A drug company sells a skin cream through drugstores in 15 districts. It would like to predict district sales, and collects some data.

District	Sales (gross of jars)	Target population (1000s persons)	Per capita income (euro)
i	y_i	x_{i1}	x_{i2}
1	162	274	2450
2	120	180	3254
3	223	375	3802
4	131	205	2838
5	67	86	2347
6	169	265	3782
7	81	98	3008
8	192	330	2450
9	116	195	2137
10	55	53	2560
11	252	430	4020
12	232	372	4427
13	144	236	2660
14	103	157	2088
15	212	370	2605

$n = 15$

A 3-d plot suggests a linear relationship

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

where ε_i is an "error term". So we should determine the plane

$$y = b_0 + b_1 x_1 + b_2 x_2$$

where b_0, b_1, b_2 are chosen to minimize the quantity

$$Q = \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_{i1} + b_1 x_{i2}) \right)^2$$

Here $Q = Q(b_0, b_1, b_2)$, and for a minimum

$$\frac{\partial Q}{\partial b_0} = 0, \quad \frac{\partial Q}{\partial b_1} = 0, \quad \frac{\partial Q}{\partial b_2} = 0 \qquad (*)$$

The normal equations $(*)$ can be expressed in matrix form:

$$B = (X^t X)^{-1} X^t Y \qquad (*)$$

where

$$B = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & & \\ 1 & x_{n1} & x_{n2} \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Equation (*) can be solved to yield

$$b_0 = 3.4526127$$
$$b_1 = 0.4960049$$
$$b_2 = 0.0091990$$

and plane

$$y = 3.45 + 0.496\, x_1 + 0.00920\, x_2 .$$

This can be used to predict sales (y) in a new district of size x_1 and income x_2.

General case: $p-1$ independent variables

Given points

$$(y_i, x_{i1}, x_{i2}, \dots, x_{i\,p-1}) \in \mathbb{R}^p$$

for $i = 1, \dots, n$ the least squares estimator is the

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_{p-1} x_{p-1}$$

with

$$B = \begin{pmatrix} b_0 \\ \vdots \\ b_{p-1} \end{pmatrix}$$ given by

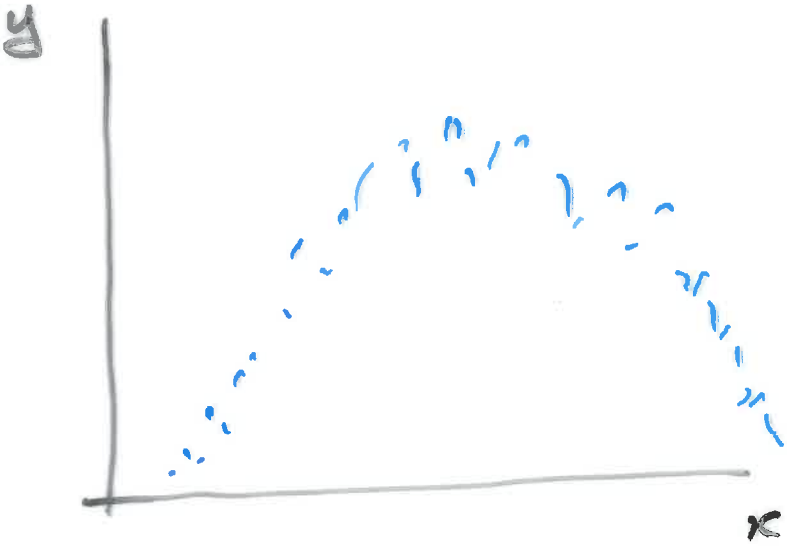
$$B = (X^t X)^{-1} X^t Y$$

where

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1\,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2\,p-2} \\ \vdots & & & & \\ 1 & x_{n1} & \cdots & \cdots & x_{n\,p-1} \end{pmatrix}.$$

Non-linear data

Suppose we have points $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ whose plot looks like



We could try finding a quadratic equation

$$y = b_0 + b_1 x + b_2 x^2$$

which is a best fit, in the least squares sense, to the data.

To do this, we construct points

$(y_1, x_1, x_1^2), (y_2, x_2, x_2^2), \ldots \in \mathbb{R}^3$

Now find the hyperplane

$$y = b_0 + b_1 x + b_2 z$$

which is the least squares fit to the data.

This ensures that

$$y = b_0 + b_1 x + b_2 x^2$$

is the quadratic which best fits, in the least squares sense, the data points in $\mathbb{R}^2$.

How good is the least squares best fit.

For simplicity consider $p = 2$, and data $(y_1, x_1), \dots, (y_n, x_n) \in \mathbb{R}^2$, and least squares fit

$$y = b_0 + b_1 x .$$

Define the fitted value

$$\hat{y}_i = b_0 + b_1 x_i$$

and the residual

$$e_i = y_i - \hat{y}_i .$$

Lemma 1

i) $\sum_{i=1}^{n} e_i = 0$, ii) $\sum_{i=1}^{n} \hat{y}_i e_i = 0$.

Proof Easy exercise using normal equations.

Define the sample mean

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n),$$

We can measure the variation in the data y_i by

$$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

the total sum of squares

We can measure the variation between the data and the fitted line by

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

the error sum of squares.

Another quantity to consider is

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

the regression sum of squares.

If the line $y = b_0 + b_1 x$ fitted the data perfectly we'd have

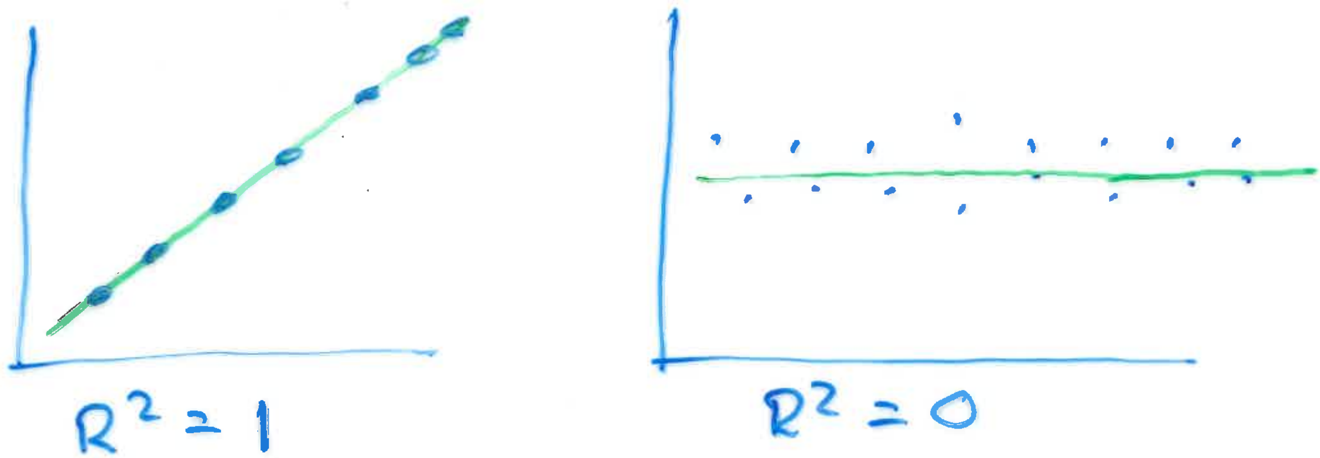
$$SSR = SSTO.$$

To measure how close to perfection we are:

Defn The coefficient of determination is

$$R^2 = \frac{SSR}{SSTO}$$

Illustrations



Typically R^2 close to 1 suggests a good fit.

<u>But</u> we can have a good fit with R^2 close to 0 in degenerate case.