REGRESSION ANALYSIS

V. Nollau
Institute of Mathematical Stochastics, Technical University of Dresden, Germany

Keywords: Linear models, least squares estimates, regression models, Gauß-Markov theorem.

Contents

Summary

The analysis of relationships is one of the most important tasks of human civilization. By learning how certain phenomena depend on others we understand our world and try to predict the consequences of our actions. The majority of such relationships we learned are based on empirical observations or need to be verified by observations, measurements or experiments. If random influences play little or no role the study of relationships is often relatively easy. On the other hand, if random effects play a role, the study of discovering relationships often requires a careful statistical analysis of the underlying data. Regression analysis includes such careful statistical methods for analyzing how one or more variables (the so-called independent variables, predictor variables or regressors) affect other variables (the so-called dependent variables or response variables).

1. Simple Regression

The general model of regression analysis which we consider in the following is a generalization of the *simple regression*, that means the description of observations (measurements) y_i (i = 1, … , n) of the dependent linear variable y by function values $\alpha_0 + \alpha_1 x^{(i)}$ of an independent variable x for $x = x^{(1)}, \ldots, x^{(n)}$ and random effects e_i (i = 1, …, n). Thus we hypothesize that we had a set of relationships of the form:

$$y_i = \alpha_0 + \alpha_1 x^{(i)} + e_i \tag{1}$$

$\left(\alpha_0,\ \alpha_1\ \in\ \mathbb{R}\right)$, where *i* = 1, …, *n* (*n* - number of observation (measurement) points). Equation (1) is called an (empirical) regression model.

Using a stochastic standard formulation a *simple regression model* is given by a set of *n* random variables Y_i of the form

$$Y_i = \alpha_0 + \alpha_1 x^{(i)} + E_i \tag{2}$$

with random errors E_i (*i* = 1, …, *n*), for which the expectations (mean values) are

$$\mathbb{E}\left(E_i\right) = 0 \tag{3}$$

and for the variances

$$\operatorname{var}\left(E_i\right) = \sigma^2\left(>0\right) \tag{4}$$

hold. Here, (3), (4) mean, that the underlying data can be fitted by a straight line with "pure" random errors of constant variance.

A measure of fit is

$$S\left(\alpha_0,\ \alpha_1\right) = \sum_{i=1}^{n}\left(Y_i - \alpha_0 - \alpha_1 x^{(i)}\right)^2 \tag{5}$$

the so-called *residual*. Obviously the smaller the residual $S\left(\alpha_0, \alpha_1\right)$ the better is the fit.

In this sense the best estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$ of the unknown parameters α_0 and α_1 are the solutions of

$$S\left(\alpha_0,\ \alpha_1\right) \to \min_{\alpha_0, \alpha_1 \in \mathbb{R}}. \tag{6}$$

This is a classical problem. Since the partial derivatives of $S\left(\alpha_0, \alpha_1\right)$ with respect to α_0 and α_1 are

$$\frac{\partial S\left(\alpha_0,\ \alpha_1\right)}{\partial \alpha_0} = -2\sum_{i=1}^{n}\left(Y_i - \alpha_0 - \alpha_1 x^{(i)}\right) \tag{7}$$

and

$$\frac{\partial S\left(\alpha_0,\ \alpha_1\right)}{\partial \alpha_1} = -2\sum_{i=1}^{n}\left(Y_i - \alpha_0 - \alpha_1 x^{(i)}\right)^2 \quad x^{(i)} \tag{8}$$

one obtains the so-called least squares estimates $\hat{\alpha}_0$, $\hat{\alpha}_1$ with

$$\hat{\alpha}_0 = \overline{Y}_{\bullet} - \hat{\alpha}_1 \overline{x}_{\bullet} \tag{9}$$

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n} x^{(i)} Y_i - n\overline{x}_{\bullet} \quad \overline{Y}_{\bullet}}{\sum_{i=1}^{n}\left(x^{(i)} - \overline{x}_{\bullet}\right)^2} \tag{10}$$

by setting

$$\overline{x}_{\bullet} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} \tag{11}$$

and

$$\overline{Y}_{\bullet} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{12}$$

if there exist at least two different values $x^{(i)} \neq x^{(i')}$, $i \neq i'$, for the independent variable *x*.

These estimates $\hat{\alpha}_0$, $\hat{\alpha}_1$ are unbiased, that means

$$\mathbb{E}\left(\hat{\alpha}_0\right) = \alpha_0 \quad \text{and} \quad \mathbb{E}\left(\hat{\alpha}_1\right) = \alpha_1 \tag{13}$$

Moreover,

$$\begin{aligned} \hat{\sigma}^2 &= \frac{S\left(\hat{\alpha}_0,\ \hat{\alpha}_1\right)}{n-2} \\ &= \frac{\sum_{i=1}^{n}\left(Y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x^{(i)}\right)^2}{n-2} \\ &= \frac{1}{n-2}\left\{\sum_{i=1}^{n}\left(Y_i - \overline{Y}_{\bullet}\right)^2 - \frac{\left[\sum_{i=1}^{n}\left(x^{(i)} - \overline{x}_{\bullet}\right)\ \left(Y_i - \overline{Y}_{\bullet}\right)\right]^2}{\sum_{i=1}^{n}\left(x^{(i)} - \overline{x}_{\bullet}\right)^2}\right\} \end{aligned} \tag{14}$$

is an unbiased estimator of σ^2 (i.e. $\mathbb{E}\left(\hat{\sigma}^{\mathbf{2}}\right) = \sigma^2$).

As noted above, when one has a good fit of the data the residuals (5) are small. Thus one can express the quality of fit by the sum of squares

$$\sum_{i=1}^{n} \tilde{E}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x^{(i)} \right)^2 \tag{15}$$

However, this term is dependent on the units in which the dependent variables $Y_i (i = 1, \ldots, n)$ are measured. Thus, if $\alpha_0 \neq 0$, a suitable measure of fit is

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \tilde{E}_i^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}_{\bullet} \right)^2} = 1 - \frac{\sum_{i=1}^{n} \left(Y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x^{(i)} \right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}_{\bullet} \right)^2} \tag{16}$$

-
-
-

Bibliography

Draper N.R. and Smith H. (1981). *Applied Regression Analysis*. Second Edition. New York: Wiley. [This book presents modern aspects of regression analysis].

Müller P.H. (ed.) (1981). *Lexikon der Stochastik*. 5. Auflage. Berlin: Akademie-Verlag. [This is a dictionary for all fields of stochastics].

Nollau V. (1979). *Statistische Analysen*. 2. Auflage. Basel und Stuttgart: Birkhäuser. [This book presents all statistical methods based on general linear models of statistics].

Seber G.A.F. (1971). *Linear Regression Analysis*. New York, Chichester, Brisbane, Toronto, Singapore: John Wiley & Sons. [This is a standard treatment of least squares fitting and hypothesis testing for the multiple linear regression model including considerations of robustness, random regressors and ANOVA, MANOVA].

Sen A. and Srivastava A. (1990). *Regression Analysis Theory, Methods, and Applications*. New York: Springer-Verlag. [This book offers an up-to-date account of the theory and methods of regression analysis].

Srivastava M.S. and Carter E.M. (1983). *Applied Multivariate Statistics*. New York, Amsterdam, Oxford: North-Holland. [This book is based on lectures on methods in current use in multivariate statistics].

Biographical Sketch

V. Nollau was born in 1941 and studied mathematics and theoretical physics at the Technical University of Dresden (Germany). He graduated in 1964, obtaining doctorate in 1966 and 1971 (Dr. habil.). From

1969 he was assistant professor at TU Dresden. His main research topics were operator theory, stochastic processes and random search. In 1972 he made the first contributions to stochastic optimization and decision processes theory. Since 1990 the author is professor for stochastic analysis and control. He wrote several text works including "*Statistische Analysen*" (Linear Models in Statistics). The author is dean of the faculty of mathematics in Dresden.