Linear regression

It models the relation between "causes" (or independent variable) and "effects" (or dependent variable).

$$(X, Y_1, \ldots, Y_k)' = (\text{effect}, \text{cause } 1, \ldots, \text{cause } k)$$

for example: ("<u>Sale Price</u>", "Sq Meter Living", "Bathrooms", "Bedrooms", "Year Built", "Zone")

effect

causes

Linear regression

The mathematical tool behind this model is the conditional expectation:

$$E(X \mid Y = y) : \mathbb{R}^k \to \mathbb{R}, \quad y = (y_1, \dots, y_k)'$$

$$E(X \mid Y = y) = \int_{\mathbb{R}} x \cdot \frac{f(x, y_1, \dots, y_k)}{f(y_1, \dots, y_k)} \, dx,$$

where the marginal density of Y is $f(y_1, \dots, y_k) = \int_{\mathbb{R}} f(x, y_1, \dots, y_k) \, dx$.

The function $y \to E(X \mid Y = y)$ is called the regression of X in Y.

Linear regression

The model of linear regression is the following:

$$E(X \mid Y = y) = \beta_0 + \sum_{j=1}^{k} y_j \beta_j ,$$

where $\beta_0, \beta_1, \dots, \beta_k$ are the regression (real) parameters.
called intercept

The regression hyperplane has the equation:

$$x = \beta_0 + \beta_1 y_1 + \dots + \beta_k y_k$$

Suppose that the statistical data are:

$$(x_i, y_{i1}, \dots, y_{ik}), \quad i = 1, \dots, n \qquad n > k$$

Linear regression

Under the assumption of linearity, the regression parameters can be determined by the Least Squares method:

$$SS(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^{n} \left(x_i - \beta_0 - \beta_1 y_{i1} - \dots - \beta_k y_{ik} \right)^2$$

$$\frac{\partial SS}{\partial \beta_0} = \frac{\partial SS}{\partial \beta_1} = \dots = \frac{\partial SS}{\partial \beta_k} = 0 \quad \Longleftrightarrow$$

Linear regression

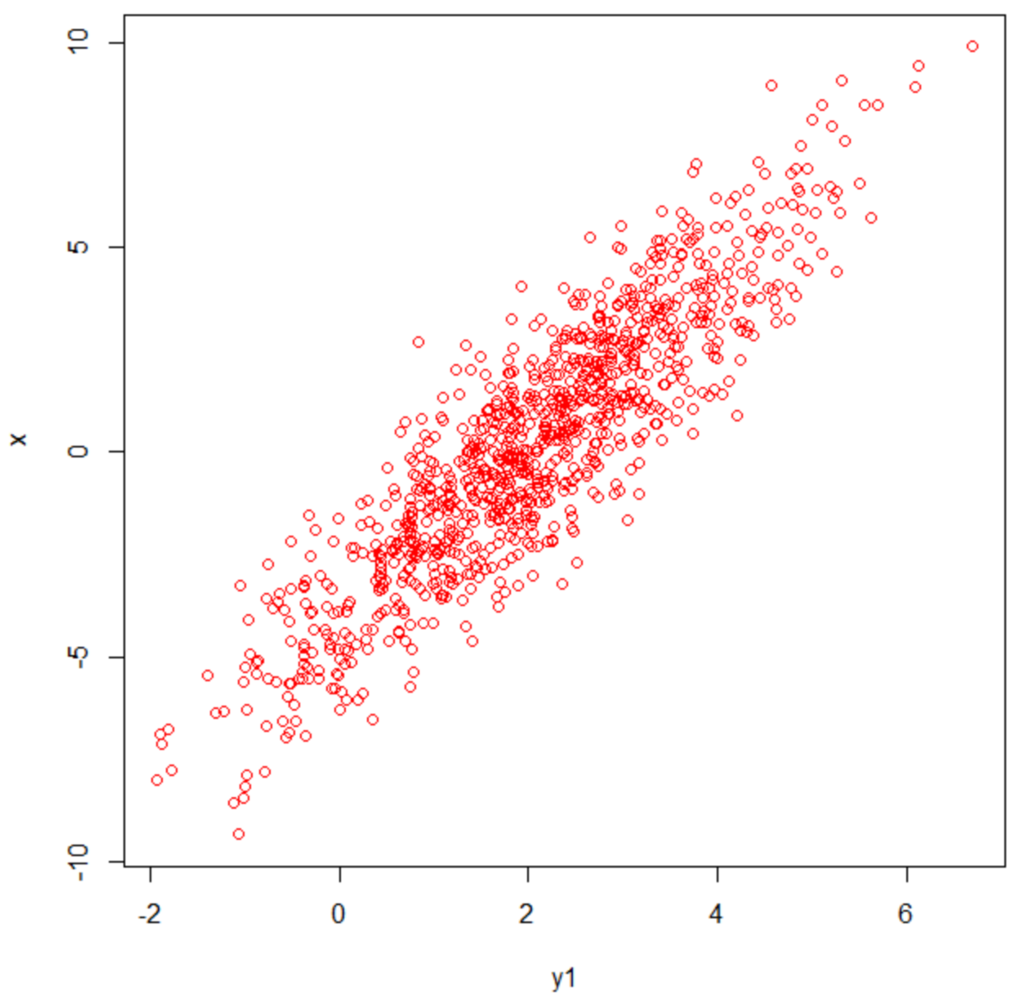
$$\begin{cases} \sum_{i=1}^{n} (x_i - \beta_0 - \beta_1 y_{i1} - \cdots - \beta_k y_{ik}) = 0 \\ \sum_{i=1}^{n} y_{ij} (x_i - \beta_0 - \beta_1 y_{i1} - \cdots - \beta_k y_{ik}) = 0, \quad j = 1, \ldots, k \end{cases} \Longrightarrow$$

$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ the LS estimators of the regression parameters (the solution is unique if $\mathrm{rank}\left(\| y_{ij} \|_{\substack{i=1,\ldots,n \\ j=1,\ldots,k}} \right) = k < n$).

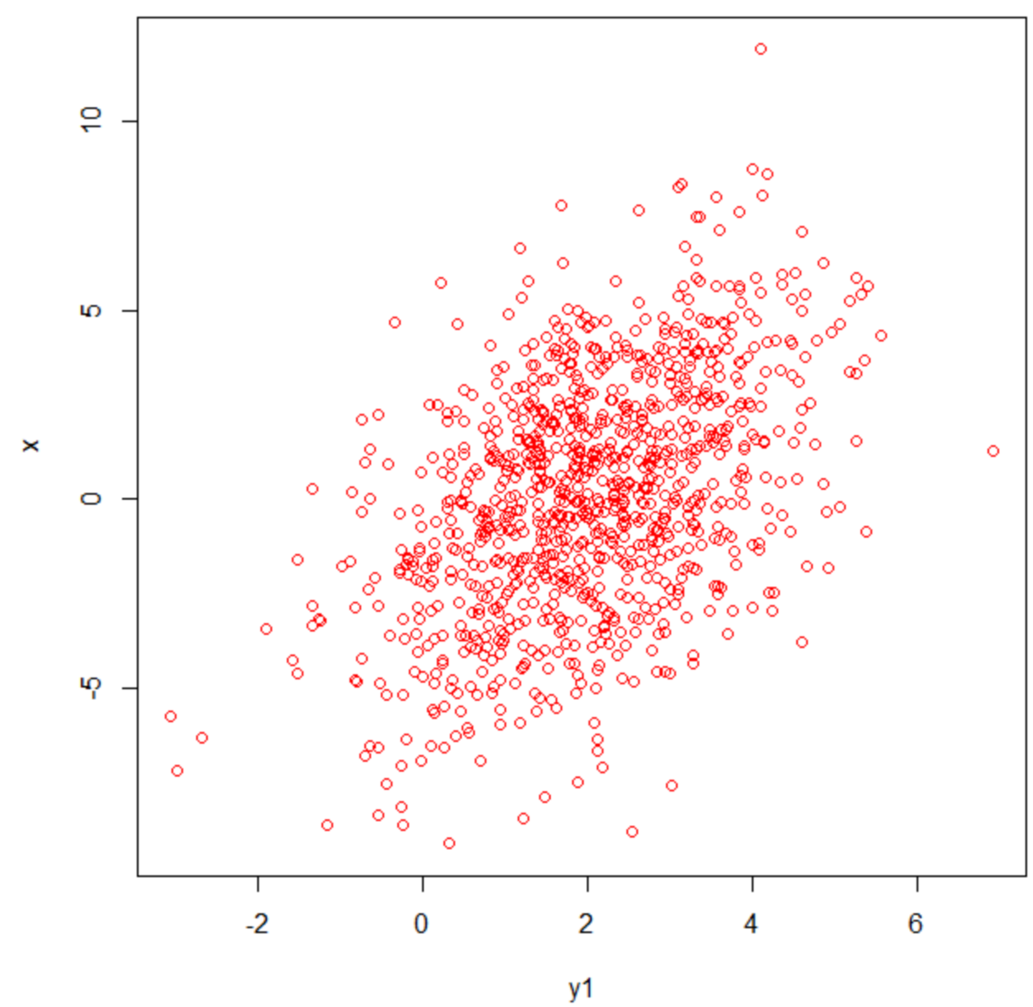
<u>Important obs.</u> This hypothesis of linearity cannot be intuitively identified by visual inspection of the data for $k \geq 2$.
In the case of multivariate normal distribution of the data, the hypothesis of linearity holds.

Linear regression

```
library(MASS)
mu=c(0,2)
Sigma=matrix(c(10,4,4,2),2,2)
X<-mvrnorm(1000,mu,Sigma)
x<-X[,1]
y1<-X[,2]
plot(y1,x,col="red")
```




```
Sigma=matrix(c(10,2,2,2),2,2)
```



Linear regression

<u>Proposition</u> Suppose that $(X, Y_1, \ldots, Y_k)'$ is a normally distributed random vector of dimension $k+1$:

$$(X, Y_1, \ldots, Y_k)' \sim N(\mu, \Sigma), \text{ where}$$

$\mu = (\mu_x, \mu_y')'$ and $\Sigma = \begin{pmatrix} \sigma_x^2 & \Sigma_{x,y} \\ \Sigma_{x,y}' & \Sigma_y \end{pmatrix}$, with $\Sigma_y = Cov(Y, Y)$,

$\Sigma_{x,y} = Cov(X, Y)$, $Y = (Y_1, \ldots, Y_k)'$.

Then, the conditional expectation $E(X \mid Y)$ is linear:

$$E(X \mid Y = y) = \Sigma_{x,y} \Sigma_y^{-1} y + (\mu_x - \Sigma_{x,y} \Sigma_y^{-1} \mu_y)$$

The regression hyperplane of X in Y can be written as:

$$x - \mu_x = \Sigma_{x,y} \Sigma_y^{-1} (y - \mu_y)$$

Linear regression

Obs. For $k=1$, $(X,Y)'$ is bidimensional.

$$(X,Y)' \sim N(\mu, \Sigma), \text{ where } \mu = (\mu_x, \mu_y)', \ \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

and $\rho = \dfrac{\sigma_{x,y}}{\sigma_x \cdot \sigma_y}$ is the correlation coefficient. The conditional expectation becomes $E(X|Y=y) = \mu_x + \rho \dfrac{\sigma_x}{\sigma_y}(y-\mu_y)$ and the regression line can be written as:

$$x - \mu_x = \rho \cdot \frac{\sigma_x}{\sigma_y}(y-\mu_y).$$

Linear regression

Statistical inference for the regression line

For the bidimensional random vector $(X, Y)'$, we consider the linear regression model

$$E(X \mid Y = y) = \beta_0 + \beta_1 y$$

with the regression line

$$x = \beta_0 + \beta_1 y.$$

Suppose that $(X_i, Y_i)'_{i=1,\dots,n}$ are observations iid $\sim (X, Y)'$ and $(x_i, y_i)'_{i=1,\dots,n}$ are the corresponding statistical data.

Linear regression

For the statistical data, we consider the linear models

$$X_i = \beta_0 + \beta_1 y_i + \xi_i, \quad i = 1, \dots, n, \text{ where } \xi_1, \dots, \xi_n$$

are i.i.d with the conditional mean 0 and the conditional variance σ^2.

We use the Least Square method to estimate the parameters β_0 and β_1.

$$SS(\beta_0, \beta_1) = \sum_{i=1}^{n} (x_i - \beta_0 - \beta_1 y_i)^2$$

$$\frac{\partial SS}{\partial \beta_0} = \frac{\partial SS}{\partial \beta_1} = 0 \iff \begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i \\ \beta_0 \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

The system has a unique solution $(\hat{\beta}_0, \hat{\beta}_1)$, assuming that not$(y_i = \bar{y}, \forall i)$.

Linear regression

Notations:

$$L_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} (x_i - \overline{x}) \cdot y_i.$$

$$L_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} (x_i - \overline{x}) \cdot x_i.$$

$$L_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - \overline{y}) \cdot y_i.$$

The unique solution is:

$$\hat{\beta}_1 = \frac{L_{xy}}{L_{yy}}, \quad \hat{\beta}_0 = \overline{x} - \hat{\beta}_1 \overline{y},$$

and the regression line is:

$$x - \overline{x} = \frac{L_{xy}}{L_{yy}} \cdot (y - \overline{y})$$

Linear regression

The LS estimators of the parameters of the regression line are:

$$\hat{\beta}_1(x_1,\dots,x_n) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{L_{yy}}$$

$$\hat{\beta}_0(x_1,\dots,x_n) = \overline{x} - \hat{\beta}_1(x_1,\dots,x_n)\,\overline{y}$$

The estimators have the following properties (in terms of the conditional distribution):

$$E(\hat{\beta}_1) = \beta_1, \quad E(\hat{\beta}_0) = \beta_0,$$

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{L_{yy}}, \quad \mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2 \cdot \overline{y}^2}{L_{yy}},$$

$$\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2\overline{y}}{L_{yy}}.$$

Linear regression

Now, we perform the analysis of variance for the linear regression ($k=1$).
The fitted (or predicted) values for the regression line are defined as:

$$\hat{x}_i = \hat{\beta}_0 + \hat{\beta}_1 y_i, \quad i = 1, \dots, n$$

We define the following sums of squares:

$$SS_t = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$SS_{regression} = \sum_{i=1}^{n} (\hat{x}_i - \overline{x})^2$$ – how far the slope of the regression line is from 0

$$SS_{residual} = \sum_{i=1}^{n} (x_i - \hat{x}_i)^2$$ – how close the sample points are to the regression line

Linear regression

For a good-fitting model, we want a high $SS_{regression}$ and a low $SS_{residual}$.

$R^2 \stackrel{def}{=} \dfrac{SS_{regression}}{SS_t}$

- if the model fits perfect then $R^2 = 1$
- if $R^2 = 0$ it means that y gives no information about x

By direct calculation, we get that:

$$SS_t = L_{xx}, \quad SS_{regression} = \frac{L_{xy}^2}{L_{yy}}, \quad SS_{residual} = L_{xx} - \frac{L_{xy}^2}{L_{yy}}$$

(hence $SS_t = SS_{regression} + SS_{residual}$)

Linear regression

Notation: $\overline{SS} = \frac{1}{\text{deg-of-freedom}} \cdot \frac{SS}{\sigma^2}$

The hypothesis $H_0: \{\beta_1 = 0\}$ can be tested by using the F-test.

<u>Proposition</u> Under the assumption of normality for $(X,Y)'$ and if the hypothesis $H_0: \{\beta_1 = 0\}$ is true, then

$$\frac{\overline{SS}_{regression}}{\overline{SS}_{residual}} = \frac{\frac{1}{1} \cdot \frac{1}{\sigma^2} \cdot SS_{regression}}{\frac{1}{n-2} \cdot \frac{1}{\sigma^2} \cdot SS_{residual}} \sim F_{1,n-2}$$

(for details see Dumitrescu & Bătătorescu, p. 223–224).

Linear regression

The test statistic for

$$H_0 : \{\beta_1 = 0\} \text{ against } H_a : \{\beta_1 \neq 0\}$$

is $\frac{\overline{SS}_{regression}}{\overline{SS}_{residual}}$, and the acceptance region is

$$W_{n;1-\varepsilon} = \left\{ (x_1, y_1, \dots, x_n, y_n) \;\middle|\; \frac{\overline{SS}_{regression}}{\overline{SS}_{residual}} \leq f_{1,n-2;1-\varepsilon} \right\},$$

where ε is the significance level and $f_{1,n-2;1-\varepsilon}$ is the $(1-\varepsilon)$ quantile of the $F_{1,n-2}$ distribution.

The hypothesis $H_0 : \{\beta_1 = 0\}$ against $H_a : \{\beta_1 \neq 0\}$ can also be tested by a T-test.

Linear regression

<u>Proposition</u> Under the assumption of normality for $(X, Y)'$ and if the hypothesis $H_0 : \{\beta_1 = 0\}$ is true, then

$$\tilde{t} \stackrel{\text{not}}{=} \frac{\hat{\beta}_1}{\sqrt{\frac{1}{n-2} \cdot SS_{residual} \cdot \frac{1}{L_{yy}}}} \sim t_{n-2}$$

The acceptance region of the T-test at the significance level ε is:

$$W'_{n;1-\varepsilon} = \left\{ (x_1, y_1, \dots, x_n, y_n) \,\middle|\, -t_{n-2;1-\varepsilon} \le \tilde{t} \le t_{n-2;1-\varepsilon} \right\},$$

where $t_{n-2;1-\varepsilon}$ is the $(1-\varepsilon)$ quantile of the t_{n-2} distribution.

Linear regression

Interval estimation for β_0, β_1

Proposition Under the assumption of normality for (X, Y), the following random variables are distributed t_{n-2}:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{1}{n-2} \cdot SS_{residual} \cdot \frac{1}{L_{yy}}}} \sim t_{n-2}$$

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{1}{n-2} \cdot SS_{residual} \cdot \left(\frac{1}{n} + \frac{\bar{y}^2}{L_{yy}}\right)}} \sim t_{n-2}$$

Linear regression

For the confidence level $1-\varepsilon$, the confidence intervals are:

$$C^{(\beta_1)}_{n;1-\varepsilon} = \left\{ -t_{n-2;1-\varepsilon} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{1}{n-2} SS_{residual} \cdot \frac{1}{L_{yy}}}} \leq t_{n-2;1-\varepsilon} \right\}$$

$$= \left\{ \hat{\beta}_1 - t_{n-2;1-\varepsilon} \cdot \sqrt{\frac{1}{n-2} \cdot SS_{residual} \cdot \frac{1}{L_{yy}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{n-2;1-\varepsilon} \cdot \sqrt{\frac{1}{n-2} \cdot SS_{residual} \cdot \frac{1}{L_{yy}}} \right\}$$

$$C^{(\beta_0)}_{n;1-\varepsilon} = \left\{ -t_{n-2;1-\varepsilon} \leq \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{1}{n-2} \cdot SS_{residual} \left(\frac{1}{n} + \frac{\bar{y}^2}{L_{yy}} \right)}} \leq t_{n-2;1-\varepsilon} \right\}$$

$$= \left\{ \ldots \leq \beta_0 \leq \ldots \right\}$$

Linear regression

The sample correlation coefficient $r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}}$.

The relationship between r and $\hat{\beta}_1$ is:

$$\hat{\beta}_1 = r \cdot \frac{\sqrt{L_{xx}}}{\sqrt{L_{yy}}}$$

We have that $-1 \leq r \leq 1$.

- $r > 0$ ($r < 0$) then X tends to increase (resp. decrease) as Y increases
- $r = 0$ then X is unrelated to Y.

if we are interested in whether there is an association between two variables, then we analyze r.
if we are interested in prediction, then we take a look at $\hat{\beta}_1$.

Linear regression

Summary (for k=1)

- we have a set of data $\{(x_1,y_1),\dots,(x_n,y_n)\}$, $x_i, y_i \in \mathbb{R}$

- we fit a linear regression model to our data – we estimate the parameters β_0 and β_1

- if we consider that $SS_{residual}$ is "small enough", then we make the assumption that the data are normally distributed.

Linear regression

- under the assumption of normality, we perform ANOVA.

This analysis is important because if $\beta_1 = 0$ then our data are uncorrelated – either the linearity hypothesis does not hold or the data are independent – in both cases, we will not apply a linear regression model for our data.

Linear regression

Comment

There are two types of variables used in statistics:

- quantitative – a measurement variable
  examples: - grades at math in a class
  - bank account balance
  - number of pets owned etc.
- qualitative or categorical: success/failure
  low/medium/high etc.

Regression = model for the relationship "causes" → "effects"

Linear regression

Depending on the type of the causes and effects, we have different situations that are differently treated:

1) The effect is qualitative; the causes are quantitative
- the typical model used is the Generalized Linear Regression (e.g. Logistic Regression, Log-Linear Regression)

2) The effect is qualitative/quantitative; the causes are qualitative - the typical model used is Factor Analysis.

3) The effect is quantitative; the causes are quantitative - the typical model used is the Linear Regression.