Simple Linear Regression

1. Introduction

Regression analysis is "a statistical methodology that utilizes the relation between two or more quantitative variables so that one variable [the *dependent* or *response* variable] can be predicted from the other, or others [the *independent* or *predictor* variables]." (textbook, pg. 2)
Examples:

- can the GPA of a student in college be predicted by their SAT score?
- what is the relationship between the area and the sale price of a house?
- does salt intake increase blood pressure? By how much?

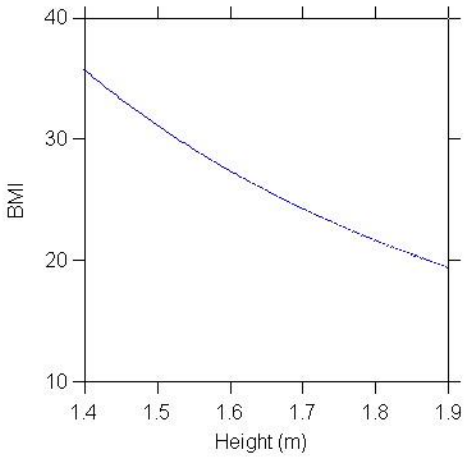
2. Functional & Statistical Relations

A *functional relation* between a dependent variable Y and an independent variable X is an exact relation: the value of Y is uniquely determined when the value of X is specified.

Examples:

- the conversion of temperature from Celsius to Fahrenheit degrees is a linear functional relation $F = 32 + (9/5)C$.

- Body Mass Index (BMI) is calculated as (weight in kilograms) divided by (height in meters) squared. For a constant weight (here 70 kilos) the relation between BMI and height is an example of a non-linear (curvilinear) functional relation.



A *statistical relation* between a dependent variable Y and an independent variable X is an inexact relation; the value of Y is not uniquely determined when the value of X is specified.

The *line* or *curve of statistical relationship* refers to the tendency of Y to vary systematically as a function of X.

3. Simple Linear Regression Model

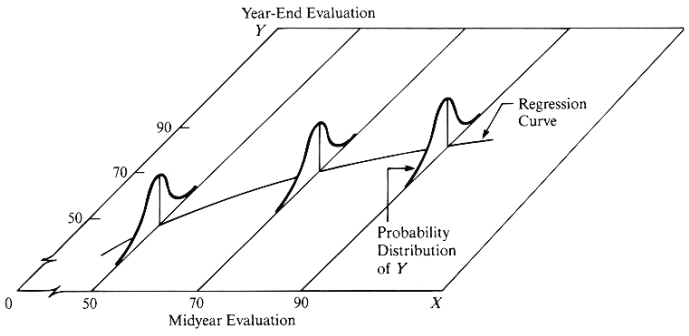
1. Population Model

The idea of a statistical relationship can be formalized as either linear or non-linear (*curvilinear*).

The regression model is the formalization of the idea of a statistical relation; it translates the idea into two components:

- the *regression function* of Y on X represents the relationship of the mean of the probability distribution of Y as a function of X; it captures the notion that Y varies systematically as a function of X (in general the regression function need not be linear)
- the *error term* represents the deviation of Y from the regression function; there is a probability distribution of Y for each level of X that represents the scatter of points around the main trend

FIGURE 1.4 Pictorial Representation of Regression Model.



When the regression function is linear, the simple linear regression model is written

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, 2, ..., n \qquad (1)$$

where

- Y_i is the value of the dependent variable for the ith observation
- X_i is the value of the independent (predictor) variable for the ith element, and is assumed to be a *known constant*
- β_0 and β_1 are parameters (or coefficients)
- ε_i is a random error term

Greek letters β_0, β_1 and ε are used for the regression coefficients and the error term to indicate that the regression model pertains to the *population* from which the sample is drawn; these parameters are not directly known and must be estimated from sample data.

The two components of a statistical relation are translated in the simple regression model as

- $E\{Y\} = \beta_0 + \beta_1 X$ is the *regression function* representing the systematic part of the model; $E\{Y\}$ represents the *expectation* of Y for a given value of the independent variable X
- ε_i is an *error term* representing the deviation of Y_i from the regression function

Model (1) is called *simple* [only 1 independent variable]

2. Assumptions on the Error Term

There are two nested sets of assumptions concerning the distribution of the error term; the second set adds the assumption of normality of the errors.

1. Distribution of Errors Unspecified

ε_i is a random error term

- with mean $E\{\varepsilon_i\} = 0$ (the expected value of each error term ε_i is 0)
- with variance $\sigma^2\{\varepsilon_i\} = \sigma^2$ (the variance of each ε_i is the same at all levels of x, and equal to σ^2, which denotes a constant number)
- such that ε_i and ε_j are uncorrelated (their covariance is zero, i.e., $\sigma\{\varepsilon_i, \varepsilon_j\} = 0$ for all i,j such that $i \neq j$)

2. Distribution of Errors Normal

To the first set add the assumption that distribution of ε_i is normal. Then the entire set of assumptions (including the normality one) can be expressed simply as

- ε_i is independent ~ $N(0,\sigma^2)$, i.e. ε_i is normally distributed with mean 0 and variance σ^2 (which implies that ε_i and ε_j are uncorrelated)

Assumption of normality of the errors is necessary to theoretically justify statistical inference especially in small samples. But most properties of least squares estimators of model parameters do not depend on the normality assumption.

3. Components of Simple Regression Model

The *regression function* or *response function* represents the systematic part of the model; it relates the expected value (or mean) $E\{Y\}$ of Y to the value of the independent variable X. The graph of the regression function is called the *regression line*. In the simple linear regression model the regression function for any value of X is

$$E\{Y_i\} = E\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \beta_0 + \beta_1 X_i + E\{\varepsilon_i\} = \beta_0 + \beta_1 X_i$$

since by assumption $E\{\varepsilon_i\} = 0$ and $\beta_0 + \beta_1 X_i$ is constant.

The parameters β_0 and β_1 are called *regression coefficients* or *regression parameters*.

The meaning of each coefficient is as follows

- the *slope* β_1 indicates the change in E{Y} per unit increase in X
- the *intercept* β_0 corresponds to the value of E{Y} at X=0

With respect to the regression model the population variance of Y_i is

$$\sigma^2\{Y_i\} = \sigma^2\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \sigma^2\{\varepsilon_i\} = \sigma^2$$

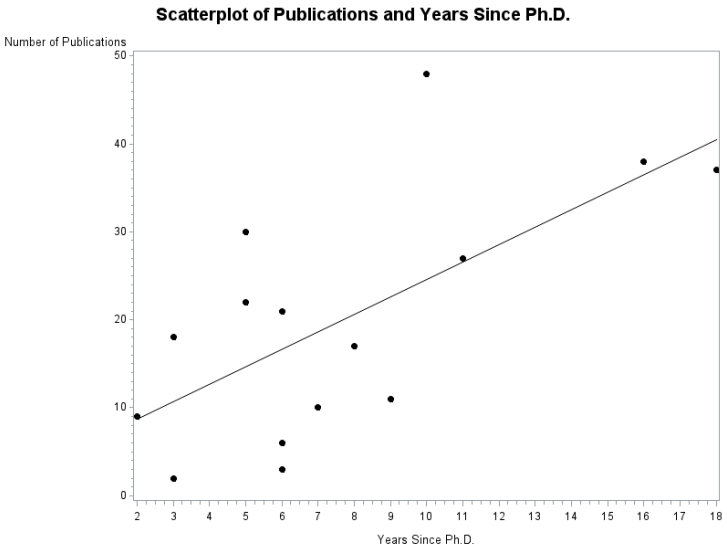
where σ^2 is the variance of ε_i. This is because ε_i is the only random variable in the expression, and the variance of the error ε_i is assumed to be the same and equal to σ^2 regardless of the value of X.

The quantities β_0 and β_1 and σ^2 are the *parameters* of the regression model; they have to be estimated from the data. (In reality one estimates β_0 and β_1 and then the estimate of σ^2 is obtained as a by-product – shown later)

4. An Example of Simple Linear Regression

Consider two variables: number of years since a professor has completed his or her Ph.D. and number of publications the professor currently has published.

A scatterplot of the data:



The solid line is the *estimated regression line*. It is represented by the equation

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

where $\hat{Y}$ (called "Y hat") represents the vertical coordinate of a point on the regression line corresponding to horizontal coordinate X. The coefficient b_0 and b_1 are calculated by the method of least squares (explained below).

The model implies that for each observation in the sample the vertical coordinate y_i of a point is given by the formula

$$Y_i = \hat{Y}_i + e_i \quad \text{or}$$

$$Y_i = b_0 + b_1 X_i + e_i$$

where e_i corresponds to the vertical deviation between the observed value Y_i and $\hat{Y}_i$ (called the *fitted value* or *predictor* of Y) implied by the regression line. e_i is called the *residual* for observation i, b_0 and b_1 are the (estimated) *regression coefficients*. Their meaning is the same as that of the population parameters, i.e.

- the *slope* b_1 measures the predicted change in Y per unit increase in X
- the *intercept* b_0 corresponds to the predicted value of $\hat{Y}$ for X=0, i.e., b_0 is the value of Y at the point where the regression line crosses the vertical line at X=0; b_0 may not be substantively meaningful if the scope of the model does not include X=0

For the Ph.D. data the slope b_1= 1.983 in the regression of Y (pubs) on X (time) means that for each additional year since completing a Ph.D., the number of total publications increases by about 2. The intercept b_0 = 4.731 means that the total number of publications for a professor who just completing his or her Ph.D. (X=0) is about 5.

Note that the simple regression model establishes an asymmetry between the dependent variable Y and independent variable X, because deviations are measured along the dependent variable dimension (usually the vertical axis). In general a different regression line is obtained if one exchange Y and X in their roles. The choice of one variable as dependent and the other as independent is a substantive choice. (Correlational models do not assume this asymmetry.)

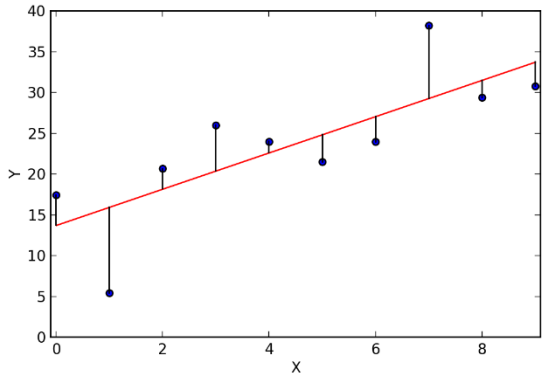
4. Least Squares (OLS)

1. Estimation of b_0 and b_1

The coefficients of the regression line (regression coefficients) are originally unknown. They can be estimated from a sample containing *n* observations on Y and X with the *method of least squares*. The method of least squares (or OLS for o*rdinary least squares*) consists in finding values for b_0 and b_1 that minimize the sum (over all observations) of the squared vertical deviations e_i of the observed value Y_i from the predicted value $\hat{Y}_i$ on the regression line. Mathematically, one wants to find the values of b_0 and b_1 that minimize the quantity Q defined as

$$Q = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

The figure below shows the vertical deviations e_i that are squared and summed up to evaluate Q.



To minimize Q one could: (1) use a "brute force" numerical search using a grid of values for b_0 and b_1 or (2) use the analytical solution, i.e. the values b_0 and b_1 that minimize Q are given by the formulas

$$b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

$$b_0 = \overline{Y} - b_1\overline{X}$$

called the *normal equations*. All the sums are over all observations (from i=1 to n). One calculates b_1 first, then b_0.

Calculating the regression coefficients by hand for the Ph.D. example:

$b_1 = (581.6667)/(293.3333) = 1.983$
$b_0 = (19.9333) - (1.983)(7.6667) = 4.731$

Note that the slope b_1 can also be calculated using the correlation coefficient, the standard deviation of X and the standard deviation of Y.

$b_1 = r(s_Y/s_X)$

$b_0 = \overline{Y} - b_1\overline{X}$

However, SAS can perform these calculations using “proc reg.”

```
proc reg data=phd;
model pubs=time;
run;
```

5. (Optional) Derivation of Least Squares Formulas and Properties of LS Residuals

1. Derivation of Formulas for b_0 and b_1 (Uses Calculus)

The sum of squared deviations

$$Q = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

can be viewed as a function of two variables, b_0 and b_1. To find the values of b_0 and b_1 that minimize Q one differentiates the function in turn with respect to b_0 and with respect to b_1, obtaining

$$\partial Q / \partial \beta_0 = -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)$$

$$\partial Q / \partial \beta_1 = -2 \sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i)$$

The values b_0 and b_1 that minimize Q are found by setting the derivatives to zero, as

$$-2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = 0$$
$$-2 \sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i) = 0$$

and solving for b_0 and b_1. Solving is done by simplifying and expanding these equations and rearranging the terms to produce the *normal equations*

$$\sum Y_i = n b_0 + b_1 \sum X_i$$
$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$$

2. Properties of OLS Residuals

Properties of predicted values $\hat{Y}_i$ and residuals e_i:

- $\Sigma e_i = 0$
- Σe_i^2 is minimum (by LS)
- $\Sigma X_i e_i = 0$
- $\Sigma \hat{Y}_i e_i = 0$

Relationships between X_i, Y_i, $\hat{Y}_i$ and e_i:

- The residuals e_i represent that part of Y_i that cannot be predicted from X_i. Therefore, e_i must be uncorrelated with X_i.
- Because $\hat{Y}_i$ is an exact linear function of X_i, the correlation between $\hat{Y}_i$ and X_i must be perfect.
- Because $\hat{Y}_i$ is an exact linear function of X_i, the correlation of $\hat{Y}_i$ with Y_i must be the same as the correlation of X_i with Y_i.

6. Standardized Regression Coefficient

1. Calculation

The *standardized regression coefficient* b_1^* is calculated as: $b_1^* = b_1(s_X/s_Y)$
i.e., b_1^* is equal to b_1 multiplied by the standard deviation of X and divided by the standard deviation of Y. Thus in the simple linear regression model the standardized regression coefficient is the same as the correlation coefficient:

$b_1^* = (s_{XY}/s_X^2)(s_X/s_Y) = s_{XY}/(s_X s_Y) = r$, but this is no longer true in the multiple regression model.

2. Interpretation

In the Ph.D. study the regression coefficient b_1 is 1.983; the standard deviations of X and Y are 1.182 and 3.569, respectively. Thus the standardized coefficient of seasonality is 1.983(1.182/3.569) = 0.657. Thus an increase of one SD in X is associated with an increase of 0.657 SD deviations of Y.

Standardized coefficients are especially useful in the multiple regression model, where they permit comparing the *relative magnitudes* of the coefficients of independent variables measured in different units (such as a variable measured in years, and another measured in thousands of dollars).

7. Causal Interpretation

Data for regression analysis comes from two kinds of sources.

1. *Observational data* are data obtained from nonexperimental studies so that values of X are not controlled. An example is life expectancy of countries as a function of literacy. Observational data do not directly offer strong support for causal interpretations.
2. *Experimental data* are measured from experimental units that are randomly assigned to *treatments*, i.e., different values of the independent variable(s) X set by the experimenter. Experimental data allow stronger causal inferences.

8. Estimation of Error Terms Variance σ^2

The error sum of squares: $SSE = \Sigma e_i^2$
The error mean square: $s^2 = MSE = SSE/(n-2)$

MSE is an estimate of the error terms variance σ^2 or $\sqrt{MSE}$ is an estimate for σ

Ex: Calculate the MSE in the Ph.D. example: MSE = 1521.5148/(15-2) = 117.0396