# Linear Regression with One Independent Variable

Notes from "Linear Statistical Models" by Kutner et al

Devere Anthony Weaver

---

## 1 Overview

These notes cover the basic ideas of regression analysis and the estimation of the parameters of the regression model.

---

## 2 Relations Between Variables

A *functional relationship* between two variables is expressed by a mathematical formula. If $X$ is the independent variable and $Y$ is the dependent, then a functional relation is of the form

$$Y = f(X).$$

When plotted, all the values of observations for $Y$ fall directly on the line of the functional relationship.

A *statistical relationship*'s observations generally don't fall directly on the curve of the relationship. When plotted, one will observe a scattering of points around the the line of the statistical relationship. This scattering represents the variation in the relationship between the two variables.

---

## 3 Regression Models an Their Uses

### 3.1 Basic Concepts

A regression model is a formal means of expressing the two essential ingredients of a statistical relation:

1. A tendency of the dependent variable $Y$ to vary with the independent variable in a systematic fashion.
2. A scattering of points around the curve of the statistical relationship.

These two characteristics are embodied in a regression model by postulating that:

1. There is a probability distribution of $Y$ for **each level** of $X$.
2. The means of these probability distributions vary in some systematic fashion with $X$ (i.e. the actual realized value of $Y$ is viewed as a random selection from this probability distribution).

This concept above means that the probability distributions have a systematic relationship to the level of $X$. This systematic relationship is called the *regression function of* $Y$ *on* $X$ and its graph is call the *regression curve*.

### 3.2 Construction of Regression Models

A central problem is that of choosing a set of independent variables that is good for the purposes of the analysis. A major consideration in this choice is the *extent to which a chosen variable contributes to reducing the remaining variation in* $Y$ after the contributions of the other independent variables that have tentatively been included in the regression model.

### 3.3 Functional Form of Regression Relation

The choice of the functional form of the regression relation is tied to the choice of the independent variables. There may be a known functional form for the relation between variables; however, in practice, it's more likely this form isn't known in advance and must be decided upon once the data have been collected and analyzed.

### 3.4 Scope of Model

Usually we'll need to restrict the coverage of the model to some interval or region of values of the independent variable(s). The scope is determined by the design of the investigation or by the range of data at hand.

### 3.5 Uses of Regression Analysis

Some uses of regression analysis include:

1. Description
2. Control
3. Prediction

---

## 4 Simple Linear Regression Model with Distribution of Error Terms Unspecified

The basic regression model, where there exists only one independent variable, is given by

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{4.1}$$

where

- $Y_i$ is the value of the response variable in the $i$th trial
- $\beta_O$ is the parameter for the intercept
- $\beta_1$ is the parameter for the slope
- $\epsilon_i$ is the random error term with mean $E[\epsilon_i] = 0$ with constant variance; each error is uncorrelated

Regression model (4.1)j is said to be *simple, linear in the parameters* and *linear in the independent variable*.

### 4.1 Important Features of the Model

The following are a list of the most important features of the model:

1. The observed value of $Y$ in the $i$th trial is the sum of two components
   - the constant term $\beta_0 + \beta_1 X_i$
   - the random term $\epsilon_i$
2. Since $E[Y_i] = 0$, it follows that

   $$E[Y_i] = E[\beta_0 + \beta_1 X_i + \epsilon_i] = \beta_0 + \beta_1 X_i + E[\epsilon_i] = \beta_0 + \beta_1 X_i.$$

   Hence, the response $Y_i$, when the level of $X$ in the $i$th trial is $X_i$, comes from a probability distribution with mean

   $$E[Y_i] = \beta_0 + \beta_1 X_i \tag{4.2}$$

   with regression function for model (4.1)

   $$E[Y] = \beta_0 + \beta_1 X. \tag{4.3}$$
3. The observed value of $Y$ in the ith trial exceeds for falls short of the value of the regression function by the error term $\epsilon_i$.

4. The error terms are assumed to have constant variance $\sigma^2$, thus the responses $Y_i$ have the same constant variance
$$Var[\beta_0 + \beta_1 X_i + \epsilon_i] = Var[\epsilon_i],$$
which is a constant term.

5. The error terms are assumed to be uncorrelated, hence the responses are also uncorrelated.

### 4.2 Meaning of Regression Parameters

As mentioned earlier, the regression coefficient $\beta_0$ is the slope of the regression line. It indicates the change in the mean of the probability distribution of $Y$ per unit increase in $X$.

The parameter $\beta_0$ is the $Y$ intercept of the regression line. If the scope of the model includes $X = 0$, then it gives the mean of the probability distribution of $Y$ at $X = 0$. If this scope of the model doesn't include $X = 0$, then the intercept doesn't have any particular meaning as a separate term.

---

## 5 Overview of Regression Analysis

In the usual piratical situation, we don't have adequate knowledge to specify the appropriate regression model in advance, so we'll need to use our observed data.

The first step is an *exploratory study of the data.* Then, based on this exploration, we can build one or more preliminary models. Each are examined for appropriateness of the data and are revised as needed until the investigator is satisfied with the suitability of a particular regression model.

After, the selected regression model can be used for making inferences about the regression parameters of the model or predictions of new observations.

---

## 6 Estimation of Regression Function

### 6.1 Method of Least Squares

For each sample observation $(X_i, Y_i)$, the method of least squares considers the deviation of $Y_i$ from its expected value
$$Y_i - (\beta_0 + \beta_1 X_i). \tag{6.1}$$
In particular, the method of least squares requires that we consider the sum of the $n$ squared deviations,
$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{6.2}$$
For the method of least squares, the estimators of $\beta_0$ and $\beta_1$ are those values $b_0$ and $b_1$ that minimize the criterion $Q$ for the given sample observations $(X_i, Y_i)$.

Graphically, the deviations correspond to the vertical distance between the responses $Y_i$ and the fitted regression line.

### 6.2 Least Squares Estimators

The estimators $b_0$ and $b_1$ that satisfy the least squares criterion can be found in two basic ways

- Using numerical search procedures to find these values.
- Analytically.

The numerical approach won't be covered here, but if interested look into *' in Numerical Analysis* by Lambers et al. In practice, statistical software implements the numerical algorithm and using it simply a matter of a function call (thank heavens).

Instead, we'll focus on the analytic approach (Note: the equations given in this document won't have any proofs nor derivations, just do them by hand as needed).

*It can be shown* (see note above) that the values $b_0$ and $b_1$ that minimize $Q$ for any particular set of sample data are given by the following simultaneous equations

$$\sum Y_i = nb_0 + b_1 \sum X_i \tag{6.3}$$

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2 \tag{6.4}$$

This pair are known as *normal equations* with our point estimators for the parameters being $b_0$ and $b_1$.

We compute the quantities $\sum Y_i$ and $\sum X_i$ using our sample observations. Then, we can solve both equations simultaneously for the estimators.

They can also be obtained directly using the following

$$b_1 = \frac{\sum X_i Y_i - \frac{\sum X_i Y_i}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \tag{6.5}$$

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \bar{Y} - b_1 \bar{X}. \tag{6.6}$$

## 6.3 Properties of Least Squares Estimators

**Theorem 1** (Gauss-Markov Theorem)**.** *Under the conditions of regression model (4.1), the least squares estimators $b_0$ and $b_1$ are unbiased and have minimum variance among all unbiased linear estimators.*

## 6.4 Point Estimation of Mean Response

Given sample estimators $b_0$ and $b_1$ of the parameters in the regression function 4.3, we estimate the regression function as follows

$$\hat{Y} = b_0 + b_1 X \tag{6.7}$$

where $\hat{Y}$ is the value of the estimated regression function at the level $X$ of the independent variable.

We'll refer to $E[Y]$ as the *mean response*. The mean response is the mean of the probability distribution of $Y$ corresponding to the level $X$. Then, $\hat{Y}$ is a point estimator of the mean response when the level of the independent variable is $X$.

By extension of the Gauss-Markov theorem, $\hat{Y}$ is an unbiased estimator of the mean response with minimum variance in the class of unbiased linear estimators. So, when we're estimating $\hat{Y}$ using sample data, we'll call it $\hat{Y}_i$ for the fitted value for the $i$th case.

$$\hat{Y}_i = b_0 + b_1 X_i \quad i = 1, \ldots, n \tag{6.8}$$

## 6.5 Residuals

The $i$th *residual* is the difference between the observed values $Y_i$ and the corresponding fitted value $\hat{Y}_i$. The $i$th residual is denoted by,

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i) = Y_i - b_0 - b_1 X_i \tag{6.9}$$

One distinction that should be made is the *model error* term versus the residual. The model error term value $\epsilon_i = Y_i - E[Y_i]$ involves the vertical deviation of $Y_i$ from the unknown true regression line and hence is unknown.

On the other hand, the residual is the vertical deviation of $Y_i$ from the fitted value $\hat{Y}_i$ on the estimated regression line and is known. As will be studied it later chapters, residuals are very useful for studying whether a given regression model is appropriate for the data at hand.

### 6.6 Properties of Fitted Regression Line

The regression line fitted by the method of least squares has a number of properties worth noting:

1. The sum of the residuals is zero:
$$\sum_{i=1}^{n} e_i = 0 \tag{6.10}$$

2. The sum of the squared residuals $\sum e_i^2$ is a minimum.

3. The sum of the observed values $Y_i$ equals the sum of the fitted values $\hat{Y}_i$:
$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i \tag{6.11}$$

   It follows from this that the mean of $\hat{Y}_i$ is the same as the mean of $Y_i$.

4. The sum of the weighted residuals is zero when the residual in the $i$th trial is weighted by the level of the independent variable in the $i$th trial:
$$\sum_{i=1}^{n} X_i e_i = 0 \tag{6.12}$$

5. The sum of weighted residuals is zero when the residual in the $i$th trial is weighted by the fitted value of the response variable for the $i$th trial.
$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0 \tag{6.13}$$

6. The regression line always goes through the point $(\bar{X}, \bar{Y})$.

---

## 7 Estimation of Error Terms Variance

Recall, the point estimator for the variance of a single population is given by

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

which is an unbiased estimator of the variance of an infinite population and we lose a degree of freedom because we had to estimate $\bar{Y}$. This sample variance is called a *mean square.*

For the regression model, the deviation of an observation $Y_i$ must be calculated around its own estimated mean $\hat{Y}_i$, hence the deviations are the residuals

$$Y_i - \hat{Y}_i = e_i$$

and the appropriate sum of squares (sum of squared deviations), denoted $SSE$ is

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2 = \sum_{i=1}^{n} e_i^2 \tag{7.1}$$

where $SSE$ stands for *error sum of squares*, *residual sum of squares*, *sum of squares error...there's just a bunch of acceptable names I guess, just know what it is conceptually and numerically.*

Observe it has $n-2$ degrees of freedom associated with it since we lost two of them from having to estimate $\beta_0$ and $\beta_1$.

Using the $SSE$, we can compute an appropriate mean square, denoted $MSE$, which will be our estimator for the variance

$$MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \tag{7.2}$$

where (again a bunch of different names) $MSE$ stands for *error mean square*, *mean square error*, *residual mean square*, just be able to recognize terms used for this value in literature.

It can be shown that $MSE$ is an unbiased estimator of the variance for the regression model. Of course, an estimator of the standard deviation is simply the positive square root of $MSE$.

---

# 8 Normal Error Regression Model

To set up interval estimates and make tests, we'll need to make a final assumption about the functional form of the distribution of the error term $\epsilon_i$. The standard assumption is that the error terms are normally distributed. This assumption of normality simplifies regression analysis and it has the same model as (4.1), except $\epsilon_i \sim N(0, \sigma^2)$.

A major reason why the normality assumption for the error terms is justifiable in many situations is that the error terms frequently represent the effects of many factors omitted explicitly from the model, that do affect the response to some extent and that vary at random without reference to the independent variable $X$.

Insofar as these random effects has a degree of mutual independence, the composite error term representing all of these factors would tend to comply with the central limit theorem and the error term distribution would approach normality as the number of factor effects becomes large.

A second reason is that the estimation and testing procedures for the error terms are based on the $t$ distribution, which is not sensitive to moderate departures from normality. Thus, unless the departures from normality are serious, particularly with respect to skewness, the actual confidence coefficients and risks of errors will be close to the levels for exact normality.

## 8.1 Estimation of Parameters by Method of Maximum Likelihood

When the functional form of the probability distribution of the error terms is specified, estimators of the parameters $\beta_0$, $\beta_1$, and $\sigma^2$ can be obtained by the *method of maximum likelihood*.

This method utilizes the joint probability distribution of the sample observations. When this joint probability distribution is viewed as a function of the parameters, given the particular sample observations, it is called the *likelihood function.*

This is an ugly function that uses the product of normal distributions and as such, I'm not going to attempt to type it up here. What should be known however, is that the values of the above parameters that maximize the likelihood function are the maximum likelihood estimators.

It turns out that the maximum likelihood estimators of $\beta_0$ and $\beta_1$ are the same estimators as provided by the method of least squares...convenient.

One difference between least squares and and ML is the estimator for the variance. The maximum likelihood estimator $\hat{\sigma}^2$ is biased

$$MSE = \frac{n}{n-2}\hat{\sigma}^2$$