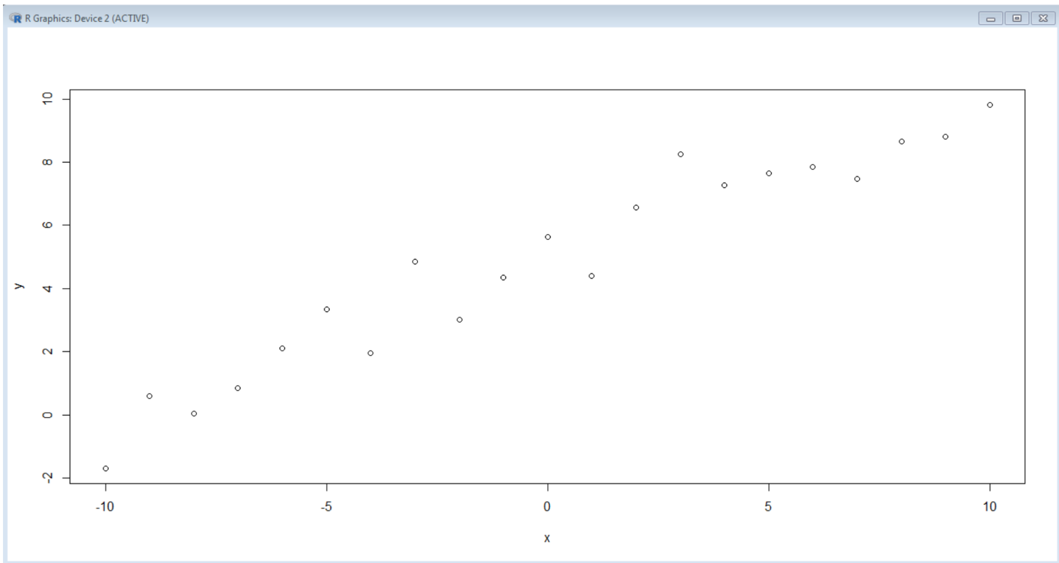
week1: Lineare Regression als deterministisches Minimierungsproblem: L^1 und L^2 Regression in einer Dimension

Wir betrachten die folgende Situation: Gegeben seien n Datenpunkte

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

in der x,y-Ebene, etwa wie in dem folgendem Bild:



Auf Grund des Plots der (x_i, y_i) vermutet man einen linearen Zusammenhang zwischen den x_i und den y_i:

$$y_i \;=\; \beta_0 \,+\, \beta_1\, x_i \,+\, \text{error/noise} \tag{1}$$

Problem: Man möchte die β_0, β_1 so wählen, dass die rechte Seite von (1) die y_i 'möglichst gut' approximieren tut. Wir müssen 'möglichst gut' mathematisch präzisieren. Es gibt mehrere Möglichkeiten, etwa:

a) L^1-Regression:

$$\sum_{i=1}^{n} \bigl|\, y_i \,-\, (\beta_0 + \beta_1 x_i) \,\bigr| \quad\to\quad \min$$

b) L^2**-Regression:**

$$\sum_{i=1}^{n} \big[\, y_i \;-\; (\beta_0 + \beta_1 x_i) \,\big]^2 \quad\to\quad \min$$

Die Namen L^1- und L^2-Regression kommen her von den L^p-Normen, oder, für Zahlenfolgen, von den ℓ^p-Normen. Diese sind folgendermassen definiert: Für $\vec{x}$, $\vec{y} \in \mathbb{R}^n$ ist

$$\|\vec{x} - \vec{y}\|_{\ell_p} \;:=\; \big\{\, \textstyle\sum_{i=1}^{n} |x_i - y_i|^p \,\big\}^{1/p}$$

also insbesondere

$$\begin{aligned} \|\vec{x} - \vec{y}\|_{\ell_1} &= \textstyle\sum_{i=1}^{n} |x_i - y_i| \\ \|\vec{x} - \vec{y}\|_{\ell_2} &= \big\{\, \textstyle\sum_{i=1}^{n} |x_i - y_i|^2 \,\big\}^{1/2} \end{aligned}$$

Das Problem (a) ist nur numerisch lösbar. Für das Problem (b) gibt es eine explizite Lösung. Wir lösen zunächst das Problem (b).

Problem (b): Wir betrachten das Minimierungsproblem

$$F(\beta_0, \beta_1) \;:=\; \sum_{i=1}^{n} \big[\, y_i \;-\; (\beta_0 + \beta_1 x_i) \,\big]^2 \quad\to\quad \min$$

Notwendige Bedingung für ein Minimum ist

$$\nabla F \;=\; \big(\, \tfrac{\partial F}{\partial \beta_0} \,,\, \tfrac{\partial F}{\partial \beta_1} \,\big) \;\stackrel{!}{=}\; (0, 0)$$

Wir bekommen

$$\begin{aligned} \tfrac{\partial F}{\partial \beta_0} &= -\,2 \sum_{i=1}^{n} \big[\, y_i \;-\; (\beta_0 + \beta_1 x_i) \,\big] \;\stackrel{!}{=}\; 0 \\ \tfrac{\partial F}{\partial \beta_1} &= -\,2 \sum_{i=1}^{n} x_i \,\big[\, y_i \;-\; (\beta_0 + \beta_1 x_i) \,\big] \;\stackrel{!}{=}\; 0 \end{aligned}$$

Das sind zwei lineare Gleichungen für die zwei Unbekannten β_0 und β_1, das können wir also lösen. Wir haben das System

$$\begin{aligned} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) &= \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} (\beta_0 x_i \;+\; \beta_1 x_i^2) &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

oder

$$\begin{aligned} n\,\beta_0 \;+\; \sum_{i=1}^{n} x_i\,\beta_1 &= \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\,\beta_0 \;+\; \sum_{i=1}^{n} x_i^2\,\beta_1 &= \sum_{i=1}^{n} y_i \end{aligned}$$

Mit den Abkürzungen

$$\begin{aligned} \vec{x} &= (x_1, x_2, \cdots, x_n) \;\in\; \mathbb{R}^n \\ \vec{y} &= (y_1, y_2, \cdots, y_n) \;\in\; \mathbb{R}^n \\ \vec{e} &= (1, 1, \cdots, 1) \;\in\; \mathbb{R}^n \end{aligned}$$

haben wir etwa

$$\vec{e}^{\,2} \;=\; 1 + 1 + \cdots + 1 \;=\; n$$

und können damit das Gleichungssystem auch folgendermassen schreiben:

$$\begin{aligned} \vec{e}\,\vec{e}\,\beta_0 \;+\; \vec{e}\,\vec{x}\,\beta_1 &= \vec{e}\,\vec{y} \\ \vec{e}\,\vec{x}\,\beta_0 \;+\; \vec{x}\,\vec{x}\,\beta_1 &= \vec{x}\,\vec{y} \end{aligned}$$

Das ist von der Form

$$A\vec{\beta} \;=\; \vec{b}$$

mit

$$A \;=\; \begin{pmatrix} \vec{e}\,\vec{e} & \vec{e}\,\vec{x} \\ \vec{e}\,\vec{x} & \vec{x}\,\vec{x} \end{pmatrix}, \qquad \vec{\beta} \;=\; \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \qquad \vec{b} \;=\; \begin{pmatrix} \vec{e}\,\vec{y} \\ \vec{x}\,\vec{y} \end{pmatrix}$$

Wir haben

$$\det A \;=\; \vec{e}^{\,2}\,\vec{x}^{\,2} \;-\; (\vec{e}\,\vec{x})^2 \;>\; 0$$

falls $\vec{x} \neq \lambda\,\vec{e}$, denn nach der Cauchy-Schwarzschen Ungleichung ist

$$|\vec{e}\,\vec{x}| \;\leq\; \|\vec{e}\|\,\|\vec{x}\|$$

mit Gleichheit nur bei linearer Abhängigkeit die wir hier ausschliessen wollen. Dann ist das A also invertierbar und wir bekommen

$$A^{-1} \;=\; \frac{1}{\det A}\begin{pmatrix} \vec{x}\,\vec{x} & -\,\vec{e}\,\vec{x} \\ -\,\vec{e}\,\vec{x} & \vec{e}\,\vec{e} \end{pmatrix}$$

Damit erhalten wir dann die folgenden Formeln für die Regressionskoeffizienten β_0 und β_1:

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \;=\; A^{-1}\vec{b} \;=\; \frac{1}{\vec{e}^{\,2}\,\vec{x}^{\,2} \;-\; (\vec{e}\,\vec{x})^2}\begin{pmatrix} \vec{x}^{\,2}\,(\vec{e}\,\vec{y}) \;-\; (\vec{e}\,\vec{x})(\vec{x}\,\vec{y}) \\ \vec{e}^{\,2}\,(\vec{x}\,\vec{y}) \;-\; (\vec{e}\,\vec{x})(\vec{e}\,\vec{y}) \end{pmatrix} \tag{2}$$

Mit Hilfe der Formeln (2) können wir also zu gegebenen Datenpunkten $(x_1, y_1), \cdots, (x_n, y_n)$ die beta's explizit berechnen. Damit ist das Problem der L^2-Regression, das Problem (b), gelöst. Die Formeln (2) sind in der R-Software schon vorimplementiert und können mit dem

```
lm( y ~ x )
```

sehr einfach berechnet werden, das schauen wir uns dann nächste Woche an. Das `lm` steht für 'linear model', wobei sich das 'linear' auf die betas bezieht, nicht auf die x, also Sie können damit auch Modelle wie $y_i = \beta_1 x_i + \beta_2 x_i^2$ fitten, das ist immer noch linear in den beta's, aber nicht in den x_i's.

Problem (a): Betrachten wir jetzt das Problem (a), die L^1-Regression. Versuchen wir, eine analoge Rechnung zu machen wie eben. Für die Betrags-Funktion

$$|x| \;=\; \begin{cases} +x & \text{fuer } x \geq 0 \\ -x & \text{fuer } x < 0 \end{cases}$$

gilt für $x \neq 0$

$$\tfrac{d}{dx}\,|x| \;=\; \begin{cases} +1 & \text{fuer } x > 0 \\ -1 & \text{fuer } x < 0 \end{cases} \;=\; \frac{x}{|x|}$$

Mit

$$F(\beta_0, \beta_1) \;=\; \sum_{i=1}^{n} \big|\, y_i \,-\, (\beta_0 + \beta_1 x_i)\,\big|$$

erhalten wir dann aus

$$\nabla F \;=\; \big(\tfrac{\partial F}{\partial \beta_0}\,,\, \tfrac{\partial F}{\partial \beta_1}\big) \;\stackrel{!}{=}\; (0,0)$$

die Gleichungen

$$\begin{aligned} \frac{\partial F}{\partial \beta_0} &= -\sum_{i=1}^{n} \frac{y_i \,-\, (\beta_0 + \beta_1 x_i)}{|\, y_i \,-\, (\beta_0 + \beta_1 x_i)\,|} \;\stackrel{!}{=}\; 0 \\ \frac{\partial F}{\partial \beta_1} &= -\sum_{i=1}^{n} x_i\, \frac{y_i \,-\, (\beta_0 + \beta_1 x_i)}{|\, y_i \,-\, (\beta_0 + \beta_1 x_i)\,|} \;\stackrel{!}{=}\; 0 \end{aligned}$$

Wir kürzen ab:

$$\varepsilon_i(\beta) \;:=\; |\, y_i \,-\, (\beta_0 + \beta_1 x_i)\,|$$

und bekommen die Gleichungen

$$\begin{aligned} \sum_{i=1}^{n} \frac{\beta_0 + \beta_1 x_i}{\varepsilon_i(\beta)} &= \sum_{i=1}^{n} \frac{y_i}{\varepsilon_i(\beta)} \\ \sum_{i=1}^{n} \frac{\beta_0 x_i \,+\, \beta_1 x_i^2}{\varepsilon_i(\beta)} &= \sum_{i=1}^{n} \frac{x_i y_i}{\varepsilon_i(\beta)} \end{aligned}$$

Mit der Definition

$$\langle \vec{x}, \vec{y} \rangle_\varepsilon \;:=\; \sum_{i=1}^{n} \frac{x_i y_i}{\varepsilon_i(\beta)}$$

können wir dann schreiben

$$\begin{aligned} \langle \vec{e}, \vec{e} \rangle_\varepsilon\, \beta_0 \,+\, \langle \vec{e}, \vec{x} \rangle_\varepsilon\, \beta_1 &= \langle \vec{e}, \vec{y} \rangle_\varepsilon \\ \langle \vec{e}, \vec{x} \rangle_\varepsilon\, \beta_0 \,+\, \langle \vec{x}, \vec{x} \rangle_\varepsilon\, \beta_1 &= \langle \vec{x}, \vec{y} \rangle_\varepsilon \end{aligned}$$

wobei der Vektor $\vec{e}$ nach wie vor durch n Einsen gegeben ist,

$$\vec{e} \;=\; (1,1,\cdots,1) \;\in\; \mathbb{R}^n$$

Das ist von der Form

$$A\vec{\beta} \;=\; \vec{b}$$

mit

$$A \;=\; \begin{pmatrix} \langle\vec{e},\vec{e}\rangle_\varepsilon & \langle\vec{e},\vec{x}\rangle_\varepsilon \\ \langle\vec{e},\vec{x}\rangle_\varepsilon & \langle\vec{x},\vec{x}\rangle_\varepsilon \end{pmatrix}, \qquad \vec{\beta} \;=\; \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \qquad \vec{b} \;=\; \begin{pmatrix} \langle\vec{e},\vec{y}\rangle_\varepsilon \\ \langle\vec{x},\vec{y}\rangle_\varepsilon \end{pmatrix}$$

wobei das A und das $\vec{b}$ aber jetzt ebenfalls von den β abhängen, über die $\varepsilon = \varepsilon(\beta)$. Wir könnten also wieder schreiben

$$\vec{\beta} \;=\; A^{-1}\vec{b}$$

oder explizit

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \;=\; \frac{1}{\langle\vec{e},\vec{e}\rangle_\varepsilon\,\langle\vec{x},\vec{x}\rangle_\varepsilon \;-\; (\langle\vec{e},\vec{x}\rangle_\varepsilon)^2}\;\begin{pmatrix} \langle\vec{x},\vec{x}\rangle_\varepsilon\,\langle\vec{e},\vec{y}\rangle_\varepsilon \;-\; \langle\vec{e},\vec{x}\rangle_\varepsilon\langle\vec{x},\vec{y}\rangle_\varepsilon \\ \langle\vec{e},\vec{e}\rangle_\varepsilon\,\langle\vec{x},\vec{y}\rangle_\varepsilon \;-\; \langle\vec{e},\vec{x}\rangle_\varepsilon\langle\vec{e},\vec{y}\rangle_\varepsilon \end{pmatrix} \tag{3}$$

Allerdings, und das ist jetzt der wesentliche Unterschied zur L^2-Regression, ist das keine explizite Formel für die Regressionskoeffizienten β_0 und β_1, da die rechte Seite von (3) ja von $\varepsilon = \varepsilon(\beta)$ und damit also auch von den β's abhängt. Man kann dann versuchen, die Gleichung (3) numerisch durch Iteration zu lösen: Man beginnt mit Startwerten für β_0 und β_1, etwa

$$\begin{aligned} \beta_0^{(0)} &:= 0 \\ \beta_1^{(0)} &:= 0 \end{aligned}$$

und berechnet dann neue beta's durch die Gleichung

$$\begin{pmatrix} \beta_0^{(k+1)} \\ \beta_1^{(k+1)} \end{pmatrix} \;=\; \frac{1}{\langle\vec{e},\vec{e}\rangle_\varepsilon\,\langle\vec{x},\vec{x}\rangle_\varepsilon \;-\; (\langle\vec{e},\vec{x}\rangle_\varepsilon)^2}\;\begin{pmatrix} \langle\vec{x},\vec{x}\rangle_\varepsilon\,\langle\vec{e},\vec{y}\rangle_\varepsilon \;-\; \langle\vec{e},\vec{x}\rangle_\varepsilon\langle\vec{x},\vec{y}\rangle_\varepsilon \\ \langle\vec{e},\vec{e}\rangle_\varepsilon\,\langle\vec{x},\vec{y}\rangle_\varepsilon \;-\; \langle\vec{e},\vec{x}\rangle_\varepsilon\langle\vec{e},\vec{y}\rangle_\varepsilon \end{pmatrix}\Bigg|_{\varepsilon=\varepsilon(\beta^{(k)})} \tag{4}$$

Also die rechte Seite berechnet man mit den alten beta's $\beta_0^{(k)}$ und $\beta_1^{(k)}$ und bekommt so neue beta's, die man dann im nächsten Iterationsschritt wieder auf der rechten Seite einsetzen kann. Es ist natürlich nicht klar, ob diese Prozedur auch konvergieren tut, für die meisten Fälle tut sie das aber, das werden wir uns dann nächste Woche mit einer R-Implementation genauer anschauen.

Die iterative Berechnungsmethode der Regressionskoeffizienten gemäss Gleichung (4) wird auch als 'iteratively reweighted least squares' bezeichnet.