STAC67H: Regression Analysis

Fall, 2014

Instructor: Jabed Tomal

Department of Computer and Mathematical Sciences
University of Toronto Scarborough
Toronto, ON
Canada

September 3, 2014

Definition:

Regression analysis: A statistical methodology that utilizes the relation between two or more quantitative variables so that a response or outcome variable can be predicted from the other, or others.

Examples:

1. Goal: Predicting sales of a product. Dependent Variable: Sales – Independent Variable: Amount of advertising expenditure.

Examples:

1. Goal: Predicting sales of a product. Dependent Variable: Sales – Independent Variable: Amount of advertising expenditure.
2. Goal: Predicting the performance of an employee on a job. Dependent Variable: Performance – Independent Variable: Aptitude tests score.

Examples:

1. Goal: Predicting sales of a product. Dependent Variable: Sales – Independent Variable: Amount of advertising expenditure.
2. Goal: Predicting the performance of an employee on a job. Dependent Variable: Performance – Independent Variable: Aptitude tests score.
3. Goal: Predicting size of the vocabulary of a child. Dependent Variable: Size of Vocabulary – Independent Variables: Age of the child and education of the parents.

Examples:

1. Goal: Predicting sales of a product. Dependent Variable: Sales – Independent Variable: Amount of advertising expenditure.
2. Goal: Predicting the performance of an employee on a job. Dependent Variable: Performance – Independent Variable: Aptitude tests score.
3. Goal: Predicting size of the vocabulary of a child. Dependent Variable: Size of Vocabulary – Independent Variables: Age of the child and education of the parents.
4. Goal: Predicting the length of hospital stay of a surgical patient. Dependent Variable: The time stay in the hospital – Independent Variable: Severity of the operation.

Functional / Mathematical Relation:

Definition: A **functional relation** between two variables is expressed by a mathematical formula:

$$Y = f(X),$$

where Y is the **dependent** variable, and X is the **independent** variable.

Example: Functional Relation

$$Y = 2X,$$

where Y is **Dollar Sales** of a product, and X is **Units Sold**. Unit price?

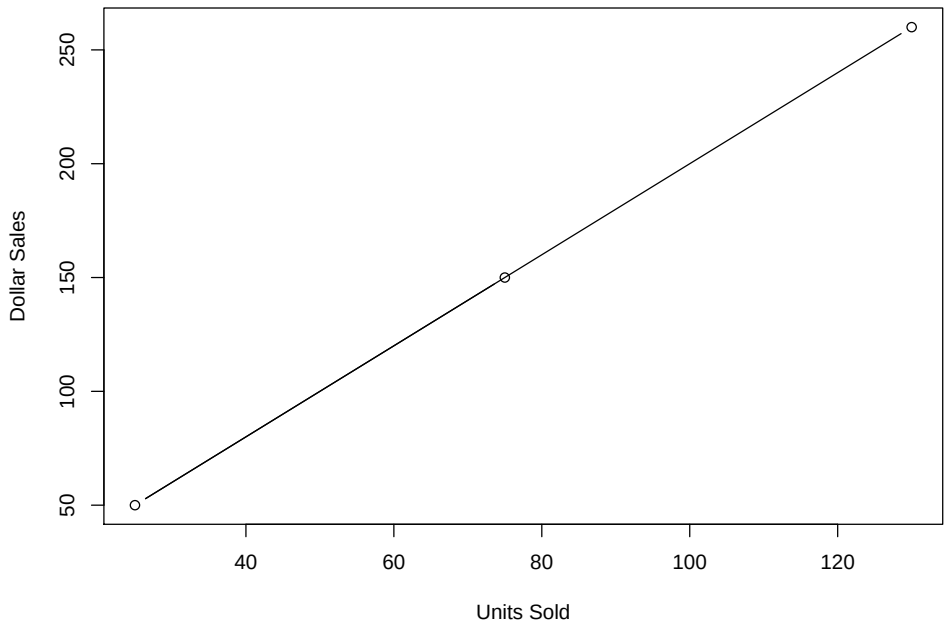

Figure: Example of Functional Relation.

Characteristics: All points fall directly on the line.

Statistical Relation

Definition: A **statistical relation** is not a perfect one: unlike a functional relation, the observations for a statistical relation do not fall directly on the curve of relationship.

Example: Performance evaluations of 10 employees were obtained at midyear and at year-end. Year-end evaluations are taken as the *dependent* or *response variable* Y, and midyear evaluations as the *independent*, *explanatory*, or *predictor variable* X.

Scatter Plot:

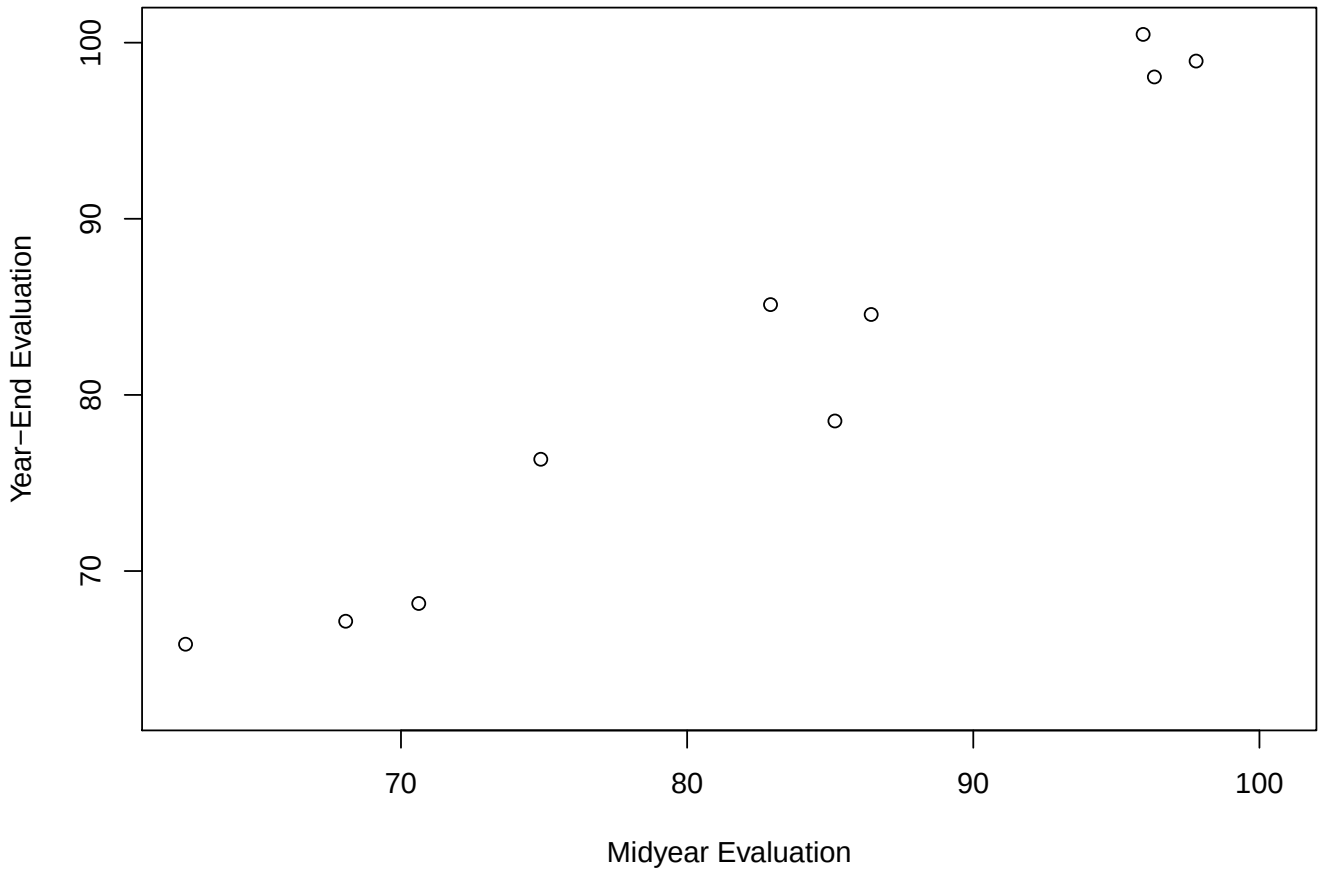

Figure: Statistical Relation between Midyear Performance Evaluation and Year-End Evaluation.

Scatter Plot and Line of Statistical Relationship:

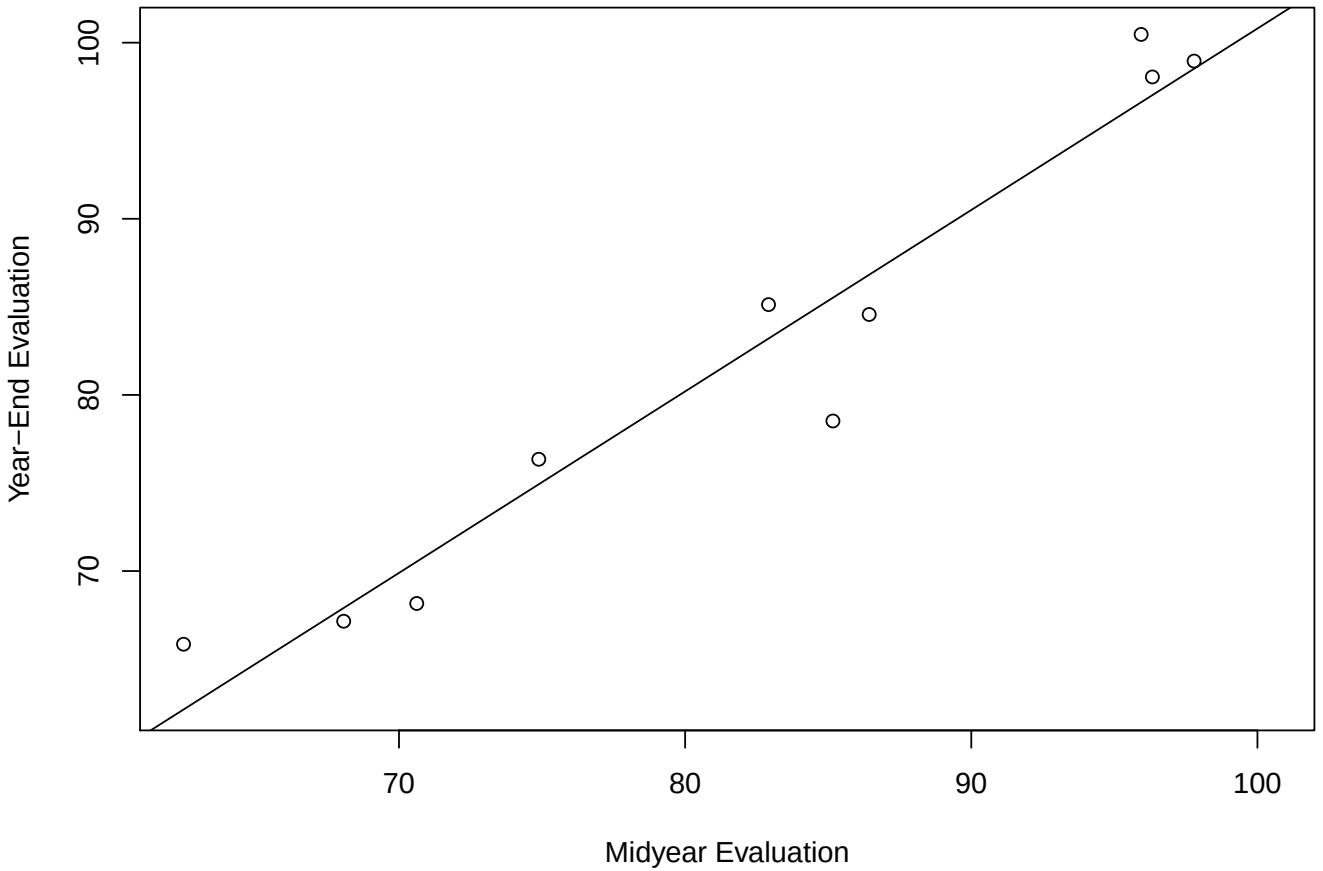

Figure: Statistical Relation between Midyear Performance Evaluation and Year-End Evaluation.

Basic Concepts:

Two essential ingredients of a statistical relation are:

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. The regression funtion of Y on X is:

$$E(Y|X) = \beta_0 + \beta_1 X,$$

where β_0 and β_1 are called the regression coefficients.

Construction of Regression Models:

- **Selection of Predictor Variables:** Only a limited number of good explanatory or predictor variables should be included in a regression model, i.e., a simple yet powerful regression model is preferred.

Construction of Regression Models:

- **Selection of Predictor Variables:** Only a limited number of good explanatory or predictor variables should be included in a regression model, i.e., a simple yet powerful regression model is preferred.
- **Functional Form of RM:** The choice of functional form of a regression model depends on the choice of predictor variables. Sometimes, relevant theory helps us finding an appropriate functional form. Often, exploratory analysis appeared handy to choose a function form of a regression model.

Construction of Regression Models:

- **Selection of Predictor Variables:** Only a limited number of good explanatory or predictor variables should be included in a regression model, i.e., a simple yet powerful regression model is preferred.
- **Functional Form of RM:** The choice of functional form of a regression model depends on the choice of predictor variables. Sometimes, relevant theory helps us finding an appropriate functional form. Often, exploratory analysis appeared handy to choose a function form of a regression model.
- **Scope of Model:** The scope of a regression model is somewhat limited to the range of predictor variable(s). The shape of the regression function substantially outside the range of predictor variable(s) would be in serious doubt because the investigation provided no evidence as to the nature of the statistical relation in that regions.

Uses of Regression Analysis:

The major purposes of regression analysis are:

1. **Description** (relation between *family income* and *expenditure*)

Uses of Regression Analysis:

The major purposes of regression analysis are:

1. **Description** (relation between *family income* and *expenditure*)
2. **Control** (relation between *profit* and *advertisement costs*)

Uses of Regression Analysis:

The major purposes of regression analysis are:

1. **Description** (relation between *family income* and *expenditure*)
2. **Control** (relation between *profit* and *advertisement costs*)
3. **Prediction** (prediction of rain tomorrow based on historical data).

Regression and Causality

A strong regression model does not necessarily imply causality.

- Example: relation between *size of vocabulary* (X) and *writing speed* (Y) of children is often found positive. Such positive relation does not indicate that Y depends on X. Here, some other variables, such as *age of the child* and *amount of education*, affect both the vocabulary (X) and writing speed (Y).