Applied Linear Statistical Models
Outline

Darshan Patel

January 5, 2020

Contents

Chapter 1

Simple Linear Regression

1.1 Linear Regression with One Predictor Variable

1.1.1 Relations between Variables

- Regression analysis is 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
- A functional relation between two variables is expressed as follows: if X denotes the independent variable and Y the dependent variable, a functional relation is of the form

  $$Y = f(X)$$

  Given a particular value of X, the function f indicates the corresponding value of Y
- A statistical relation, unlike a functional relation, is not a perfect one; in general, the observations for a statistical relation do not fall directly on he curve of relationship
- Statistical relations can be highly useful, even though they do not have the exactitude of a functional relation

1.1.2 Regression Models and their Uses

- A regression model is a formal means of expressing the two essential ingredients of a statistical relation:
  - A tendency of the response variable Y to vary with the predictor variable X in a systematic fashion
  - A scattering of points around the curve of statistical relationship
- These two characteristics are embodied in a regression model by postulating that:
  - There is a probability distribution of Y for each level of X
  - The means of these probability distributions vary in some systematic fashion with X
- The systematic relationship between X and Y is called the regression function of Y on X; the graph of the regression function is called the regression curve

- Regression models may differ in the form of the regression function (linear, curvilinear), in the shape of the probability distribution of Y (symmetrical, skewed), and in other ways
- Regression models may contain more than one predictor variable
- Since reality must be reduced to manageable proportions whenever models are constructed, only a limited number of explanatory or predictor variables can, or should, be included in a regression model for any situation of interest
- A central problem in many exploratory studies is therefore that of choosing, for a regression model, a set of predictor variables that is "good" in some sense for the purposes of the analysis
- The choice of the functional form of the regression relation is tied to the choice of the predictor variables; sometimes, relevant theory may indicate the appropriate functional form
- More frequently, the functional form of the regression relation is not known in advance and must be decided upon empirically once the data have been collected
- Linear or quadratic regression functions are often used as satisfactory first approximations to regression functions of unknown nature
- In formulating a regression model, the coverage is usually restricted to some interval or region of values of the predictor variable(s) which is determined either by the design of the investigation or by the range of data at hand
- Regression analysis serves three major purposes: description, control and prediction
- The existence of a statistical relation between the response variable Y and the explanatory or predictor variable X does not imply in any way that Y depends casually on X

1.1.3 Simple Linear Regression Model with Distribution of Error Terms Unspecified

- A basic regression model where there is only one predictor variable and the regression function is linear can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where

Y_i is the value of the response variable in the ith trial

β_0 and β_1 are parameters

X_i is a known constant, namely, the value of the predictor variable in the ith trial

ε_i is a random error with mean $\mathrm{E}[\varepsilon_i] = 0$ and variance $\mathrm{Var}[\varepsilon_i] = \sigma^2$; ε_i and ε_j are uncorrelated so that their covariance is zero (i.e., $\mathrm{Cov}[\varepsilon_i,\ \varepsilon_j] = 0$ for all $i, j; i \neq j$)

$i = 1, \ldots, n$

- This regression model is said to be simple, linear in its parameters, and linear in the predictor variable

- Important Features of the Model

  1. The response Y_i in the ith trial is the sum of two components: (1) the constant term $\beta_0 + \beta_1 X_i$ and (2) the random term ε_i; hence Y_i is a random variable
  2. Since $\mathrm{E}[\varepsilon_i] = 0$, then

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

  Thus, the response Y_i, when the level of X in the ith trial is X_i, comes from a probability distribution whose mean is

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

  3. The response Y_i in the ith trial exceeds or falls short of the value of the regression function by the error term amount ε_i
  4. The error terms ε_i are assumed to have constant variance σ^2 and so the responses Y_i have the same constant variance

  $$\mathrm{Var}[Y_i] = \sigma^2$$

- The parameters β_0 and β_1 in the regression model are called regression coefficients
- The parameter β_1 is the slope of the regression line (indicating the change in the mean of the probability distribution of Y per unit change in X)
- The parameter β_0 is the Y intercept of the regression line
- When the scope of the model includes $X = 0$, β_0 gives the mean of the probability distribution of Y at $X = 0$; when the scope of the model does not cover $X = 0$, β_0 does not have any particular meaning as a separate term in the regression model
- The simple linear regression model can be written equivalently as follows: let X_0 be a constant identically equal to 1, then

  $$Y_i = \beta_0 X_0 + \beta_1 X_i + \varepsilon_i \text{ where } X_0 \equiv 1$$

  This version of the model associates an X variable with each regression coefficient
- An alternative modification is to use for the predictor variable the deviation $X_i - \overline{X}$ rather than X_i, then

  $$\begin{aligned} Y_i &= \beta_0 + \beta_1(X_i - \overline{X}) + \beta_1 \overline{X} + \varepsilon_i \\ &= (\beta_0 + \beta_1 \overline{X}) + \beta_1(X_i - \overline{X}) + \varepsilon_i \\ &= \beta_0^* + \beta_1(X_9 - \overline{X}) + \varepsilon_i \end{aligned}$$

  Thus this alternative model is

  $$Y_i = \beta_0^* + \beta_1(X_i - \overline{X}) + \varepsilon_i$$

  where

  $$\beta_0^* = \beta_0 + \beta_1 \overline{X}$$

1.1.4 Data for Regression Analysis

- Data for regression analysis may be obtained from nonexperimental or experimental studies
- Observational data are data obtained from nonexperimental studies; such studies do not control he explanatory or predictor variable(s) of interest
- A major limitation of observational data is that they often do not provide adequate information about cause and effect relationships
- Frequently, it is possible to conduct a controlled experiment to provide data from which the regression parameters can be estimated
- When control over the explanatory variable(s) is exercised through random assignments, the resulting experimental data provide much stronger information about cause and effect relationships than do observational data; the reason is that randomization tends to balance out the effects of any other variables that might affect the response variable
- In the terminology of experimental design, a treatment is the object being measured and the experimental units are the subjects of the study, from whom the treatment is done on and measured; control over the explanatory variable(s) then consists of assigning a treatment to each of the experimental units by means of randomization
- The most basic type of statistical design for making randomized assignments of treatments to experimental units (or vice versa) is the completely randomized design; with this design, the assignments are made completely at random
- This complete randomization provides that all combinations of experimental units assigned to the different treatments are equally likely, which implies that every experimental unit has an equal change to receive any one of the treatment
- A completely randomized design is particularly useful when the experimental units are quite homogeneous; this design is very flexible; it accommodates any number of treatments and permits different sample sizes for different treatments
- Its chief disadvantage is that, when the experimental units are heterogeneous, this design is not as efficient as some other statistical designs

1.1.5 Overview of Steps in Regression Analysis

- The regression models given in the following chapters can be used either for observational data or for experimental data from a completely randomized design (regression analysis can also utilize data from other types of experimental designs, bu the regression models presented here will need to be modified)
- Typical Strategy for Regression Analysis
    1. Start
    2. Exploratory data analysis
    3. Develop one or more tentative regression models
    4. Is one or more of the regression models suitable for the data at hand?

  - If yes, continue
  - If no, revise regression models and/or develop new ones and answer the question again
5. Identify most suitable model
6. Make inferences on basis of regression model
7. Stop

1.1.6 Estimation of Regression Function

- The observational or experimental data to be used for estimating the parameters of the regression function consist of observations on the explanatory or predictor variable X and the corresponding observations on the response variable Y; for each trial, there is an X observation and a Y observation; denote the (X, Y) observations for he first trial as (X_1, Y_1), for the second trial as (X_2, Y_2), and in general for the ith trial as (X_i, Y_i), where $i = 1, \ldots, n$
- For the observations (X_i, Y_i) for each case, the method of least squares considers the deviation of Y_i from its expected value $Y_i - (\beta_0 + \beta_1 X_i)$; in particular, the method of least squares requires considering the sum of the n squared deviations; this criterion is denoted by Q:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

  According to the method of least squares, the estimators of β_0 and β_1 are those values b_0 and b_1, respectively, that minimize the criterion Q for the given sample observations $(X_1, Y_1), \ldots, (X_n, Y_n)$
- The estimators b_0 and b_1 that satisfy the least squares criterion can be found in two basic ways:
  - Numerical search procedures can be used hat evaluate in a systematic fashion the least squares criterion for different estimates b_0 and b_1 until the ones that minimize Q are found
  - Analytical procedures can often be used to find the values of b_0 and b_1 that minimize Q; this is feasible when the regression model is not mathematically complex
- Using the analytical approach, the values b_0 and b_1 that minimize Q for the simple linear regression model are given by the following simultaneous equations:

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

  These two equations are called normal equations; b_0 and b_1 are called point estimators of β_0 and β_1 respectively
- The normal equations can be solved simultaneously for b_0 and b_1:

$$b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$
$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \overline{Y} - b_1 \overline{X}$$

  where $\overline{X}$ and $\overline{Y}$ are the means of the X_i and Y_i observations respectively

- Derivation of above result: for given sample observations (X_i, Y_i), the quantity Q is a function of β_0 and β_1; the values of β_0 and β_1 that minimize Q can be derived by differentiating Q with respect to β_0 and β_1:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i)$$
$$\frac{\partial Q}{\partial \beta_1} = -2 \sum X_i (Y_i - \beta_0 - \beta_1 X_i)$$

  Setting these partial derivatives to zero, using b_0 and b_1 to denote the particular values of β_0 and β_1 that minimize Q and simplifying, the following is obtained

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

  Expanding this, the following is true:

$$\sum Y_i - n b_0 - b_1 \sum X_i = 0$$
$$\sum X_i Y_i - b_0 \sum X_i - b_2 \sum X_i^2 = 0$$

  By rearranging terms, the normal equations are obtained

- Gauss-Markov Theorem: Under the conditions of the simple linear regression model the least squares estimators b_0 and b_1, as given above, are unbiased and have minimum variance among all unbiased linear estimators

- This theorem states first that b_0 and b_1 are unbiased estimators and so

$$\mathrm{E}[b_0] = \beta_0 \quad \mathrm{E}[b_1] = \beta_1$$

  so that neither estimator tends to overestimate or underestimate systematically

- Second, the theorem states that the estimators b_0 and b_1 are more precise (o.e., their sampling distributions are less variable) than any other estimators belonging to the class of unbiased estimators that are linear functions of the observations $Y_1, \ldots, Y_n$; the estimators b_0 and b_1 are such linear functions of the Y_i

- Given sample estimators b_0 and b_1 of the parameters in the regression function, $\mathrm{E}[Y] = \beta_0 + \beta_1 X$, the regression function is estimates as

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

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

- A value of the response variable is called a response while $\mathrm{E}[Y]$ is called the mean response; thus, the mean response stands for the mean of the probability distribution of Y corresponding to the level X of the predictor variable

- $\hat{Y}$ is a point estimator of the mean response when the level of the predictor variable is X
- As an extension of the Gauss-Markov Theorem, $\hat{Y}$ is an unbiased estimator of $\mathrm{E}[Y]$, with minimum variance in the class of unbiased linear estimators
- Let $\hat{Y}_i$ be the fitted value for the ith case

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

  Thus the fitted value $\hat{Y}_i$ is to viewed in distinction to the observed value Y_i
- The ith residual is the difference between the observed value Y_i and the corresponding fitted value $\hat{Y}_i$; this residual is denoted by e_i and is defined as follows:

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

- For the simple linear regression model, the residual e_i becomes

$$e_i = Y_i - (b_0 + b_1 X_i) = Y_i - b_0 - b_1 X_i$$

- The model error term value $\varepsilon_i = Y_i - \mathrm{E}[Y_i]$ involves the vertical deviation of Y_i from the unknown true regression line hence is unknown; the residual $e_i = Y_i - \hat{Y}_i$ is the vertical deviation of Y_i from the fitted value $\hat{Y}_i$ on the estimated regression line, and it is known
- Properties of Fitted Regression Line
  1. The sum of the residuals is zero
  $$\sum_{i=1}^{n} e_i = 0$$
  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 $\hat{Y}_i$:
  $$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$$
  4. The sum of the weighted residuals is zero when he residual in the ith trial is weighted by the level of the predictor variable in the ith trial:
  $$\sum_{i=1}^{n} X_i e_i = 0$$
  5. The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the fitted value of the response variable for the ith trial:
  $$\sum_{i=1}^{n} \hat{Y}_i e_i = 0$$
  6. The regression line always goes through the point $(\overline{X}, \overline{Y})$

1.1.7 Estimation of Error Terms Variance σ^2

- The variance σ^2 of the error terms ε_i in the regression model needs to be estimated to obtain an indication of the variability of the probability distribution of Y
- The variance σ^2 of a single population is estimated by the sample variance s^2 as follows:

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

  where the sum is called a sum of squares and $n-1$ is the degrees of freedom; this number is $n-1$ because one degree of freedom is lost by using $\overline{Y}$ as an estimate of the unknown population mean μ
- This estimator is the usual sample variance which is an unbiased estimator of the variance σ^2 of an infinite population
- The sample variance is often called a mean square because a sum of squares has been divided by the appropriate number of degrees of freedom
- For the regression model, the variance for each observation Y_i is σ^2, the same as that of each error term ε_i; a sum of squared deviations are needed to be calculated but note that the Y_i now come from different probability distributions with different means that depend upon the level X_i; thus, the deviation of an observation Y_i must be calculated around its own estimated mean $\hat{Y}_i$
- The deviations are the residuals

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

  and the appropriate sum of squares, denoted by SSE, is

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

  where SSE stands for error sum of squares or residual sum of squares
- The SSE has $n-2$ degrees of freedom because both β_0 and β_1 had to be estimated in obtaining the estimated means $\hat{Y}_i$
- The appropriate mean square, denoted by MSE or s^2 is

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

  where MSE stands for error mean square or residual mean square
- The MSE is an unbiased estimator for σ^2 for a regression model

$$\text{E}\left[\text{MSE}\right] = \sigma^2$$

- An estimator of the standard deviation σ is simply $s = \sqrt{\text{MSE}}$, the positive square root of the MSE

1.1.8 Normal Error Regression Model

- The normal error regression model is as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where

Y_i is the observed response in the ith trial

X_i is a known constant, the level of the predictor variable in the ith trial

β_0 and β_1 are parameters

ε_i are independent $N(0, \sigma^2)$

$i = 1, \ldots, n$

- The symbol $N(0, \sigma^2)$ stands for normally distributed with mean 0 and variance σ^2
- The normal error model is the same as the regression model with unspecified error distribution, except this one assumes that the errors ε_i are normally distributed
- Since the errors are normally distributed, the assumption of uncorrelatedness of the ε_i in the regression model becomes one of independence in the normal error model
- This model implies that the Y_i are independent normal random variables, with mean $\mathrm{E}[Y_i] = \beta_0 + \beta_1 X_i$ and variance σ^2
- The normality assumption for the error term is justifiable in many situations because the error terms frequently represent the effects of factors omitted from the model that affect the response to some extent and that vary at random without reference to the variable X
- A second reason why the normality assumption of the error terms is frequently justifiable is that the estimation and testing procedures are based on the t distribution and are usually only sensitive to large departures from normality
- When the functional form of the probability distribution of the error terms is specified, estimators of the parameters β_0, β_1 and σ^2 can be obtained using the method of maximum likelihood; this method chooses as estimates those values of the parameters that are most consistent with the sample data
- The method of maximum likelihood uses the product of the densities as the measure of consistency of the parameter value with the sample data; the product is called the likelihood value of the parameter value and is denoted by $L(\cdot)$ where $\cdot$ is the parameter being estimated; if the value of $\cdot$ is consistent with the sample data, the densities will be relatively large and so will be the likelihood value; if the value of $\cdot$ is not consistent with the data, the densities will be small and the product $L(\cdot)$ will be small
- The method of maximum likelihood chooses as the maximum likelihood estimate that value of $\cdot$ for which the likelihood value is largest; there are two methods of finding the estimates: by a systematic numerical search or by use of an analytical solution
- The product of the densities viewed as a function of the unknown parameters is called the likelihood function

- In general, the density of an observation Y_i for the normal error regression model is as follows, utilizing the fact that $\mathrm{E}[Y_i] = \beta_0 + \beta_1 X_i$ and $\mathrm{Var}[Y_i] = \sigma^2$:

$$f_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right]$$

- The likelihood function for n observations $Y_1, \ldots, Y_n$ is the product of the individual densities; since the variance σ^2 of the error terms is usually unknown, the likelihood function is a function of three parameters β_0, β_1 and σ^2

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right] \end{aligned}$$

- The values of β_0, β_1 and σ^2 that maximize this likelihood function are the maximum likelihood estimators and are denoted by $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$ respectively; these estimators are calculated analytically and are as follows:

Parameter	Maximum Likelihood Estimator
β_0	$\hat{\beta}_0 = b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \overline{Y} - b_1\overline{X}$
β_1	$\hat{\beta}_1 = b_1 = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2}$
σ^2	$\hat{\sigma}^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n}$

- Thus, the maximum likelihood estimators of β_0 and β_1 are the same estimators as those provided by the methods of least squares; the maximum likelihood estimator $\hat{\sigma}^2$ is biased and ordinarily the unbiased MSE or s^2 is used

- The unbiased MSE or s^2 differs but slightly from the maximum likelihood estimator $\hat{\sigma}^2$, especially if n is not small

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

- Since the maximum likelihood estimators of $\hat{\beta}_0$ and $\hat{\beta}_1$ are the same as the least squares estimators b_0 and b_1, they have the properties of all least squares estimators:
  - There are unbiased
  - They have minimum variance among all unbiased linear estimators
- In addition, the maximum likelihood estimators b_0 and b_1 for the normal error regression model have other desirable properties:
  - They are consistent
  - They are sufficient
  - They are minimum variance unbiased; that is, they have minimum variance in the class of all unbiased estimators (linear or otherwise)

- Derivation of maximum likelihood estimators: take partial derivatives of L with respect to β_0, β_1 and σ^2, equating each of the partials to zero and solving the system of equations obtained; work with $\log_e L$ rather than L since both are maximized for the same values of β_0, β_1 and σ^2:

$$\log L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(Y_i - \beta_0 - \beta_1 X_i)^2$$

The partial derivatives are as shown:

$$\frac{\partial \log L}{\partial \beta_0} = \frac{1}{\sigma^2}\sum(Y_i - \beta_0 - \beta_1 X_i)$$
$$\frac{\partial \log L}{\partial \beta_1} = \frac{1}{\sigma^2}\sum X_i(Y_i - \beta_0 - \beta_1 X_i)$$
$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(Y_i - \beta_0 - \beta_1 X_i)^2$$

Setting these partial derivatives equal to zero and replacing β_0, β_1 and σ^2 by the estimators $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$, and after some simplifications:

$$\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$
$$\sum X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$
$$\frac{\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n} = \hat{\sigma}^2$$

The first two equations are identical to the earlier least squares normal equations and the last one is the biased estimator of σ^2 as given earlier

1.2 Inferences in Regression and Correlation Analysis

Throughout this chapter (excluding Section 2.11), and in the remainder of Part I unless otherwise stated, assume that the normal error regression model is applicable. This model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where:

β_0 and β_1 are parameters

X_i are known constants

ε_i are independent $N(0, \sigma^2)$

1.2.1 Inferences Concerning β_1

- At times, tests concerning β_1 are of interest, particularly one of the form:

$$H_0: \ \beta_1 = 0$$
$$H_A: \ \beta_1 \neq 0$$

- The reason for interest in testing whether or not $\beta_1 = 0$ is that, when $\beta_1 = 0$, there is no linear association between Y and X

- When $\beta_1 = 0$, the regression line is horizontal and the means of the probability distributions of Y are therefore all equal, namely:

$$\mathrm{E}[Y] = \beta_0 + (0)X = \beta_0$$

- $\beta_1 = 0$ for the normal error regression model also implies that there is no relation of any type between Y and X since the probability distributions of Y are then identical at all levels of X

- The point estimator b_1 is as follows:

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

- The sampling distribution of b_1 refers to the different values of b_1 that would be obtained with repeated sampling when the levels of the predictor variable X are held constant from sample to sample

- For the normal error regression model, the sampling distribution of b_1 is normal, with mean and variance: $\mathrm{E}[b_1] = \beta_1$ and $\mathrm{Var}[b_1] = \frac{\sigma^2}{\sum(X_i - \overline{X})^2}$

- b_1 can be expressed as follows:

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

  where $k_i = \frac{X_i - \overline{X}}{\sum(X_i - \overline{X})^2}$; Observe that the k_i are a function of the X_i are therefore are fixed quantities since the X_i are fixed; hence, b_1 is a linear combination of the Y_i where the coefficients are solely a function of the fixed X_i

- The coefficients k_i have a number of interesting properties

$$\sum k_i = 0 \qquad \sum k_i X_i = 1 \qquad \sum k_i^2 = \frac{1}{\sum(X_i - \overline{X})^2}$$

- To show that b_1 is a linear combination of the Y_i with coefficients k_i, first prove that

$$\sum(X_i - \overline{X})(Y_i - \overline{Y}) = \sum(X_i - \overline{X})Y_i$$

  This follows since

$$\sum(X_i - \overline{X})(Y_i - \overline{Y}) = \sum(X_i - \overline{X})Y_i - \sum(X_i - \overline{X})\overline{Y}$$

  But $\sum(X_i. - \overline{X})\overline{Y} = \overline{Y}\sum(X_i - \overline{X}) = 0$ since $\sum(X_i - \overline{X}) = 0$. Now

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

- The normality of the sampling distribution of b_1 follows at once from the fact that b_1 is a linear combination of the Y_i; the Y_i are independently, normally distributed; note that a linear combination of independent normal random variables is normally distributed

- The unbiasedness of the point estimator b_1, stated in the Gauss-Markov theorem, can be proved as follows:

$$\begin{aligned} \mathrm{E}[b_1] &= \mathrm{E}\left[\sum k_i Y_i\right] = \sum k_i \mathrm{E}[Y_i] = \sum k_i(\beta_0 + \beta_1 X_i) \\ &= \beta_0 \sum k_i + \beta_1 \sum k_i X_i \end{aligned}$$

  and so $\mathrm{E}[b_1] = \beta_1$

- The variance of b_1 can be derived as follows:

$$\begin{aligned} \mathrm{Var}[b_1] &= \mathrm{Var}\left[\sum k_i Y_i\right] = \sum k_i^2 \mathrm{Var}[Y_i] \\ &= \sum k_i^2 \sigma^2 = \sigma^2 \sum k_i^2 \\ &= \frac{\sigma^2}{\sum (X_i - \overline{X})^2} \end{aligned}$$

- The variance of the sampling distribution of b_1 can be estimated by replacing σ^2 with MSE, the unbiased estimator of σ^2

$$s^2[b_1] = \frac{\mathrm{MSE}}{\sum (X_i - \overline{X})^2}$$

  which is an unbiased estimator of $\mathrm{Var}[b_1]$

- Since b_1 is normally distributed, the standardized statistic $\frac{b_1 - \beta_1}{\sigma[b_1]}$ is a standard normal variable

- When a statistic is standardized but the denominator is an estimated standard deviation rather than the true standard deviation, it is called a studentized statistic

- Theorem:

$$\frac{b_1 - \beta_1}{s[b_1]} \text{ is distributed as } t_{n-2} \text{ for the normal error regression model}$$

- Proof: Note that $\frac{\mathrm{SSE}}{\sigma^2}$ is distributed as χ^2 with $n-2$ degrees of freedom and is independent of b_0 and b_1. First rewrite $(b_1 - \beta_1)/s[b_1]$ as follows:

$$\frac{b_1 - \beta_1}{\sigma[b_1]} \Big/ \frac{s[b_1]}{\sigma[b_1]}$$

  The numerator is a standard normal variable z; now,

$$\begin{aligned} \frac{s^2[b_1]}{\sigma^2[b_1]} &= \frac{\frac{\mathrm{MSE}}{\sum (X_i - \overline{X})^2}}{\frac{\sigma^2}{\sum (X_i - \overline{X})^2}} = \frac{\mathrm{MSE}}{\sigma^2} = \frac{\frac{\mathrm{SSE}}{n-2}}{\sigma^2} \\ &= \frac{\mathrm{SSE}}{\sigma^2(n-2)} \sim \frac{\chi^2_{n-2}}{n-2} \end{aligned}$$

  where the symbol $\sim$ stands for "is distributed as"; hence

$$\frac{b_1 - \beta_1}{s[b_1]} \sim \frac{z}{\sqrt{\frac{\chi^2_{n-2}}{n-2}}}$$

But z and χ^2 are independent since z is a function of b_1 and b_1 is independent of $\text{SSE}/\sigma^2 \sim \chi^2$ and so

$$\frac{b_1 - \beta_1}{s[b_1] \sim t_{n-2}}$$

- Since $(b_1 - \beta_1)/s[b_1]$ follows a t distribution, the following can be said

$$\mathbb{P}\left(t_{\frac{\alpha}{2},n-2} \leq \frac{b_1 - \beta_1}{s[b_1]} \leq t_{1-\frac{\alpha}{2},n-2}\right) = 1 - \alpha$$

  $t_{\frac{\alpha}{2},n-2}$ denotes the $(\alpha/2)100$ percentile of the t distribution with $n-2$ degrees of freedom

- The $1-\alpha$ confidence limits for β_1 are

$$b_1 \pm t_{1-\frac{\alpha}{2},n-2} s[b_1]$$

- This is derived from the following: because of the symmetry of the t distribution around its mean 0, it follows that

$$t_{\frac{\alpha}{2},n-2} = -t_{1-\frac{\alpha}{2},n-2}$$

  and by rearranging the probability statement,

$$\mathbb{P}\left(b_1 - t_{1-\frac{\alpha}{2},n-2} s[b_1] \leq \beta_1 \leq b_1 + t_{1-\frac{\alpha}{2},n-2} s[b_1]\right) = 1 - \alpha$$

- Two-Sided T Test: Let the null and alternative hypotheses be

$$H_0: \ \beta_1 = 0$$
$$H_A: \ \beta_1 \neq 0$$

  An explicit test of the alternatives is based on the test statistic

$$t^* = \frac{b_1}{s[b_1]}$$

  The decision rule with this test statistic for controlling the level of significance at α is:

  - If $|t^*| \leq t_{1-\frac{\alpha}{2},n-2}$, conclude H_0 (fail to reject H_0)
  - If $|t^*| > t_{1-\frac{\alpha}{2},n-2}$, conclude H_A (reject H_0)

- When the test of whether or not $\beta_1 = 0$ leads to the conclusion that $\beta_1 \neq 0$, the association between Y and X is sometimes described to be a linear statistical association

- The two-sided P-value is obtained by first finding the one-sided P-value and then multiplying by 2; if it is less than α, then conclude H_A (or reject H_0) else conclude H_0 (or fail to reject H_0)

- One-Sided T Test: Let the null and alternative hypotheses be:

$$H_0: \ \beta_1 \leq 0$$
$$H_A: \ \beta_1 > 0$$

  The decision rule based on this test statistic would be:

- If $t^* \leq t_{1-\alpha,n-2}$, conclude H_0 (fail to reject H_0)
- If $t^* > t_{1-\alpha,n-2}$, conclude H_A (reject H_0)

- Occasionally, it is desired to test whether or not β_1 equals some specified nonzero value v; the alternatives now are

$$H_0 : \ \beta_1 = v$$
$$H_A : \ \beta_1 \neq v$$

and the appropriate test statistic is

$$t^* = \frac{b_1 - v}{s[b_1]}$$

The decision rule remains the same

1.2.2 Inferences Concerning β_0

- Inferences concerning β_0 only occur when the scope of the model includes $X = 0$
- The point estimator b_0 is as follows:

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

- The sampling distribution of b_0 refers to the different values of b_0 that would be obtained with repeated sampling when the levels of the predictor variable X are held constant from sample to sample
- For the normal error regression model, the sampling distribution of b_0 is normal with mean and variance

$$\mathrm{E}\left[b_0\right] = \beta_0 \qquad \mathrm{Var}\left[b_0\right] = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right]$$

- The normality of the sampling distribution of b_0 follows because b_0 is a linear combination of the observations Y_i and the mean and variance of the sampling distribution of b_0 can be derived as before for b_1
- An estimator of $\mathrm{Var}\left[b_0\right]$ is obtained by replacing σ^2 by its point estimator MSE

$$s^2[b_0] = \mathrm{MSE}\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right]$$

The positive square root, $s[b_0]$ is an estimator of $\sigma\left[b_0\right]$

- Theorem:

$$\frac{b_) - \beta_0}{s[b_0]} \text{ is distributed as } t_{n-2} \text{ for the normal error regression model}$$

- The $1 - \alpha$ confidence limits for β_0 are obtained in the same manner as those for β_1 and are:

$$b_0 \pm t_{1-\frac{\alpha}{2},n-2}s[b_0]$$

1.2.3 Some Considerations on Making Inferences Concerning β_0 and β_1

- If the probability distribution of Y are not exactly normal but do not depart seriously, the sampling distributions of b_0 and b_1 will be approximately normal and the use of the t distribution will provide approximately the specified confidence coefficient or level of significance
- Even if the distribution of Y are far from normal, the estimators b_0 and b_1 generally have the property of asymptotically normality - their distributions approach normality under very general conditions as the sample size increases
- For large samples, the t value is replaced by the z value for the standard normal distribution
- Since the regression model assumes that the X_i are known constants, the confidence coefficients and risks of errors are interpreted with respect to taking repeated samples in which the X observations are kept at the same levels as in the observed sample; for example, concerning a confidence interval for β_1, the coefficient is interpreted to mean that if many independent samples are taken where the levels of X are the same as in the data set and a α confidence interval is constructed for each sample, α percent of the intervals will contain the true value of β_1
- Variances of b_1 and b_0 are affected by the spacing of the X levels in the observed data, as indicated by the use of n and σ^2 in the formulas; for example, the greater is the spread in the X levels, the larger is the quantity $\sum(X_i - \overline{X})^2$ and the smaller is the variance of b_1
- The power of tests on β_0 and β_1 is the probability that the test correctly rejects the null hypothesis (concluding H_A)
- For example, using the hypothesis test concerning β_1 where

$$H_0: \ \beta_1 = v$$
$$H_A: \ \beta_1 \neq v$$

the test statistic computed is

$$t^* = \frac{b_1 - v}{s[b_1]}$$

and the decision rule for level of significance α is

  - If $|t^*| \leq t_{1-\frac{\alpha}{2},n-2}$, conclude H_0 (fail to reject H_0)
  - If $|t^*| > t_{1-\frac{\alpha}{2},n-2}$, conclude H_A (reject H_0)

The power of test is the probability that the decision rule will lead to conclusion H_A when H_A in fact holds; specifically, the power of the test is given by

$$\text{Power} = \mathbb{P}\left(|t^*| > t_{1-\frac{\alpha}{2},n-2} | \delta\right)$$

where δ is the noncentrality measure, i.e., a measure of how far the true value of β_1 is from a given value v

$$\delta = \frac{|\beta_1 - v|}{\sigma\left[b_1\right]}$$

1.2.4 Interval Estimation of $\mathrm{E}[Y_h]$

- Let X_h denote the level of X for which the mean response is to be estimated; it may be a value which occurred in the sample or it may be some other value of the predictor variable within the scope of the model; the mean response when $X = X_h$ is denoted by $\mathrm{E}[Y_h]$
- The point estimator $\hat{Y}_h$ of $\mathrm{E}[Y_h]$ is $\hat{Y}_h = b_0 + b_1 X_h$
- The sampling distribution of $\hat{Y}_h$ refers to the different values of $\hat{Y}_h$ that would be obtained if repeated samples were selected, each holding the levels of the predictor variable X constant, and calculating $\hat{Y}_h$ for each sample
- For the normal error regression model, the sampling distribution of $\hat{Y}_h$ is normal with mean and variance

$$\mathrm{E}\left[\hat{Y}_h\right] = \mathrm{E}[Y_h] \qquad \mathrm{Var}\left[\hat{Y}_h\right] = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right]$$

- The normality of the sampling distribution of $\hat{Y}_h$ follows directly from the fact that $\hat{Y}_h$ is a linear combination of the observations Y_i
- $\hat{Y}_h$ is an unbiased estimator of $\mathrm{E}[Y_h]$

$$\mathrm{E}\left[\hat{Y}_h\right] = \mathrm{E}[b_0 + b_1 X_h] = \mathrm{E}[b_0] + X_h \mathrm{E}[b_1] = \beta_0 + \beta_1 X_h$$

- The variability of the sampling distribution of $\hat{Y}_h$ is affected by how far X_h is from $\overline{X}$, through the term $(X_h - \overline{X})^2$; the further from $\overline{X}$ is X_h, the greater the quantity $(X_h - \overline{X})^2$ and the larger is the variance of $\hat{Y}_h$
- The estimated variance of $\hat{Y}_h$ is

$$s^2[\hat{Y}_h] = \mathrm{MSE}\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right]$$

  The estimated standard deviation of $\hat{Y}_h$ is then $s[\hat{Y}_h]$, the positive square root of $s^2[\hat{Y}_h]$
- When $X_h = 0$, the variance of $\hat{Y}_h$ is reduced to the variance of b_0 since $\hat{Y}_h = b_0 + b_1 X_h = b_0 + b_1(0) = b_0$
- To derive $\sigma\left[\hat{Y}_h\right]$, first show that b_1 and $\overline{Y}$ are uncorrelated and hence, for the regression model, independent: $\mathrm{Cov}\left[\overline{Y},\ b_1\right] = 0$, where the LHS denotes the covariance between the two; now,

$$\overline{Y} = \sum\left(\frac{1}{n}\right) Y_i \qquad b_1 = \sum k_i Y_i$$

  where k_i is defined as before; now, knowing that Y_i are independent random variables,

$$\mathrm{Cov}\left[\overline{Y},\ b_1\right] = \sum\left(\frac{1}{n}\right) k_i \sigma^2[Y_i] = \frac{\sigma^2}{n}\sum k_i$$

  but $\sum k_i = 0$ and so the covariance is 0; to find the variance of $\hat{Y}_h$, use the alternative form of the estimator

$$\mathrm{Var}\left[\hat{Y}_h\right] = \mathrm{Var}\left[\overline{Y} - b_1(X_h - \overline{X})\right]$$

Since $\overline{Y}$ and b_1 are independent and X_h and $\overline{X}$ are constants, then

$$\text{Var}\left[\hat{Y}_h\right] = \text{Var}\left[\overline{Y}\right] + (X_h - \overline{X})^2 \text{Var}\left[b_1\right]$$

Since

$$\text{Var}\left[\overline{Y}\right]. = \frac{\text{Var}\left[Y_i\right]}{n} = \frac{\sigma^2}{n}$$

and so

$$\text{Var}\left[\hat{Y}_h\right] = \frac{\sigma^2}{n} + (X_h - \overline{X})^2 \frac{\sigma^2}{\sum (X_i - \overline{X})^2}$$

- Theorem:

  $\frac{\hat{Y}_h - \text{E}\left[Y_h\right]}{s[\hat{Y}_h]}$ is distributed as t_{n-2} for the regression model

  All inferences concerning $\text{E}\left[Y_h\right]$ are carried out in the usual fashion with the t distribution

- A confidence interval for $\text{E}\left[Y_h\right]$ is constructed in the standard fashion as follows

$$\hat{Y}_h \pm t_{1-\frac{\alpha}{2}, n-2} s[\hat{Y}_h]$$

- Since the X_i are known constants in the regression model, the interpretation of confidence intervals and risks of errors in inferences on the mean response is in terms of taking repeated samples in which the X observations are at the same levels as in the actual study

- For given sample results, the variance of $\hat{Y}_h$ is smallest when $X_h = \overline{X}$; thus, in an experiment to estimate the mean response at a particular level X_h of the predictor variable, the precision of the estimate will be greatest if (everything else remaining equal) the observations on X are spaced so that $\overline{X} = X_h$

- The usual relationship between confidence intervals and tests applies in inferences concerning the mean response; thus, the two-sided confidence limits can be utilized for two-sided tests concerning the mean response at X_h; alternatively, a regular decision rule can be set up

- The confidence limits for a mean response $\text{E}\left[Y_h\right]$ are not sensitive to moderate departures from the assumption that the error terms are normally distributed

- Confidence limits apply when a single mean response is to be estimated from the study

1.2.5 Prediction of New Observation

- A new observation on Y to be predicted is viewed as a result of a new trial, independent of the trials on which the regression analysis is based; denote the level of X for the new trial as X_h and the new observation on Y as $Y_{h(\text{new})}$

- In the estimation of the mean response $\text{E}\left[Y_h\right]$, the mean of the distribution of Y is estimated; in the prediction of a new response $Y_{h(\text{new})}$, an individual outcome drawn from the distribution of Y is predicted

- The basic idea of a prediction interval is to choose a range in the distribution of Y wherein most of the observations will fall and then to declare that the next observation will fall in this range; the usefulness of the prediction interval depends on the width of the interval and the needs for precision by the user

- Assume that all regression parameters of the normal error regression model are known, then the $1-\alpha$ prediction limits for $Y_{h(\text{new})}$ are

$$\mathrm{E}[Y_h] \pm z_{1-\frac{\alpha}{2}}\sigma$$

In centering the limits around $\mathrm{E}[Y_h]$, the narrowest interval consistent with the specified probability of a correct prediction is obtained

- When the regression parameters are unknown, the mean of the distribution of Y is estimated by $\hat{Y}_h$ as usual and the variance of the distribution of Y is estimated by the MSE but the prediction limit above with the parameters replaced by the corresponding point estimators cannot be used since the mean $\mathrm{E}[Y_h]$ itself is estimated by a confidence interval, making the location of the distribution of Y uncertain

- Prediction limits for $Y_{h(\text{new})}$ must take account of the variation in possible location of the distribution of Y and the variation within the probability distribution of Y

- Prediction limits for a new observations $Y_{h(\text{new})}$ at a given level X_h are obtained by the following theorem:

$$\frac{Y_{h(\text{new})} - \hat{Y}_h}{s[\text{pred}]} \text{ is distributed as } t(n-2) \text{ for the normal error regression model}$$

Note that the studentized statistic uses the point estimator $\hat{Y}_h$ in the numerator rather than the true mean $\mathrm{E}[Y]_h$ because the true mean is unknown

- Thus, when the regression parameters are unknown, the $1-\alpha$ prediction limits for a new observation $Y_{h(\text{new})}$ are

$$\hat{Y}_h \pm t_{1-\frac{\alpha}{2},n-2}s[\text{pred}]$$

- The variance of this prediction error can be obtained by utilizing the independence of the new observation $Y_{h(\text{new})}$ and the original n sample cases on which $\hat{Y}_h$ is based

$$\mathrm{Var}[\text{pred}] = \mathrm{Var}\left[Y_{h(\text{new})} - \hat{Y}_h\right] = \mathrm{Var}\left[Y_{h(\text{new})}\right] + \mathrm{Var}\left[\hat{Y}_h\right] = \sigma^2 + \mathrm{Var}\left[\hat{Y}_h\right]$$

The first term is the variance of the distribution of Y at $X = X_h$ while the second term is the variance of the sampling distribution of $\hat{Y}_h$

- An unbiased estimator of $\mathrm{Var}[\text{pred}]$ is

$$s^2[\text{pred}] = \text{MSE} + s^2[\hat{Y}_h]$$

which can be expressed as

$$s^2[\text{pred}] = \text{MSE}\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right]$$

- The prediction interval for $Y_{h(\text{new})}$ is wider than the confidence interval for $\mathrm{E}[Y_h]$ because both the variability in $\hat{Y}_h$ from sample to sample and the variation within the probability distribution of Y is encountered

- The prediction interval is wider the further X_h is from $\overline{X}$ since the estimate of the mean $\hat{Y}_h$ is less precise as X_h is located farther away from $\overline{X}$
- The prediction limits for a mean response $\mathrm{E}[Y_h]$ are sensitive to departures from normality of the error terms distributions
- The confidence coefficient for the prediction limits refers to the taking of repeated samples based on the same set of X values, and calculating prediction limits for $Y_{h(\text{new})}$ for each sample
- Prediction limits apply for a single prediction based on the sample data
- Prediction intervals resemble confidence intervals but differ conceptually; a confidence interval represents an inference on a parameter and is an interval that is intended to cover the value of the parameter; a prediction interval is a statement about the value to be taken by a random variable, the new observation $Y_{h(\text{new})}$
- Suppose the mean of m new observations on Y for a given level of the predictor variable is to be predicted, then the mean of the new Y observations to be predicted is denoted $\overline{Y}_{h(\text{new})}$ and the appropriate $1-\alpha$ prediction limits are, assuming that the new Y observations are independent:

  $$\hat{Y}_h \pm t_{1-\frac{\alpha}{2},n-2} s[\text{predmean}]$$

  where

  $$s^2[\text{predmean}] = \frac{\text{MSE}}{m} + s^2[\hat{Y}_h]$$

  or equivalently

  $$s^2[\text{predmean}] = \text{MSE}\left[\frac{1}{m} + \frac{1}{n}. + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right]$$

  Note that the variance $s^2[\text{predmean}]$ has two components: (1) the variance of the mean of m observations from the probability distribution of Y at $X = X_h$ and (2) the variance of the sampling distribution of $\hat{Y}_h$
- The prediction limits for predicting m new observations on Y are narrower than those for predicting for a single new observation on Y because it involve a prediction of the mean response for m new observations

1.2.6 Confidence Band for Regression Line

- A confidence band for the entire regression line $\mathrm{E}[Y] = \beta_0 + \beta_1 X$ allows one to determine the appropriateness of a fitted regression function
- The Working-Hotelling $1-\alpha$ confidence band for the regression line has the following two boundary values at any level X_h:

  $$\hat{Y}_h \pm W s\left[\hat{Y}_h\right]$$

  where

  $$W^2 = 2F_{1-\alpha,n-2}$$

  Here $F_{1-\alpha,n-2}$ denotes the density of the F distribution at $1-\alpha$ confidence with $n-2$ degrees of freedom; this formula for the boundary values is of exactly the same form as the one for the confidence limits for the mean response at X_h, except that the t multiple has been replaced by the W multiple

- The boundary points of the confidence band for the regression line are wider apart the farther X_h is from the mean $\overline{X}$ of the X observations; the W multiple will be larger than the t multiple because the confidence band must encompass the entire regression line, whereas the confidence limits for $\mathrm{E}[Y_h]$ at X_h apply only at the single level X_h
- The boundary values of the confidence band for the regression line define a hyperbola, as seen by replacing $\hat{Y}_h$ and $s[\hat{Y}_h]$ by their definitions

$$b_0 + b_1 X \pm W\sqrt{\mathrm{MSE}}\left[\frac{1}{n} + \frac{(X-\overline{X})^2}{\sum(X_i-\overline{X})^2}\right]^{\frac{1}{2}}$$

- The boundary values of the confidence band for the regression line at any value X_h often are not substantially wider than the confidence limits for the mean response at that single X_h level; with the somewhat wider limits for the entire regression line, one is able to draw conclusions about any and all mean responses for the entire regression line and not just about the mean response at a given X level
- The confidence band applies to the entire regression line over all real-numbered values of X from $-\infty$ to ∞; the confidence coefficient indicates the proportion of time that the estimating procedure will yield a band that covers the entire line, in a long series of samples in which the X observations are kept at the same level as in the actual study; in applications, the confidence band is ignored for that part of the regression line which is not of interest in the problem at hand
- The confidence coefficient for a limited segment of the band of interest is somewhat higher than $1-\alpha$, so $1-\alpha$ serves then as a lower bound to the confident coefficient

1.2.7 Analysis of Variance Approach to Regression Analysis

- The analysis of variance approach is based on the partitioning of sums of squares and degrees of freedom associated with the response variable Y
- Variation is conventionally measured in terms of the deviations of the Y_i around their mean $\hat{Y}$:

$$Y_i - \overline{Y}$$

- The measure of total variation, denoted by SSTO (total sum of squares), is the sum of the squared deviations

$$\mathrm{SSTO} = \sum(Y_i - \overline{Y})^2$$

  If all Y_i observations are the same, $\mathrm{SSTO} = 0$; the greater the variation among the Y_i observations, the larger is SSTO
- When the predictor variable X is utilized, the variation reflecting the uncertainty concerning the variable Y is that of the Y_i observations around the fitted regression line:

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

- The measure of variation in the Y_i observations that is present when the predictor variable X is taken into account is the sum of the squared deviations

$$\mathrm{SSE} = \sum(Y_i - \hat{Y}_i)^2$$

where SSE denotes error sum of squares; if all Y_i observations fall on the fitted regression line, SSE = 0; the greater the variation of the Y_i observations around the fitted regression line, the larger is SSE

- Another important deviations is squared deviations

$$\hat{Y}_i - \overline{Y}$$

SSR, or regression sum of squares, is a sum of squared deviations

$$\text{SSR} = \sum(\hat{Y}_i - \overline{Y})^2$$

Each deviation is simply the difference between the fitted value on the regression line and the mean of the fitted values $\overline{Y}$

- If the regression line is horizontal so that $\hat{Y}_i - \overline{Y} \equiv 0$, then SSR = 0; otherwise SSR is positive
- SSR may be considered a measure of that part of the variability of the Y_i which is associated with the regression line; the larger SSR is in relation to SSTO, the greater is the effect of the regression relation in accounting for the total variation in the Y_i observations
- The total deviation $Y_i - \overline{Y}$, used in the measure of the total variation of the observations Y_i without taking the predictor variable into account, can be decomposed into two components:

$$\underbrace{Y_i - \overline{Y}}_{\text{total deviation}} = \underbrace{\hat{Y}_i - \overline{Y}}_{\text{deviation of fitted regression value around mean}} + \underbrace{Y_i - \hat{Y}_i}_{\text{deviation around fitted regression line}}$$

These two components are: the deviation of the fitted value $\hat{Y}_i$ around the mean $\overline{Y}$ and the deviation of the observation Y_i around the fitted regression line

- This relationship can be summarized as

$$\text{SSTO} = \text{SSR} + \text{SSE}$$
$$\sum(Y_i - \overline{Y})^2 = \sum(\hat{Y}_i - \overline{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$

- Proof;

$$\begin{aligned}
\sum(Y_i - \overline{Y})^2 &= \sum[(\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)]^2 \\
&= \sum[(\hat{Y}_i - \overline{Y})^2 + (Y_i - \hat{Y}_i)^2 + 2(\hat{Y}_i - \overline{Y})(Y_i - \hat{Y}_i)] \\
&= \sum(\hat{Y}_i - \overline{Y})^2 + \sum(Y_i - \hat{Y}_i)^2 + 2\sum(\hat{Y}_i - \overline{Y})(Y_i - \hat{Y}_i) \\
&= \sum(\hat{Y}_i - \overline{Y})^2 + \sum(Y_i - \hat{Y}_i)^2 + 2\sum\hat{Y}_i(Y_i - \hat{Y}_i) - 2\overline{Y}\sum(Y_i - \hat{Y}_i) \\
&= \sum(\hat{Y}_i - \overline{Y})^2 + \sum(Y_i - \hat{Y}_i)^2 + 0 - 0 \\
&= \sum(\hat{Y}_i - \overline{Y})^2 + \sum(Y_i - \hat{Y}_i)^2
\end{aligned}$$

- There are $n - 1$ degrees of freedom associated with SSTO; one degree is lost because the deviations $Y_i - \overline{Y}$ are subject to one constraint: they must sum to zero; equivalently, one degree is lost because the sample mean $\overline{Y}$ is used to estimate the population mean

- There are $n-2$ degrees of freedom associated with SSE because two parameters are estimated in obtaining the fitted values $\hat{Y}_i$
- SSR has one degree of freedom associated with it; although there are n deviations $\hat{Y}_i - \overline{Y}$, all fitted values $\hat{Y}_i$ are calculated from the same regression line; two degrees of freedom are associated with a regression line (corresponding to the intercept and slope); one of the degrees is lost because the deviations $\hat{Y}_i - \overline{Y}$ are subject to a constraint: they must sum to zero
- Note that the degrees of freedom are additive

$$n - 1 = 1 + (n - 2)$$

- A sum of squares divided by its associated degrees of freedom is called a mean square (MS)
- The regression mean square, MSR, is

$$\text{MSR} = \frac{\text{SSR}}{1} = \text{SSR}$$

and the error mean square (MSE) is

$$\text{MSE} = \frac{\text{SSE}}{n-2}$$

- Note that mean squares are not additive
- The breakdown of the total sum of squares and associated degrees of freedom are displayed in the form of an analysis of variance table (ANOVA table)

Source of variation	SS	df	MS	E [MS]
Regression	$\text{SSR} = \sum(\hat{Y}_i - \overline{Y})^2$	1	$\text{MSR} = \frac{\text{SSR}}{1}$	$\sigma^2 + \beta_1^2 \sum(X_i - \overline{X})^2$
Error	$\text{SSE} = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$\text{MSE} = \frac{\text{SSE}}{n-2}$	σ^2
Total	$\text{SSTO} = \sum(Y_i - \overline{Y})^2$	$n-1$	-	-

ANOVA Table for Simple Linear Regression

- The total sum of squares can be decomposed as follows:

$$\text{SSTO} = \sum(Y_i - \overline{Y})^2 = \sum Y_i^2 - n\overline{Y}^2$$

In a modified ANOVA table, the total uncorrected sum of squares, denoted by SSTOU, is defined as

$$\text{SSTOU} = \sum Y_I^2$$

and the correction for the mean sum of squares, denoted by SS(correction for mean) is

$$\text{SS(correction for mean)} = n\overline{Y}^2$$

Source of variation	SS	df	MS
Regression	$\text{SSR} = \sum(\hat{Y}_i - \overline{Y})^2$	1	$\text{MSR} = \frac{\text{SSR}}{1}$
Error	$\text{SSE} = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$\text{MSE} = \frac{\text{SSE}}{n-2}$
Total	$\text{SSTO} = \sum(Y_i - \overline{Y})^2$	$n-1$	-
Correction for mean	$\text{SS(correction for mean)} = n\overline{Y}^2$	1	-
Total, uncorrected	$\text{SSTOU} = \sum Y_i^2$	n	-

Modified ANOVA Table for Simple Linear Regression

- The expected value of a mean square is the mean of its sampling distribution; it tells what is being estimated by the mean square

$$\begin{aligned} \text{E}[\text{ SE}] &= \sigma^2 \\ \text{E}[\text{MSR}] &= \sigma^2 + \beta_1^2 \sum(X_i - \overline{X})^2 \end{aligned}$$

- The mean of the sampling distribution of MSE is σ^2 whether or not X and Y are linearly related, i.e., whether or not $\beta_1 = 0$; the mean of the sampling distribution of MSR is also σ^2 when $\beta_1 = 0$; hence when $\beta_1 = 0$, the sampling distribution of MSR and MSE are located identically and MSR and MSE will tend to be of the same order of magnitude; when $\beta_1 \neq 0$, the mean of the sampling distribution of MSR is located to the right of that of MSE and hence MSR will tend to be larger than MSE

- For the simple linear regression case, the analysis of variance provides a test for

$$\begin{aligned} H_0 &: \ \beta_1 = 0 \\ H_A &: \ \beta_1 \neq 0 \end{aligned}$$

  The test statistic for the analysis of variance is denoted by F^*

$$F^* = \frac{\text{MSR}}{\text{MSE}}$$

  This suggests that large values of F^* support H_A and values of F^* near 1 support H_0

- Cochran's Theorem: If all n observations Y_i come from the same normal distribution with mean μ and variance σ^2, and SSTO is decomposed into k sums of squares SS_r, each with degrees of freedom df_r, then the $\frac{\text{SS}_r}{\sigma^2}$ terms are independent χ^2 variables with df_r degrees of freedom if

$$\sum_{r=1}^{k} \text{df}_r = n - 1$$

  If $\beta_1 = 0$, so that all Y_i have the same mean $\mu = \beta_0$ and the same variance σ^2, $\frac{\text{SSE}}{\sigma^2}$ and $\frac{\text{SSR}}{\sigma^2}$ are independent χ^2 variables

- The test statistic can be written as follows:

$$F^* = \frac{\frac{\text{SSR}}{\sigma^2}}{1} \Big/ \frac{\frac{\text{SSE}}{\sigma^2}}{n-2} = \frac{\text{MSR}}{\text{MSE}}$$

But by Cochran's theorem, when H_0 holds:

$$F^* \sim \frac{\chi^2_1}{1} / \frac{\chi^2_{n-2}}{n-2} \text{ when } H_0 \text{ holds}$$

where the χ^2 variables are independent; thus when H_0 holds, F^* is the ratio of two independent χ^2 variables, each divided by its degrees of freedom; this is the definition of an F random variable

- Thus when H_0 holds, F^* follows the $F_{1,n-2}$ distribution
- Even if $\beta_1 \neq 0$, SSR and and SSE are independent and $\frac{\text{SSE}}{\sigma^2} \sim \chi^2$; however, the condition that both $\frac{\text{SSR}}{\sigma^2}$ and $\frac{\text{SSE}}{\sigma^2}$ are χ^2 random variables requires $\beta_1 = 0$
- Since the test is upper tail and $F^* \sim F_{1,n-2}$ when H_0 holds, the decision rule is as follows when the risk of a Type 1 error is to be controlled at α
  - If $F^* \leq F_{1-\alpha,n-2}$, conclude H_0
  - If $F^* > F_{1-\alpha,n-2}$, conclude H_A

  where $F_{1-\alpha,n-2}$ is the $(1-\alpha)100$ percentile of the appropriate F distribution
- For a given α level, the F test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is equivalent algebraically to the two-tailed t test; recall that $\text{SSR} = b_1^2 \sum (X_i - \overline{X})^2$, then

$$F^* = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{b_1^2 \sum (X_i - \overline{X})^2}{\text{MSE}}$$

  But since $s2[b_1] = \text{MSE} / \sum (X_i - \overline{X})^2$,

$$F^* = \frac{b_1^2}{s^2[b_1]} = \left(\frac{b_1}{s[b_1]} \right)^2 = (t^*)^2$$

  The last step follows because the t^* statistic for testing whether or not $\beta_1 = 0$ is $t^* = b_1 / s[b_1]$
- The following relation between the required percentiles of the t and F distributions for the tests exists:

$$[t_{1-\frac{\alpha}{2},n-2}]^2 = F_{1-\alpha,1,n-2}$$

- Thus at any given α level, either the t test or the F test for testing $\beta_1 = 0$ vs $\beta_1 \neq 0$ can be used; whenever one test leads to H_0, so will the other and corresponding for H_A; the t test, however, is more flexible since it can be used for one-sided alternatives involving $\beta_1 (\leq \geq) 0$ versus $\beta_1 (> <) 0$, while the F test cannot

1.2.8 General Linear Test Approach

- The analysis of variance test of $\beta_1 = 0$ vs $\beta_1 \neq 0$ is an example of the general test for a linear statistical model
- The general linear test approach for a simple linear regression model involves three steps

- First, for the simple linear regression model, the full model, or the normal error regression model, is obtained

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{full model}$$

This full model is fit and the error sum of squares is obtained (SSE(F))

$$\text{SSE(F)} = \sum [Y_i - (b_0 + b_1 X_i)]^2 = \sum (Y_I - \hat{Y}_i)^2 = \text{SSE}$$

Thus for the full model, the error sum of squares is simply SSE

- Next, consider H_0:

$$H_0 : \ \beta_1 = 0$$
$$H_A : \ \beta_1 \neq 0$$

The model when H_0 holds is called the reduced or restricted model; when $\beta_1 = 0$, the model reduces to

$$Y_i = \beta_0 + \varepsilon_i \quad \text{reduced model}$$

This reduced model is fit and the error sum of squares is obtained (SSE(R))

$$\text{SSE(R)} = \sum (Y_i - b_0)^2 = \sum (Y_i - \overline{Y})^2 = \text{SSTO}$$

- It can be shown that SSE(F) never is greater than SSE(R) because the more parameters there are in the model, the better one can fit the data and the smaller are the deviations around the fitted regression function
- When SSE(F) is not much less than SSE(R), using the full model does not account for much more of the variability of the Y_i than does the reduced model, in which case the data suggest that the reduced model is adequate (i.e., that H_0 holds)
- A large difference would suggest that H_A holds because the additional parameters in the model do help to reduce substantially the variation of the observations Y_i around the fitted regression function
- The actual test statistic is a function of SSE(R) − SSE(F):

$$F^* = \frac{\text{SSE(R)} - \text{SSE(F)}}{\text{df}_R - \text{df}_F} \Big/ \frac{\text{SSE(F)}}{\text{df}_F}$$

which follows the F distribution when H_0 holds; the degrees of freedom df_R and df_F are those associated with the reduced and full model sums of squares, respectively

- The decision rule is:
  - If $F^* \leq F_{1-\alpha, \text{df}_R - \text{df}_F, \text{df}_F}$, conclude H_0
  - If $F^* > F_{1-\alpha, \text{df}_R - \text{df}_F, \text{df}_F}$, conclude H_A
- For testing whether or not $\beta_1 = 0$, the following is stated:

$$\begin{aligned} \text{SSE(R)} &= \text{SSTO} & \text{SSE(F)} &= \text{SSE} \\ \text{df}_F &= n - 1 & \text{df}_F &= n - 2 \end{aligned}$$

and thus

$$F^* = \frac{\text{SSTO} - \text{SSE}}{(n-1) - (n-2)} \Big/ \frac{\text{SSE}}{n-2} = \frac{\text{SSR}}{1} \Big/ \frac{\text{SSE}}{n-2} = \frac{\text{MSR}}{\text{MSE}}$$

which is identical to the analysis of variance test statistic

- The general linear test approach can be used for highly complex tests of linear statistical models, as well as for simple tests; the basic steps in summary form are:

  1. Fit the full model and obtain the error sum of squares SSE(F)
  2. Fit the reduced model under H_0 and obtain the error sum of squares SSE(R)
  3. Compute the test statistic and use the decision rule

1.2.9 Descriptive Measures of Linear Association between X and Y

- SSTO is a measure of the uncertainty in predicting Y when X is not considered; similarly, SSE measures the variation in the Y_i when a regression model utilizing the predictor variable X is employed

- A natural measure of the effect of X in reducing the variation in Y, i.e., in reducing the uncertainty in predicting Y, is to express the reduction in variation (SSTO − SSE = SSR) as a proportion of the total variation

$$R^2 = \frac{\text{SSR}}{\text{SSTO}} = 1 - \frac{\text{SSE}}{\text{SSTO}}$$

  The measure R^2 is called the coefficient of determination; since $0 \leq \text{SSE} \leq \text{SSTO}$, it follows that

$$0 \leq R^2 \leq 1$$

- R^2 can be interpreted as the proportionate reduction of total variation associated with the use of the predictor variable X; thus the larger R^2 is, the more variation of Y is reduced by introducing the predictor variable X

- When all observations fall on the fitted regression line, then SSE = 0 and $R^2 = 1$; the predictor variable X accounts for all variation in the observations Y_i

- When the fitted regression line is horizontal so that $b_1 = 0$ and $\hat{Y}_i = \overline{Y}$, then SSE = SSTO and $R^2 = 0$l the predictor variable X is of no help in reducing the variation in the observations Y_i

- In practice, R^2 is somewhere between 0 and 1; the closer it is to 1, the greater is said to be the degree of association between X and Y

- It is not true that a high coefficient of determination indicates that useful predictions can be made; it is not true that a high coefficient of determination indicates that the estimated regression line is a good fit; it is not true that a coefficient of determination near zero indicates that X and Y are not related

- Note that R^2 measures only a relative reduction from SSTO and provides no information about absolute precision for estimating a mean response or predicting a new observation; R^2 measures the degree of linear association between X and Y

- A measure of linear association between Y and X when both Y and X are random is the coefficient of correlation; this measure is the signed square root of R^2

$$r = \pm\sqrt{R^2}$$

  A plus or minus sign is attached to this measure according to whether the slope of the fitted regression line is positive or negative; thus, the range of r is: $-1 \leq r \leq 1$

- The value taken by R^2 in a given sample tends to be affected by the spacing of the X observations; SSE is not affected systematically by the spacing of the X_i since, for the normal error regression model, $\text{Var}[Y_i] = \sigma^2$ at all X levels; however, the wider the spacing of the X_i in the sample when $b_1 \neq 0$, the greater will tend to be the spread of the observed Y_i around $\overline{Y}$ and hence the greater SSTO will be; consequently, the wider the X_i are spaced, the higher R^2 will tend to be
- The regression sum of squares SSR is often called the "explained variation" in Y and the residual sum of squares SSE is called the "unexplained variation"; the coefficient R^2 is then interpreted in terms of the proportion of the total variation in Y (SSTO) which has been "explained" by X; remember that in a regression model, there is no implication that Y necessarily depends on X in a causal or explanatory sense
- Regression models do not contain a parameter to be estimated by R^2 or r; these are simply descriptive measures of the degree of linear association between X and Y in the sample observations that may, or may not, be useful in any instance

1.2.10 Considerations in Applying Regression Analysis

- Regression analysis is used to make inferences for the future
- The validity of the regression application depends upon whether basic causal conditions in the period ahead will be similar to those in existence during the period upon which the regression analysis is based; this caution applies whether mean responses are to be estimated, new observations predicted or regression parameters estimated
- In predicting new observations on Y, the predictor variable X itself often has to be predicted
- If the X level does not fall far beyond the range of the predictor variable observations, one may have reasonable confidence int he application of the regression analysis; on the other hand, if the X level falls far beyond the range of past data, extreme caution should be excises since one cannot be sure that the regression function that fits the past data is appropriate over the wider range of the predictor variable
- A statistical test that leads to the conclusion that $\beta_1 \neq 0$ does not establish a cause and effect relation between the predictor and response variables; with nonexperimental data, both the X and Y variables may be simultaneously influenced by other variables not in the regression model; on the other hand, the existence of a regression relation in controlled experiments is often good evidence of a cause and effect relation
- There are special problems when one wants to estimate several mean responses or predict several new observations for different levels of the predictor variable; the confidence coefficients for the limits given before for estimating a mean response and for the prediction limits for a new observation apply only for a single level of X for a given sample
- When observations on the predictor variable X are subject to measure errors, the resulting parameter estimates are generally no longer unbiased

1.2.11 Normal Correlation Models

- The normal correlation model for the case of two variables is based on the bivariate normal distribution

- Two variable Y_1 and Y_2 are jointly normally distributed if the density function of their joint distribution is that for the bivariate normal distribution

$$f(Y_1, Y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} \exp\left[-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho_{12}\left(\frac{Y_1-\mu_1}{\sigma_1}\right)\left(\frac{Y_2-\mu_2}{\sigma_2}\right) + \left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2\right]\right]$$

- If Y_1 and Y_2 are jointly normally distributed, it can be shown that their marginal distributions have the following characteristics:
  - The marginal distribution of Y_1 is normal with mean μ_1 and standard deviation σ_1:

$$f_1(Y_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2\right]$$

  - The marginal distribution of Y_2 is normal with mean μ_2 and standard deviation σ_2:

$$f_2(Y_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2\right]$$

- If Y_1 and Y_2 are each normally distributed, they need not be jointly normally distributed
- The five parameters of the bivariate normal density function have the following meaning:
  - μ_1 and σ_1 are the mean and standard deviation of the marginal distribution of Y_1
  - μ_2 and σ_2 are the mean and standard deviation of the marginal distribution of Y_2
  - ρ_{12} is the coefficient of correlation between the random variables Y_1 and Y_2

$$\rho_{12} = \rho[Y_1, Y_2] = \frac{\sigma_{12}}{\sigma_1\sigma_2}$$

    Here, σ_1 and σ_2 denote the standard deviations of Y_1 and Y_2 and σ_{12} denotes the covariance $\text{Cov}[Y_1,\ Y_2]$ between Y_1 and Y_2

$$\sigma_{12} = \text{Cov}[Y_1,\ Y_2] = \text{E}[(Y_1-\mu_1)(Y_2-\mu_2)]$$

    Note that $\sigma_{12} \equiv \sigma_{21}$ and $\rho_{12} \equiv \rho_{21}$

- If Y_1 and Y_2 are independent, $\sigma_{12} = 0$ and so $\rho_{12} = 0$; if Y_1 and Y_2 are positively related, σ_{12} is positive and so is ρ_{12}; if Y_1 and Y_2 are negatively related, ρ_{12} is negative and so is ρ_{12}
- The coefficient of correlation ρ_{12} can take on any value between -1 and 1 inclusive; it assumes 1 is the linear relation between Y_1 and Y_2 is perfectly positive (direct) and -1 if it is perfectly negative (inverse)
- One principle use of a bivariate correlation model is to make conditional inferences regarding one variable, given the other variable

- The density function of the conditional probability distribution of Y_1 for any given value of Y_2 is denoted by $f(Y_1|Y_2)$ and defined as follows:

$$f(Y_1|Y_2) = \frac{f(Y_1, Y_2)}{f_2(Y_2)}$$

where $f(Y_1, Y_2)$ is the joint density function of Y_1 and Y_2 and $f_2(Y_2)$ is the marginal density function of Y_2

- When Y_1 and Y_2 are jointly normally distributed, the conditional probability distribution of Y_1 for any given value of Y_2 is normal with mean $\alpha_{1|2} + \beta_{12}Y_2$ and standard deviation $\sigma_{1|2}$ and its density function is

$$f(Y_1|Y_2) = \frac{1}{\sqrt{2\pi}\sigma_{1|2}} \exp\left[-\frac{1}{2}\left(\frac{Y_1 - \alpha_{1|2} - \beta_{12}Y_2}{\sigma_{1|2}}\right)^2\right]$$

The parameters $\alpha_{1|2}$, β_{12} and $\sigma_{1|2}$ of the conditional probability distribution of Y_1 are functions of the parameters of the joint probability distribution as follows

$$\alpha_{1|2} = \mu_1 - \mu_2\rho_{12}\frac{\sigma_1}{\sigma_2}$$
$$\beta_{12} = \rho_{12}\frac{\sigma_1}{\sigma_2}$$
$$\sigma^2_{1|2} = \sigma^2_1(1 - \rho^2_{12})$$

The parameter $\alpha_{1|2}$ is the intercept of the line of regression of Y_1 on Y_2 and the parameter β_{12} is the slope of this line

- The conditional distribution of Y_1, given Y_2, is equivalent to the normal error regression model

- The conditional probability distribution of Y_2 for any given value of Y_1 is normal with mean $\alpha_{2|1} + \beta_{21}Y_1$ and standard deviation $\sigma_{2|1}$ and its density function is

$$f(Y_2|Y_1) = \frac{1}{\sqrt{2\pi}\sigma_{2|1}} \exp\left[-\frac{1}{2}\left(\frac{Y_2 - \alpha_{2|1} - \beta_{21}Y_1}{\sigma_{2|1}}\right)^2\right]$$

The parameters $\alpha_{2|1}$, β_{21} and $\sigma_{2|1}$ of the conditional probability distributions of Y_2 are functions of the parameters of the joint probability distribution as follows

$$\alpha_{2|1} = \mu_2 - \mu_1\rho_{12}\frac{\sigma_2}{\sigma_1}$$
$$\beta_{21} = \rho_{12}\frac{\sigma_2}{\sigma_1}$$
$$\sigma^2_{2|1} = \sigma^2_2(1 - \rho^2_{12})$$

- The conditional probability distribution of Y_1 for any given value of Y_2 is normal

- The means of the conditional probability distributions of Y_1 fall on a straight line and hence are a linear function of Y_2

$$\mathrm{E}[Y_1|Y_2] = \alpha_{1|2} + \beta_{12}Y_2$$

Here $\alpha_{1|2}$ is the intercept parameter and β_{12} the slope parameter; thus the relation between the conditional means and Y_2 is given by a linear regression function

- All conditional probability distributions of Y_1 have the same standard deviation $\sigma_{1|2}$
- Suppose a random sample of observations (Y_1, Y_2) was to be selected from a bivariate normal population and conditional inferences about Y_1, given Y_2, was to be made, then the normal error regression model is entire applicable because the Y_1 observations are independent and the Y_1 observations when Y_2 is considered given or fixed are normally distributed with mean $\mathrm{E}[Y_1|Y_2] = \alpha_{1|2} + \beta_{12}Y_2$ and constant variance $\sigma^2_{1|2}$
- All conditional inferences with these correlation models can be made by means of the usual regression methods
- If Y_1 and Y_2 are not bivariate normal, but if $Y_1 = Y$ and $Y_2 = X$ are random variables, then all results on estimation, testing and prediction obtained the regression model apply if the following conditions hold:
  - The conditional distributions of the Y_i, given X_i, are normal and independent, with conditional means $\beta_0 + \beta_1 X_i$ and conditional variance σ^2
  - The X_i are independent random variables whose probability distribution $g(X_i)$ does not involve the parameters β_0, β_1 and σ^2

  These conditions require only that the regression model is appropriate for each conditional distribution of Y_i and that the probability distribution of the X_i does not involve the repression parameters
- Two distinct regressions are involved in a bivariate normal model, that of Y_1 on Y_2 when Y_2 is fixed and that of Y_2 on Y_1 when Y_1 is fixed; in general, the two regression lines are not the same
- When interval estimates for the conditional correlation models are obtained, the confidence coefficient refers to repeated samples where pairs of observations (Y_1, Y_2) are obtained from the bivariate normal distribution
- A principal use of the bivariate normal correlation model is to study the relationship between two variables; in a bivariate normal model, the parameter ρ_{12} provides information about the degree of the linear relationship between the two variables Y_1 and Y_2
- The maximum likelihood estimator of ρ_{12}, denoted by r_{12}, is given by

$$r_{12} = \frac{\sum(Y_{i1} - \overline{Y}_1)(Y_{i2} - \overline{Y}_2)}{[\sum(Y_{i1} - \overline{Y}_1)^2 \sum(Y_{i2} - \overline{Y})^2]^{\frac{1}{2}}}$$

  This estimator is called the Pearson product-moment correlation coefficient; if is a biased estimator of ρ_{12} (unless $\rho_{12} = 0$ or 1), but the bias is small when n is large
- The range of r_{12} is $-1 \leq r_{12} \leq 1$; generally, values of r_{12} near 1 indicate a strong positive (direct) linear association between Y_1 and Y_2 whereas values of r_{12} near -1 indicate a strong negative (indirect) linear association; values of r_{12} near 0 indicate little or no linear association between Y_1 and Y_2
- When the population is bivariate normal, it is desired to test whether the coefficient of correlation is zero

$$H_0 : \ \rho_{12} = 0$$
$$H_A : \ \rho_{12} \neq 0$$

When Y_1 and Y_2 are jointly normally distributed, $\rho_{12} = 0$ implies that Y_1 and Y_2 are independent

- The test statistic for testing these hypotheses is

$$t^* = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}}$$

If H_0 holds, t^* follows the t_{n-2} distribution; the appropriate decision rule to control the Type I error at α is

  - If $|t^*| \leq t_{1-\frac{\alpha}{2},n-2}$, conclude H_0
  - If $|t^*| > t_{1-\frac{\alpha}{2},n-2}$, conclude H_A

- Since the sampling distribution of r_{12} is complicated with $\rho_{12} \neq 0$, interval estimation of ρ_{12} is usually carried out by means of an approximate procedure based on a Fisher z transformation

$$z' = \frac{1}{2}\log_e\left(\frac{1+r_{12}}{1-r_{12}}\right)$$

When n is large, the distribution of z' is approximately normal with approximate mean and variance:

$$\mathrm{E}\left[z'\right] = \xi = \frac{1}{2}\log_e\left(\frac{1+\rho_{12}}{1-\rho_{12}}\right)$$

$$\mathrm{Var}\left[z'\right] = \frac{1}{n-3}$$

- When the sample size is large, the standardized statistic:

$$\frac{z'-\xi}{s\left[z'\right]}$$

is approximately a standard normal variable; therefore, approximate $1-\alpha$ confidence limits for ξ are

$$z' \pm z_{1-\frac{\alpha}{2}} s\left[z'\right]$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})100$ percentile of the standard normal distribution; the $1-\alpha$ confidence limits for ρ_{12} are then obtained by transforming the limits on ξ using the appropriate mean of z' above

- A confidence interval for ρ_{12} can be employed to test whether or not ρ_{12} has a specified value by noting whether or not the specified value falls within the confidence limits

- It can be shown that the square of the coefficient of correlation, ρ_{12}^2, measures the relative reduction in the variability of Y_2 associated with the use of variable Y_1; note that

$$\sigma_{1|2}^2 = \sigma_1^2(1-\rho_{12}^2) \qquad \rho_{2|1}^2 = \sigma_2^2(1-\rho_{12}^2)$$

Then these expressions can be rewritten as

$$\rho_{12}^2 = \frac{\sigma_1^2 - \sigma_{1|2}^2}{\sigma_1^2} = \frac{\sigma_2^2 - \sigma_{2|1}^2}{\sigma_2^2}$$

ρ_{12}^2 measures how much smaller relatively is the variability in the conditional distributions of Y_1, for any given level of Y_2, than is the variability in the marginal distribution of Y_1; thus, ρ_{12}^2 measures the relative reduction in the variability of Y_1 associated with the use of Y_2; correspondingly, ρ_{12}^2 also measures the relative reduction in the variability of Y_2 associated with the use of variable Y_1

- The limits of ρ_{12}^2 are $0 \leq \rho_{12}^2 \leq 1$; the limiting value $\rho_{12}^2 = 0$ occurs when Y_1 and Y_2 are independent, so that the variances of each variable in the conditional probability distributions are then no smaller than the variance in the marginal distribution; the limiting value $\rho_{12}^2 = 1$ occurs when there is no variability in the conditional probability distributions for each variable, so perfect predictions of either variable can be made from each other

- The interpretation of ρ_{12}^2 as measuring the relative reduction in the conditional variances as compared with the marginal variance is valid for the case of a bivariate normal population, but not for many other bivariate populations

- Confidence limits for ρ_{12}^2 can be obtained by squaring the respective confidence limits for ρ_{12}, provided the latter limits do not differ in sign

- When the joint distribution of two random variables Y_1 and Y_2 differs considerably from the bivariate normal distribution, transformations of the variables Y_1 and Y_2 may be sought to make the joint distribution of the transformed variables approximately bivariate normal

- When no appropriate transformations can be found, a nonparametric rank correlation procedure may be useful for making inferences about the association between Y_1 and Y_2

- The Spearman rank correlation coefficient is calculated as follows: first the observations on Y_1 are expressed in ranks from 1 to n and denoted by R_{i1}; similarly, the observations on Y_2 are ranked, denoted by R_{i2}; the Spearman rank correlation coefficient, r_s , is then defined as the ordinary Pearson product-moment correlation coefficient based on the rank data:

$$r_s = \frac{\sum(R_{i1} - \overline{R}_1)(R_{i2} - \overline{R}_2)}{[\sum(R_{i1} - \overline{R}_1)^2 \sum(R_{i2} - \overline{R}_2)^2]^{\frac{1}{2}}}$$

  Here $\overline{R}_1$ is the mean of the ranks R_{i1} and $\overline{R}_2$ is the mean of the ranks R_{i2}

- Note that

$$\overline{R}_1 = \overline{R}_2 = \frac{n+1}{2}$$

  since the ranks are the integers $1, \ldots, n$

- The Spearman rank correlation coefficient takes on values between -1 and 1 inclusive: $-1 \leq r_s \leq 1$; the coefficient r_s equals 1 when the ranks for Y_1 are identical to those for Y_2; in that case, there is perfect association between the ranks for the two variables; the coefficient r_s equals -1 when the case with rank 1 for Y_1 has rank n for Y_2, the case with rank 2 for Y_1 has rank $n-1$ for Y_2 and so on; here, there is perfect inverse association between the ranks for the two variables; when there is little, if any, association between the ranks of Y_1 and Y_2, the Spearman rank correlation coefficient tends to have a value near zero

- The Spearman rank correlation coefficient can be used the test the alternatives:

  - H_0: There is no association between Y_1 and Y_2

  - H_A: There is an association between Y_1 and Y_2

  A two-sided test is conducted here since H_A includes either positive or negative association

- When the alternative H_A is:

$$H_A: \text{ There is positive (negative) association between } Y_1 \text{ and } Y_2$$

  an upper-tail (power-tail) one-sided test is conducted

- The probability distribution of r_s under H_0 is based on the condition that, for any ranking of Y_1, all rankings of Y_2 are equally likely when there is no association between Y_1 and Y_2

- When the sample size n exceeds 10, the test can be carried out approximately by using the following test statistic

$$t^* = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

  based on the t distribution with $n-2$ degrees of freedom

- Another nonparametric rank procedure similar to Spearman's r_s is Kendall's τ; this statistic also measures how far the rankings of Y_1 and Y_2 differ from each other, but in a somewhat different way than the Spearman rank correlation coefficient

1.3 Diagnostics and Remedial Measures

1.3.1 Diagnostics for Predictor Variable

1.3.2 Residuals

1.3.3 Diagnostics for Residuals

1.3.4 Overview for Tests Involving Residuals

1.3.5 Correlation Test for Normality

1.3.6 Tests for Constancy of Error Variance

1.3.7 F Test for Lack of Fit

1.3.8 Overview of Remedial Measures

1.3.9 Transformations

1.3.10 Exploration of Shape of Regression Function

1.3.11 Case Example - Plutonium

1.4 Simultaneous Inferences and Other Topics in Regression Analysis

1.4.1 Joint Estimation of β_0 and β_1

1.4.2 Simultaneous Estimation of Mean Responses

1.4.3 Simultaneous Prediction Intervals for New Observations

1.4.4 Regression through Origin

1.4.5 Effects of Measurement Errors

1.4.6 Inverse Predictions

1.4.7 Choice of X Levels

1.5 Matrix Approach to Simple Linear Regression Analysis

1.5.1 Matrices

1.5.2 Matrix Addition and Subtraction

1.5.3 Matrix Multiplication

1.5.4 Special Types of Matrices

1.5.5 Linear Dependence and Rank of Matrix

1.5.6 Inverse of a Matrix

1.5.7 Some Basic Results for Matrices

1.5.8 Random Vectors and Matrices

1.5.9 Simple Linear Regression Model in Matrix Terms

1.5.10 Least Squares Estimation of Regression Parameters

1.5.11 Fitted Values and Residuals

1.5.12 Analysis of Variance Results