# LINEAR REGRESSION MODELS

**F. Tibaldi**
*Limburgs Universitair Centrum, Hasselt University, Belgium*

**Keywords:** Simple linear regression, multiple regression, estimation, inference, diagnostics

## Contents





## Summary

The basic linear regression model is introduced, followed by estimation and inferential methods. Diagnostic and remedial measures are discussed. Then, the method and its inferential procedures is generalized to the multiple linear regression setting, i.e., the context where the response variable is explained in terms of several rather than one explanatory variable.

## 1. Introduction

Understanding relationships among sets of variables is a basic problem in statistical science. In the late nineteenth century, Sir Francis Galton made a fundamental contribution to understanding multivariate relationships by introducing regression analysis. In one dataset, described in his 1885 presidential address before the Anthropological Section of the British Association of the Advancement of Sciences, Galton linked the distribution of children's heights to their parents'. Galton showed not only that each distribution was approximately normal but also that the joint distribution could be described as a bivariate normal. Thus, the conditional distribution of adult children's height, could also be described by using a normal distribution. As a by-product of his analysis, Galton observed that “tall parents tend to have tall children although not as tall as the parents” (and vice versa for short children). From this, he incorrectly inferred that children would “regress to mediocrity” in subsequent

generations, hence suggesting the term that has become known as *regression analysis*. Several authors have given insightful and entertaining accounts of the terminology use of Galton as well as of other contributors to statistical science.

Regression analysis has developed into the most widely applied statistical methodology. It is an important component of multivariate analysis because it allows researchers to focus on the effects of explanatory variables. To illustrate, in the Galton dataset of family heights, regression allows the analyst to describe the effect of parents' height on a child's adult height.

## 2. Simple Linear Regression model

Regression in its simplest form, is a technique for modeling a relationship between two variables. This, of course, can be extended to multiple variables.

### 2.1. The Model

The simple linear regression model can be stated as follows

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ \ i = 1, \ldots, n \qquad (1)$$

where

- $Y_i$ is the *response* (dependent variable) for the *i*th trial (subject, sample,…);
- $x_i$ is the value of the independent variable (predictor, regressor,…) in the *i*th trial;
- the error term $\varepsilon_i$ represents the residuals, assumed to be independent random variables having a normal distribution with mean zero and constant variance $\sigma^2$. In other words,
    - $\mathrm{E}(\varepsilon_i) = 0$,
    - $\mathrm{Var}(\varepsilon_i) = \sigma^2$ (homoscedastic errors),
    - $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ with $i \neq j$ (uncorrelated errors),
- the unknown parameters $\beta_0$ and $\beta_1$, also called the *regression coefficients*, need to be estimated. In Section 2.2 we will outline a method to obtain estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for $\beta_0$ and $\beta_1$.

The simple linear regression model (1) is called a *statistical model* and needs to be distinguished from a so-called *deterministic model*. The “law of gravity” in physics, for example, is a deterministic model that assumes an ideal setting where the response variable varies in a completely prescribed way according to a perfect mathematical function of the independent variables. Statistical models allow for the possibility of error (variability) in describing a relationship. We also need to distinguish between *observational data* and *experimental data*. The first type of data is obtained without controlling the independent variable. A major limitation of this kind of data is that they often do not provide adequate information about causal relationships. One always should investigate whether other independent variables might explain causal relationships more directly. When control is exercised over the independent variable,

the resulting experimental data provide much stronger information about causal relationships. In a completely randomized design, treatments are assigned to each of the experimental units completely at random. Randomization tends to balance out the effects of any other variable that might affect the response.

### 2.2. Estimation

The *regression coefficients,* $\beta_0$ and $\beta_1$, are traditionally estimated using least squares. Such estimators, usually denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, are defined as the minimizer of

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2. \tag{2}$$

Differentiating $Q(\beta_0, \beta_1)$ with respect to $\beta_0$ and $\beta_1$ and setting these partial derivatives equal to zero leads to the normal equations that, once solved, yield:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \text{ and } \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3}$$

We can use the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ to estimate the regression function $\mathrm{E}(Y) = \beta_0 + \beta_1 x$ by $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$. We call $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ the *i*th fitted value and $e_i = Y_i - \hat{Y}_i$ the *i*th residual. Residuals play a very important role in studying whether a given regression model is appropriate for the data at hand.

Next, we propose an estimator for the variance parameter $\sigma^2$. Recall that, based on a sample of independent normally distributed random variables $Z_1, \ldots, Z_n$, $S^2 = \sum_{i=1}^{n} (Z_i - \bar{Z})^2 / (n-1)$ is an unbiased estimator for $\sigma^2$. Now, in the regression model (1), each $Y_i$ has its own mean $\beta_0 + \beta_1 x_i$, which can be estimated by the fitted value $\hat{Y}_i$. Hence, the deviation from the mean is now represented by the residual $e_i$ and the appropriate sum of squares $SSE = \sum_{i=1}^{n} e_i^2$ with $n-2$ degrees of freedom is used to obtain an unbiased estimator of $\sigma^2$ by $MSE = SSE/(n-2)$.

### 2.3. Inference

To set up interval estimates and test procedures, we need to specify the error distribution. In the normal error regression model we extend (1) with the assumption that $\varepsilon_i$ are independent zero-mean normally distributed with variance $\sigma^2$. Therefore, the normal regression model can be formulated as $Y_i$ are independent $N(\beta_0 + \beta_i x_i, \sigma^2)$.

In the next section we will explain how inferences can be made under the assumption of the linear regression model. To do so, we will use the following results.

- If $SSE$ is the sum of squares defined in Section 2.2, then $SSE/\sigma^2 \sim \chi^2_{(n-2)}$.
- $SSE$ and $(\hat{\beta}_0, \hat{\beta}_1)$ are independent.
- $\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\hat{\beta}_0))$ where $\sigma^2(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$.
- $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2(\beta_1)\right)$ where $\sigma^2(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

The variances $\sigma^2(\hat{\beta}_0)$ and $\sigma^2(\hat{\beta}_1)$ contain the unknown parameter $\sigma^2$ and therefore have to be estimated. Using the fact that $\hat{\sigma}^2 = MSE$ we have

$$\hat{\sigma}^2(\hat{\beta}_0) = MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \qquad \text{and} \qquad \hat{\sigma}^2(\hat{\beta}_1) = \frac{MSE}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Then,

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}^2(\hat{\beta}_j)} \sim t(n-2), \qquad j = 1, 2.$$

### 2.3.1. Inferences about the Regression Coefficients

Mainly hypothesis tests regarding $\beta_1$ are of importance, with particular emphasis on

$$H_0 : \beta_1 = 0 \qquad \text{versus} \qquad H_1 : \beta_1 \neq 0.$$

Indeed, $\beta_1 = 0$ indicates that there is no linear association between the response $Y$ and the independent variable $X$. If $\beta_1 = 0$ the linear model simplifies to $Y_i$ being independent $N(\beta_0, \sigma^2)$, which implies not only that there is no linear association between response and independent variable but also that there is not relation of any type between them. In contrast there are only infrequent occasions when we wish to make inferences concerning $\beta_0$.

Using the distributional results we can then construct $(1-\alpha)\times 100\%$ confidence intervals for $\beta_0$ and $\beta_1$ in the following way:

$$\hat{\beta}_0 \pm t(1-\alpha/2; n-2)\hat{\sigma}(\hat{\beta}_0)$$

and

$$\hat{\beta}_1 \pm t(1-\alpha/2; n-2)\hat{\sigma}(\hat{\beta}_1).$$

Test concerning the parameters of this model can be set up in a standard fashion using the $t$ distribution. The decision rule for the two-sided alternative $H_0 : \beta_1 = 0$ is

$$\text{if } \left|\frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)}\right| \le t(1-\alpha/2; n-2) \text{ do not reject } H_0 : \beta_1 = 0,$$

$$\text{if } \left|\frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)}\right| > t(1-\alpha/2; n-2) \text{ reject } H_0 : \beta_1 = 0,$$

This rule can also be used for the one-sided alternative.

If the response distribution is not exactly normal but does not seriously depart from normal, the distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ will still be approximately normal. If the response distribution is far from normal, one can use the asymptotic normality of $\hat{\beta}_0$ and $\hat{\beta}_1$: their distributions approach normality as the sample size increases. Thus, with sufficiently large samples, the confidence intervals and decision rules given earlier still apply with $t$-percentiles replaced by normal percentiles.

-
-
-



**Bibliography**

Berry, W.D. and Feldman, S. (1992) *Multiple Regression In Practice*. Quantitative Applications in Social Sciences. Beverly Hills: Sage University Paper. [Basic text on regression.]

Draper, N. (1998) *Applied Regression Analysis*. New York: John Wiley. [Standard text on regression.]

Fox, J. (1997) *Applied Regression Analysis, Linear Models, and Regression Methods*. Thousand Oaks, California: Sage Publications. [Reference text on regression.]

Galton, F. (1885). Regression towards mediocrity in heredity stature. *Journal of Anthropological Institute*, **15**, 246-263. [Historic perspective on regression.]

Mosteller, F. and Tukey, J.W. (1977) *Data Analysis and Regression*. Reading, MA: Addison-Wesley. [Perspective on regression.]

Neter, J., Kutner, M., and Nachtsheim, C. (1996) *Applied Linear Statistical Models* (4th ed.) Irwin: Chicago. [Classical introductory text to the subject.]

Ronald, C. (1998) *Analysis of Variance, Design and Regression: Applied Statistical Methods*.

Boca Raton: Chapman and Hall. [Standard linear models text.]

Sokal, R.R. and Rohlf, F.J. (1993) Biometry: *The Principles and Practice of Statistics in Biological Research*. (3rd ed.) New York: Freeman. [Particular focus on biology.]

Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge: Harvard University Press. [Entertaining perspective on the history of statistics.]

Weisberg, S. (1980) *Applied Linear Regression*. New York: John Wiley. [Regression from an applied point of view.]

**Biographical Sketch**

**Fabián S. Tibaldi** holds a B.S. degree in mathematics from the University of Buenos Aires (Argentina) and Master of Science and Ph.D. degrees in Biostatistics from the Limburgs Universitair Centrum (Belgium). His research interests focus on methods for correlated and hierarchical normally distributed and survival data, with applications in the evaluation of surrogate endpoints in controlled clinical trials

and in population genetics. He has extensive experience in survey methodology and worked in official statistics, both in Argentina and in Belgium.