Ignacio Cascos Fernández
Department of Statistics
Universidad Carlos III de Madrid

Linear Regression

Statistics — 2014–2015

1 Introduction

Given two variables with some causal relationship, a *regression model* aims to predict the values of one variable from the other. In general, if we aim to predict Y in terms of X, a regression model is of the type

$$y = f(x) + u\,,$$

where f is some function and u the random error. Substituting x by some prescribed value x_0, it is possible to predict the value of variable Y when $X = x_0$ as $f(x_0)$.

2 Simple linear regression

2.1 Model

When $f(x)$ is the equation of a straight line, we have the simple linear regression model

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

The components of such model are:

1. Variables
   - y: dependent (or response) variable
   - x: independent (or predictor or regressor or covariate or explanatory) variable

2. Parameters (regression coefficients)

   - β_0: intercept
   - β_1: slope or gradient

3. Error term: u, which is random.

In practice, we have n data points, represented by pairs (x_i, y_i), and for each of them

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

2.2 Assumptions

The simple linear regression model is based on the five assumptions below.

1. *Linearity.* There is a linear relationship between x and y.
2. *Homogeneity.* $\mathbb{E}[u_i] = 0$ for all i.
3. *Homocedasticity.* $\mathrm{Var}[u_i] = \sigma^2$, the variability of the error term is the same for all x_is.
4. *Independence.* If $i \neq j$, then u_i and u_j are independent.
5. *Normality.* Each u_i is normally distributed, specifically $u_i \sim \mathrm{N}(0, \sigma)$.

2.3 Parameter estimation

The regression coefficients β_0 and β_1 are commonly unknown and have to be estimated form the data. Assume their estimators are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, then the predicted (or fitted or adjusted) value of y for a given x is denoted by $\hat{y}$,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x\,.$$

For each of the original data point (x_i, y_i) there is an adjusted value $\hat{y}_i$ and a *residual error* obtained as

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

Minimizing the addition of the squared residuals $(\sum_i e_i^2)$, we obtain the so-called *Least Squares Estimators* of the regression coefficients.

$$\hat{\beta}_1 = \frac{S_{X,Y}}{S_X^2} \qquad ; \qquad \hat{\beta}_0 = \overline{y} - \frac{S_{X,Y}}{S_X^2}\overline{x}\,,$$

where $\overline{x}$ and $\overline{y}$ are the sample means of X and Y, $S_{X,Y}$ is the sample covariance and S_X^2 is the sample variance of X.

In conclusion, the regression line contains the point $(\overline{x}, \overline{y})$ and has slope $\hat{\beta}_1$,

$$\hat{y} - \overline{y} = \frac{S_{X,Y}}{S_x^2}(x - \overline{x})$$

The last parameter that has to be estimated is the variance of the error term (σ^2) which is estimated by means of the residual variance, $S^2(e) = \sum_i e_i^2/(n-2)$.

2.4 Inference

The estimators of both coefficients are unbiased and normally distributed, while their variances depend on σ^2. Their estimated standard errors are obtained by substituting σ^2 by the residual variance on the expression of their respective variances. The distribution of each of the two estimators is

$$\frac{\hat{\beta}_i - \beta_i}{S(\hat{\beta}_i)} \sim t_{n-2} \qquad \text{for } i = 0, 1\,.$$

The null hypothesis of the two-sided test $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$ (sig. level α) is rejected if $|\hat{\beta}_i / S(\hat{\beta}_i)| > t_{n-2,\alpha/2}$. Furthermore, if $\alpha = 0.05$ and n is large, then $t_{n-2,\alpha/2} = 1.96 \approx 2$.

2.5 Sum of squares identity

The variability of the response variable (y) is commonly quantified by means of the Total Sum of Squares. A part of it (Regression Sum of Squares) is explained by the regression model, while other part remains unexplained after the regression is performed (Error Sum of Squares)

SS_T = Total Sum of Squares: $\sum_i (y_i - \overline{y})^2$

SS_R = Regression Sum of Squares: $\sum_i (\hat{y}_i - \overline{y})^2$

SS_E = Error Sum of Squares: $\sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$

$$SS_T = SS_R + SS_E$$

The R-squared coefficient is commonly given as a percentage and interpreted as the percentage of variability of the response variable explained by the model

$$R^2 = \frac{SS_R}{SS_T} = \frac{\sum_i (\hat{y}_i - \overline{y})^2}{\sum_i (y_i - \overline{y})^2}\,.$$

2.6 Model violations

Diagnostic graphs are used in order to check for possible violations of the model assumptions. Specifically, the linearity and homocedasticity assumptions are checked with a scatter plot of residuals vs. fitted-values (which should not show a functional form and the spread of the residuals should be approximately the same for all values of the fitted variable). The independence assumption is considered to hold true whenever there is no correlation with time, and this is checked by means of a scatter plot of residuals vs. time. Finally normality tests, or normal QQ-plots of the residuals are used to check the normality assumption.

2.7 Numerical interpretation of the coefficients

Among all possible transformations of the data, the logarithmic is the most frequent one.

- $y = \beta_0 + \beta_1 x$ (no transformation applied). When x is enlarged 1 unit, y enlarges β_1 units.
- $\log(y) = \beta_0 + \beta_1 x$ (logarithmic transformation on y). When x is enlarged 1 unit, y approximately enlarges by $100\beta_1\%$.
- $\log(y) = \beta_0 + \beta_1 \log(x)$ (logarithmic transformation on x and y). When x is enlarged by 1%, y approximately enlarges by $\beta_1\%$.
- $y = \beta_0 + \beta_1 \log(x)$ (logarithmic transformation on x). When x is enlarged by 1%, y approximately enlarges $\beta_1/100$ units.

3 Multiple linear regression

In case there are several explanatory variables $x_1, \ldots, x_k$, we can include all of them in the model in order to obtain the best possible prediction of y,

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

Assumptions. The assumptions of the multiple linear regression model are identical to the ones of the simple linear regression model. The existing linear relationship is now to be interpreted as data points approximately lying on a hyperplane.

In order to estimate the regression coefficients, the sample size (n) should be greater than $k + 1$ and all explanatory variables should be linearly independent.

3.1 Matrix approach

The regression model can be written in its matrix form as

$$Y = X\beta + U\,,$$

where

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} ; X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} ; \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} ; U = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} .$$

Matrix X is called *design matrix* and U is assumed to follow a multivariate normal distribution with mean vector located at the origin of coordinates and covariance matrix equal to $\sigma^2 I_n$, with I_n the $n \times n$ identity matrix.

The Least Squares Estimator of vector β is denoted by $\hat{\beta}$ and corresponds to

$$\hat{\beta} = \left(X^t X\right)^{-1} X^t Y\,.$$

Mathematically, multiple linear regression consists on projecting the vector with the observations of the response variable, Y, on the vector subspace spanned by the columns of the design matrix. The projection will be another vector $\hat{Y}$ containing the fitted-values of the response variable.

The variance of the model (σ^2) is estimated by means of the residual variance

$$S^2(e) = \frac{\sum_i e_i^2}{n-k-1}\,.$$

Inference. As in simple linear regression, the estimators of all coefficients are unbiased and normally distributed, while their variances depend on σ^2. Their estimated standard errors are obtained by substituting σ^2 by the residual variance on the expression of their respective variances. The distribution of each of the estimators is now

$$\frac{\hat{\beta}_i - \beta_i}{S(\hat{\beta}_i)} \sim t_{n-k-1} \qquad \text{for } i = 0, \ldots, k\,.$$

The null hypothesis of the two-sided test $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$ (sig. level α) is rejected if $|\hat{\beta}_i / S(\hat{\beta}_i)| > t_{n-k-1,\alpha/2}$. Furthermore, if $\alpha = 0.05$ and n is much greater than k, then $t_{n-k-1,\alpha/2} = 1.96 \approx 2\,.$

Sum of Squares identity. Identical to the simple linear regression model.

3.2 Adjusted R-squared and ANOVA test

A new coefficient, called adjusted-R^2 and commonly denoted by $\overline{R}^2$, is introduced in multiple linear regression. This coefficient is similar to the R^2, but penalizes models with a large number of regressors involved in them,

$$\overline{R}^2 = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)}\,.$$

The so-called ANOVA test is used to test whether there is some linear relation between the response variable and the regressors by comparing the Regression Sum of Squares with the Error Sum of Squares. In particular, the ANOVA test is

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \\ &H_1 : \text{ some } \beta_i \neq 0\,, \quad i = 1, \ldots, k\,. \end{aligned}$$

Model violations. Identical to the simple linear regression model.

3.3 Multicollinearity

Multicollinearity is a common phenomenon in multiple linear regression. It arises whenever the regressors are highly correlated, and it constitutes a big problem since it makes the estimates of the regression coefficients change erratically in response to small changes in the data. We detect multicollinearity when one or more independent variables that contributed significantly to their respective simple linear regression models stop being significant on the multiple one. In practice, if for a given sample size, we sufficiently increase the number of regressors, multicollinearity will almost always arise.

When building a multiple linear regression model, there are several strategies to select which regressors should be included in it, but none of them is completely satisfactory.

The best possible model is selected among the ones all whose independent variables contributed significantly to it. Among them, we select the model with the highest adjusted R-squared coefficient.

3.4 Dummy variables

If a sample contains observations from several populations (groups), they can be encoded by means of an *indicator* or dummy variable. Such variable will assume value 1 for the individuals from the given population and 0 for the remaining individuals. If z is a dummy variable that assumes value 1 on the individuals from population A, the regression model

$$y = \beta_0 + \beta_1 x + \beta_2 z + u$$

will have intercept $\beta_0+\beta_2$ on the individuals from population A, and intercept β_0 on the remaining individuals. Similarly, we can modify the gradient of the regression model, depending on the group an individual belongs to.

If there are individuals from s different populations, then the needed number of dummy variables is $s-1$. In case we introduce s dummy variables, the model will not be solved.