Linear Regression

Prof. Dr. Josef F. Bürgler

Studiengang Informatik
Hochschule Luzern, Informatik

I.BA_ML

Introduction

Topic: Linear regression

Goals: You understand how to estimate one dependent feature based on the knowledge of one or more independent features. You are able to judge, how accurate this estimate is.

Results: Using linear regression You are able to estimate the price of a house, based on various influencing features like number of (bath-) rooms, number of floors, etc.

Further steps: We'll start with a simple example with one independent variable and then proceed to more complex examples.

Content

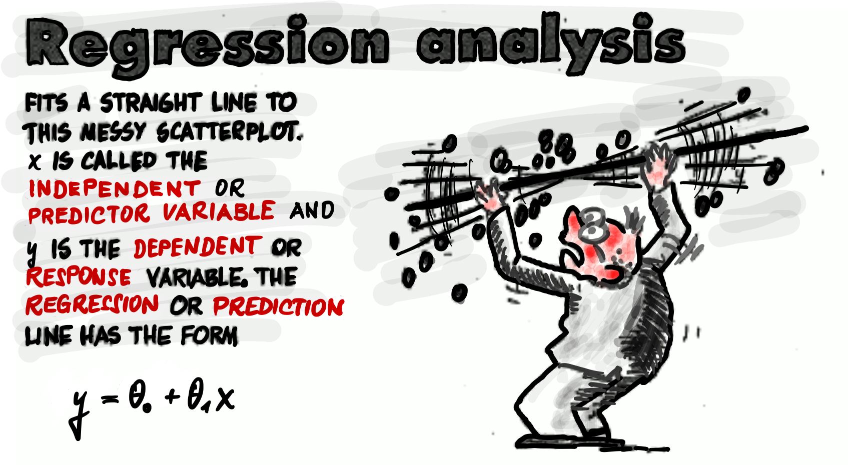
Übersicht

Motivation

Examples

- How does the blood pressure p of a person depend on the age a of this person. What's the relationship between p and a? Can we find a function f, such that $p = f(a)$?
- How does the body mass m of a male person depend on its height h? Is there a linear dependence? Can we find a function f, such that $m = f(h)$?
- Driving speed v is related to the gas mileage m, i.e. as driving speed increases, we would expect gas mileage to decrease.
- Does the price p of a house linearly depend on the number of (bath-) rooms x_1, floorsize x_2, number of stores x_3, location x_4, etc.? Can we find a function f, such that $p = f(x_1, x_2, x_3, x_4, \dots)$?

Motivation (cont.)

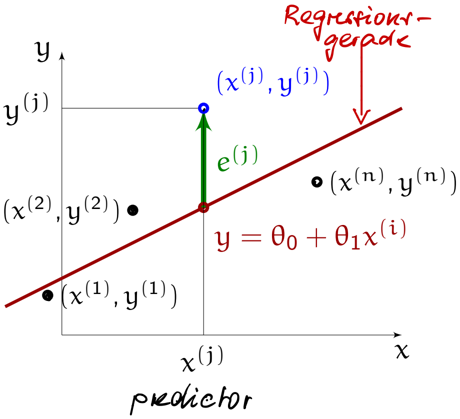


Regression and Machine Learning

- Regression is a very powerful technique for prediction in Machine Learning.
- Regression is a supervised learning type algorithm.
- The **prediction of continuous outcomes** (target values), based on a number of predictor (explanatory) variables is called **regression analysis**.
- Linear Regression assumes a linear relation between predictor variables and target, response or estimated variables.
- Simple linear regression models have only one predictor.
- Regression models attempt to minimize the distance measured vertically between the observation point and the model line (or curve).
- The length of the line segment is called residual, modeling error, or simply error.
- The negative and positive errors should each cancel out. We want zero overall error. Many lines will satisfy this criterion; but we want the best line!
- Regression analysis can also be applied to problems where the dependence is nonlinear.

The best fitting line

Let's assume we are given the n data points $(x^{(1)}, y^{(1)})$, $(x^{(2)}, y^{(2)})$, ..., $(x^{(n)}, y^{(n)})$. Let's assume the first variable corresponds to a sample of the independent random variable X and the second to the corresponding sample of the dependent random variable Y. We want to know how (if at all) Y depends on X.



Let's determine a straight line, the **hypothesis** (or the model), of the form

$$y = h_{\theta_0,\theta_1}(x) \; = \; \theta_0 + \theta_1 x$$

Each sample point $(x^{(j)}, y^{(j)})$ has the vertical distance from the straight line, i.e. the **error** (also called **residual**)

$$e^{(j)} = y^{(j)} - \underbrace{\left(\theta_0 + \theta_1 x^{(j)}\right)}_{y(x^{(j)})}, \; j = 1, \ldots, n$$

The best fitting line — the least squares method

Which line fits the sample points best? How do we have to choose the **parameters** of the model, θ_0 and θ_1? Idea: let's minimize the sum of the squares of the errors, i.e. let's minimize the **cost function**

$$J(\theta_0, \theta_1) = \frac{1}{2n} \sum_{j=1}^{n} \left(e^{(j)}\right)^2 = \frac{1}{2n} \sum_{j=1}^{n} \left[y^{(j)} - \overbrace{h_{\theta_0,\theta_1}(x^{(j)})}^{\theta_0 + \theta_1 x^{(j)}}\right]^2$$

From (multivariable) calculus we know, that a necessary condition for $J(\theta_0, \theta_1)$ to be minimal is that the gradient of J (with respect to the parameters θ_0 and θ_1) vanishes, i.e.

$$\frac{\partial J}{\partial \theta_0} = 0 \quad \text{and} \quad \frac{\partial J}{\partial \theta_1} = 0$$

It can be shown, that this leads to the following regression line formula

$$y - \overline{y} = \theta_1 (x - \overline{x})$$

where $\overline{x}$ and $\overline{y}$ are the well known means of the x- and y-values of our sample and θ_1 is the **regression coefficient** of the sample, given by

$$\theta_1 = \frac{S_{xy}}{S_{xx}}. \quad \text{Furthermore} \quad \theta_0 = \overline{y} - \theta_1 \overline{x}.$$

The best fitting line (cont.)

The **regression coefficient** θ_1 can be computed using the following sums:

$$S_{xy} = \sum_{j=1}^{n} (x^{(j)} - \overline{x})(y^{(j)} - \overline{y}), \quad \text{and} \quad S_{xx} = \sum_{j=1}^{n} (x^{(j)} - \overline{x})^2.$$

Example (Pressure dependence of volume decrease of leather)

The following pairs of numbers give the decrese of volume y (in %) of leather under the pressure x in (MPa): $(4, 2.3)$, $(6, 4.1)$, $(8, 5.7)$ and $(10, 6.9)$.

Compute the best fitting line, i.e. the regression line.

Solution: $\overline{x} = 7$, $\overline{y} = \frac{19}{4} = 4.75$, $S_{xx} = 20$, $S_{xy} = 15.4$, $\theta_1 = 0.77$, $\theta_0 = -0.64$, $y = 0.77x - 0.64$.

Proof of the least squares method

- The least squares estimation of $y^{(i)}$ is $\underline{\hat{y}^{(i)} = \theta_0 + \theta_1 x^{(i)}}$ (geschätzter y-Wert)
- Therefore the error in the i-th observation is

$$e^{(i)} = y^{(i)} \underline{- \hat{y}^{(i)}} = y^{(i)} \underline{- \theta_0 - \theta_1 x^{(i)}}$$

- Now we use the fact, that $\underline{\theta_0 = \overline{y} - \theta_1 \overline{x}}$ and find

$$e^{(i)} = y^{(i)} \underline{- \overline{y} + \theta_1 \overline{x}} - \theta_1 x^{(i)} = \left(y^{(i)} - \overline{y}\right) - \theta_1 \left(x^{(i)} - \overline{x}\right)$$

- We want to minimize the sum of the squared errors

$$\begin{aligned} SSE &= \sum_{i=1}^{n} \left(e^{(i)}\right)^2 = \sum_{i=1}^{n} \left[\left(y^{(i)} - \overline{y}\right) - \theta_1 \left(x^{(i)} - \overline{x}\right)\right]^2 \\ &= \sum_{i=1}^{n} \left(y^{(i)} - \overline{y}\right)^2 - 2\theta_1 \sum_{i=1}^{n} \left(y^{(i)} - \overline{y}\right)\left(x^{(i)} - \overline{x}\right) + \theta_1^2 \sum_{i=1}^{n} \left(x^{(i)} - \overline{x}\right)^2 \\ &= S_{yy} - 2\theta_1 S_{xy} + \theta_1^2 S_{xx} = SSE(\theta_1) \end{aligned}$$

Proof of the least squares method (cont.)

- Therefore we want

$$\frac{d}{d\theta_1}\text{SSE} = -2S_{xy} + 2\theta_1 S_{xx} = 0 \quad \text{which implies} \quad \theta_1 = \frac{S_{xy}}{S_{xx}}$$

- We can rewrite this as follows

$$\theta_1 = \frac{\sum_{i=1}^{n}\left(y^{(i)} - \bar{y}\right)\left(x^{(i)} - \bar{x}\right)}{\sum_{i=1}^{n}\left(x^{(i)} - \bar{x}\right)^2} = \frac{S_{xy}}{S_{xx}} = \frac{s_{xy}}{s_x^2}$$

where we have used the **sample covariance** of the x- and y-values

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y^{(i)} - \bar{y}\right)\left(x^{(i)} - \bar{x}\right)$$

and die **sample variance** of the x-values

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x^{(i)} - \bar{x}\right)^2$$

Coefficient of Determination (R-squared)

- The sum of squared errors without regression is called the **total sum of squares** (SST). We have

$$\begin{aligned}
\mathrm{SST} = \sum_{i=1}^{n} \left(y^{(i)} - \bar{y}\right)^2 &= \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)} + \hat{y}^{(i)} - \bar{y}\right)^2 \\
&= \sum_{i=1}^{n} \left[\left(y^{(i)} - \hat{y}^{(i)}\right) + \left(\hat{y}^{(i)} - \bar{y}\right)\right]^2 \\
&= \underbrace{\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}_{SSE} + \underbrace{2\sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)\left(\hat{y}^{(i)} - \bar{y}\right)}_{0} + \underbrace{\sum_{i=1}^{n} \left(\hat{y}^{(i)} - \bar{y}\right)^2}_{SSR}
\end{aligned}$$

- We can show, that the middle term vanishes. The first term represents the **sum of the squared errors** (SSE) and the last term is the **sum of squares explained by regression** (SSR). Therefore we can write

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$$

Coefficient of Determination (cont.)

- The fraction of the total variation that is explained by the regression determines the goodness of the regression and is called the **Coefficient of Determination** or **R-squared**:

$$0 \leq R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} \leq 1$$

- The higher the value of R^2, the better the regression, i.e. the better regression describes the data!
- $R^2 = 1$ means a perfect fit.
- $R^2 = 0$ means no fit at all.
- We'll show later, that **Coefficient of Determination** (R^2) is equal to the **Squared Correlation Coefficient** (r^2), i.e. $R^2 = r^2$.

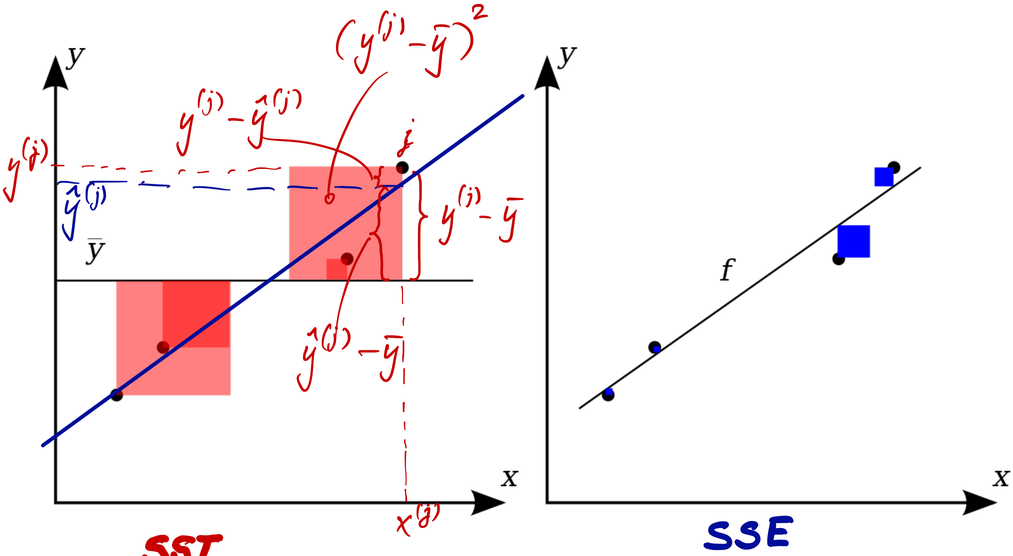
R-squared

Example (Pressure dependence ... (cont.))

Compute SSE, SST and R^2.

Solution: $SSE = 0.092$, $SST = 11.95$, and $R^2 = 0.9923$.

Coefficient of Determination - graphical visualization



$$R^2 = 1 - \frac{SSE}{SST}$$

(Pearson) Correlation Coefficient

The Pearson Correlation Coefficient r is a measure of the correlation between two quantitive variables X and Y. It is defined by

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}} = \frac{s_{xy}}{s_x s_y}.$$

Here we used the **sample variances**

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \qquad \text{and} \quad s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

and the **sample covariance**

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Note the difference between capital S_{xy}, as used above, and small caps s_{xy} (the latter is the former divided by $n-1$).

(Pearson) Correlation Coefficient (cont.)

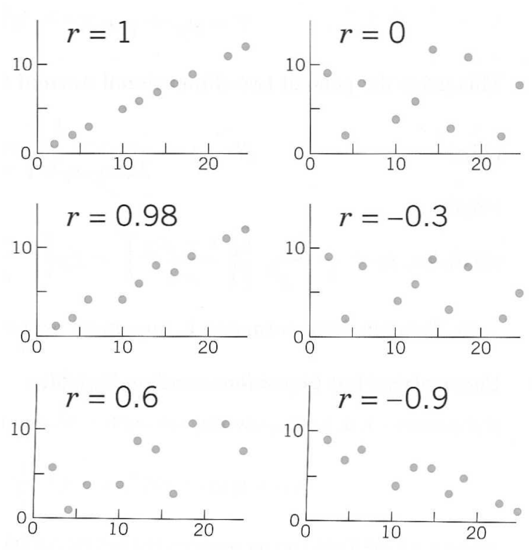
- The correlation coefficient measures the strength of the linear relationship between two variables x and y.
- Correlation always lays between -1 and 1, i.e. $-1 \leqslant r \leqslant 1$.
- Points that fall on a straight line with positive slope have a correlation of 1.
- Points that fall on a straight line with negative slope have a correlation of -1.
- Points that are not linearly related have a correlation of 0.
- The farther the correlation is from 0, the stronger the linear relationship.
- The correlation does not change if we change units of measurement, i.e. if we standarize the variables.

Example (Pressure dependence . . . (cont.))

Compute the Pearson correlation coefficient r.

Solution: $r = 0.9961$.

(Pearson) Correlation Coefficient (cont.)



For $r = 1$ the samples perfectly lie on a line.

For $|r|$ close to one, they lie close to a line.

For r close to zero, the samples are totally uncorrelated, i.e. the lie in a *cloud*.

Depending on the absolute value of r we classify

$$\begin{aligned} |r| = 0 &\rightarrow \text{uncorrelated,} \\ 0 < |r| \leqslant 0.5 &\rightarrow \text{weakly correlated,} \\ 0.5 < |r| \leqslant 0.8 &\rightarrow \text{correlated,} \\ 0.8 < |r| \leqslant 1 &\rightarrow \text{strongly correlated.} \end{aligned}$$

Correlation Analysis

- Since $\mathtt{SSE} = \sum_{i=1}^{n}\left(y^{(i)}\right)^2 + \theta_0 \sum_{i=1}^{n} y^{(i)} - \theta_1 \sum_{i=1}^{n} x^{(i)} y^{(i)}$ can only be obtained after calculating two regression parameters from the data, $\mathtt{SSE}$ has $n-2$ degrees of freedom.
- $\mathtt{SST}$ has $n-1$ degrees of freedom, since one parameter must be calculated from the data before $\mathtt{SST}$ can be computed.
- $\mathtt{SSR}$ has 1 degree of freedom, since $\mathtt{SSR} = \mathtt{SST} - \mathtt{SSE}$ and the correspondig equation for the degrees is $\deg(\mathtt{SSR}) = (n-1) - (n-2) = 1$.
- The **mean square error** ($\mathtt{MSE}$) is defined by

$$\mathtt{MSE} = \frac{\mathtt{SSE}}{(n-2)}$$

- and its **standard deviation** is the square root of $\mathtt{MSE}$.
- The regression coefficients θ_0 and θ_1 are estimates from a single sample of size n. Using another sample would lead to different regression coefficients. Let's assume the β_0 and β_1 are the true parameters of the population. That is $y = \beta_0 + \beta_1 x$

Correlation Analysis

- The standard deviations of the parameters θ_0 and θ_1 of the sample are

$$s_{\theta_0} = \sqrt{\mathrm{MSE}} \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} \left(x^{(i)}\right)^2 - n\overline{x}^2} \right]^{1/2} \quad : \text{std-Abw. von } \theta_0$$

95% conf. interval $\rightarrow \alpha = 0.05$

$$s_{\theta_1} = \frac{\sqrt{\mathrm{MSE}}}{\left[\sum_{i=1}^{n} \left(x^{(i)}\right)^2 - n\overline{x}^2 \right]^{1/2}} \quad : \text{Std-Abw. von } \theta_1$$

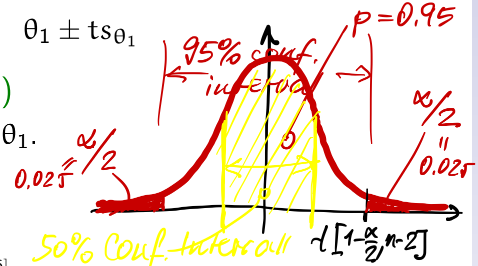
- The $100(1-\alpha)\%$ confidence intervals for θ_0 and θ_1 can be computed using $t[1-\alpha/2; n-2]$, i.e. the $1-\alpha/2$ quantile of a Student-t variate with $n-2$ degrees of freedom. The confidence intervals are:

0.975

$$\theta_0 \pm t s_{\theta_0} \quad \text{and} \quad \theta_1 \pm t s_{\theta_1}$$

Example (Pressure dependence ... (cont.))

Compute the convidence intervals for θ_0 and θ_1.



Solution: $t_q = 4.302$, $\mathbf{Conf}(\theta_0) = [-2.16, 0.876]$ $\mathbf{Conf}(\theta_1) = [0.564, 0.976]$.

Disk I/O and CPU-Time

Example (Disk I/O and CPU-Time)

The number of disk I/O's and processor times of 7 programs were measured as $(14, 2)$, $(16, 5)$, $(27, 7)$, $(42, 9)$, $(83, 20)$, $(50, 13)$, $(39, 10)$.

Compute the regression line, the coefficient of determination R^2, the Pearson correlation coefficient r

Solution: $\theta_0 = -0.0083$, $\theta_1 = 0.2438$, **CPU-time** $= -0.0083 + 0.2438\,(\#$ **Disk I/O's**$)$, $\mathtt{SSE} = 5.87$, $\mathtt{SST} = 205.71$, $R^2 = 0.9715$ **(regression explains 97% of CPU-time's variation)**, $\mathtt{MSE} = 1.17$, $s_{\theta_0} = 0.8311$, $s_{\theta_1} = 0.0187$, **with** $t[0.95; 5] = 2.015$ **we find** $\mathbf{Conf}(\theta_0) = [-1.683, 1.666]$ **and** $\mathbf{Conf}(\theta_1) = [0.2061, 0.2814]$.

Skin Cancer Mortality versus State Latitude (US, 1950)

The response variable y is the mortality due to skin cancer (number of deaths per 10 million people) and the predictor variable x is the latitude (degrees North) at the center of each of 49 states in the US at 1950. Use the file SkinCancerMortalityUSA1950.txt. From ILIAS.

State	Lat	Mort	Ocean	Long
Alabama	33.0	219	1	87.0
Arizona	34.5	160	0	112.0
Arkansas	35.0	170	0	92.5
California	37.5	182	1	119.5
Colorado	39.0	149	0	105.5
Connecticut	41.8	159	1	72.8
...				
...				
...				
WestVirginia	38.8	136	0	80.8
Wisconsin	44.5	110	0	90.2
Wyoming	43.0	134	0	107.5

- Is there a linear relationship between the Mortality and the Latitute of a US-state?
- Compute the corresponding regression line in the form $\hat{y} = \theta_0 + \theta_1 x$.
- Compute SSE and SST.
- Compute R-squared and the correlation coefficient r.
- Compute MSE.
- Compute 95% confidence intervals for θ_0 and θ_1.

Übersicht

Standardization or rescaling variables

If the independent variables are of vastly different size and dispersity, standardize them. This is done as follows: Given the original, independent (random) variable X and the sample $x_1, x_2, x_3, \ldots, x_n$ we

- compute the sample mean and the sample standard deviation

$$\hat{\mu}_x = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}_x = s_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_x)^2$$

- and then the standardized (or normalized) variable

$$x_i' = \frac{1}{\hat{\sigma}_x}(x_i - \hat{\mu}_x) = \frac{1}{s_x}(x_i - \overline{x})$$

The standardized (or normalized) variable $X' = \frac{1}{\hat{\sigma}_x}(X - \hat{\mu}_X)$ has mean zero (0) and standard deviation one (1).

Note: the sample mean and standard deviation are always marked with the hat symbol (ˆ) to distinguish it from the real mean and standard deviation of the underlying probability distribution (which we don't know). The subscript x is meant to specify the variable under considaration.

Standardization or rescaling variables (Cont.)

Example

Example with python!

End of example with python!

Nonlinearity

In Regression the following assumptions have been made:

- The true relationship between the response variable y and the predictor variable x is linear.
- The predictor variable x is non-stochastic and is measured withour error.
- The model errors $e^{(i)}$ are statistically independent
- and identically distributed (i.i.d) with zero mean and a constant deviation.

Some (visual) Tests:

- A good visual test of the validity of these assumptions is the scatter plot of $e^{(i)}$ versus the predicted response $\hat{y}^{(i)}$. The error should not substantially change with $\hat{y}^{(i)}$.
- Plot the residuals as a function of the number of experiments n. The residual should not depend on n.
- Prepare a normal quantile-quantile plot of errors. If it is linear, the assumptions are satisfied.

Übersicht

Motivation

Suppose we want to predict the weight of a weightlifter based on the training hours per week and the delivery of protein.

i	y	$x^{(1)}$	$x^{(2)}$
1	93	2	1.1
2	106	2	1.9
3	146	4	2
4	140	5	1.5
5	151	6	1.3
6	158	7	2.1
7	130	4	1.8
8	159	5	2.5

Description:

i number of observation

y weight in kg

$x^{(1)}$ Training h/Week

$x^{(2)}$ Supply of protein g/kg/d

We assume the relation (hypotesis, model)

$$y = h_{\boldsymbol{\theta}}(x_1, x_2) \; = \; \theta_0 + \theta_1 x_1 + \theta_2 x_2.$$

The parameters of our model are θ_0, θ_1, θ_2 which we abreviated using $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)$.

Motivation (cont.)

$$140 = 1 \cdot \theta_0 + 5 \cdot \theta_1 + 1.5\,\theta_2 + e^{(4)}$$

The model is

$$\underbrace{\begin{bmatrix} 93 \\ 106 \\ 146 \\ 140 \\ 151 \\ 158 \\ 130 \\ 159 \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & 2 & 1.1 \\ 1 & 2 & 1.9 \\ 1 & 4 & 2 \\ 1 & 5 & 1.5 \\ 1 & 6 & 1.3 \\ 1 & 7 & 2.1 \\ 1 & 4 & 1.8 \\ 1 & 5 & 2.5 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}}_{\theta} + \overbrace{\begin{bmatrix} e^{(1)} \\ e^{(2)} \\ e^{(3)} \\ e^{(4)} \\ e^{(5)} \\ e^{(6)} \\ e^{(7)} \\ e^{(8)} \end{bmatrix}}^{e}$$

short: $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}$.

It can be shown that the solution which minimizes the sum of the squared errors is given by:

$$\boldsymbol{\theta} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} = \begin{bmatrix} 55.7 \\ 11.1 \\ 17.5 \end{bmatrix}$$

Therefore the model is $y = 55.7 + 11.1x_1 + 17.5x_2$.

Multiple Regression (cont.)

Given n data points, i.e. training examples $(\mathbf{x}^{(i)}, y^{(i)})$, $(i = 1, 2, \ldots, n)$ where $\mathbf{x}^{(i)} = [1, x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)}]$ are m predictors (plus one dummy variable), i.e. features (independent variables) and $y^{(i)}$ is the response or estimated (dependent) variable. Here the model is

$$y^{(i)} = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_m x_m^{(i)} + e^{(i)} = \sum_{k=0}^{n} \theta_k x_k^{(i)} + e^{(i)}, \ 1 \leqslant i \leqslant n.$$

Using the parameter vector $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots, \theta_m]^T$ and the (extended) predictor vector $\mathbf{x}^{(i)} = [1, x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)}]$ we can write the last equation as

(mx1)x1-matrix

Skalar

$$y^{(i)} = \mathbf{x}^{(i)}\boldsymbol{\theta} + e^{(i)}.$$

Skalar

1x(m+1)-matrix

where the first product is the usual scalar product of the extended predictor vector and the parameter vector.

Multiple Regression (cont.)

n x 1 - matrix

Using $\mathbf{y} = [y^{(1)}, y^{(2)}, y^{(3)}, \ldots, y^{(n)}]^\intercal$, $\boldsymbol{\theta}$, $\mathbf{e} = [e^{(1)}, e^{(2)}, e^{(3)}, \ldots, e^{(n)}]^\intercal$ and

n x 1 - matrix

$$\mathbf{X} = \begin{bmatrix} - & \mathbf{x}^{(1)} & - \\ - & \mathbf{x}^{(2)} & - \\ & \vdots & \\ - & \mathbf{x}^{(n)} & - \end{bmatrix}$$

vorherige Zeilenvektoren

we can write

$$\overbrace{\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}}^{y} = \overbrace{\begin{bmatrix} - & \mathbf{x}^{(1)} & - \\ - & \mathbf{x}^{(2)} & - \\ & \vdots & \\ - & \mathbf{x}^{(n)} & - \end{bmatrix}}^{X} \overbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}}^{\theta} + \overbrace{\begin{bmatrix} e^{(1)} \\ e^{(2)} \\ e^{(3)} \\ \vdots \\ e^{(n)} \end{bmatrix}}^{e}$$

Or short $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}$ which implies $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\theta}$.

Multiple Regression (cont.)

(cost)

The target function is

$$
\begin{aligned}
\mathrm{J}(\boldsymbol{\theta}) &= \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{e}^{(i)}\right)^2 \;=\; \frac{1}{2}\mathbf{e}^{\mathsf{T}}\mathbf{e} \;=\; \frac{1}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}\right)^{\mathsf{T}}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}\right) \\
&= \frac{1}{2}\left[\mathbf{y}^{\mathsf{T}}\mathbf{y}-\boldsymbol{\theta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y}-\mathbf{y}^{\mathsf{T}}\mathbf{X}\boldsymbol{\theta}+\boldsymbol{\theta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\theta}\right] \\
&= \frac{1}{2}\left[\mathbf{y}^{\mathsf{T}}\mathbf{y}-2\mathbf{y}^{\mathsf{T}}\mathbf{X}\boldsymbol{\theta}+\boldsymbol{\theta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\theta}\right]
\end{aligned}
$$

where we used the fact, that $\boldsymbol{\theta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y} = \left(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y}\right)^{\mathsf{T}} \;=\; \mathbf{y}^{\mathsf{T}}\mathbf{X}\boldsymbol{\theta}$ because this quantity is a number. A necessary condition for $\mathrm{J}(\boldsymbol{\theta})$ to be minimal with respect to the variation of $\boldsymbol{\theta}$ is

$$
\frac{\partial}{\partial\boldsymbol{\theta}}\mathrm{J}(\boldsymbol{\theta}) \;=\; \mathbf{0} \quad\Longleftrightarrow\quad \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)\boldsymbol{\theta} \;=\; \mathbf{X}^{\mathsf{T}}\mathbf{y}.
$$

If $\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}$ exists, the solution finally is $\boldsymbol{\theta} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$. Proof follows!

Multiple Regression (cont.)

Without proof we state, that for a symmetric matrix $\mathbf{A}$:

$$\frac{\partial}{\partial \mathbf{x}} \left(\mathbf{x}^\mathsf{T} \mathbf{A} \mathbf{x}\right) = 2\mathbf{x}^\mathsf{T} \mathbf{A}$$

and for any constant vector $\mathbf{c}$:

$$\frac{\partial}{\partial \mathbf{x}} \left(\mathbf{c}^\mathsf{T} \mathbf{x}\right) = \mathbf{c}^\mathsf{T}.$$

Therefore

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{J}\left(\boldsymbol{\theta}\right) = \text{-}2\mathbf{y}^\mathsf{T} \mathbf{X} + 2\boldsymbol{\theta}^\mathsf{T} \left(\mathbf{X}^\mathsf{T} \mathbf{X}\right)$$

The transpose of this is the conjecture.

If $\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}$ exists, we can write $\boldsymbol{\theta} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} \mathbf{X}^\mathsf{T}\mathbf{y}$ (In Octave/Matlab: `pinv(X'*X)*X'*y`).

Multiple Regression (cont.)

Example

Solve the corresponding example in the exercises!

Exploring and visualizing datasets

See the corresponding Jupyter Notebook!

(Sample) covariance versus (sample) correlation

The sample covariance between a pair of standardized features (or predictors) is in fact their sample correlation coefficient. To show this, let's first standardize the features x and y:

$$x' = \frac{x - \overline{x}}{s_x} \quad \text{and} \quad y' = \frac{y - \overline{y}}{s_y}$$

Here $\overline{x}$ and $\overline{y}$ are the sample means and s_x and s_y are the sample standard deviations of x and y, respectively. Since the sample covariance of x and y is given by

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x^{(i)} - \overline{x})(y^{(i)} - \overline{y})$$

the sample covariance between the standarized features (which have means equal to zero) is

$$s'_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x^{(i)} - \overline{x}}{s_x} - 0 \right) \left(\frac{y^{(i)} - \overline{y}}{s_y} - 0 \right)$$

Covariance versus Correlation

$$
\begin{aligned}
s'_{xy} &= \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x^{(i)} - \overline{x}}{s_x} \right) \left(\frac{y^{(i)} - \overline{y}}{s_y} \right) \\
&= \frac{1}{s_x s_y} \frac{1}{n-1} \sum_{i=1}^{n} \left(x^{(i)} - \overline{x} \right) \left(y^{(i)} - \overline{y} \right) \\
&= \frac{s_{xy}}{s_x s_y} \\
&= r \\
&= \text{Corr. Coeff.}
\end{aligned}
$$

Conclusion

- Students are able to perform simple linear regression.
- Students know about the goodness of the linear regression fit.
- Students understand the coefficient of determination or R-squared.
- Students know the relation between R-squared and the correlation coefficient.
- Students can compute the confidence intervals for the regression coefficients θ_0 and θ_1.
- Students know how to compute multiple regression
- Students know how to do this by hand and by using python inside a Jupyter notebook.

I'm happy to answer Your

Questions