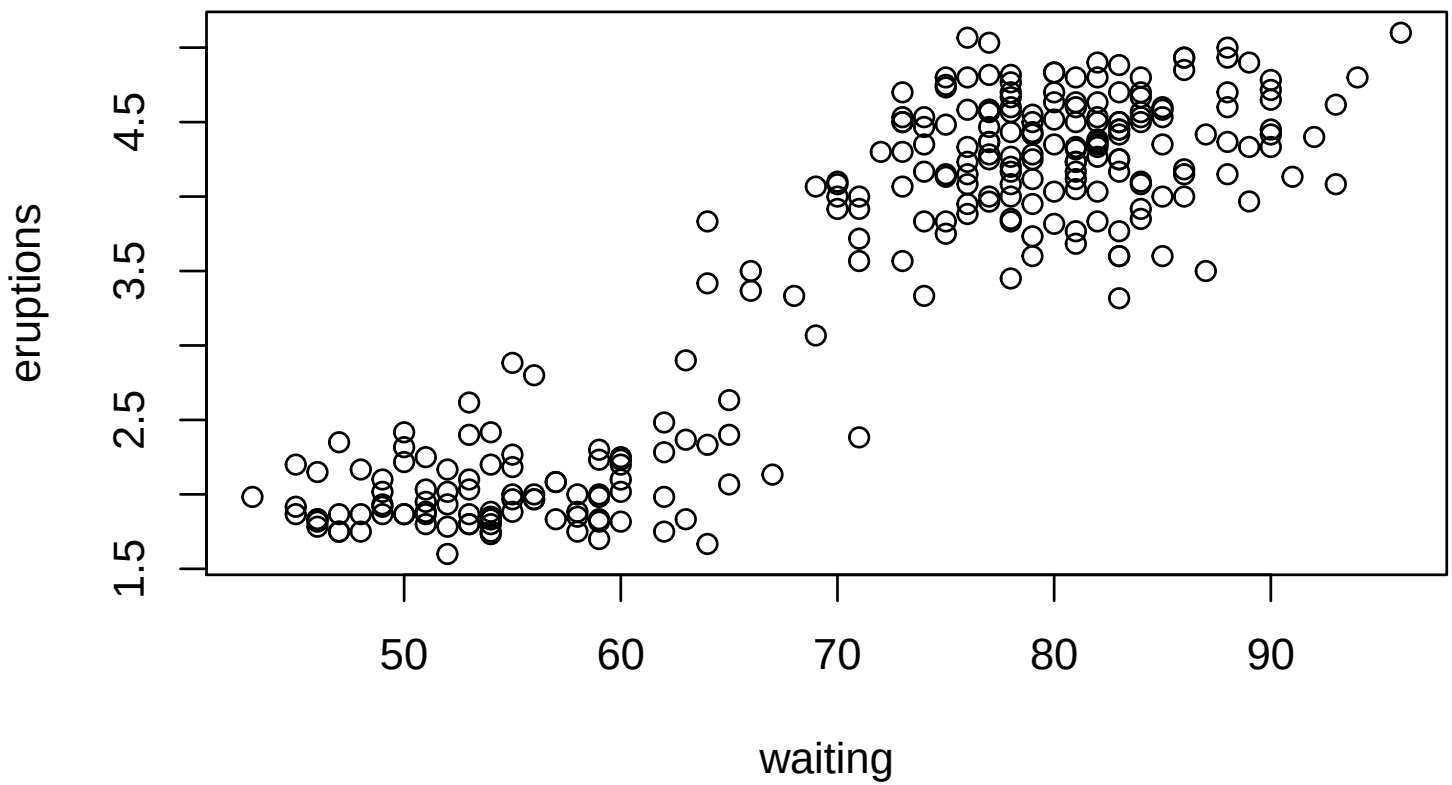
Chapter 1. The Simple Regression Model

Prof. dr. Thomas Neyens

Course notes: Prof. dr. Mia Hubert & Prof. dr. Stefan Van Aelst

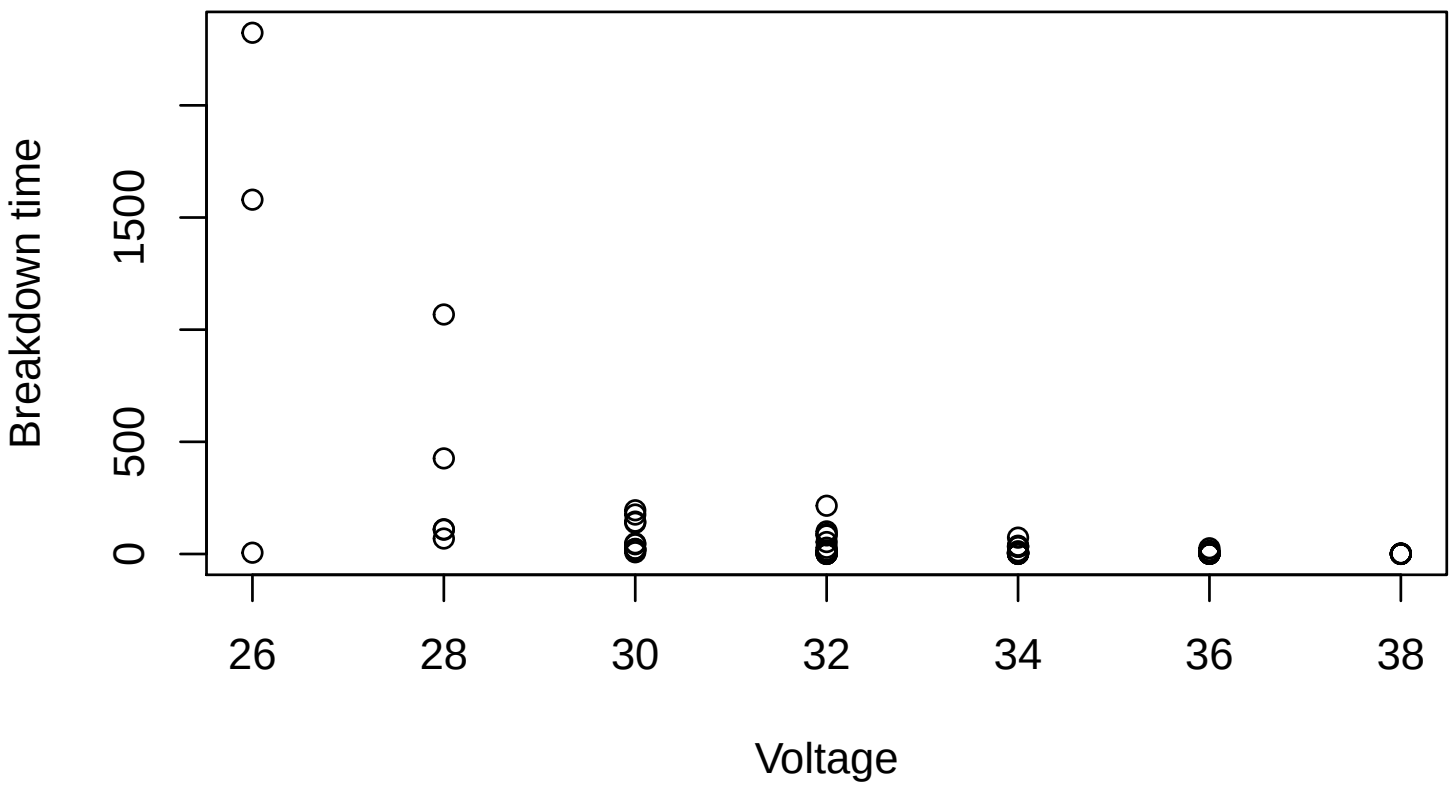
A relationship between two variables

Example: Geyser eruption length vs. waiting time since last eruption



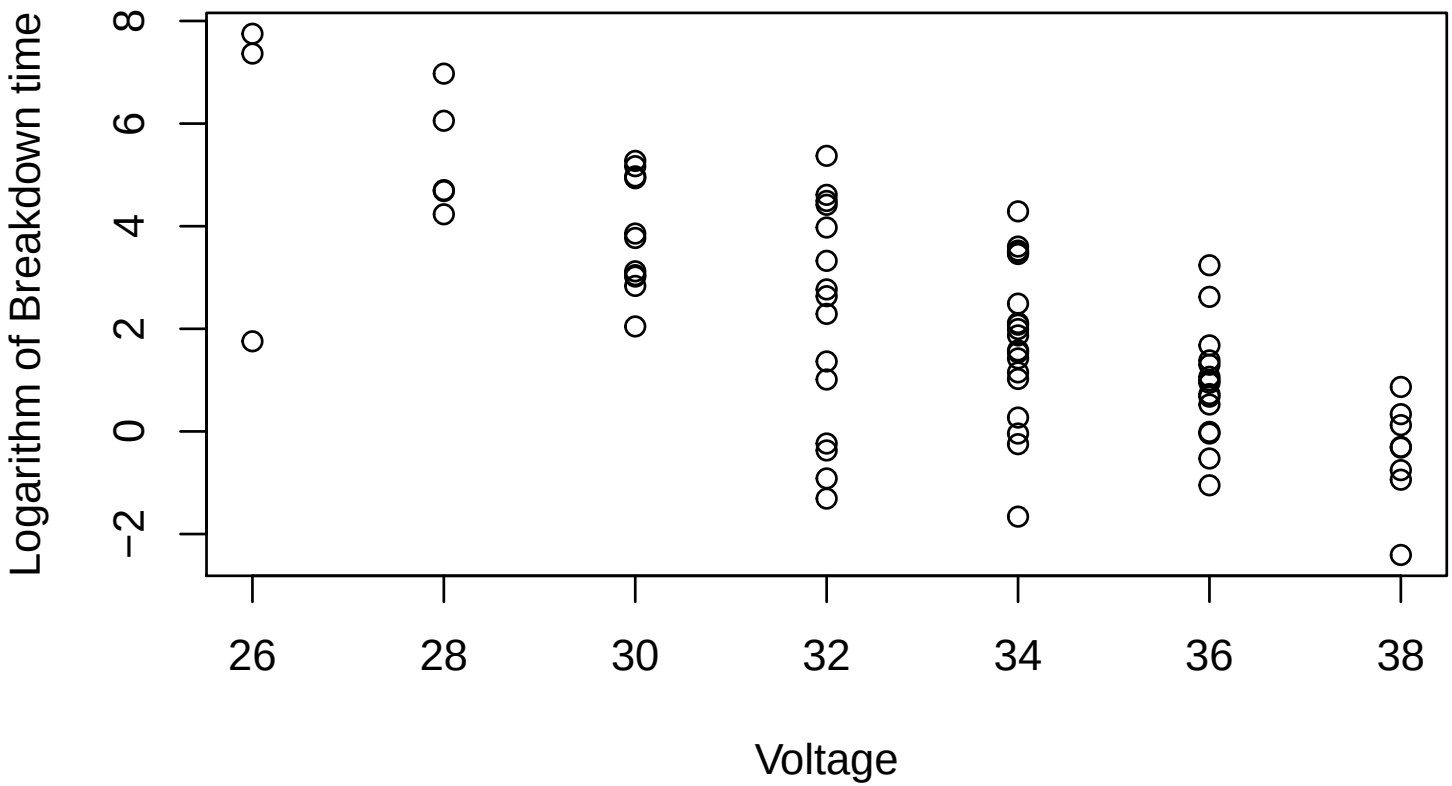
A relationship between two variables

Example: time until breaking of fluid's insulating property vs. applied voltage



A relationship between two variables

Example: **ln(**time until breaking of fluid's insulating property**)** vs. applied voltage



Important difference between both examples

- Geyser example: the recorded values for both the waiting and eruption times are observed values $\rightarrow$ X and Y variable are **random variables**.
- Fluid example: the voltage dose is chosen by the experimenters and the breakdown time is recorded for these fixed doses. $\rightarrow$ Y is a **random variable**, X is **not**.

Regression modelling as discussed next can be used for both types of data under suitable conditions.

The simple linear model

The simple linear model is given by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for $i = 1, \ldots, n$. The parameter β_0 is called the *intercept*, whereas β_1 is called the regression *slope*.

- y_i (x_i) are not necessarily values of the observed response (predictor) variable Y, but can be values for any suitable function $g(Y)$ ($f(X)$).
- We assume that X does not contain any random effect or measurement error (difficult to satisfy in observational studies).

The simple linear model

For the error term, we assume that the Gauss-Markov conditions are satisfied which means that

$$\begin{aligned} \mathrm{E}[\epsilon_i] &= 0 \\ \mathrm{Var}[\epsilon_i] &= \sigma^2 \\ \mathrm{E}[\epsilon_i \epsilon_j] &= 0 \text{ for all } i \neq j \end{aligned}$$

for $i = 1, \ldots, n$.

The simple linear model

A detail: in fact, when X is not set by design and is treated as a random variable, the assumptions are conditioned on X

$$\begin{aligned} \mathrm{E}[\epsilon_i|X] &= 0 \\ \mathrm{Var}[\epsilon_i|X] &= \sigma^2 \\ \mathrm{E}[\epsilon_i\epsilon_j|X] &= 0 \text{ for all } i \neq j \end{aligned}$$

for $i = 1, \ldots, n$.

So in the case that X is random, it is also assumed that the errors ϵ_i are independent of X.

The simple linear model

As the ϵ_i are random variables with zero mean, also Y is a random variable that satisfies:

$$\mathrm{E}[Y|X] = \beta_0 + \beta_1 X$$

Conditionally on the observed values for X, this can also be written as:

$$\mathrm{E}[Y|X = x_i] = \beta_0 + \beta_1 x_i$$

The simple linear model

For $X = 0$, $\mathrm{E}[Y|X = 0] = \beta_0$

- The intercept of the model can be interpreted as the expected response when X equals zero.
- Or, β_0 is the mean of the distribution of Y at $X = 0$.

The simple linear model

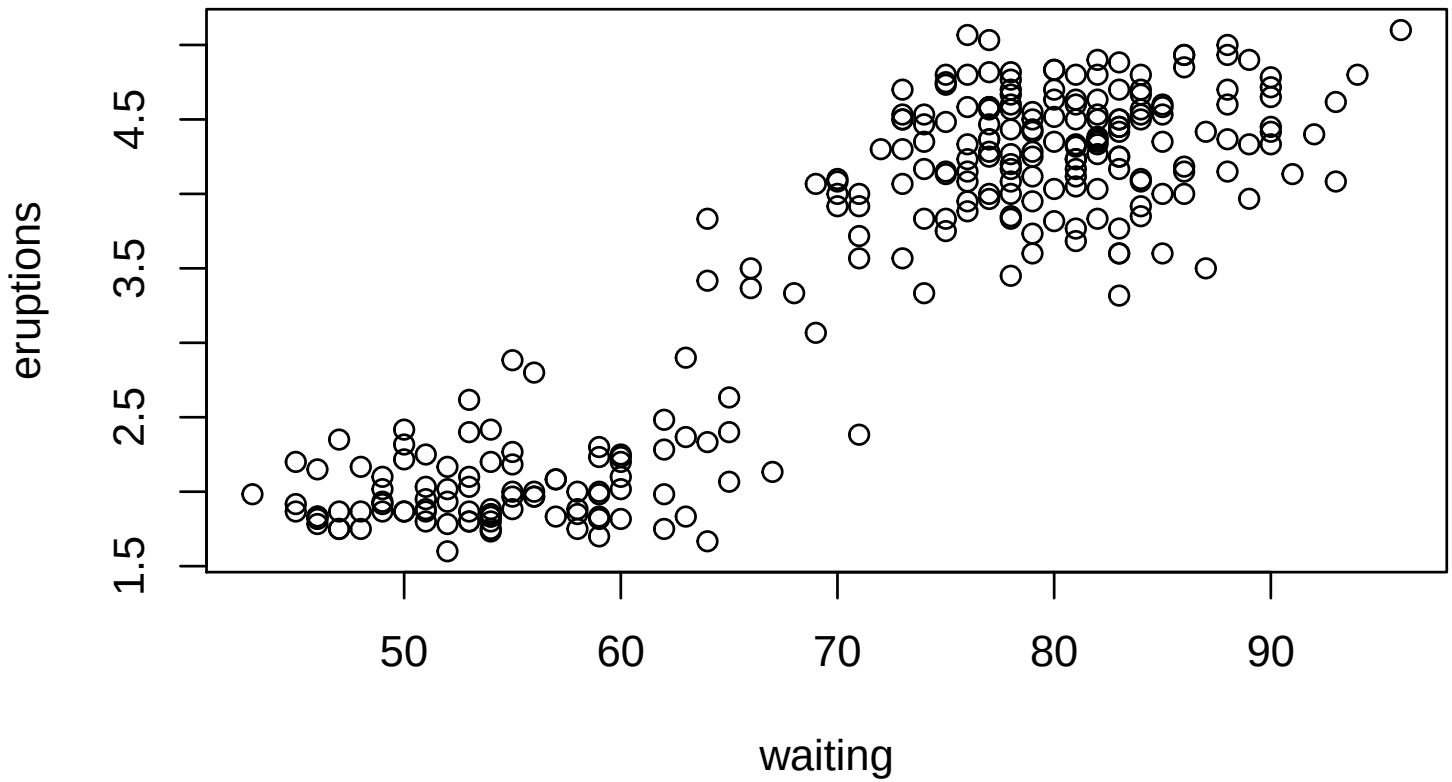
If we go from x to $x+1$, the expected response increases from $\mathrm{E}[Y|X=x] = \beta_0 + \beta_1 x$ to $\mathrm{E}[Y|X=x+1] = \beta_0 + \beta_1(x+1)$. Therefore, we find that

$$\beta_1 = \mathrm{E}[Y|X=x+1] - \mathrm{E}[Y|X=x]$$

- The slope β_1 can be interpreted as the change in the expected response Y if X increases by one unit.
- Or, if X increases one unit, then Y is expected to change by β_1 units on average.

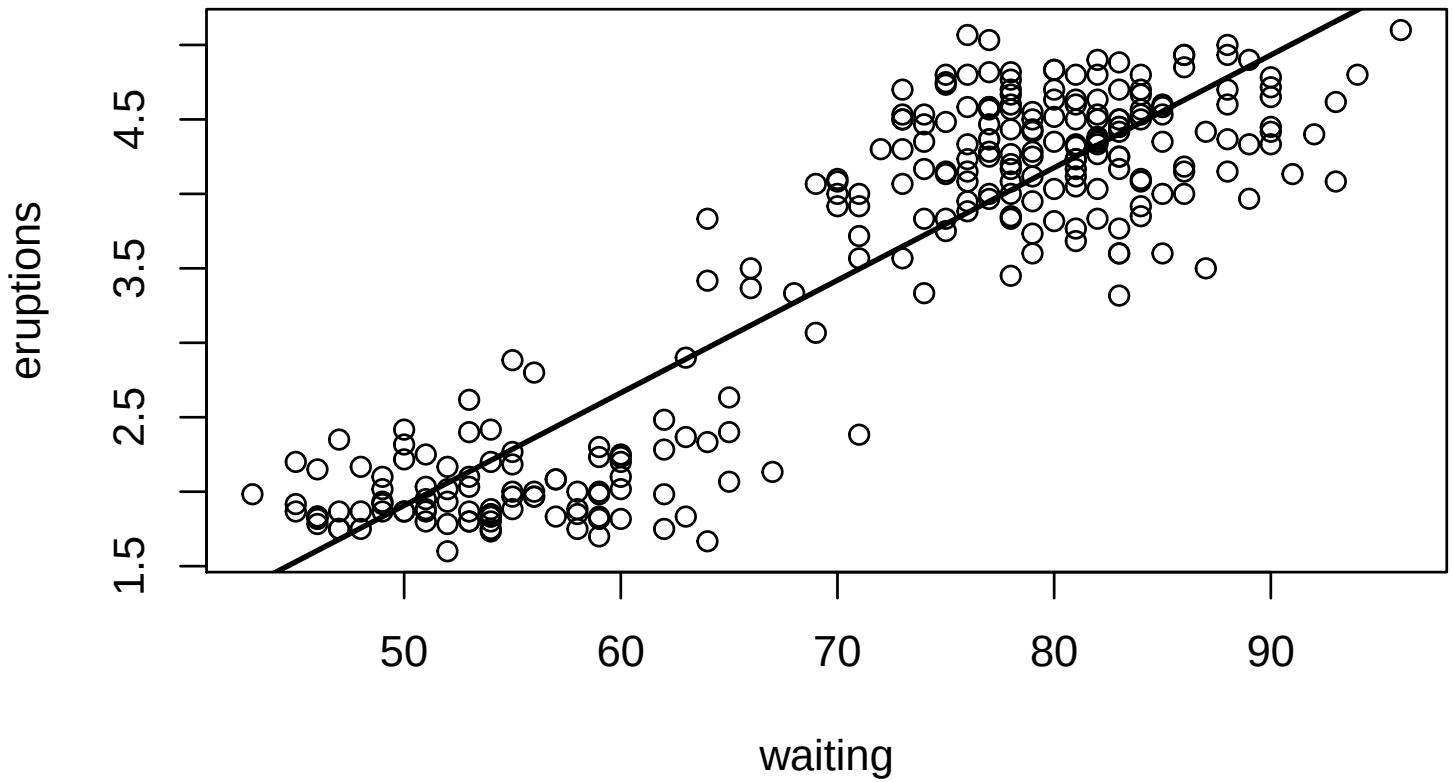
The least squares estimator

Intuitively, which (straight) line fits best to these data points?



The least squares estimator

"The one such that all points lie as close as possible to the line."



Always remember...

The simple linear model contains **three (not two!) parameters** that we should find estimates for: β_0, β_1 and σ.

The least squares estimator

- A natural strategy: estimate the regression parameters such that the corresponding linear function fits the available data points as well as possible.
- Or, the estimation method should aim to keep the errors as small as possible.
- The errors corresponding to any parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by

$$e_i(\hat{\beta}_0, \hat{\beta}_1) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i; \qquad i = 1, \ldots, n.$$

The least squares estimator

- To avoid that large positive and large negative errors can cancel each other out, we estimate the parameters β_0 and β_1 by minimizing the sum of the squared errors:

$$(\hat{\beta}_{0,\mathrm{LS}}, \hat{\beta}_{1,\mathrm{LS}}) = \mathrm{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{n} e_i^2(\beta_0, \beta_1)$$

$$= \mathrm{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

  This estimator is called the **least squares estimator**.

- The LS estimator can be solved analytically and it has some good (optimal) statistical properties that will be discussed later.

The least squares estimator

- $\sum_{i=1}^{n} e_i^2(\beta_0, \beta_1)$ is called the *objective function* or loss function of the least squares estimator.
- Differentiating this objective function $L(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$ with respect to β_0 and β_1 and setting these derivatives equal to zero, yields the normal equations

$$\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] = 0 \quad (1)$$

$$\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] x_i = 0. \quad (2)$$

The least squares estimator

- The least squares estimators $\hat{\beta}_{0,\mathrm{LS}}$ and $\hat{\beta}_{1,\mathrm{LS}}$ for the simple regression model are the solution of this system of equations.
- From (1) we find that

$$\sum_{i=1}^{n}[y_i - (\hat{\beta}_{0,\mathrm{LS}} + \hat{\beta}_{1,\mathrm{LS}} x_i)] = \sum_{i=1}^{n}(y_i) - n\hat{\beta}_{0,\mathrm{LS}} - \hat{\beta}_{1,\mathrm{LS}} \sum_{i=1}^{n}(x_i) = 0$$

$$\hat{\beta}_{0,\mathrm{LS}} = \frac{\sum_{i=1}^{n}(y_i)}{n} - \hat{\beta}_{1,\mathrm{LS}} \frac{\sum_{i=1}^{n}(x_i)}{n} = \bar{y}_n - \hat{\beta}_{1,\mathrm{LS}} \bar{x}_n. \quad (3)$$

- The second equation can be replaced by

$$\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_1} - \bar{x}_n \frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0} = 0$$

which leads to the equation

$$\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)](x_i - \bar{x}_n) = 0.$$

The least squares estimator

- By substituting result (3) into this equation, we obtain that

$$\sum_{i=1}^{n}[(y_i - \bar{y}_n) - \hat{\beta}_{1,\mathrm{LS}}(x_i - \bar{x}_n)](x_i - \bar{x}_n) = 0.$$

  Solving this equation yields

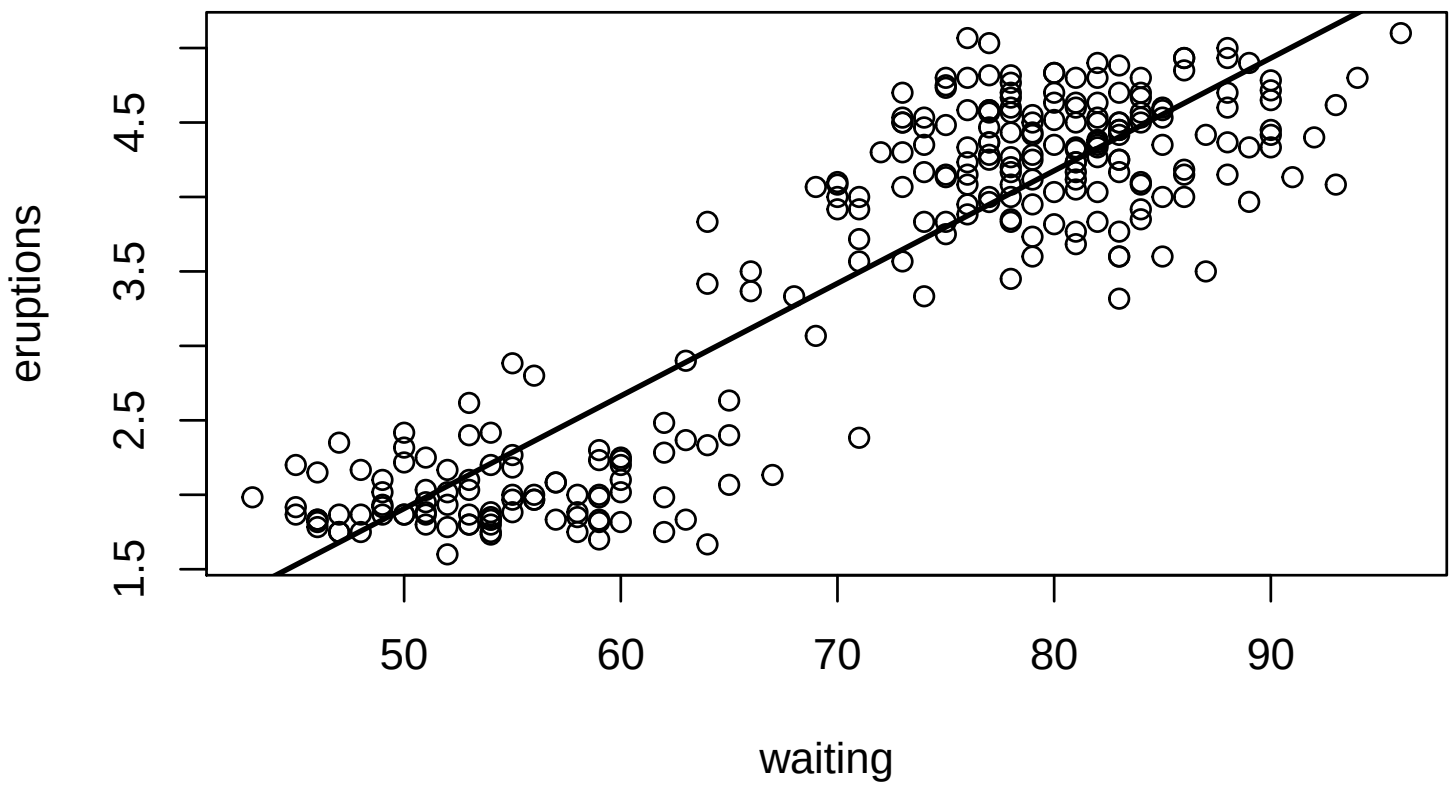
$$\hat{\beta}_{1,\mathrm{LS}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} = \frac{\mathrm{cov}(X, Y)}{s_X^2} = \mathrm{cor}(X, Y)\frac{s_Y}{s_X}, \tag{4}$$

  where s_X en s_Y are the sample standard deviations of the variables X and Y, and $\mathrm{cov}(X, Y)$ and $\mathrm{cor}(X, Y)$ are resp. the sample covariance and sample correlation between X and Y.

- Note that for the existence of the least squares estimator it is required that $s_X^2 > 0$.

The geyser example

$$\text{Average eruption time} = -1.87 + 0.076 * \text{Waiting time}.$$



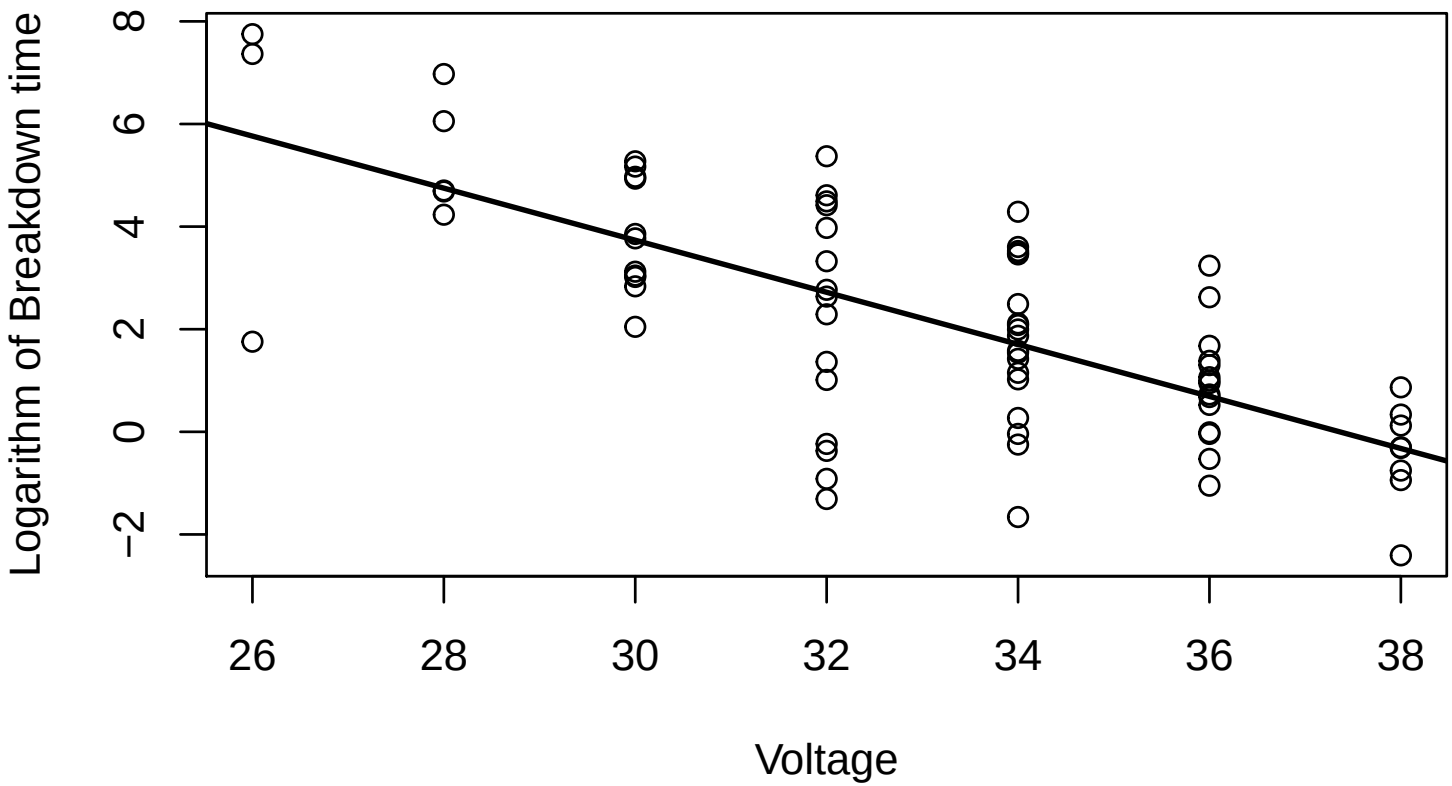
The geyser example

$$\text{Average eruption time} = -1.87 + 0.076 * \text{Waiting time}.$$

- The intercept has a negative sign which is physically not a meaningful value.
- A regression model cannot be used to reliably predict events beyond the range of information.

The fluid example

$$\text{Average } \log(\text{Breakdown time}) = 18.96 - 0.51 * \text{Voltage}.$$



Logarithmic transformation of the outcome

$$\text{Average } \log(\text{Breakdown time}) = 18.96 - 0.51 * \text{Voltage}.$$

- In the Fluid example, the interpretation of the regression coefficients in terms of the transformed response variable remains as before.
- This cannot easily be transformed into an interpretation in terms of the originally measured response variable (since $\mathrm{E}[\log(Y)] \neq \log(\mathrm{E}[Y])$).

Logarithmic transformation of the outcome

- if we assume that the error distribution is symmetric around its mean zero, the mean coincides with the median of Y at each $X = x$:

$$\mathrm{E}[Y|X = x] = \mathrm{med}[Y|X = x]$$

- Hence the regression line also models

$$\mathrm{med}[Y|X] = \beta_0 + \beta_1 X$$

- If $Y = \log(\tilde{Y})$

$$\mathrm{med}[Y|X] = \mathrm{med}[\log(\tilde{Y})|X] = \log(\mathrm{med}[\tilde{Y}|X]) = \beta_0 + \beta_1 X$$

- Or equivalently,

$$\mathrm{med}[\tilde{Y}|X] = \exp(\beta_0)\exp(\beta_1 X).$$

Logarithmic transformation of the outcome

- Moreover, we find that

$$\frac{\text{med}[\tilde{Y}|X = x + 1]}{\text{med}[\tilde{Y}|X = x]} = \exp(\beta_1)$$

- Or

$$\text{med}[\tilde{Y}|X = x + 1] = \exp(\beta_1)\text{med}[\tilde{Y}|X = x]$$

- $\exp(\beta_1)$ is the multiplicative change of the median of the measured response $\tilde{Y}$ if the predictor X increases by one unit.
- Fluid example: we find that $\exp(-0.51) = 0.60$ so with every unit increase in voltage the median breakdown point is only 60% of what it was before

Logarithmic transformation of the predictor

- Consider a linear model of the form

$$E[Y|X] = \beta_0 + \beta_1 \log(X)$$

- What can we say about Y if the original X is changed?
- We find for any $c > 0$ that

$$E[Y|X = cx] = \beta_0 + \beta_1 \log(cx) = \beta_0 + \beta_1 \log(c) + \beta_1 \log(x)$$

- such that

$$E[Y|X = cx] - E[Y|X = x] = \beta_1 \log(c)$$

- With every doubling of the value of X, the expected response Y changes by the value $\beta_1 \log(2)$.

What's next?

- Check model adequacy and measure precision of parameter estimates.
- Test whether X does have an effect on the response Y.
- Predict the expected or individual response value for given values of X.
- Such questions can be answered by confidence intervals and hypothesis tests.
- This requires assumptions about the distribution of the errors ϵ_i (often assumed to follow a normal distribution).
- Is the normal assumption sufficiently reasonable to validate the resulting inference?