Machine Learning

1. Linear Regression

Lars Schmidt-Thieme

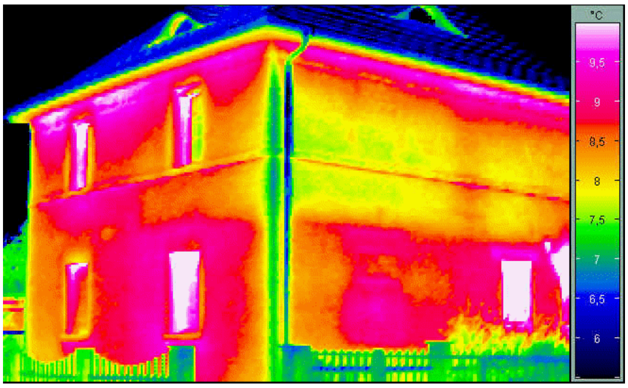
Information Systems and Machine Learning Lab (ISMLL)
University of Hildesheim, Germany
http://www.ismll.uni-hildesheim.de

Machine Learning

Lars Schmidt-Thieme, Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim, Germany,
Course on Machine Learning, winter term 2006 1/58

Example

Example: how does gas consumption depend on external temperature? (Whiteside, 1960s).



weekly measurements of

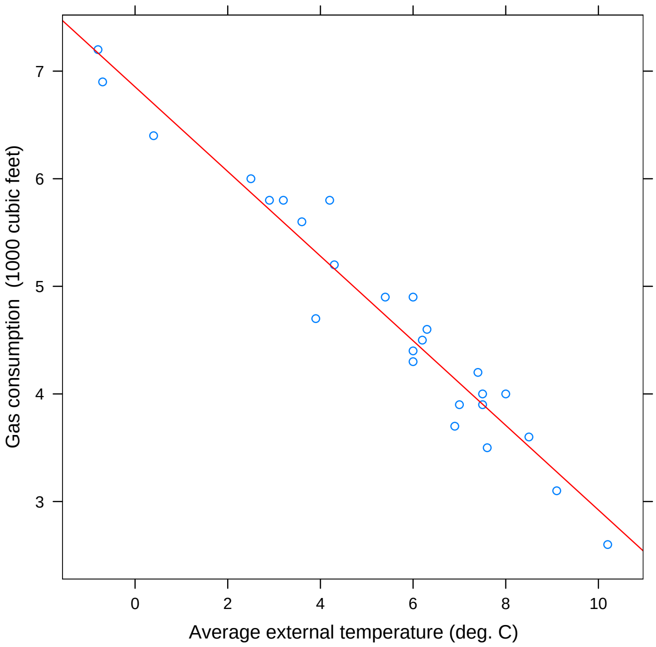
- average external temperature
- total gas consumption (in 1000 cubic feets)

A third variable encodes two heating seasons, before and after wall insulation.

How does gas consumption depend on external temperature?

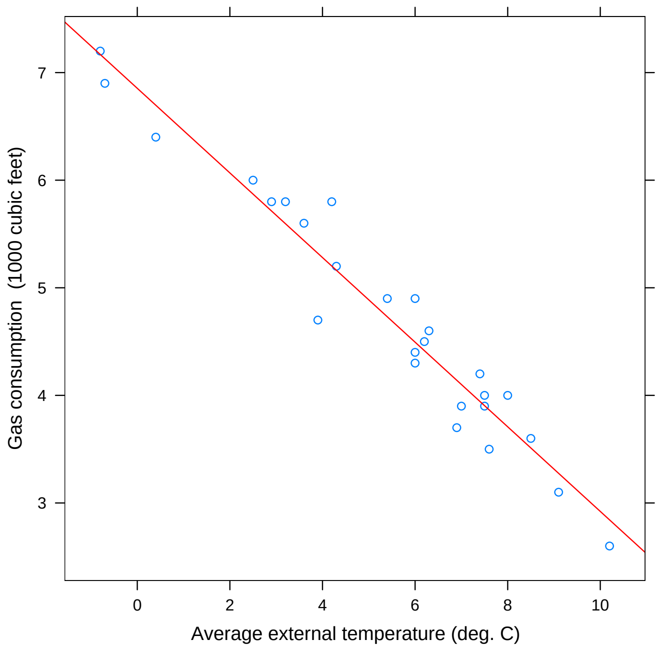
How much gas is needed for a given termperature ?

Example

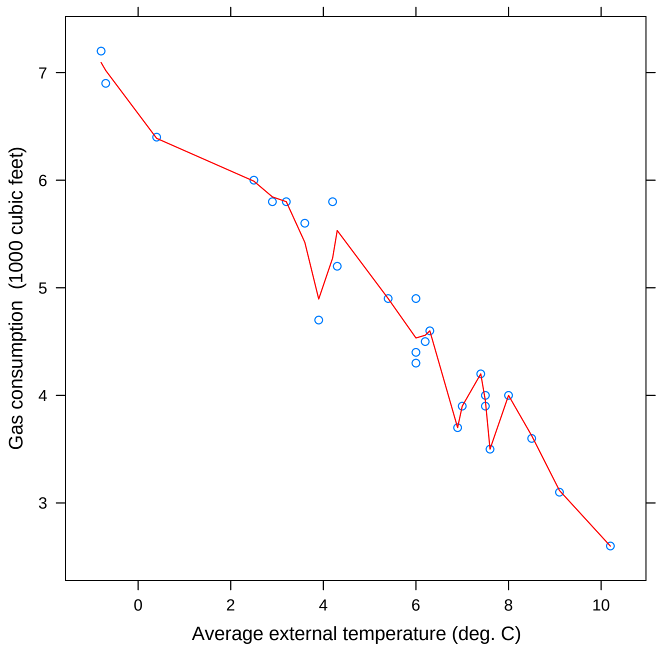


linear model

Example



linear model



more flexible model

Variable Types and Coding

The most common variable types:

numerical / interval-scaled / quantitative
where differences and quotients etc. are meaningful,
usually with domain $\mathcal{X} := \mathbb{R}$,
e.g., temperature, size, weight.

nominal / discret / categorical / qualitative / factor
where differences and quotients are not defined,
usually with a finite, enumerated domain,
e.g., $\mathcal{X} := \{\text{red}, \text{green}, \text{blue}\}$
or $\mathcal{X} := \{\text{a}, \text{b}, \text{c}, \ldots, \text{y}, \text{z}\}$.

ordinal / ordered categorical
where levels are ordered, but differences and quotients are not defined,
usually with a finite, enumerated domain,
e.g., $\mathcal{X} := \{\text{small}, \text{medium}, \text{large}\}$

Variable Types and Coding

Nominals are usually encoded as binary **dummy variables**:

$$\delta_{x_0}(X) := \begin{cases} 1, & \text{if } X = x_0, \\ 0, & \text{else} \end{cases}$$

one for each $x_0 \in X$ (but one).

Example: $\mathcal{X} := \{\text{red}, \text{green}, \text{blue}\}$

Replace

one variable X with 3 levels: $\text{red}, \text{green}, \text{blue}$

by

two variables $\delta_{\text{red}}(X)$ and $\delta_{\text{green}}(X)$ with 2 levels each: 0, 1

X	$\delta_{\text{red}}(X)$	$\delta_{\text{green}}(X)$
red	1	0
green	0	1
blue	0	0
—	1	1

The Regression Problem Formally

Let

$X_1, X_2, \dots, X_p$ be random variables called **predictors** (or **inputs**, **covariates**).
Let $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_p$ be their domains.
We write shortly

$$X := (X_1, X_2, \dots, X_p)$$

for the vector of random predictor variables and

$$\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times X_p$$

for its domain.

Y be a random variable called **target** (or **output**, **response**).
Let $\mathcal{Y}$ be its domain.

$\mathcal{D} \subseteq \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be a (multi)set of instances of the unknown joint distribution $p(X, Y)$ of predictors and target called **data**.
$\mathcal{D}$ is often written as enumeration

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

The Regression Problem Formally

The task of regression and classification is
to predict Y based on X,
i.e., to estimate

$$r(x) := E(Y \mid X = x) = \int y\, p(y|x) dx$$

based on data (called **regression function**).

If Y is numerical, the task is called **regression**.

If Y is nominal, the task is called **classification**.

Simple Linear Regression Model

Make it simple:

- the predictor X is simple, i.e., one-dimensional ($X = X_1$).

- $r(x)$ is assumed to be linear:

$$r(x) = \beta_0 + \beta_1 x$$

- assume that the variance does not depend on x:

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad E(\epsilon|x) = 0, V(\epsilon|x) = \sigma^2$$

- 3 parameters:
  β_0 **intercept** (sometimes also called bias)
  β_1 **slope**
  σ^2 **variance**

Simple Linear Regression Model

parameter estimates

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$$

fitted line

$$\hat{r}(x) := \hat{\beta}_0 + \hat{\beta}_1 x$$

predicted / fitted values

$$\hat{y}_i := \hat{r}(x_i)$$

residuals

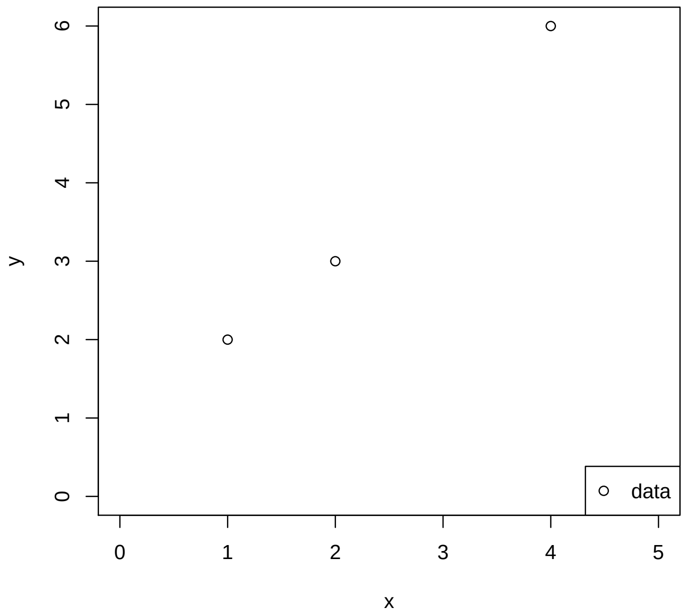
$$\hat{\epsilon}_i := y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

residual sums of squares (RSS)

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

How to estimate the parameters?

Example:
Given the data $\mathcal{D} := \{(1,2),(2,3),(4,6)\}$, predict a value for $x = 3$.

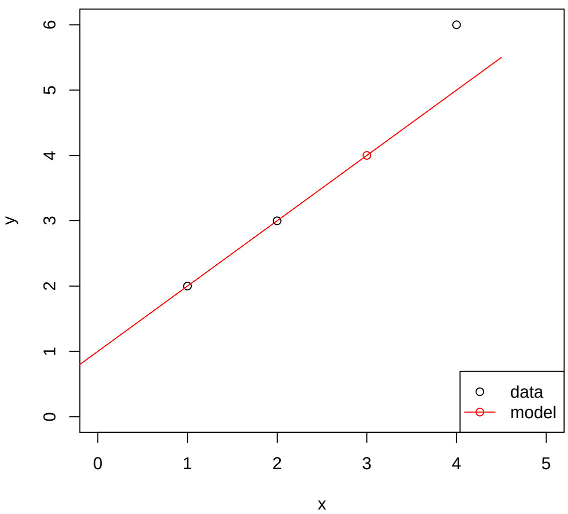


How to estimate the parameters?

Example:
Given the data $\mathcal{D} := \{(1,2),(2,3),(4,6)\}$, predict a value for $x = 3$.

Line through first two points:

$$\hat\beta_1 = \frac{y_2 - y_1}{x_2 - x_1} = 1$$
$$\hat\beta_0 = y_1 - \hat\beta_1 x_1 = 1$$



RSS:

i	y_i	$\hat y_i$	$(y_i - \hat y_i)^2$
1	2	2	0
2	3	3	0
3	6	5	1
$\sum$			1

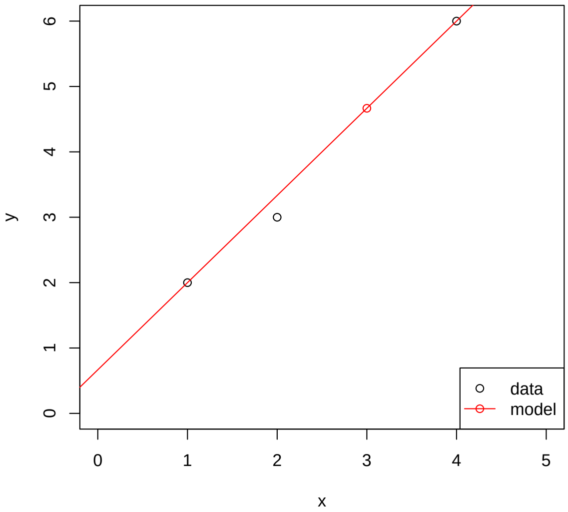
$\hat r(3) = 4$

How to estimate the parameters?

Example:
Given the data $\mathcal{D} := \{(1,2), (2,3), (4,6)\}$, predict a value for $x = 3$.

Line through first and last point:

$$\hat\beta_1 = \frac{y_3 - y_1}{x_3 - x_1} = 4/3 = 1.333$$
$$\hat\beta_0 = y_1 - \hat\beta_1 x_1 = 2/3 = 0.667$$



RSS:

i	y_i	$\hat y_i$	$(y_i - \hat y_i)^2$
1	2	2	0
2	3	3.333	0.111
3	6	6	0
$\sum$			0.111

$\hat r(3) = 4.667$

Least Squares Estimates / Definition

In principle, there are many different methods to estimate the parameters $\hat\beta_0, \hat\beta_1$ and $\hat\sigma^2$ from data — depending on the properties the solution should have.

The **least squares estimates** are those parameters that minimize

$$\text{RSS} = \sum_{i=1}^{n} \hat\epsilon_i = \sum_{i=1}^{n} (y_i - \hat y_i)^2 = \sum_{i=1}^{n} (y_i - (\hat\beta_0 + \hat\beta_1 x_i))^2$$

They can be written in closed form as follows:

$$\hat\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar x)(y_i - \bar y)}{\sum_{i=1}^n (x_i - \bar x)^2}$$
$$\hat\beta_0 = \bar y - \hat\beta_1 \bar x$$
$$\hat\sigma^2 = \frac{1}{n-2} \sum_{i=1}^{n} \epsilon_i^2$$

Least Squares Estimates / Proof

Proof (1/2):

$$\text{RSS} = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

$$\frac{\partial\,\text{RSS}}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))(-1) \overset{!}{=} 0$$

$$\implies \quad n\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \hat{\beta}_1 x_i$$

Least Squares Estimates / Proof

Proof (2/2):

$$\begin{aligned}
\text{RSS} &= \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \\
&= \sum_{i=1}^{n} (y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i)^2 \\
&= \sum_{i=1}^{n} (y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x}))^2
\end{aligned}$$

$$\frac{\partial\,\text{RSS}}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} 2(y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x}))(-1)(x_i - \bar{x}) \overset{!}{=} 0$$

$$\implies \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

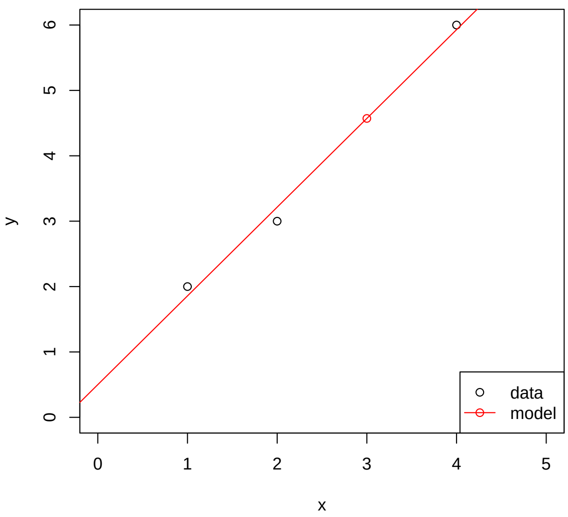
Least Squares Estimates / Example

Example:
Given the data $\mathcal{D} := \{(1,2),(2,3),(4,6)\}$, predict a value for $x = 3$.
Assume simple linear model.
$\bar{x} = 7/3$, $\bar{y} = 11/3$.

i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	$-4/3$	$-5/3$	$16/9$	$20/9$
2	$-1/3$	$-2/3$	$1/9$	$2/9$
3	$5/3$	$7/3$	$25/9$	$35/9$
$\sum$			$42/9$	$57/9$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 57/42 = 1.357$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{11}{3} - \frac{57}{42} \cdot \frac{7}{3} = \frac{63}{126} = 0.5$$



Least Squares Estimates / Example

Example:
Given the data $\mathcal{D} := \{(1,2),(2,3),(4,6)\}$, predict a value for $x = 3$.
Assume simple linear model.

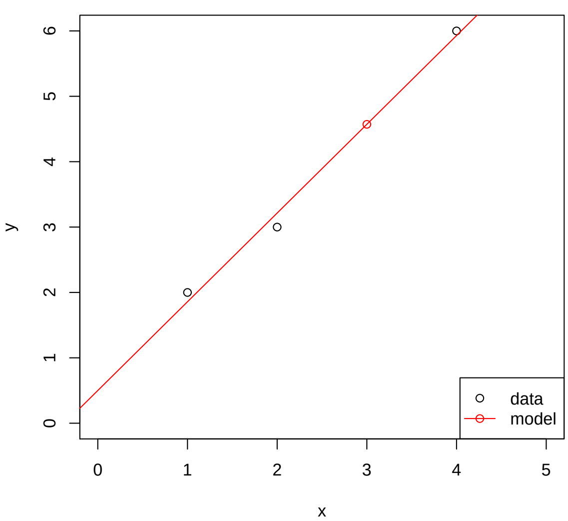
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 57/42 = 1.357$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{11}{3} - \frac{57}{42} \cdot \frac{7}{3} = \frac{63}{126} = 0.5$$

RSS:

i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	2	1.857	0.020
2	3	3.214	0.046
3	6	5.929	0.005
$\sum$			0.071

$\hat{r}(3) = 4.571$



A Generative Model

So far we assumed the model

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad E(\epsilon|x) = 0, V(\epsilon|x) = \sigma^2$$

where we required some properties of the errors, but not its exact distribution.

If we make assumptions about its distribution, e.g.,

$$\epsilon|x \sim \mathcal{N}(0, \sigma^2)$$

and thus

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 X, \sigma^2)$$

we can sample from this model.

Maximum Likelihood Estimates (MLE)

Let $\hat{p}(X, Y \mid \theta)$ be a joint probability density function for X and Y with parameters θ.

Likelihood:

$$L_{\mathcal{D}}(\theta) := \prod_{i=1}^{n} \hat{p}(x_i, y_i \mid \theta)$$

The likelihood describes the probabilty of the data.

The **maximum likelihood estimates (MLE)** are those parameters that maximize the likelihood.

Least Squares Estimates and Maximum Likelihood Estimates

Likelihood:

$$L_{\mathcal{D}}(\hat\beta_0, \hat\beta_1, \hat\sigma^2) := \prod_{i=1}^n \hat p(x_i, y_i) = \prod_{i=1}^n \hat p(y_i \,|\, x_i) p(x_i) = \prod_{i=1}^n \hat p(y_i \,|\, x_i) \prod_{i=1}^n p(x_i)$$

Conditional likelihood:

$$L_{\mathcal{D}}^{\text{cond}}(\hat\beta_0, \hat\beta_1, \hat\sigma^2) := \prod_{i=1}^n \hat p(y_i \,|\, x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\hat\sigma} e^{-\frac{(y_i-\hat y_i)^2}{2\hat\sigma^2}} = \frac{1}{\sqrt{2\pi}^n \hat\sigma^n} e^{\frac{1}{-2\hat\sigma^2}\sum_{i=1}^n (y_i-\hat y_i)^2}$$

Conditional log-likelihood:

$$\log L_{\mathcal{D}}^{\text{cond}}(\hat\beta_0, \hat\beta_1, \hat\sigma^2) \propto -n \log \hat\sigma - \frac{1}{2\hat\sigma^2} \sum_{i=1}^n (y_i - \hat y_i)^2$$

$\Longrightarrow$ if we assume normality, the maximum likelihood estimates are just the minimal least squares estimates.

Implementation Details

```
simple-regression(𝒟) :
sx := 0, sy := 0
for i = 1, ..., n do
    sx := sx + x_i
    sy := sy + y_i
od
x̄ := sx/n, ȳ := sy/n
a := 0, b := 0
for i = 1, ..., n do
    a := a + (x_i − x̄)(y_i − ȳ)
    b := b + (x_i − x̄)^2
od
β_1 := a/b
β_0 := ŷ − β_1 x̂
return (β_0, β_1)
```

Implementation Details

naive:

simple-regression($\mathcal{D}$) :
sx $:= 0$, sy $:= 0$
for $i = 1, \ldots, n$ **do**
    sx $:=$ sx $+ x_i$
    sy $:=$ sy $+ y_i$
od
$\bar{x} :=$ sx$/n$, $\bar{y} :=$ sy$/n$
$a := 0, b := 0$
for $i = 1, \ldots, n$ **do**
    $a := a + (x_i - \bar{x})(y_i - \bar{y})$
    $b := b + (x_i - \bar{x})^2$
od
$\beta_1 := a/b$
$\beta_0 := \hat{y} - \beta_1 \hat{x}$
return (β_0, β_1)

single loop:

simple-regression($\mathcal{D}$) :
sx $:= 0$, sy $:= 0$, sxx $:= 0$, syy $:= 0$, sxy $:= 0$
for $i = 1, \ldots, n$ **do**
    sx $:=$ sx $+ x_i$
    sy $:=$ sy $+ y_i$
    sxx $:=$ sxx $+ x_i^2$
    syy $:=$ syy $+ y_i^2$
    sxy $:=$ sxy $+ x_i y_i$
od
$\beta_1 := (n \cdot \text{sxy} - \text{sx} \cdot \text{sy})/(n \cdot \text{sxx} - \text{sx} \cdot \text{sx})$
$\beta_0 := (\text{sy} - \beta_1 \cdot \text{sx})/n$
return (β_0, β_1)

1. The Regression Problem

2. Simple Linear Regression

3. Multiple Regression

4. Variable Interactions

5. Model Selection

6. Case Weights

Several predictors

Several predictor variables $X_1, X_2, \ldots, X_p$:

$$\begin{aligned} Y =& \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_P X_P + \epsilon \\ =& \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon \end{aligned}$$

with $p+1$ parameters $\beta_0, \beta_1, \ldots, \beta_p$.

Linear form

Several predictor variables $X_1, X_2, \ldots, X_p$:

$$\begin{aligned} Y =& \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon \\ =& \langle \beta, X \rangle + \epsilon \end{aligned}$$

where

$$\beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad X := \begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_p \end{pmatrix},$$

Thus, the intercept is handled like any other parameter, for the artificial constant variable $X_0 \equiv 1$.

Simultaneous equations for the whole dataset

For the whole dataset $(x_1, y_1), \ldots, (x_n, y_n)$:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where

$$\mathbf{Y} := \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{pmatrix}, \quad \epsilon := \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

Least squares estimates

Least squares estimates $\hat{\beta}$ minimize

$$||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = ||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2$$

The least squares estimates $\hat{\beta}$ are computed via

$$\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{Y}$$

Proof:

$$||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2 = \langle \mathbf{Y} - \mathbf{X}\hat{\beta}, \mathbf{Y} - \mathbf{X}\hat{\beta} \rangle$$

$$\frac{\partial(\ldots)}{\partial\hat{\beta}} = 2\langle -\mathbf{X}, \mathbf{Y} - \mathbf{X}\hat{\beta} \rangle = -2(\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{X}\hat{\beta}) \stackrel{!}{=} 0$$

How to compute least squares estimates $\hat{\beta}$

Solve the $p \times p$ system of linear equations

$$X^T X \hat{\beta} = X^T Y$$

i.e., $Ax = b$ (with $A := X^T X, b = X^T Y, x = \hat{\beta}$).

There are several numerical methods available:

1. Gaussian elimination
2. Cholesky decomposition
3. QR decomposition

How to compute least squares estimates $\hat{\beta}$ / Example

Given is the following data:

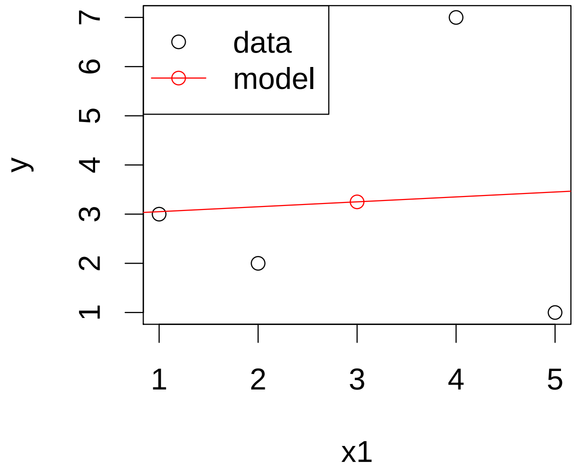
x_1	x_2	y
1	2	3
2	3	2
4	1	7
5	5	1

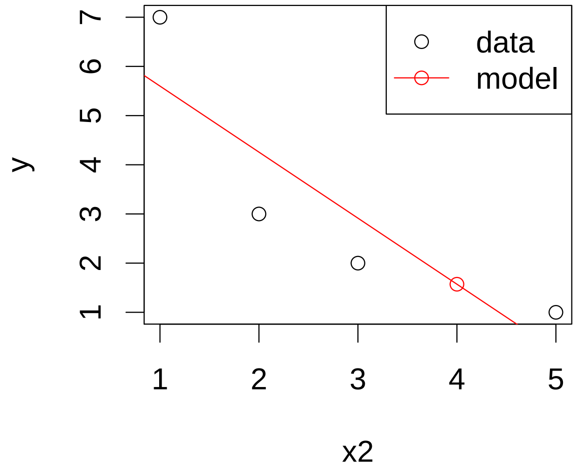
Predict a y value for $x_1 = 3, x_2 = 4$.

How to compute least squares estimates $\hat{\beta}$ / Example

$$Y = \beta_0 + \beta_1 X_1 + \epsilon = 2.95 + 0.1 X_1 + \epsilon$$

$$Y = \beta_0 + \beta_2 X_2 + \epsilon = 6.943 - 1.343 X_2 + \epsilon$$





$$\hat{y}(x_1 = 3) = 3.25$$

$$\hat{y}(x_2 = 4) = 1.571$$

How to compute least squares estimates $\hat{\beta}$ / Example

Now fit

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

to the data:

x_1	x_2	y
1	2	3
2	3	2
4	1	7
5	5	1

$$X = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 4 & 1 \\ 1 & 5 & 5 \end{pmatrix}, \quad Y = \begin{pmatrix} 3 \\ 2 \\ 7 \\ 1 \end{pmatrix}$$

$$X^T X = \begin{pmatrix} 4 & 12 & 11 \\ 12 & 46 & 37 \\ 11 & 37 & 39 \end{pmatrix}, \quad X^T Y = \begin{pmatrix} 13 \\ 40 \\ 24 \end{pmatrix}$$

How to compute least squares estimates $\hat{\beta}$ / Example

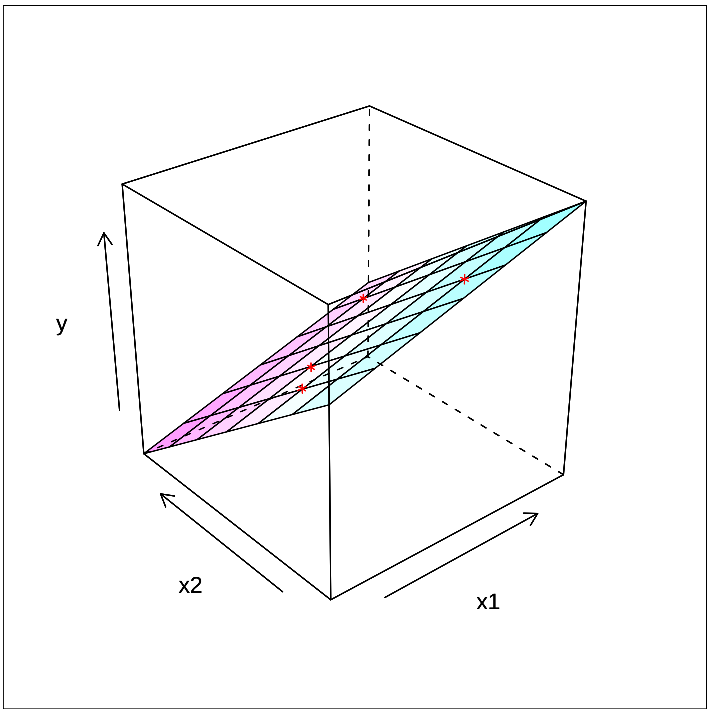
$$\left(\begin{array}{ccc|c} 4 & 12 & 11 & 13 \\ 12 & 46 & 37 & 40 \\ 11 & 37 & 39 & 24 \end{array}\right) \sim \left(\begin{array}{ccc|c} 4 & 12 & 11 & 13 \\ 0 & 10 & 4 & 1 \\ 0 & 16 & 35 & -47 \end{array}\right) \sim \left(\begin{array}{ccc|c} 4 & 12 & 11 & 13 \\ 0 & 10 & 4 & 1 \\ 0 & 0 & 143 & -243 \end{array}\right)$$

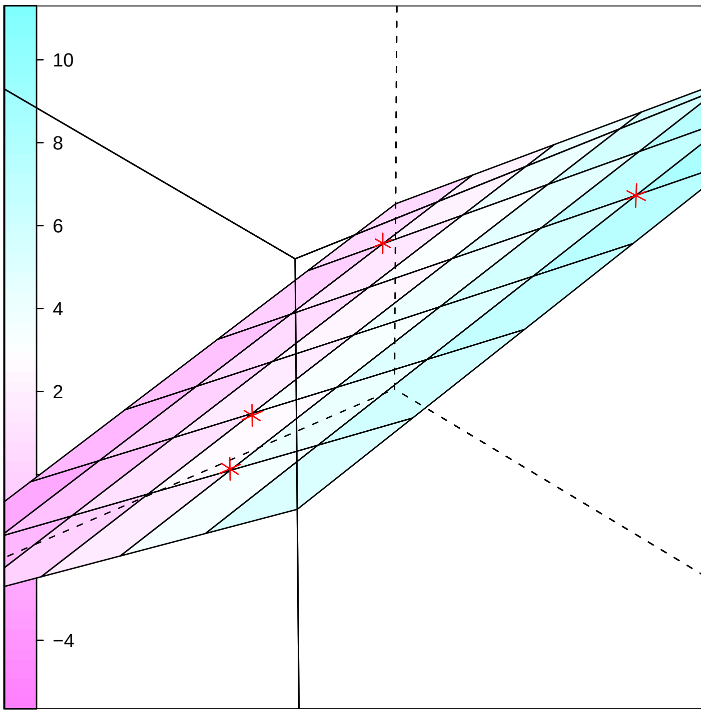
$$\sim \left(\begin{array}{ccc|c} 4 & 12 & 11 & 13 \\ 0 & 1430 & 0 & 1115 \\ 0 & 0 & 143 & -243 \end{array}\right) \sim \left(\begin{array}{ccc|c} 286 & 0 & 0 & 1597 \\ 0 & 1430 & 0 & 1115 \\ 0 & 0 & 143 & -243 \end{array}\right)$$

i.e.,

$$\hat{\beta} = \begin{pmatrix} 1597/286 \\ 1115/1430 \\ -243/143 \end{pmatrix} \approx \begin{pmatrix} 5.583 \\ 0.779 \\ -1.699 \end{pmatrix}$$

How to compute least squares estimates $\hat{\beta}$ / Example



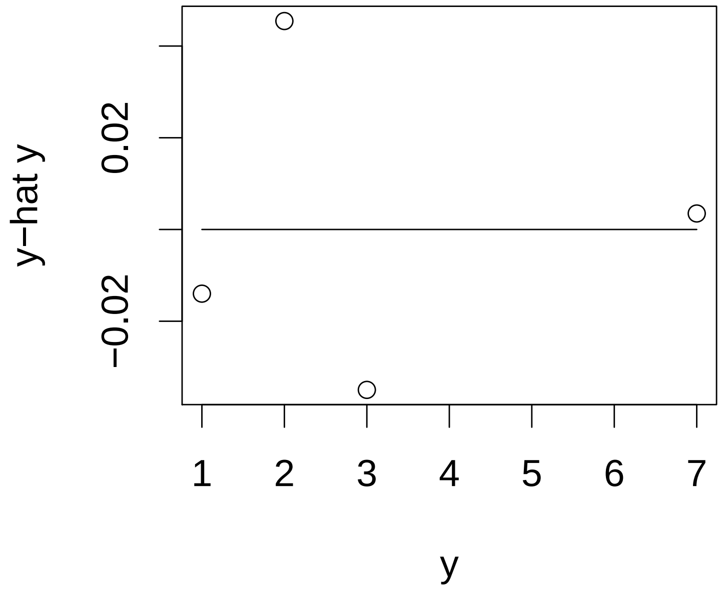


How to compute least squares estimates $\hat{\beta}$ / Example

To visually assess the model fit, a plot

$$\text{residuals } \hat{\epsilon} = y - \hat{y} \text{ vs. true values } y$$

can be plotted:



The Normal Distribution (also Gaussian)

written as:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

with parameters:
μ mean,
σ standard deviance.
probability density function (pdf):

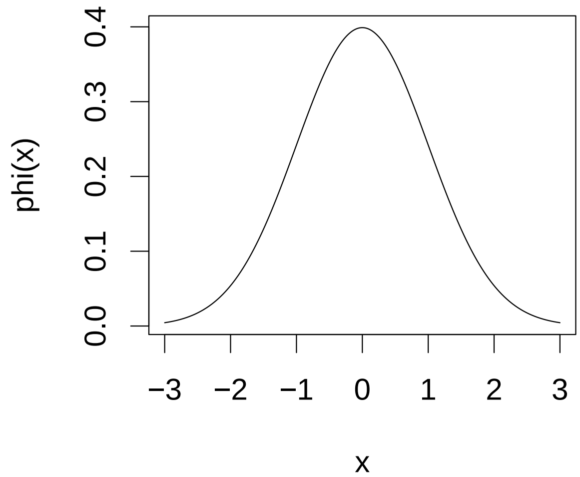
$$\phi(x) := \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

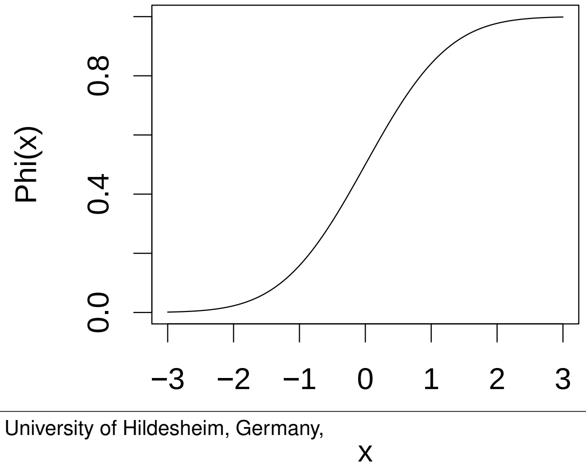
cummulative density function (cdf):

$$\Phi(x) := \int_{-\infty}^{x} \phi(x) dx$$

Φ^{-1} is called **quantile function.**

Φ and Φ^{-1} have no analytical form, but have to computed numerically.





The t Distribution

written as:

$$X \sim t_p$$

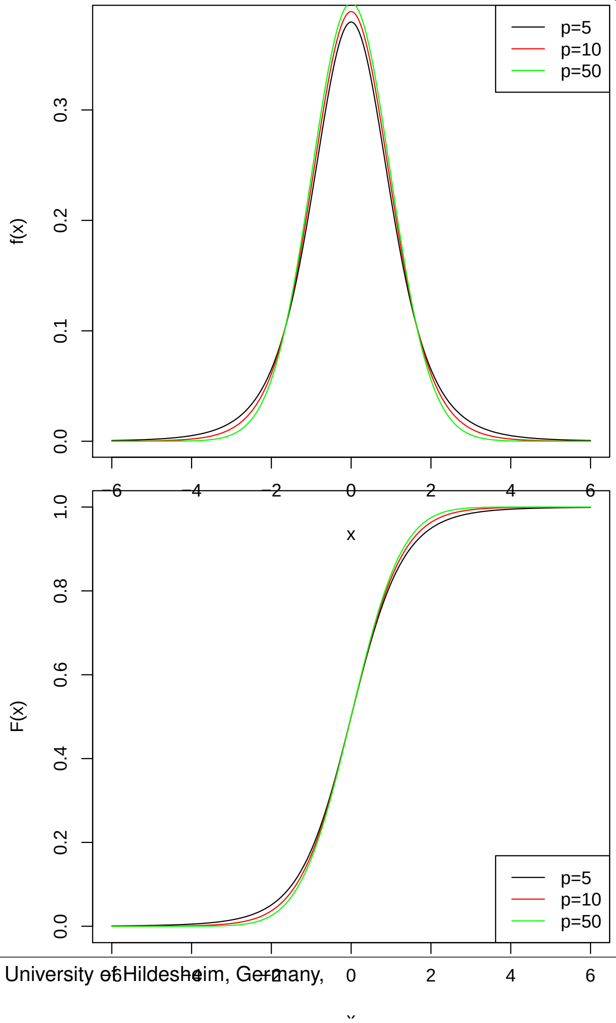
with parameter:

p degrees of freedom.

probability density function (pdf):

$$p(x) := \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})}(1 + \frac{x^2}{p})^{-\frac{p+1}{2}}$$

$$t_p \overset{p \to \infty}{\longrightarrow} \mathcal{N}(0,1)$$



The χ^2 Distribution

written as:

$$X \sim \chi^2_p$$

with parameter:

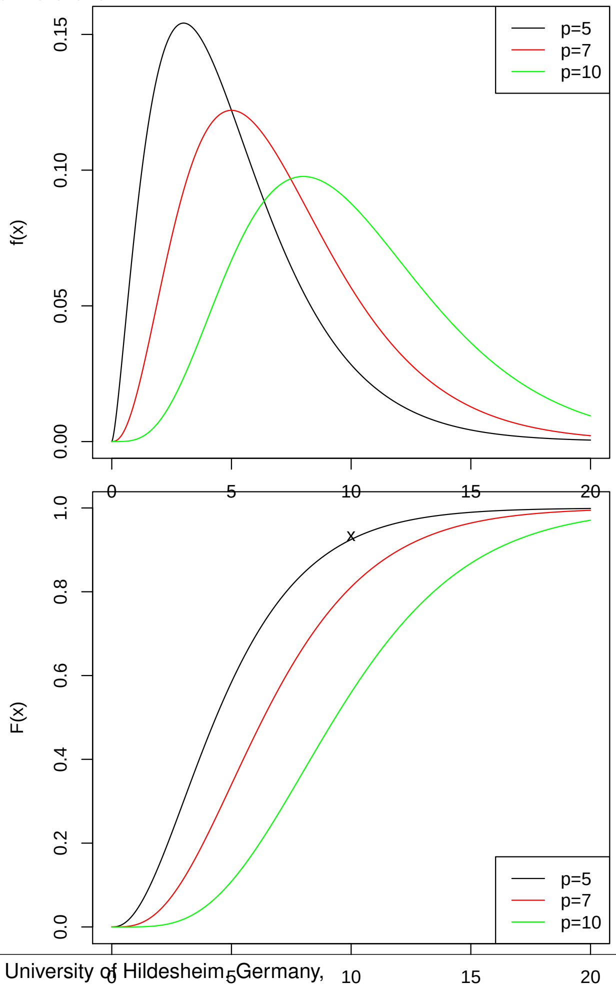
p degrees of freedom.

probability density function (pdf):

$$p(x) := \frac{1}{\Gamma(p/2)2^{p/2}} x^{\frac{p}{2}-1} e^{-\frac{x}{2}}$$

If $X_1, \ldots, X_p \sim \mathcal{N}(0,1)$, then

$$Y := \sum_{i=1}^{p} X_i^2 \sim \chi^2_p$$



Parameter Variance

$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is an unbiased estimator for β (i.e., $E(\hat{\beta}) = \beta$).
Its variance is

$$V(\hat{\beta}) = (X^T X)^{-1}\sigma^2$$

proof:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \epsilon) = \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$$

As $E(\epsilon) = 0$: $E(\hat{\beta}) = \beta$

$$\begin{aligned} V(\hat{\beta}) &= E((\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))^T) \\ &= E((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\epsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \end{aligned}$$

Parameter Variance

An unbiased estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y - \hat{y})^2$$

If $\epsilon \sim \mathcal{N}(0, \sigma^2)$, then

$$\hat{\beta} \sim \mathcal{N}(\beta, (X^T X)^{-1}\sigma^2)$$

Furthermore

$$(n-p)\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-p}$$

Parameter Variance / Standardized coefficient

standardized coefficient (“z-score”):

$$z_i := \frac{\hat{\beta}_i}{\widehat{\text{se}}(\hat{\beta}_i)}, \quad \text{with } \widehat{\text{se}}^2(\hat{\beta}_i) \text{ the } i\text{-th diagonal element of } (X^T X)^{-1}\hat{\sigma}^2$$

z_i would be $z_i \sim \mathcal{N}(0,1)$ if σ is known.
With estimated $\hat{\sigma}$ it is $z_i \sim t_{n-p}$.

The Wald test for $H_0 : \beta_i = 0$ with size α is:

$$\text{reject } H_0 \text{ if } |z_i| = |\frac{\hat{\beta}_i}{\widehat{\text{se}}(\hat{\beta}_i)}| > F^{-1}_{t_{n-p}}(1 - \frac{\alpha}{2})$$

i.e., its p-value is

$$p\text{-value}(H_0 : \beta_i = 0) = 1 - F_{t_{n-p}}(|z_i|) = 2(1 - F_{t_{n-p}}(|\frac{\hat{\beta}_i}{\widehat{\text{se}}(\hat{\beta}_i)}|))$$

and small p-values such as 0.01 and 0.05 are good.

Parameter Variance / Confidence interval

The $1 - \alpha$ confidence interval for β_i:

$$\beta_i \pm F^{-1}_{t_{n-p}}(1 - \frac{\alpha}{2})\widehat{\text{se}}(\hat{\beta}_i)$$

For large n, $F_{t_{n-p}}$ converges to the standard normal cdf Φ.

As $\Phi^{-1}(1 - \frac{0.05}{2}) \approx 1.95996 \approx 2$, the rule-of-thumb for a 5% confidence interval is

$$\beta_i \pm 2\widehat{\text{se}}(\hat{\beta}_i)$$

Parameter Variance / Example

We have already fitted

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$
$$= 5.583 + 0.779 X_1 - 1.699 X_2$$

to the data:

x_1	x_2	y	$\hat{y}$	$\hat{\epsilon}^2 = (y - \hat{y})^2$
1	2	3	2.965	0.00122
2	3	2	2.045	0.00207
4	1	7	7.003	0.0000122
5	5	1	0.986	0.000196
RSS				0.00350

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \frac{1}{4-3} 0.00350 = 0.00350$$

$$(X^T X)^{-1} \hat{\sigma}^2 = \begin{pmatrix} 0.00520 & -0.00075 & -0.00076 \\ -0.00075 & 0.00043 & -0.00020 \\ -0.00076 & -0.00020 & 0.00049 \end{pmatrix}$$

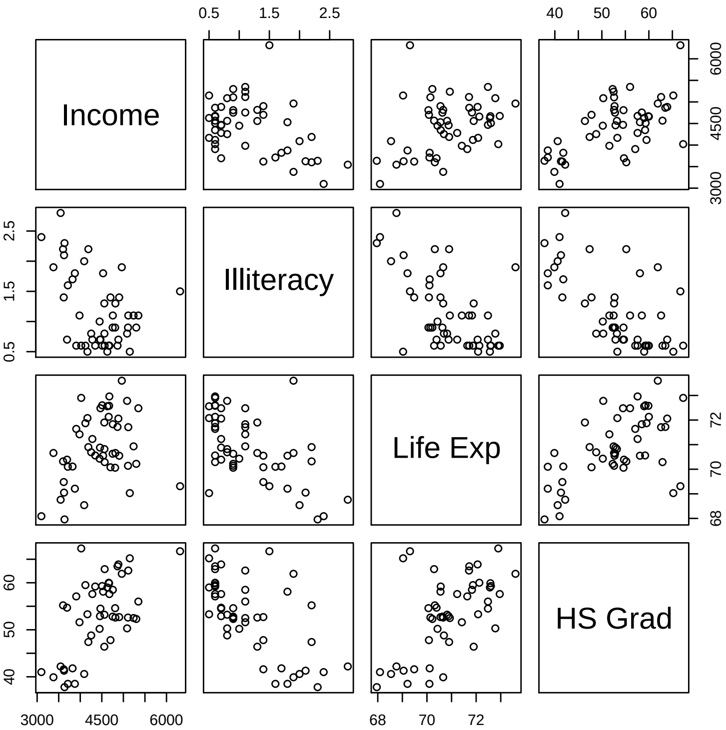
covariate	$\hat{\beta}_i$	$\widehat{\text{se}}(\hat{\beta}_i)$	z-score	p-value
(intercept)	5.583	0.0721	77.5	0.0082
X_1	0.779	0.0207	37.7	0.0169
X_2	−1.699	0.0221	−76.8	0.0083

Parameter Variance / Example 2

Example: sociographic data of the 50 US states in 1977.

state dataset:

- income (per capita, 1974),
- illiteracy (percent of population, 1970),
- life expectancy (in years, 1969–71),
- percent high-school graduates (1970).
- population (July 1, 1975)
- murder rate per 100,000 population (1976)
- mean number of days with minimum temperature below freezing (1931–1960) in capital or large city
- land area in square miles



Parameter Variance / Example 2

$$\text{Murder} = \beta_0 + \beta_1 \text{Population} + \beta_2 \text{Income} + \beta_3 \text{Illiteracy} + \beta_4 \text{LifeExp} + \beta_5 \text{HSGrad} + \beta_6 \text{Frost} + \beta_7 \text{Area}$$

$n = 50$ states, $p = 8$ parameters, $n - p = 42$ degrees of freedom.

Least squares estimators:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.222e+02  1.789e+01   6.831 2.54e-08 ***
Population   1.880e-04  6.474e-05   2.905  0.00584 **
Income      -1.592e-04  5.725e-04  -0.278  0.78232
Illiteracy   1.373e+00  8.322e-01   1.650  0.10641
`Life Exp`  -1.655e+00  2.562e-01  -6.459 8.68e-08 ***
`HS Grad`    3.234e-02  5.725e-02   0.565  0.57519
Frost       -1.288e-02  7.392e-03  -1.743  0.08867 .
Area         5.967e-06  3.801e-06   1.570  0.12391
```

Need for higher orders

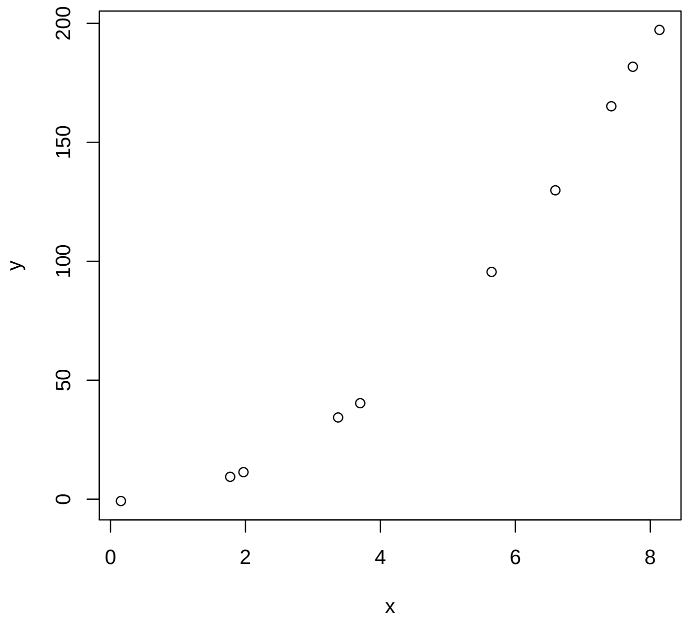
Assume a target variable does not depend linearly on a predictor variable, but say quadratic.

Example: way length vs. duration of a moving object with constant acceleration a.

$$s(t) = \frac{1}{2}at^2 + \epsilon$$

Can we catch such a dependency?

Can we catch it with a linear model?



Need for general transformations

To describe many phenomena, even more complex functions of the input variables are needed.

Example: the number of cells n vs. duration of growth t:

$$n = \beta e^{\alpha t} + \epsilon$$

n does not depend on t directly, but on $e^{\alpha t}$ (with a known α).

Need for variable interactions

In a linear model with two predictors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Y depends on both, X_1 and X_2.

But changes in X_1 will affect Y the same way, regardless of X_2.

There are problems where X_2 mediates or influences the way X_1 affects Y, e.g. : the way length s of a moving object vs. its constant velocity v and duraction t:

$$s = vt + \epsilon$$

Then an additional $1s$ duration will increase the way length not in a uniform way (regardless of the velocity), but a little for small velocities and a lot for large velocities.

v and t are said to interact: y does not depend only on each predictor separately, but also on their product.

Derived variables

All these cases can be handled by looking at **derived variables**, i.e., instead of

$$\begin{aligned} Y =&\beta_0 + \beta_1 X_1^2 + \epsilon \\ Y =&\beta_0 + \beta_1 e^{\alpha X_1} + \epsilon \\ Y =&\beta_0 + \beta_1 X_1 \cdot X_2 + \epsilon \end{aligned}$$

one looks at

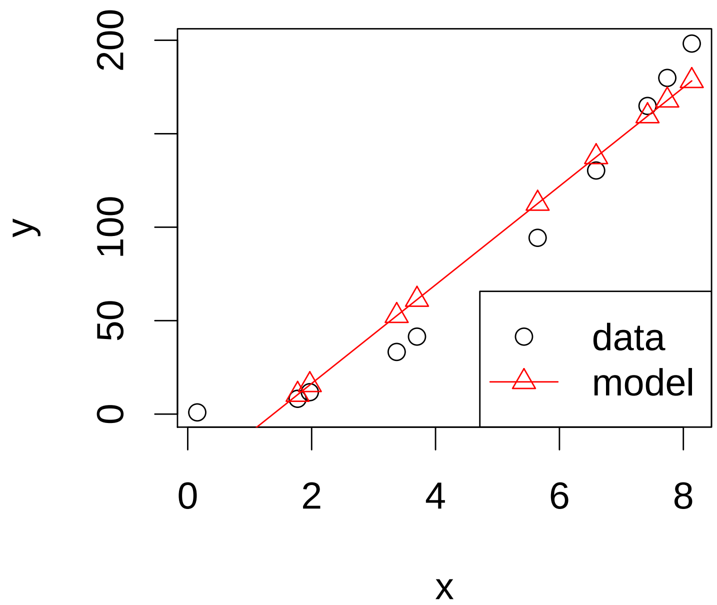
$$Y =\beta_0 + \beta_1 X_1' + \epsilon$$

with

$$\begin{aligned} X_1' :=&X_1^2 \\ X_1' :=&e^{\alpha X_1} \\ X_1' :=&X_1 \cdot X_2 \end{aligned}$$

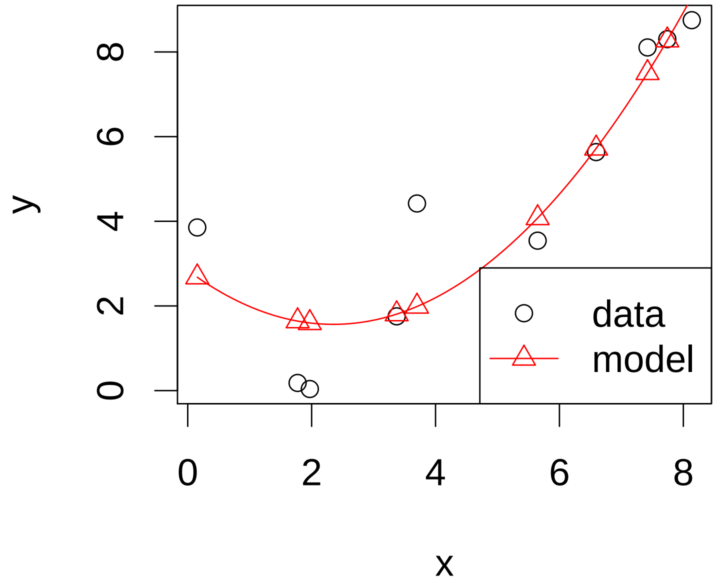
Derived variables are computed before the fitting process and taken into account either additional to the original variables or instead of.

Underfitting



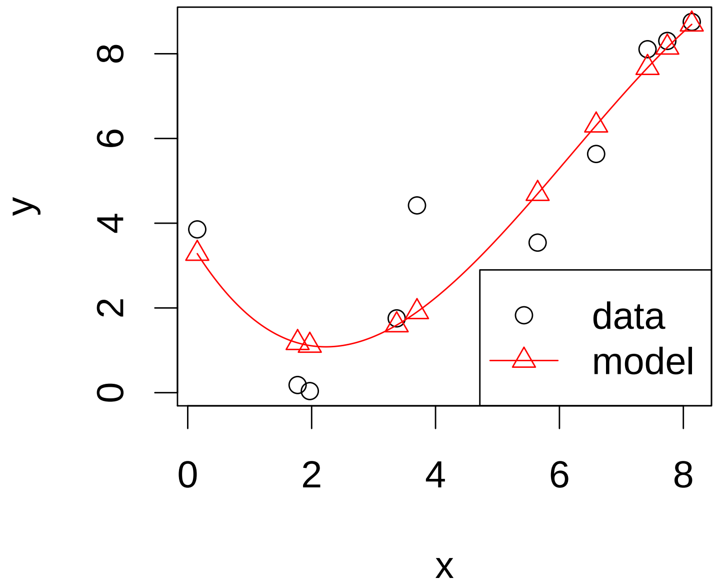
If a model does not well explain the data,
e.g., if the true model is quadratic, but we try to fit a linear model,
one says, the model **underfits**.

Overfitting / Fitting Polynomials of High Degree



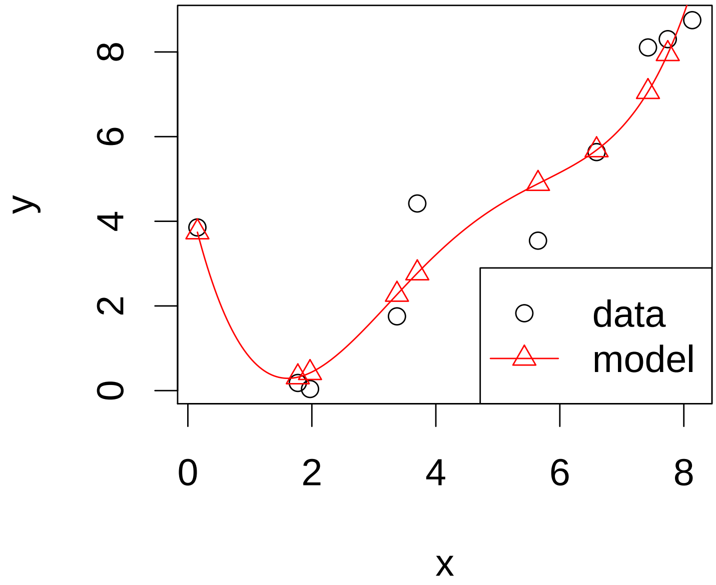
Machine Learning / 5. Model Selection

Overfitting / Fitting Polynomials of High Degree



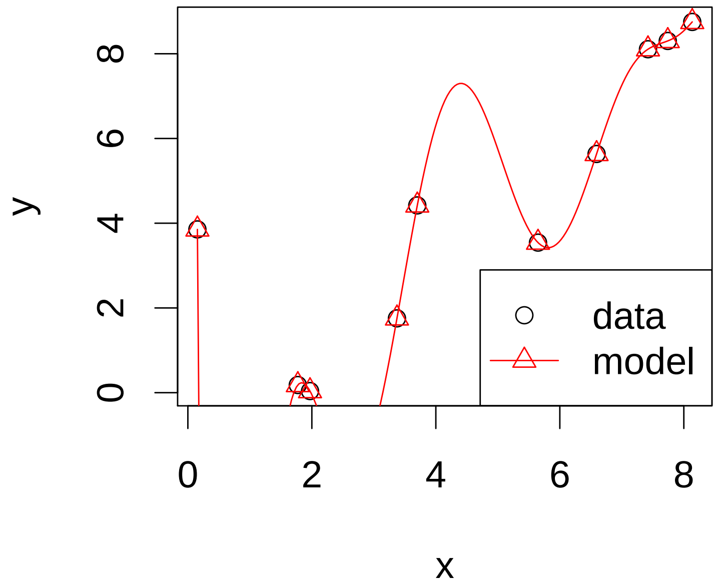
Lars Schmidt-Thieme, Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim, Germany,
Course on Machine Learning, winter term 2006
48/58

Overfitting / Fitting Polynomials of High Degree



Machine Learning / 5. Model Selection

Overfitting / Fitting Polynomials of High Degree



Lars Schmidt-Thieme, Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim, Germany,
Course on Machine Learning, winter term 2006
48/58

Overfitting / Fitting Polynomials of High Degree

If to data

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

consisting of n points we fit

$$X = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{n-1} X_{n-1}$$

i.e., a polynomial with degree $n-1$, then this results in an interpolation of the data points
(if there are no repeated measurements, i.e., points with the same X_1.)

As the polynomial

$$r(X) = \sum_{i=1}^{n} y_i \prod_{j \neq i} \frac{X - x_j}{x_i - x_j}$$

is of this type, and has minimal RSS $= 0$.

Model Selection Measures

Model selection means: we have a set of models, e.g.,

$$Y = \sum_{i=0}^{p-1} \beta_i X_i$$

indexed by p (i.e., one model for each value of p),
make a choice which model **describes** the data best.

If we just look at **fit measures** such as RSS, then the larger p the better the fit
as the model with p parameters can be **reparametrized** in a model with $p' > p$ parameters by setting

$$\beta'_i = \begin{cases} \beta_i, & \text{for } i \leq p \\ 0, & \text{for } i > p \end{cases}$$

Model Selection Measures

One uses **model selection measures** of type

$$\text{model selection measure} = \text{lack of fit} + \text{complexity}$$

The smaller the lack of fit, the better the model.

The smaller the complexity, the simpler and thus better the model.

The model selection measure tries to find a trade-off between fit and complexity.

Model Selection Measures

Akaike Information Criterion (AIC): (maximize)

$$\text{AIC} = \log L - p$$

Bayes Information Criterion (BIC) /
Bayes-Schwarz Information Criterion: (maximize)

$$\text{BIC} = \log L - \frac{p}{2} \log n$$

Shrinkage

Model selection operates by

- fitting models for a set of models with varying complexity and then picking the "best one" ex post,
- omitting some parameters completely (i.e., forcing them to be 0)

shrinkage operates by

- including a penalty term directly in the model equation and
- favoring small parameter values in general.

Shrinkage / Ridge Regression

Ridge regression: minimize

$$\begin{aligned} \text{RSS}_\lambda(\hat\beta) =& \text{RSS}(\hat\beta) + \lambda \sum_{i=1}^p \hat\beta^2 \\ =& \langle \mathbf{y} - \mathbf{X}\hat\beta, \mathbf{y} - \mathbf{X}\hat\beta \rangle + \lambda \sum_{i=1}^p \hat\beta^2 \\ \Rightarrow \hat\beta =& (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

with $\lambda \geq 0$ a complexity parameter.

As

- solutions of ridge regression are not equivariant under scaling of the predictors, and as
- it does not make sense to include a constraint for the parameter of the intercept

data is normalized before ridge regression:

$$x'_{i,j} := \frac{x_{i,j} - \bar{x}_{.,j}}{\hat\sigma(x_{.,j})}$$

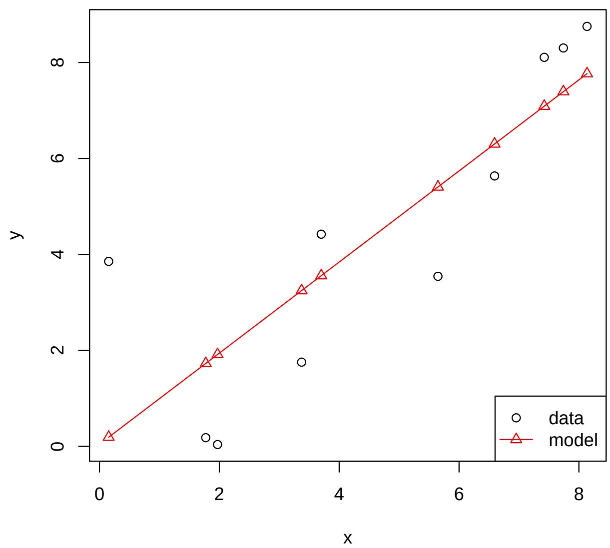
Cases of Different Importance

Sometimes different cases are of different importance, e.g., if their measurements are of different accurracy or reliability.

Example: assume the left most point is known to be measured with lower reliability.

Thus, the model does not need to fit to this point as it needs to do to the other points.

I.e., residuals of this point should get lower weight than the others.



Case Weights

In such situations, each case (x_i, y_i) is assigned a **case weight** $w_i \geq 0$:

- the higher the weight, the more important the case.
- cases with weight 0 should be treated as if they have been discarded from the data set.

Case weights can be managed as an additional pseudo-variable w in applications.

Weighted Least Squares Estimates

Formally, one tries to minimize the **weighted residual sum of squares**

$$\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 = ||\mathbf{W}^{\frac{1}{2}}(\mathbf{y} - \hat{\mathbf{y}})||^2$$

with

$$\mathbf{W} := \begin{pmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_n \end{pmatrix}$$

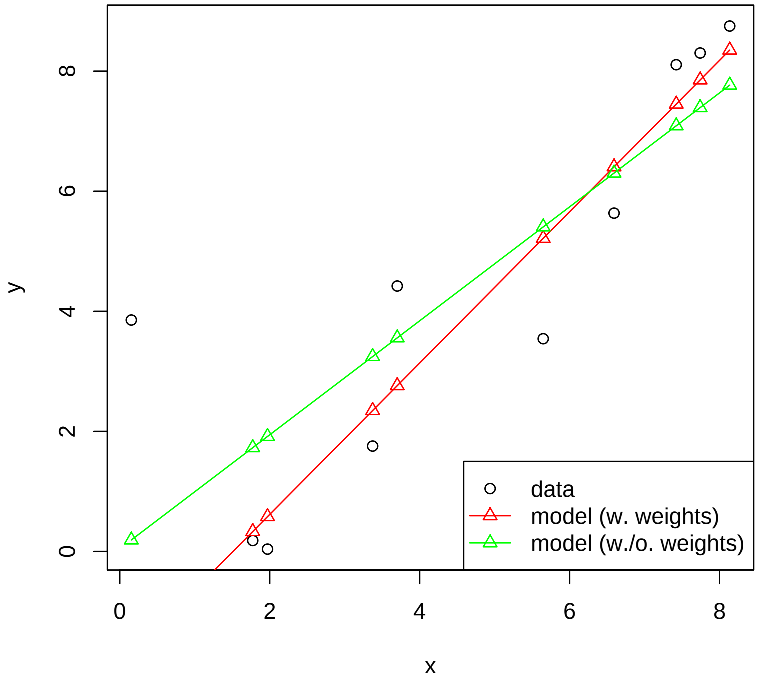
The same argument as for the unweighted case results in the **weighted least squares estimates**

$$\mathbf{X}^T \mathbf{W} \mathbf{X} \hat{\beta} = \mathbf{X}^T \mathbf{W} \mathbf{y}$$

Weighted Least Squares Estimates / Example

Do downweight the left most point, we assign case weights as follows:

w	x	y
1	5.65	3.54
1	3.37	1.75
1	1.97	0.04
1	3.70	4.42
0.1	0.15	3.85
1	8.14	8.75
1	7.42	8.11
1	6.59	5.64
1	1.77	0.18
1	7.74	8.30



Summary

- For regression, **linear models** of type $Y = \langle X, \beta \rangle + \epsilon$ can be used to predict a quantitative Y based on several (quantitative) X.
- The **ordinary least squares estimates (OLS)** are the parameters with minimal residual sum of squares (RSS). They coincide with the **maximum likelihood estimates (MLE)**.
- OLS estimates can be computed by solving the **system of linear equations** $\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{Y}$.
- The **variance of the OLS estimates** can be computed likewise ($(\mathbf{X}^T\mathbf{X})^{-1}\hat{\sigma}^2$).
- For deciding about inclusion of predictors as well as of powers and interactions of predictors in a model, **model selection measures** (AIC, BIC) and different search strategies such as forward and backward search are available.