MS-C1620 Statistical inference

7 Linear regression I

Pekka Pere

Department of Mathematics and Systems Analysis
School of Science
Aalto University

Academic year 2023–2024
Period III–IV

Contents

Regression analysis

The aim in regression analysis is to study how the value of a response ("dependent variable") changes when the values of one or more explanatory variables ("independent variables", covariates) are varied.

The name regression comes from the concept of *regression toward the mean* stating that, given independent replications, extremal values tend to be followed by average sized ones.

Regression analysis, examples

- Does the number of violent crimes depend on alcohol consumption and if it does, how strong is this dependence?
- Does statistics exam score depend on hours slept on the night prior to the exam and if it does, how strong is this dependence?
- Does salary depend on education level and if it does, how strong is this dependence?
- Does a parent's smoking have an effect on the height of a child and if it does, how strong is this dependence?
- Do crime rates depend on income inequality level and if yes, how strong is this dependence?

Regression analysis, objectives

Possible aims in regression analysis are for example:

- Description of the dependence between the explanatory and dependent variables. What is the type of the relationship? How strong is the dependence?
- Predicting the values of the dependent variable.
- Controlling the values of the dependent variable.

Simple linear regression

We begin by discussing the simplest (but still extremely useful!) form of regression, linear regression, starting with the case of single explanatory variable.

Simple linear regression, assumptions

- Consider n observations (pairs) $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ of (x, y). Assume, for simplicity, that the values x_i are non-random (otherwise we need an assumption of *exogeneity*).
- Assume that the values y_i depend linearly on the value x_i:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the regression coefficients β_0 and β_1 are unknown constants.

Simple linear model

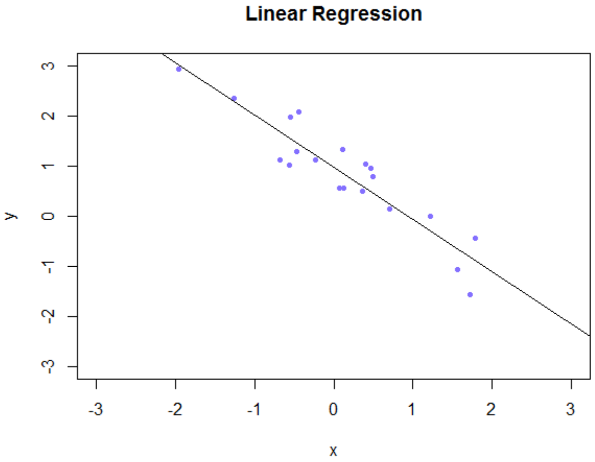

Figure: When the values of the variable x increase, the values of the variable y decrease linearly.

Simple linear regression

The simple linear regression model is usually coupled with the following additional assumptions.

Simple linear regression, assumptions, continued

- The expected value of the errors is $\mathrm{E}[\varepsilon_i] = 0$ for all $i = 1, \ldots, n$.
- The errors have the same variance $\mathrm{Var}[\varepsilon_i] = \sigma^2$.
- The errors are uncorrelated i.e. $\rho(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$.
- The errors are i.i.d. *(a stronger version of the previous two assumptions)*.

Simple linear regression

Under the previous assumptions, the random variables y_i have the following properties:

- Expected value: $\mathrm{E}[y_i] = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n,$
- Variance: $\mathrm{Var}(y_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2.$
- Correlation: $\rho(y_i, y_j) = 0, \quad i \neq j.$
- If we chose to assume that the errors are i.i.d., then y_i are independent of each other.

Simple linear regression, parameters

The linear model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

has three unknown parameters: regression coefficients β_0, β_1 and the error variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

These parameters are usually unknown and have to be *estimated* from the observations.

Under the assumption that $\mathrm{E}[\varepsilon_i] = 0$, for all $i = 1, \ldots, n$, the simple linear model can be given as

$$y_i = \mathrm{E}[y_i] + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\mathrm{E}[y_i] = \beta_0 + \beta_1 x_i$ is the systematic part and ε_i is the random part of the model.

Simple linear regression, parameter interpretation

The systematic part

$$\mathrm{E}[y_i] = \beta_0 + \beta_1 x_i$$

of the linear model defines the regression line

$$"y = \beta_0 + \beta_1 x,"$$

where β_0 (intercept) is the intersection of the regression line and the y-axis and β_1 is the slope of the regression line.

- The intercept β_0 tells the expected value of the response when the explanatory variable x has the value zero.
- The slope β_1 tells how much the expected value of the response variable y changes when the explanatory variable x grows by one unit.
- The error variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$ describes the magnitude of the random deviations of the observed values from the regression line.

Contents

Simple linear regression, objective

The aim in (simple) linear regression analysis is to find estimates for the regression coefficients β_0 and β_1.

The estimates $\hat{\beta}_0, \hat{\beta}_1$ should be chosen such that the fitted values/predictions,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

best match the observations in some suitable sense.

Numerous ways of choosing the "best" estimates exist and the most popular of these is the method of least squares.

The method of least squares

- In the method of least squares we choose the estimates by minimizing the sum of squared differences between the observations y_i and the fitted values $\hat{y}_i$,

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2.$$

- The solutions are

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \hat{\rho}(x,y)\frac{s_y}{s_x} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x},$$

where $s_x, s_y, s_{xy}, \hat{\rho}(x,y)$ are the sample standard deviations, the sample covariance and the sample correlation of x and y.

The estimated regression line

- The least squares estimates give an estimated regression line

$$
\begin{aligned}
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\
&= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\rho}(x,y)\frac{s_y}{s_x} x_i \\
&= \bar{y} + \hat{\rho}(x,y)\frac{s_y}{s_x}(x_i - \bar{x})
\end{aligned}
$$

- The slope (up or down) of the line is determined by the correlation between the two variables:
  - If $\hat{\rho}(x,y) > 0$, the line is increasing.
  - If $\hat{\rho}(x,y) < 0$, the line is decreasing.
  - If $\hat{\rho}(x,y) = 0$, the line is horizontal.

Contents

Fitted values and residuals

- Recall that the fitted value of the variable y_i, i.e. the value given to the variable y by the regression line at points x_i, is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, n.$$

- The residual $\hat{\varepsilon}_i$ of the estimated model is the difference

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n$$

  between the observed value y_i (of the variable y) and fitted value $\hat{y}_i$.
- The smaller the residuals of the estimated model are, the better the regression model explains the observed values of the response variable.

Example

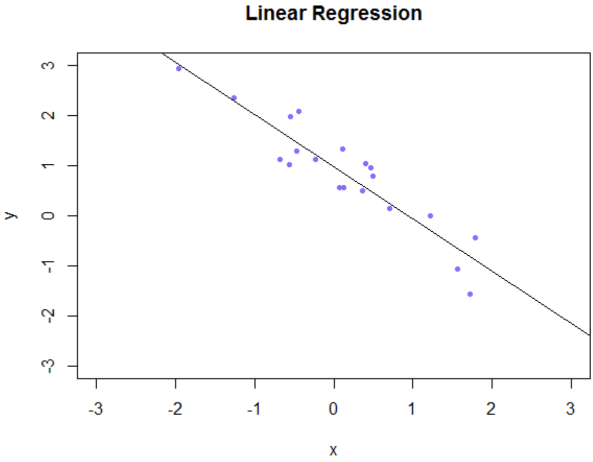

Figure: Estimated regression line minimizes the squared sum of the residuals.

Residual mean square estimation

Under the regression assumptions, an unbiased estimate for the error variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$ is given by

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Coefficient of determination

- Coefficient of determination (also known as "R-squared") gives a single number with which to assess the accuracy of the model fit.
- Coefficient of determination is defined as

$$R^2 = 1 - \frac{SSE}{SST} = (\hat{\rho}(y, \hat{y}))^2,$$

  where

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{and} \quad SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

  measure the variation of the data "before" and "after" fitting the model.
- If SSE is small compared to SST, the model has managed to *explain* a large proportion of the variance in the data.
- We always have $0 \leq R^2 \leq 1$.

Properties of the coefficient of determination

The following conditions are equivalent:

- The coefficient of determination $R^2 = 1$.
- All the residuals vanish, $\hat{\varepsilon}_i = 0, \quad i = 1, \dots, n$.
- All the observations (x_i, y_i) lie on the same line.
- The sample correlation coefficient $\hat{\rho}(x, y) = \pm 1$.
- The regression model explains the variation of the observed values of the response y completely.

The following conditions are equivalent:

- The coefficient of determination $R^2 = 0$.
- The regression coefficient $\hat{\beta}_1 = 0$.
- The sample correlation coefficient $\hat{\rho}(x, y) = 0$.
- The regression model completely fails in explaining the variation of the observed values of the dependent variable y.

Model diagnostics

We will discuss the checking of the model assumptions ("*linear model diagnostics*") next time, when we have first defined multiple linear regression.

Contents

Inference for model parameters

We next go discuss *confidence intervals* and *hypothesis tests* for the intercept β_0 and slope β_1 of the simple linear regression model.

In addition to our earlier assumptions, the following results assume that

Simple linear regression, assumptions, continued

- The errors ε_i are i.i.d.
- The errors ε_i are normally distributed.

The assumption of normality can be replaced by a *large enough* sample size.

Slope, hypothesis test

The following test is used to test whether the slope parameter β_1 of the simple linear model equals a given value (most often zero).

Slope test, assumptions

(The assumptions of slides 8 and 23.)

Slope test, hypotheses

$$H_0 : \beta_1 = \beta_1^0 \quad H_1 : \beta_1 \neq \beta_1^0.$$

Slope, hypothesis test

Slope test, test statistic

- The t-test statistic,

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{\frac{\hat{\sigma}}{\sqrt{n-1}s_x}},$$

  where $\hat{\sigma} = \sqrt{\widehat{\mathrm{Var}}(\varepsilon_i)}$ (see slide 18) and s_x is the sample standard deviation of x, has under H_0 Student's t-distribution with $n-2$ degrees of freedom.
- Under H_0, the expected value of t is 0 and **large absolute values** of the test statistic suggest that the null hypothesis H_0 does not hold.

Slope, confidence interval

A $(1-\alpha)100\%$ confidence interval for the slope β_1 of the regression line is given as

$$\left(\hat{\beta}_1 - t_{1-\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{n-1}s_x}, \hat{\beta}_1 + t_{1-\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{n-1}s_x}\right),$$

where $t_{1-\alpha/2}(n-2)$ is the $(1-\alpha/2)$-quantile of the $t(n-2)$-distribution.

Intercept, hypothesis test

Testing whether the intercept parameter β_0 equals a given value is also sometimes of interest.

Intercept test, assumptions

(The assumptions of slides 8 and 23.)

Intercept test, hypotheses

$$H_0 : \beta_0 = \beta_0^0 \quad H_1 : \beta_0 \neq \beta_0^0.$$

Intercept, hypothesis test

Intercept test, test statistic

- The $t-$test statistic
$$t = \frac{\hat{\beta}_0 - \beta_0^0}{\frac{\hat{\sigma}\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{n(n-1)}s_x}},$$
where $\hat{\sigma} = \sqrt{\widehat{\mathrm{Var}}(\varepsilon_i)}$ (see slide 18) and s_x is the sample standard deviation of x, has under H_0 Student's t-distribution with $n-2$ degrees of freedom.
- Under H_0, the expected value of t is 0 and **large absolute values** of the test statistic suggest, that the null hypothesis H_0 does not hold.

Intercept, confidence interval

A $(1-\alpha)100\%$ confidence interval for the intercept β_0 of the regression line is given as

$$\left(\hat{\beta}_0 - t_{1-\alpha/2}(n-2)\frac{\hat{\sigma}\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{n(n-1)}s_x}, \hat{\beta}_0 + t_{1-\alpha/2}(n-2)\frac{\hat{\sigma}\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{n(n-1)}s_x}\right),$$

where $t_{1-\alpha/2}(n-2)$ is the $(1-\alpha/2)$-quantile of the $t(n-2)$-distribution.