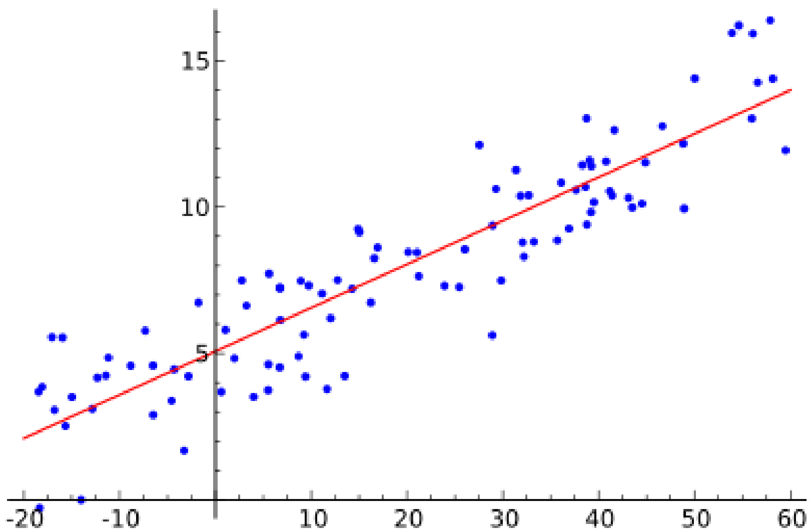
1 Simple linear regression

Example 1 If you are thinking of buying a house in Sweden you may be interested in the relationship between the price of a house and its rateable value. If you plot the prices of sold houses against their rateable values it would seem as if though there is a linear relation between them. □



Figur 1: In a linear regression a line is fit to represent data.

1.1 The model for simple linear regression

In general you will be given n pairs of values

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

where x_1, x_2, $\ldots$, x_n are given numbers, the (**regressors**), also known as exogenous variables, explanatory variables, covariates, input variables, predictor variables, or independent variables, and y_1, y_2, $\ldots$, y_n are observations of

$$Y_i = Y(x_i) = \underbrace{\alpha + \beta x_i}_{=\mu_i} + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

with independent **residuals** $\varepsilon_i \in N(0, \sigma)$. The variable Y_i is known as the regressand, endogenous variable, response variable, measured variable, criterion variable, or dependent variable. The line $y = \alpha + \beta x$ is the **theoretical regression line** and gives the relationship between the expectations of the Y_i's and the x_i's.

1.2 LS estimates

How do you know what the regression line looks like, i.e. is there some way of estimating the parameters α, β (and σ)? The LS estimates of α and β minimize

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (y_i - E[Y_i])^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

If you solve

$$
\begin{aligned}
\frac{\partial}{\partial \alpha} Q(\alpha, \beta) &= \sum_{i=1}^{n} (-2)(y_i - \alpha - \beta x_i) = 0 \\
\frac{\partial}{\partial \beta} Q(\alpha, \beta) &= \sum_{i=1}^{n} (-2x_i)(y_i - \alpha - \beta x_i) = 0
\end{aligned}
$$

then you will obtain the LS estimates

$$
(\beta)^*_{obs} = \frac{s_{xy}}{s_{xx}}, \qquad (\alpha)^*_{obs} = \bar{y} - (\beta)^*_{obs}\bar{x},
$$

where

$$
s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \qquad s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2.
$$

The estimates for α and β are the same regardless of whether σ is known or unknown. The minimum value for $Q(\alpha, \beta)$ is given by the *residual sum of squares*

$$
Q_0 = Q((\alpha)^*_{obs}, (\beta)^*_{obs}) = \sum_{i=1}^{n} (y_i - (\alpha)^*_{obs} - (\beta)^*_{obs} x_i)^2 = s_{yy} - \frac{s_{xy}^2}{s_{xx}}
$$

where

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

You can use Q_0 to estimate the standard deviation σ, or more precisely σ^2. The ML estimate for σ^2 is given by Q_0/n, but it is not unbiased. The bias corrected ML estimate is given by

$$
(\sigma^2)^*_{obs} = s^2 = \frac{Q_0}{n-2}.
$$

1.3 The distributions of the sample variables

Note that $(\alpha)^*$ and $(\beta)^*$ are *linear* in the Y_i's, since we have that

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

where we have used that $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$ (just as $\sum_{i=1}^{n} (y_i - \bar{y}) = 0$). Given this you see that the sample variables are *normally distributed*. You can show that (do this!)

$$
(\alpha)^* \in N\left(\alpha, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}\right), \qquad (\beta)^* \in N\left(\beta, \frac{\sigma}{\sqrt{s_{xx}}}\right).
$$

From this you can see that it is easier to estimate α and β if the x_i's are spread out in the sense thats_{xx} is big. It can also be shown that

$$
\frac{(n-2)S^2}{\sigma^2} \in \chi^2(n-2).
$$

Finally it holds that $(\sigma^2)^* = S^2$ is*independent* of $(\alpha)^*$ and $(\beta)^*$.

1.4 Confidence intervals

By using the λ- or t-method you can derive a two-sided confidence interval with confidence level $1-p$

$$I_\beta = \begin{cases} ((\beta)^*_{obs} \pm \lambda_{p/2} D) & \text{if } \sigma \text{ is known } (D = \sigma/\sqrt{s_{xx}}), \\ ((\beta)^*_{obs} \pm t_{p/2}(n-2)d) & \text{if } \sigma \text{ is unknown } (d = s/\sqrt{s_{xx}}) \end{cases}$$

for the slope β. This can be used together with the confidence method to test the hypothesis that $\beta = 0$, i.e. to determine if y depends on x at all.
Given some value $x = x_0$ we are often interested in estimating the expectation

$$\mu_0 = \alpha + \beta x_0$$

i.e. of finding the corresponding point on the theoretical regression line. As an estimate you simple use

$$(\mu_0)^*_{obs} = (\alpha)^*_{obs} + (\beta)^*_{obs} x_0$$

which is described by

$$\mu_0^* = (\alpha)^* + (\beta)^* x_0 \in N\left(\mu_0, \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}}\right)$$

Remark 1 For an x_0 which deviates a lot from $\bar{x}$ the variance of the estimate will become large. In general you should be careful if using the model for x-values which lie outside of the interval in which your observations x_i lie, since the linear relationship may not apply then.

Confidence intervals can as usual be obtained using either theλ- or the t-method (depending on whether σ is known or unknown). In the case when σ is unknown a confidence interval confidence level $1-p$ is given by

$$I_{\mu_0} = \left(\mu^*_{obs} \pm t_{p/2}(n-2)s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}}\right) \qquad (1-p)$$

1.5 Extensions

- *Multiple linear regression*:

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon,$$

- *General linear models* also called multivariate linear models (the response is a vector),
- *Generalized linear models* is a framework for modeling response variables that are bounded or discrete, for instance
  - Poisson regression for count data.
  - Logistic regression and probit regression for binary data.
- Lasso regression, ridge regression
- SF2930 Regression analysis, 7.5 hp!

Referenser

[1] Blom, G., Enger, J., Englund, G., Grandell, J., och Holst, L., (2005). Sannolikhetsteori och statistikteori med tillämpningar.

[2] Blom, Gunnar, (1989). Probability and Statistics. Theory and Applications.