Applied Linear Statistical Models
Outline

Darshan Patel

January 18, 2021

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

- Diagnostic information about the predictor variable tell if there are any outlying X values that could influence the appropriateness of the fitted regression function

- A dot plot is helpful when the number of observations in the data set is not large

- A sequence plot is useful when the data is obtained in a sequence, such as over time or for adjacent geographic areas

- A stem and leaf plot provides information similar to a frequency histogram but by displaying last digits explicitly

- A box plot shows the minimum, maximum, first quartile, third quartile and the median of the data set; this visualization is particularly helpful when there are many observations in the data set

1.3.2 Residuals

- Direct diagnostic plots for the response variable Y are ordinarily not too useful in regression analysis because the values of the observations on the response variable are a function of the level of the predictor variable

- The residual e_i is the difference between the observed value Y_i and the fitted value $\hat{Y}_i$

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

  The residual may be regarded as the observed error, in distinction to the unknown true error ε_i in the regression model

$$\varepsilon_i = Y_i - \mathrm{E}[Y_i]$$

- If the model is appropriate for the data at hand, the observed residuals e_i should then reflect the properties assumed for the ε_i
- The mean of the n residuals e_i for the simple linear regression model is

$$\overline{e} = \frac{\sum e_i}{n} = 0$$

where $\overline{e}$ denotes the mean of the residuals; since $\overline{e}$ is always 0, it provides no information as to whether the true errors ε_i have expected value $\mathrm{E}[\varepsilon_i] = 0$

- The variance of the n residuals e_i for the simple linear regression model is

$$s^2 = \frac{\sum (e_i - \overline{e})^2}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{\mathrm{SSE}}{n-2} = \mathrm{MSE}$$

If the model is appropriate, the MSE is an unbiased estimator of the variance of the error terms σ^2

- The residuals for the regression model are subjected to two constraints: the sum of the e_i must be 0 and the products $X_i e_i$ must sum to 0
- When the sample size is large in comparison to the number of parameters in the regression model, the dependency effect among the residuals e_i is relatively unimportant and can be ignored for most purposes
- It is helpful to consider the standardization of the residuals for residual analysis as it can identify outlying observations
- The semistudentized residual is defined as follows:

$$e_i^* = \frac{e_i - \overline{e}}{\sqrt{\mathrm{MSE}}} = \frac{e_i}{\sqrt{\mathrm{MSE}}}$$

- The following can be noted as departures from the simple linear regression model
  - The regression function is not linear
  - The error terms do not have constant variance
  - The error terms are not independent
  - The model fits all but one or a few outlier observations
  - The error terms are not normally distributed
  - One or several important predictor variables have been omitted from the model

1.3.3 Diagnostics for Residuals

- The following plots of residuals (or semistudentized residuals) can be utilized to diagnose any departures from the simple linear regression model
  - Plot of residuals against predictor variable
  - Plot of absolute or squared residuals against predictor variable
  - Plot of residuals against fitted values

  - Plot of residuals against time or other sequence
  - Plots of residuals against omitted predictor variables
  - Box plot of residuals
  - Normal probability plot of residuals

- Whether a linear regression function is appropriate for the data being analyzed can be studied from a residual plot against the predictor variable, or equivalently, from a residual plot against the fitted values

- Nonlinearity of the regression function can also be studied from a scatter plot, but this plot is not always as effective as a residual plot

- In general, the residual plot is to be preferred over the scatter plot due to its advantages

  - The residual plot can be used for examining other facets of the aptness of the model
  - When the scaling of the scatter plot places Y_i observations close to the fitted values $\hat{Y}_i$, it becomes difficult to study the appropriateness of a linear regression function from the scatter plot

- Note: A plot of residuals against the fitted values $\hat{Y}$ provides equivalent information as a plot of residuals against X for the simple linear regression model because the fitted values $\hat{Y}_i$ are a linear function of the values X_i for the predictor variable and thus, only the X scale values, not the basic pattern of the plotted points, are affected by whether the residual plot is against the X_i or the $\hat{Y}_i$

- Plots of the residuals against the predictor variable or against the fitted values are not only helpful to study whether a linear regression function is appropriate but also to examine whether the variance of the error terms is constant

- Plots of the absolute values of the residuals or of the squared residuals against the predictor variable X or against the fitted values $\hat{Y}$ are also useful for diagnosing nonconstancy of the error variance since the signs of the residuals are not meaningful for examining the constancy of the error variance

- These plots are especially useful when there are not many cases in the data set because plotting of either the absolute or squared residuals places all of the information on changing magnitudes of the residuals above the horizontal zero line so that one can more readily see whether the magnitude of the residuals (irrespective of sign) is changing with the level of X or $\hat{Y}$

- Residual outliers can be identified from residual plots against X or $\hat{Y}$, as well as from box plots, stem and leaf plots and dot plots of the residuals

- Plotting of semistudentized residuals is helpful for distinguishing outlying observations, since it then becomes easy to identify residuals that lie many standard deviations away from zero

- A rough rule of thumb when the number of cases is large is to consider semistudentized residuals with absolute value of four or more to be outliers

- Outliers can create great difficulty - a major reason for discarding it is that under the least squares method, a fitted line may be pulled disproportionately toward an outlying observation because the sum of the squared deviations is minimized, causing a misleading fit if indeed the outlying observation resulted from a mistake or other extraneous cause
- When a linear regression model is fitted to a data set with a small number of cases and an outlier is present, the fitted regression can be so distorted by the outlier that the residual plot may improperly suggest a lack of fit for the linear regression model, in addition to flagging the model
- Whenever data are obtained in a time sequence or some other type of sequence, it is a good idea to prepare a sequence plot of the residuals to see if there is any correlation between error terms that are near each other in the sequence
- When the error terms are independent, the residuals in a sequence plot are to be expected to fluctuate in a more or less random pattern around the base line 0; lack of randomness can take the form of too much or too little alternation of points around the zero line
- Small departures from normality do not create any serious problems; major departures should be of concern, such as the normality of the error terms
- A box plot of the residuals is helpful for obtaining summary information about the symmetry of the residuals and about possible outliers; a histogram, dot plot or stem and leaf plot of the residuals can also be useful for detecting gross departures from normality
- When the number of cases is reasonably large, it can be useful to compare actual frequencies of the residuals against expected frequencies under normality
- A normal probability plot of the residuals can also shed some information about the normality of error terms; here each residual is plotted against its expected value under normality; a plot that is nearly linear suggests agreement with normality, whereas a plot that departs substantially from linearity suggests that the error distribution is not normal
  - To find the expected values of the ordered residuals under normality, note that the expected value of the error terms for the regression model is zero and that the standard deviation of the error terms is estimated by $\sqrt{\text{MSE}}$
  - For a normal random variable with mean 0 and estimated standard deviation $\sqrt{\text{MSE}}$, a good approximation of the expected value of the kth smallest observation in a random sample of n is
    $$\sqrt{\text{MSE}}\left(z\left(\frac{k-.375}{n+.25}\right)\right)$$
    where $z(A)$ is the $(A)100$ percentile of the standard normal distribution
- When the distribution of the error terms departs substantially from normality:
  - If the error term distribution is highly skewed to the right in the normal probability plot, the plot is shaped concave-upward
  - If the error term distribution is highly skewed to the left in the normal probability plot, the plot is shaped concave-downward

  - If the error term distribution is symmetrical but has heavy tails in the normal probability plot, one side is concave-upward while the other is concave-downward; here the distribution has higher probabilities in the tails than a normal distribution

- The analysis for model departures with respect to normality is more difficult than that for other types of departure because random variation can be particularly mischievous when studying the nature of a probability distribution unless the sample size is quite large; even worse, other types of departures can and do affect the distribution of the residuals
- Residuals should also be plotted against variables omitted from the model that might have important effects on the response to determine whether there are any other key variables that could provide important additional descriptive and predictive power to the model
- Note that in actuality, several types of model departures may occur together
- Although graphic analysis of residuals is only an informal method of analysis, in many cases it suffices for examining the aptness of a model
- Model misspecification due to either nonlinearity or the omission of important predictor variables tends to be serious, leading to biased estimates of the regression parameters and error variance
- Nonconstancy of error variance tends to be less serious, leading to less efficient estimates and invalid error variance estimates
- The presence of outliers can be serious for smaller data sets when their influence is large
- The nonindependence of error terms results in estimators that are unbiased but whose variances are seriously biased

1.3.4 Overview for Tests Involving Residuals

- A runs test is frequently used to test for lack of randomness in the residuals arranged in time order; another, specifically designed for lack of randomness in least squares residuals, is the Durbin-Watson test
- When a residual plot gives the impression that the variance may be increasing or decreasing in a systematic manner related to X or $\mathrm{E}[Y]$, a simple test is based on the rank correlation between the absolute values of the residuals and the corresponding values of the predictor variable; two other simple tests for constancy of the error variance are the Brown-Forsythe test and the Breusch-Pagan test
- A simple test for identifying an outlier observation involves fitting a new regression line to the other $n-1$ observations; the suspect observation can now be regarded as a new observation; the probability that in n observations, a deviation from the fitted line as great as that of the outlier will be obtained by chance can be calculated; if this probability is sufficiently small, the outlier can be rejected as not having come from the same population as the other $n-1$ observations; otherwise, the outlier is retained
- Goodness of fit tests can be used for examining the normality of the error terms, such as the chi-square test or the Kolmogorov-Smirnov test and its modification, as well as the Lilliefors test; a simple test based on the normal probability plot of the residuals is also useful

1.3.5 Correlation Test for Normality

- A formal test for normality of the error terms can be conducted by calculating the coefficient of correlation between the residuals e_i and their expected values under normality
- A high value of the correlation coefficient is indicative of normality
- If the observed coefficient of correlation is at least as large as the tabled value of the expected value for a given α level, one can conclude that the error terms are reasonably normally distributed
- The correlation test for normality shown here is simpler than the Shapiro-Wilk test, which can be viewed as being based approximately also on the coefficient of correlation between the ordered residuals and their expected values under normality

1.3.6 Tests for Constancy of Error Variance

- The Brown-Forsythe test and the Breusch-Pagan test can both be used to ascertain whether the error terms have constant variance
- Brown-Forsythe Test
  - The Brown-Forsythe test does not depend on normality of the error terms and is thus robust against serious departures from normality, in the sense that the nominal significance level remains approximately correct when the error terms have equal variances even if the distribution of the error terms is far from normal
  - The Brown-Forsythe test is applicable to simple linear regression when the variance of the error terms either increases or decreases with X; the sample size needs to be large enough so that the dependencies among the residuals can be ignored
  - The test is based on the variability of the residuals; the larger the error variance, the larger the variability of the residuals will tend to be
  - To conduct the Brown-Forsythe test, divide the data set into two groups, according to the level of X, so that one group consists of cases where the X level is comparatively low and the other group consists of cases where the X level is comparatively high
  - If the error variance is either increasing or decreasing with X, the residuals in one group will tend to be more variable than those in the other group; equivalently, the absolute deviations of the residuals around their group mean will tend to be larger for one group than for the other group
  - The Brown-Forsythe test consists simply of the two-sample t test based on the test statistic to determine whether the mean of the absolute deviations of the residuals around the median for one group differs significantly from the mean absolute deviations os the residuals around the median of the other group
  - Let e_{i1} denote the ith residual for group 1 and e_{i2} the ith residual for group 2 and let n_1 and n_2 denote the sample sizes of the two groups such that $n = n_1 + n_2$
  - Let $\overline{e}_1$ and $\overline{e}_2$ be the medians of the residuals in the two groups; the Brown-Forsythe test uses the absolute deviations of the residuals around their group median, to be denoted by d_{i1} and d_{i2}:

$$d_{i1} = |e_{i1} - \overline{e}_1| \qquad d_{i2} = |e_{i2} - \overline{e}_2|$$

- The two-sample t test statistic is

$$t^*_{BF} = \frac{\overline{d}_1 - \overline{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\overline{d}_1$ and $\overline{d}_2$ are the sample means of the d_{i1} and d_{i2} respectively and the pooled variance s^2 is

$$s^2 = \frac{\sum(d_{1i} - \overline{d}_1)^2 + \sum(d_{i2} - \overline{d}_2)^2}{n-2}$$

- If the error terms have constant variance and n_1 and n_2 are sufficiently large, t^*_{BF} follows the t_{n-2} distribution; large absolute values of t^*_{BF} indicates that the error terms do not have constant variance
- If the data set contains many cases, the two-sample t test for constancy of error variance can be conducted after dividing the cases into three or four groups, according to the level of X, and using the two extreme groups
- A robust test for constancy of the error variance is desirable because nonnormality and lack of constant variance go hand in hand

- Breutch-Pagan Test
  - The Breutch-Pagan test, a large sample test, assumes that the error terms are independent and normally distributed and that the variance of the error terms ε_i, denoted by σ_i^2, is related to the level of X such that

$$\log_e \sigma_i^2 = \gamma_0 + \gamma_1 X_i$$

This implies that σ_i^2 either increases or decreases with the level of X, depending on the sign of γ_1

  - Constancy of error variance corresponds to $\gamma_1 = 0$
  - The test of $H_0 : \gamma_1 = 0$ vs $H_A : \gamma_1 \neq 0$ is carried out by means of regressing the squared residuals e_i^2 against X_i in the usual manner and obtaining the regression sum of squares, SSR^*
  - The test statistic χ^2_{BP} is as follows:

$$\chi^2_{BP} = \frac{\text{SSR}^*}{2} \Big/ \left(\frac{\text{SSE}}{n}\right)^2$$

where SSR^* is the regression sum of squares when regressing e^2 on X and SSE is the error sum of squares when regressing Y on X

  - If $H_0 : \gamma_1 = 0$ holds and n is reasonably large, χ^2_{BP} follows the χ^2_1 distribution
  - Large values of of χ^2_{BP} lead to the conclusion H_A, that the error variance is not constant
  - The Breusch-Pagan test can be modified to allow for different relationships between the error variance and the level of X
  - The test statistic was developed independently by Cook and Weisberg and the test is sometimes referred to as the Cook-Weisberg test

1.3.7 F Test for Lack of Fit

- The F test for lack of fit is for ascertaining whether a linear regression function is a good fit for the data
- The lack of fit test assumes that the observations Y for given X are (1) independent and (2) normally distributed, and that (3) the distributions of Y have the same variance σ^2
- The lack of fit test requires repeat observations at one or more X levels
- Repeat trials for the same level of the predictor variable are called replications and the resulting observations are called replicates
- The different X levels in a study, whether or not replicated observations are present, are denoted as $X_1, \ldots, X_c$; the number of replicates for the jth level of X is denoted as n_j and therefore the total number of observations n is

  $$n = \sum_{j=1}^{c} n_j$$

  The observed value of the response variable for the ith replicate for the jth level of X is denoted by Y_{ij}, where $i = 1, \ldots, n_j$ and $j = 1, \ldots, c$
- The general linear test approach begins with the specification of the full model, making the same assumptions as the simple linear regression model except for assuming a linear regression function, the subject of the test

  $$Y_{ij} = \mu_i + \varepsilon_{ij}$$

  where μ_j are parameters $j = 1, \ldots, c$ and ε_{ij} are independent $N(0, \sigma^2)$
- Since the error terms have expectation zero, it follows that $\mathrm{E}[Y_{ij}] = \mu_j$ and so the parameter μ_j is the mean response when $X = X_j$
- Note that the full model here makes no restrictions on the mean μ_j, whereas in the regression model, the mean responses are linear related to X (i.e., $\mathrm{E}[Y] = \beta_0 + \beta_1 X$)
- To fit the full model to the data, the least squares or maximum likelihood estimators for the parameters μ_j is needed, which is simply the sample means $\overline{Y}_j$
- The estimated expected value for observation Y_{ij} is $\overline{Y}_j$ and the error sum of squares for the full model therefore is

  $$\mathrm{SSE}(F) = \sum_j \sum_i (Y_{ij} - \overline{Y}_j)^2 = \mathrm{SSPE}$$

  This is called the pure error sum of squares, or SSPE
- Note that SSPE is made up of the sums of squared deviations at each X level; at level $X = X_j$, this sum of squared deviations is $\sum_i (Y_{ij} - \overline{Y}_j)^2$
- Note that any X level with no replications make no contribution to SSPE because $\overline{Y}_j = Y_{1j}$
- The degrees of freedom associated with SSPE is the sum of the component degrees of freedom:

  $$df_F = \sum_j (n_j - 1) = \sum_j n_j - c = n - c$$

- The general linear test approach also requires consideration of the reduced model under H_0; the alternatives are
$$H_0 : \mathrm{E}[Y] = \beta_0 + \beta_1 X$$
$$H_A : \mathrm{E}[Y] \neq \beta_0 + \beta_1 X$$
and so H_0 postulates that μ_j in the full model is linear related to X_j: $\mu_j = \beta_0 + \beta_1 X_j$

- The reduced model under H_0 is
$$Y_{ij} = \beta_0 + \beta_1 X_j + \varepsilon_{ij}$$

- The error sum of squares for the reduced model is the usual error sum of squares SSE
$$\mathrm{SSE}(R) = \sum\sum(Y_{ij} - (b_0 + b_1 X_j)^2) = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2 = \mathrm{SSE}$$

- The degrees of freedom associated with SSE(R) is
$$df_R = n - 2$$

- The general linear test statistic here is
$$F^* = \frac{\mathrm{SSE}(R) - \mathrm{SSE}(F)}{df_R - df_F} \Big/ \frac{\mathrm{SSE}(F)}{df_F} = \frac{\mathrm{SSE} - \mathrm{SSPE}}{(n-2) - (n-c)} \Big/ \frac{\mathrm{SSPE}}{n-c}$$

- The difference between the two error sums of squares is called the lack of fit sum of squares and is denoted by SSLF
$$\mathrm{SSLF} = \mathrm{SSE} - \mathrm{SSPE}$$

- The test statistic can then be expressed as
$$F^* = \frac{\mathrm{SSLF}}{c-2} \Big/ \frac{\mathrm{SSPE}}{n-c} = \frac{\mathrm{MSLF}}{\mathrm{MSPE}}$$
where MSLF denotes the lack of fit mean square and MSPE denotes the pure error mean square

- Large values of F^* lead to rejection of H_0 in the general linear test; the decision rules are as follows:
$$\text{If } F^* \leq F_{1-\alpha,c-2,n-c}, \text{ conclude } H_0$$
$$\text{If } F^* > F_{1-\alpha,c-2,n-c}, \text{ conclude } H_A$$

- The error sum of squares SSE can be decomposed as follows
$$\mathrm{SSE} = \mathrm{SSPE} + \mathrm{SSLF}$$
which follows from the identity
$$\underbrace{Y_{ij} - \hat{Y}_{ij}}_{\text{error deviation}} = \underbrace{Y_{ij} - \overline{Y}_j}_{\text{pure error deviation}} + \underbrace{\mathrm{Var}[Y]_j - \hat{Y}_{ij}}_{\text{lack of fit deviation}}$$
This identity shows that the error deviations in SSE are made up of a pure error component and a lack of fit component

- When this is squared and summed over all observations, the following is achieved

$$\sum\sum(Y_{ij} - \hat{Y}_{ij})^2 = \sum\sum(Y_{ij} - \overline{Y}_j)^2 + \sum\sum(\overline{Y}_j - \hat{Y}_{ij})^2$$
$$\text{SSE} = \text{SSPE} + \text{SSLF}$$

  Thus the lack of fit sum of squares can then be defined as

$$\text{SSLF} = \sum\sum(\overline{Y}_j - \hat{Y}_{ij})^2$$

  Since all Y_{ij} observations at the level X_j have the same fitted value, denoted by $\hat{Y}_j$, this can be stated equivalently as

$$\text{SSLF} = \sum_j n_j(\overline{Y}_j - \hat{Y}_j)^2$$

- If the linear regression function is appropriate, then the means $\overline{Y}_j$ will be near the fitted values $\hat{Y}_j$ calculated from the estimated linear regression function and SSLF will be small; if the linear regression function is not appropriate, the means $\overline{Y}_j$ will not be near the fitted values calculated from the estimated linear regression function and SSLF will be large

- There are $c-2$ degrees of freedom associated with SSLF because there are c means $\overline{Y}_j$ in the sum of squares and 2 degrees of freedom are lost in estimating the parameters β_0 and β_1 of the linear regression function to obtain the fitted values $\hat{Y}_j$

- An ANOVA table can be constructed for the decomposition of SSE

Source of Variation	Sum of Squares	df	Mean Squares
Regression	$\text{SSR} = \sum\sum(\hat{Y}_{ij} - \overline{Y})^2$	1	$\text{MSR} = \frac{\text{SSR}}{1}$
Error	$\text{SSE} = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$\text{MSE} = \frac{\text{SSE}}{n-2}$
Lack of Fit	$\text{SSLF} = \sum\sum(\overline{Y}_j - \hat{Y}_{ij})^2$	$c-2$	$\text{MSLF} = \frac{\text{SSLF}}{c-2}$
Pure error	$\text{SSPE} = \sum\sum(Y_{ij} - \overline{Y}_j)^2$	$n-c$	$\text{MSPE} = \frac{\text{SSPE}}{n-c}$
Total	$\text{SSTO} = \sum\sum(Y_{ij} - \overline{Y})^2$	$n-1$	

- It can be shown that the mean squares MSPE and MSLF have the following expectations when testing whether the regression function is linear

$$\text{E}[\text{MSPE}] = \sigma^2$$
$$\text{E}[\text{MSLF}] = \sigma^2 + \frac{\sum n_j(\mu_j - (\beta_0 + \beta_1 X_j))^2}{c-2}$$

- The terminology "error sum of squares" and "error mean square" is not precise when the regression function under test in H_0 is not the true function since the error sum of squares and error mean square then reflect the effects of both the lack of fit and the variability of the error terms

- Note that when concluding H_0, or that $\beta_1 = 0$, it is to say that there is no linear association between X and Y, not "no relation" between the two variables

- The general linear test approach can be used to test the appropriateness of other regression functions; only the degrees of freedom for SSLF needs be be modified; in general, $c-p$ degrees of freedom are associated with SSLF, where p is the number of parameters in the regression function

- The alternative H_A in the test hypotheses includes all regression functions other a linear one
- When concluding that the employed model in H_0 is appropriate, the usual practice is to use the error mean square MSE as an estimator of σ^2 in preference to the pure error mean square MSPE, since the former contains more degrees of freedom
- Observations at the same level of X are genuine repeats only if they involve independent trials with respect to the error term
- When no replications are present in a data set, an approximate test of lack of fit can be conducted if there are some cases at adjacent X levels for which the mean responses are quite close to each other; such adjacent cases are grouped together and treated as pseudoreplicates, and the test for lack of fit is then carried out using these groupings of adjacent cases

1.3.8 Overview of Remedial Measures

- If the simple linear regression model is not appropriate for a data set, there are two basic choices:
  - Abandon the regression model and develop and use a more appropriate model - may lead to more complex procedures for estimating the parameters
  - Employ some transformation on the data so that the regression model is appropriate for the transformed data - can lead to relatively simple methods of estimation and may involve fewer parameters than a complex model
- When the regression function is not linear, a direct approach is to modify the regression model by altering the nature of the regression function such as by using a quadratic or exponential regression function
- When the error variance is not constant but varies in a systematic fashion, a direct approach is to modify the model to allow this and use the method of weighed least squares to obtain the estimators of the parameters; transformations can also be effective in stabilizing the variance
- When the error terms are correlated, a direct remedial measure is to work with a model that calls for correlated error terms; a simple remedial transformation that is often helpful is to work with first differences
- Lack of normality and nonconstant error variance frequently go hand in hand; it is often the case that the same transformation that helps stabilize the variance is also helpful in approximately normalizing the error terms
- When residual analysis indicates that an important predictor variable has been omitted from the model, the solution is to modify the model to include that predictor variable
- When outlying observations are present, use of the least squares and maximum likelihood estimators for the regression model may lead to serious distortions in the estimated regression function; when the outlying observations should not be discarded, it may be desirable to use an estimation procedure that places less emphasis on such outlying observations

1.3.9 Transformations

- When the distribution of the error terms is reasonably close to a normal distribution and the error terms have approximately constant variance, transformations on X should attempted because a transformation on Y may materially change the shape of the distribution of the error terms from the normal distribution and may also lead to substantially differing error term variances
- When the error variance is constant and the regression pattern is
  - increasing and concave down, consider $X' = \log_{10} X$ or $X' = \sqrt{X}$
  - increasing and concave up, consider $X' = X^2$ or $X' = \exp(X)$
  - decreasing, consider $X' = 1/X$ or $X' = \exp(-X)$
- If some of the X data are near zero and the reciprocal transformation is desired, the origin can be shifted by using the transformation $X' = 1/(X+k)$ where k is an appropriately chosen constant
- Scatter plots and residuals plots based on each transformation can be prepared and analyzed to decide which transformation is most effective
- Unequal error variances and nonnormality of the error terms frequently appear together and to remedy these departures from the simple linear regression model, a transformation on Y is needed since the shapes and spreads of the distributions of Y need to be changed, and can also help to linearize a curvilinear regression relation
- When the error variance is nonconstant and its pattern is
  - increasing and concave down, consider $Y' = \sqrt{Y}$
  - decreasing and concave up, consider $Y' = \log_{10} Y$
  - increasing at constant slope, consider $Y' = 1/Y$
- When Y may be negative, it may be desirable to introduce a constant into a transformation of Y; in the case of the logarithmic transformation, shift the origin in Y and make all observations positive by using $Y' = \log_{10}(Y + k)$ where k is an appropriately chosen transformation
- Several alternative transformations of Y may be tried, as well as some simultaneous transformations on X; scatter plots and residual plots should be prepared to determine the most effective transformation(s)
- When unequal error variances are present but the regression relation is linear, a transformation on Y may not be sufficient since it may be change the linear linear relationship to a curvilinear and thus a transformation on X may also be required; weighted least squares is another option in this case
- The Box-Cox procedure can automatically identify a transformation from the family of power transformations on Y which is most appropriate for correcting skewness of the distributions of error terms, unequal error variances and nonlinearity of the regression function
- The family of power transformations is of the form:

  $$Y' = Y^{\lambda}$$

  where λ is a parameter to be determined from the data

- The family of power transformations encompasses the following simple transformations

$\lambda = 2$	$\lambda = .5$	$\lambda = 0$	$\lambda = -.5$	$\lambda = -1.0$
$Y' = Y^2$	$Y' = \sqrt{Y}$	$Y' = \log_e(Y)$	$Y' = \frac{1}{\sqrt{Y}}$	$Y' = \frac{1}{Y}$

- The normal error regression model with the response variable a member of the family transformations becomes

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

- The Box-Cox procedure uses the method of maximum likelihood to estimate λ as well as the other parameters β_0, β_1 and σ^2; thus the Box-Cox procedure identifies $\hat{\lambda}$, the maximum likelihood estimate of λ to use in the power transformation

- A simple procedure for obtaining $\hat{\lambda}$ involves a numerical search in a range of potential λ values; for each λ value, the Y_i^λ observations are first standardized so that the magnitude of the error sum of squares does not depend on the value of λ

$$W_i = \begin{cases} K_i(Y_i^\lambda - 1) & \lambda \neq 0 \\ K_2(\log_e(Y_i)) & \lambda = 0 \end{cases}$$

where

$$K_2 = \left(\prod_{i=1}^{n} Y_i\right)^{\frac{1}{n}}$$

$$K_1 = \frac{1}{\lambda K_2^{\lambda-1}}$$

Note that K_2 is the geometric mean of the Y_i observations

- Once the standardized observations W_i have been obtained for a given λ value, they are regressed on the predictor variable X and the error sum of squares SSE is obtained; the maximum likelihood estimate $\hat{\lambda}$ is that value of λ for which SSE is a minimum

- If desired, a finer search can be conducted in the neighborhood of the λ value that minimizes SSE but the Box-Cox procedure is only used to provide a guide for selecting a transformation, so overly precise results are not needed

- At times, theoretical or a priori considerations can be utilized to help in choosing an appropriate transformation

- After a transformation has been tentatively selected, residual plots and other analyses described earlier need to be employed to ascertain that the simple linear regression model is appropriate for the transformed data

- When transformed models are employed, the estimators β_0 and β_1 obtained by least squares have the least squares properties with respect to the transformed observations, not the original ones

- The maximum likelihood estimate of λ with the Box-Cox procedure is subject to sampling variability; in addition, the error sum of squares SSE is often fairly stable in a neighborhood around the estimate and thus it is reasonable to use a nearby λ value for which the power transformation is easy to understand

- When the Box-Cox procedure leads to a λ value near 1, no transformation of Y may be needed

1.3.10 Exploration of Shape of Regression Function

- Scatter plots often indicate readily the nature of the regression function; at other times, the scatter plot is complex and it becomes difficult to see the nature of the regression relationship, if any, from the plot
- It is helpful to explore the nature of the regression relationship by fitting a smoothed curve, or nonparametric regression curve, without any constraints on the regression function; these are useful not only for exploring regression relationships but also for confirming the nature of the regression function when the scatter plot visually suggests the nature of the regression relationship
- The method of moving averages uses the mean of the Y observations for adjacent time periods to obtain smoothed values; special procedures are required for obtaining smoothed values at two ends of the series; the larger the successive neighborhoods used for obtaining the smoothed values, the smoother the curve will be
- The method of running medians is similar to the method of moving averages, except that the median is used the average measure in order to reduce the influence of outlying observations
- With the method of running methods as well as with the moving average method, successive smoothing of the smoothed values and other refinements may be undertaken to provide a suitable smoothed curve
- When the X values are not equally spaced apart, a smoothing method such as band regression can be useful; this method divides the data set into a number of groups or "bands" consisting of adjacent cases according to their X levels; for each band, the median X value and the median Y value are calculated and the points defined by the pairs of these median values are then connected by straight lines
- The lowess method, a more refined nonparametric method than band regression, obtains a smoothed curve by fitting successively linear regression functions in local neighborhoods
- The name lowess stands for "locally weighted regression scatter plot smoothing"
- This method obtains the smoothed Y value at a given X by fitting a linear regression to the data in the neighborhood of the X value and then using the fitted value at X as the smoothed value
- Smoothed values at each end of the X range are also obtained by the lowess procedure
- The lowess method uses a number of refinements in obtaining the final smoothed values to improve the smoothing and to make the procedure robust to outlying observations

    1. The linear regression is weighted to give cases further from the middle X level in each neighborhood smaller weights
    2. To make the procedure robust to outlying observations, the linear regression fitting is repeated, with the weights revised so that cases that had large residuals in the first fitting receive smaller weights in the second fitting
    3. To improve the robustness of the procedure further, step 2 is repeated one or more times by revising the weights according to the size of the residuals in the latest fitting

- To implement the lowess procedure, one must choose the size of the successive neighborhoods to be used when fitting each linear regression as well as the weight function that gives less weight to neighborhood cases with X values far from each center X level and another weight function that gives less weight to cases with large residuals; finally, the number of iterations to make the procedure robust must be chosen
- In practice, two iterations appear to be sufficient to provide robustness
- The weight functions suggested by Cleveland appear to be adequate for many circumstances; hence, the primary choice to be made for a particular application is the size of the successive neighborhoods
- The larger the size of the successive neighborhoods, the smoother the function but the greater the danger that the smoothing will lose essential features of the regression relationship
- Smoothed curves are useful not only in the exploratory stages when a regression model is selected but they are also helpful in confirming the regression function chosen
- To confirm: the smoothed curve is plotted together with the confidence band for the fitted regression function; if the smoothed curve falls within the confidence band, there is supporting evidence of the appropriateness of the fitted regression function
- Note that smoothed curves, such as the lowess curve, do not provude an analytical expression for the functional form of the regression relationship; they only suggest the shape of the regression curve
- The lowess procedure is not restricted to fitting linear regression functions in each neighborhood; higher-degree polynomials can also be utilized with this method
- Smoothed curves are also useful when examining residual plots to ascertain whether the residuals (or the absolute or squared residals) follow some relationship with X or $\hat{Y}$

1.4 Simultaneous Inferences and Other Topics in Regression Analysis

1.4.1 Joint Estimation of β_0 and β_1

- Analysis of data frequently requires a series of estimates (or tests) being correct; the set of estimates (or tests) of interest are called the family of estimates (or tests)
- A statement confidence coefficient indicates the proportion of correct estimates that are obtained when repeated samples are selected and the specified confidence interval is calculated for each sample
- A family confidence coefficient indicates the proportion of families of estimates that are entirely correct when repeated samples are selected and the specified confidence intervals for the entire family are calculated for each sample
- A family confidence coefficient corresponds to the probability, in advance of sampling, that the entire family of statements will be correct

- The Bonferroni procedure for developing joint confidence intervals for β_0 and β_1 with a specified family confidence coefficient can be used to assure that the entire set of estimates is correct; each statement confidence coefficient is adjusted to be higher than $1-\alpha$ so that the family confidence coefficient is at least $1-\alpha$
- The ordinary confidence limits for β_0 and β_1 with statement confidence coefficients $1-\alpha$ each are:
$$b_0 \pm t_{1-\frac{\alpha}{2},n-2}s[b_0]$$
$$b_1 \pm t_{1-\frac{\alpha}{2},n-2}s[b_1]$$
- Let A_1 be the event that the first confidence interval does not cover β_0 and A_2 be the event that the second confidence interval does not cover β_1. then the Bonferroni inequality is
$$P(\overline{A}_1 \cap \overline{A}_2) \geq 1 - P(A_1) - 1 - P(A_2) = 1 - \alpha - \alpha = 1 - 2\alpha$$
- If β_0 and β_1 are separately estimated with 95% confidence intervals, the Bonferroni inequality guarantees a family confidence coefficient of at least 90% that both intervals based on the same sample are correct
- To obtain a family confidence coefficient of at least $1-\alpha$ for estimating β_0 and β_1, estimate β_0 and β_1 separately with statement confidence coefficients of $1-\frac{\alpha}{2}$, yielding the Bonferroni bound $1-\frac{\alpha}{2}-\frac{\alpha}{2}=1-\alpha$
- The $1-\alpha$ family confidence limits for β_0 and β_1 for the simple linear regression model by the Bonferroni procedure are:
$$b_0 \pm Bs[b_0] \quad b_1 \pm Bs[b_1]$$
where
$$B = t_{1-\frac{\alpha}{4},n-2}$$
and b_0, b_1, $s[b_0]$ and $s[b_1]$ are defined as
$$b_1 = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2}$$
$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1\sum X_i\right) = \overline{Y} - b_1\overline{X}$$
$$s^2[b_1] = \frac{MSE}{\sum(X_i-\overline{X})^2}$$
$$s^2[b_0] = MSE\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i-\overline{X})^2}\right]$$
Note that a statement confidence coefficient of $1-\frac{\alpha}{2}$ requires the $1-\frac{\alpha}{4}100$ percentile of the t distribution for a two-sided confidence interval
- The Bonferroni $1-\alpha$ family confidence coefficient is actually a lower bound on the true (but unknown) family confidence coefficient and are frequently specified at lower levels than when a single estimate is made

- The Bonferroni inequality can be extended to g simultaneous confidence intervals with family confidence coefficient $1-\alpha$

$$P\left(\cap_{i=1}^{g}\overline{A}_i\right) \geq 1-g\alpha$$

If g interval estimates are desired with family confidence coefficient $1-\alpha$, constructing each interval estimate with statement confidence coefficient $1-\frac{\alpha}{g}$ will suffice

- For a given family confidence coefficient, the larger the number of confidence intervals in the family, the greater becomes the multiple B, which may make some or all of the confidence intervals too wide to be helpful
- It is not necessary with the Bonferroni procedure that the confidence intervals have the same statement confidence coefficient; different statement confidence coefficients, depending on the importance of each estimate, can be used
- Joint confidence intervals can be used directly for testing
- The estimators b_0 and b_1 are usually correlated, but the Bonferroni simultaneous confidence limits only recognize this correlation by means of the bound on the family confidence coefficient; it can be shown that the covariance between b_0 and b_1 is

$$\sigma[b_0, b_1] = -\overline{X}\sigma^2[b_1]$$

If $\overline{X}$ is positive, b_0 and b_1 are negatively correlated, implying that if the estimate b_1 is too high, the estimate b_0 is likely to be too low, and vice versa

1.4.2 Simultaneous Estimation of Mean Responses

- The mean responses at a number of X levels often need to be estimated from the same sample data; to assure that the all of the estimates of mean responses are correct, a family confidence coefficient is used
- A family confidence coefficient is needed to estimate several mean responses, even though all estimates are based on the same fitted regression line, because the separate interval estimates of $E[Y_h]$ at the different X_h levels need not all be correct or all be incorrect; the combination of sampling errors in b_0 and b_1 may be such that the interval estimates of $E[Y_h]$ will be correct over some range of X levels and incorrect elsewhere
- The Working-Hotelling procedure, based on the confidence band for the regression line, contains the mean responses at all X levels; to obtain simultaneous confidence intervals for the mean responses at selected X levels, use the boundary values for the X levels of interest
- The simultaneous confidence limits for g mean responses $E[Y_h]$ for the simple linear regression model, with the Working-Hotelling procedure, is

$$\hat{Y}_h \pm Ws[\hat{Y}_h]$$

where

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

and $\hat{Y}_h$ and $s[\hat{Y}_h]$ defined as

$$\hat{Y}_h = b_0 + b_1X_h$$

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

- To construct a family of confidence intervals for mean responses at different X levels with the Bonferroni procedure, calculate the usual confidence limits for a single mean response $E[Y_h]$ and adjust the statement confidence coefficient to yield the specified family confidence coefficient
- When $E[Y_h]$ is to be estimated for g levels X_h with family confidence coefficient $1-\alpha$, the Bonferroni confidence limits for the simple linear regression model is

$$\hat{Y}_h \pm Bs[\hat{Y}_h]$$

where

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

where g is the number of confidence intervals in the family

- In cases where the number of statements is small, the Bonferroni confidence limits may be tighter than the Working-Hoteling confidence limits; for larger families, the Working-Hotelling confidence limits will always be the tighter one, since W stays the same for any number of statements in the family whereas B becomes larger as the number of statements increases
- Both the Working-Hotelling and Bonferroni procedures provide lower bounds to the actual family confidence coefficient
- The levels of the predictor variable for which the mean response is to be estimated are sometimes not known in advance and so the levels of interest are determined via speculation; when this is the case, it is better to use the Working-Hotelling procedure because the family for this procedure encompasses all possible levels of X

1.4.3 Simultaneous Prediction Intervals for New Observations

- There are two procedures for making simultaneous predictions of g new observations on Y in g independent trials at g different levels of X
- The simultaneous prediction limits for g predictions with the Scheffé procedure with family confidence coefficient $1-\alpha$ are

$$\hat{Y}_h \pm Ss[pred]$$

where

$$S^2 = gF_{1-\alpha,g,n-2}$$

and

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

- The simultaneous prediction limits for g predictions with the Bonferroni procedure with family confidence coefficient $1-\alpha$ are

$$\hat{Y}_h \pm Bs[pred]$$

where

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

- The Scheffé procedure uses the F distribution while the Boneferroni procedure uses the t distribution

- The S and B multiples can be evaluated in advance to see which procedure produces prediction limits
- Simultaneous prediction intervals for g new observations on Y at g different levels of X with a $1-\alpha$ family confidence coefficient are wider than the corresponding single prediction intervals; when the number of simultaneous predictions is not large, however, the difference int he width is only moderate
- Note that both the B and S multiples for simultaneous predictions become larger as g increases; this contrasts with simultaneous estimation of mean responses where the B multiple becomes larger but not the W multiple; when g is large, both the B and S multiples for simultaneous predictions may become so large that the prediction intervals will be too wide to be helpful

1.4.4 Regression through Origin

- When the regression function is known to be linear and go through the orifgn at $(0,0)$, the normal error model becomes
$$Y_i = \beta_1 X_i + \varepsilon_i$$
where β_1 is a parameter, X_i are known constants and ε_i are independent $N(0,\sigma^2)$
- Note that in this model, $\beta_0 = 0$
- The regression function for the normal error model here is
$$E[Y] = \beta_1 X$$
which is a straight line through the origin, with slopes β_1
- The least squares estimator of β_1 is obtained by minimizing
$$Q = \sum (Y_i - \beta_1 X_i)^2$$
with respect to β_1, resulting in the normal equation
$$\sum X_i(Y_i - b_1 X_i) = 0$$
leading to the point estimator
$$b_1 = \frac{\sum X_i Y_i}{\sum X_i^2}$$
The estimate b_1 is also the maximum likelihood estimator for the normal error regression model
- The fitted value $\hat{Y}_i$ for the ith case is
$$\hat{Y}_i = b_1 X_i$$
and the ith residual is defined as the difference between the observed and fitted values
$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 X_i$$

- An unbiased estimator of the error variance σ^2 for the regression model is

$$s^2 = MSE = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-1} = \frac{\sum e_i^2}{n-1}$$

The reason for the denominator $n-1$ is that only one degree of freedom is lost in estimating the single parameter in the regression function

- Confidence limits for β_1, $E[Y_h]$ and a new observation $Y_{h(\text{new})}$ for the regression model with $\beta_0 = 0$ are shown below

Estimate of	Estimated Variance	Confidence Limits
β_1	$s^2[b_1] = \frac{MSE}{\sum X_I^2}$	$b_1 \pm ts[b_1]$
$E[Y_h]$	$s^2[\hat{Y}_h] = \frac{X_h^2 MSE}{\sum X_i^2}$	$\hat{Y}_h \pm ts[\hat{Y}_h]$
$Y_{h(\text{new})}$	$s^2[pred] = MSE\left(1 + \frac{X_h^2}{\sum X_i^2}\right)$	$\hat{Y}_h \pm ts[pred]$

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

- In using regression-through-the-origin model, the residuals must be interpreted with case because they do not sum to zero usually

- The only constraint on the residuals, from the normal equation, is of the form $\sum X_i e_i = 0$ and so, in a residual plot, the residuals will usually not be balanced around the zero line

- The sum of the squared residuals $SSE = \sum e_i^2$ for this type of regression may exceed the total sum of squares $SSTO = \sum(Y_i - \overline{Y})^2$; this can occur when the data form a curvilinear pattern or a linear pattern with an intercept away from the origin, and so the coefficient of determination ($R^2 = 1 - \frac{SSE}{SSTO}$) may become negative; consequently, the coefficient of determination R^2 has no clear meaning for regression through the origin

- It is generally a safe practice not to use regression-through-the-origin model and instead use the intercept regression model so that regression diagnostics can be done; even when it is known that the regression function must go through the origin, the function may not be linear or the variance of the error terms may not be constant

- If the regression line does go through the origin under the intercept model, b_0 will differ from 0 only by a small sampling error, and unless the sample size is very small, use of the intercept regression model has no disadvantages of any consequence

- In interval estimation of $E[Y_h]$ or prediction of $Y_{h(\text{new})}$, with regression through the origin, note that the intervals widen the further X_h is from the origin; this happens because the value of the true regression function is known precisely at the origin and so the effect of the sampling error in the slope b_1 becomes increasingly important the farther X_h is from the origin

- Since with regression through the origin, only one parameter β_1 must be estimated, simultaneous estimation methods are not required to make a family of statements about several mean responses; for a given confidence coefficient $1-\alpha$, confidence limits can be used repeatedly with the given sample results for different levels of X to generate a family of statements for which the family confidence coefficient is still $1-\alpha$

- The ANOVA tables for regression through the origin is based on $SSTOU = \sum Y_i^2$ (total uncorrected sum of squares), $SSRU = \sum \hat{Y}_i^2 = b_1^2 \sum X_i^2$ (uncorrected sum of squares) and $SSE = \sum (Y_i - b_1 X_i)^2$; it can be shown that these sums of squares are additive

$$SSTOU = SSRU + SSE$$

1.4.5 Effects of Measurement Errors

- When random measurement errors are present in the observations on the response variable Y, no new problems are created when these errors are uncorrelated and not biased (positive and negative measurement errors tend to cancel out)
- The model error term will always reflect the composite effects of a large number of factors not considered in the model, one of which now would be the random variation due to inaccuracy in the process of measuring Y
- When the observations on the predictor variable X are subject to measurement errors, a different regression model must be used
- Let the measurement error δ_i be defined as

$$\delta_i = X_i^* - X_i$$

  and so the regression model becomes

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = \beta_0 + \beta_1 (X_i^* - \delta_i) + \varepsilon_i$$

  This can be rewritten as

$$Y_i = \beta_0 + \beta_1 X_i^* + (\varepsilon_i - \beta_1 \delta_i)$$

  This may look like an ordinary regression model, with predictor variable X^* and error term $\varepsilon - \beta_1 \delta$, but it is not, since the predictor variable observation X_i^* is a random variable that is correlated with the error term $\varepsilon_i - \beta_1 \delta_i$
- Assume that $E[\delta_i] = 0$, $E[\varepsilon_i] = 0$ and $E[\delta_i \varepsilon_i] = 0$; then $E[X_i^*] = E[X_i + \delta_i] = X_i$, the model error terms ε_i have expectation 0, and the measurement error δ_i is not correlated with the model error ε_i; with this, the covariance between the observations X_i^* and the random terms $\varepsilon_i - \beta_1 \delta_i$ under the 3 conditions given, which imply that $E[X_i^*] = X_i$ and $E[\varepsilon_i - \beta_1 \delta_i] = 0$ is

$$\begin{aligned}
\sigma[X_i^*, \varepsilon_i - \beta_1 \delta_i] &= E(X_i^* - E[X_i^*])((\varepsilon_i - \beta_1 \delta_i) - E[\varepsilon_i - \beta_1 \delta_i]) \\
&= E[(X_i^* - X_i)(\varepsilon_i - \beta_1 \delta_i)] \\
&= E[\delta_i (\varepsilon_i - \beta_1 \delta_i)] \\
&= E[\delta_i \varepsilon_i - \beta_1 \delta_i^2] \\
&= -\beta_1 \sigma^2[\delta_i]
\end{aligned}$$

  The last result occurs because $E[\delta_i \varepsilon_i] = 0$ and $E[\delta_i^2] = \sigma^2[\delta_i]$ since $E[\delta_i] = 0$
- The covariance above is not zero whenever there is a linear regression relation between X and Y

- If it is assumed that the response Y and the random predictor variable X^* follow a bivariate normal distribution, then the conditional distribution of the Y_i given X_i^*, are normal and independent with conditional mean

$$E[Y_i \mid X_i^*] = \beta_0^* + \beta_1^* X_i^*$$

and conditional variance $\sigma^2_{Y|X^*}$; it can also be shown that

$$\beta_1^* = \beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_Y^2}$$

where σ_X^2 is the variance of X and σ_Y^2 is the variance of Y; hence, the least squares slope estimate from fitting Y on X^* is not an estimate of β_1, but an estimate of $\beta_1^* \leq \beta_1$, which is too small on average, with the magnitude of the bias dependent upon the relative sizes of σ_X^2 and σ_Y^2

- Another approach is to use additional variables that are known to be related to the true value of X but not to the errors of measurement δ, called instrumental variables since they are used as an instrument in studying the relation between X and Y

- Under the Berkson model, measurement errors in X are no problem when the observation X_i^* is a fixed quantity but the unobserved true value X_i is a random variable

- Under the Berkson model, let the measurement error be

$$\delta_i = X_i^* - X_i$$

where there is no constraint on the relation between X_i^* and δ_i, since X_i^* is a fixed quantity, and assume $E[\delta_i] = 0$; then the model under the Berkson case is

$$Y_i = \beta_0 + \beta_1 X_i^* + (\varepsilon_i - \beta_1 \delta_i)$$

Here the expected value of the error term, $E[\varepsilon_i - \beta_1 \delta_i] = 0$ since $E[\varepsilon_i] = 0$ and $E[\delta_i] = 0$; however, $\varepsilon_i - \beta_1 \delta_i$ is now uncorrelated with X_i^* since X_i^* is a constant for the Berkson case

- Least squares procedures can be applied for the Berkson case without modification and the estimators b_0 and b_1 will be unbiased; if the standard normality and constant variance assumptions are made for the errors $\varepsilon_i - \beta_1 \delta_i$, then the usual tests and interval estimates can be utilized

1.4.6 Inverse Predictions

- An inverse prediction is made using a regression model of Y on X where a prediction of the value of X is made, after a new observation Y is found

- In inverse predictions, the regression model is assumed as

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

and the estimated regression function based on n observations is obtained as follows

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

- A new observation $Y_{h(\text{new})}$ becomes available and it is desired to estimate the level of $X_{h(\text{new})}$ that gave rise to this new observation, as follows

$$\hat{X}_{h(\text{new})} = \frac{Y_{h(\text{new})} - b_0}{b_1} \quad b_1 \neq 0$$

  where $\hat{X}_{h(\text{new})}$ denotes the point estimator of the new level $X_{h(\text{new})}$; it can be shown that the estimator $\hat{X}_{h(\text{new})}$ is the maximum likelihood estimator of $X_{h(\text{new})}$ for the normal error regression model

- Approximate $1 - \alpha$ confidence limits for $X_{h(\text{new})}$ are

$$\hat{X}_{h(\text{new})} \pm t_{1-\frac{\alpha}{2},n-2} s[\text{predX}]$$

  where

$$s^2[\text{predX}] = \frac{MSE}{b_1^2}\left[1 + \frac{1}{n} + \frac{(\hat{X}_{h(\text{new})} - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right]$$

- The inverse prediction problem is also known as a calibration problem since it is applicable when inexpensive, quick and approximate measurements Y are related to precise, often expensive, and time-consuming measurements X based on n observations; the resulting regression model is then used to estimate the precise measurement $X_{h(\text{new})}$ for a new approximate measurement $Y_{h(\text{new})}$

- The approximate confidence interval is appropriate if the quantity

$$\frac{(t_{1-\frac{\alpha}{2},n-2})^2 MSE}{b_1^2 \sum(X_i - \overline{X})^2}$$

  is small

- Simultaneous prediction intervals based on g different new observed measurements $Y_{h(\text{new})}$, with a $1 - \alpha$ family confidence coefficient, are easily obtained by using either the Bonferroni or the Scheffé procedures; the value of $t_{1-\frac{\alpha}{2},n-2}$ is replaced by either $B = t_{1-\frac{\alpha}{2g},n-2}$ or $S = \sqrt{gF_{1-\alpha,g,n-2}}$

- The inverse prediction problem is controversial among statisticians, where some suggest that inverse predictions be made in direct fashion by regressing X on Y (inverse regression)

1.4.7 Choice of X Levels

- When regression data are obtained by experiment, the levels of X at which observations on Y are to be taken are under the control of the experimenter; among other things, the experimenter will have to consider:
  - How many levels of X should be investigated?
  - What shall the two extreme levels be?
  - How shall the other levels of X, if any, be spaced?
  - How many observations should be taken at each level of X?

- In many cases, the main objective in regression analysis is to predict one or more new observations or to estimate one of more mean responses; when the regression function is curvilinear, the main objective may be to locate the maximum or minimum mean responses; at still other times, the main purpose is to determine the nature of the regression function
- If the main purpose of the regression analysis is to estimate the slope β_1, the variance of b_1 is minimized if $\sum(X_i - \overline{X})^2$ is maximized, which is accomplished by using two levels of X, at the two extremes for the scope of the model, and placing half of the observations at each of the two levels
- If the main purpose of the regression analysis is to estimate the intercept β_0, the number and placement of levels does not affect the variance of b_0 as long as $\overline{X} = 0$
- From "Planning of Experiments", by D. R. Cox (1958),
  - Use two levels when the object is primarily to examine whether or not the predictor variable has an effect and in which direction that effect is
  - Use three levels whenever a description of the response curve by its slope and curvature is likely to be adequate
  - Use four levels if further examination of the shape of the response curve is important
  - Use more than four levels when it is required to estimate the detailed shape of the response curve, or when the curve is expected to rise to an asymptotic value, or in general to show features not adequately described by slope and curvature
  - In these last cases, it is generally satisfactory to use equally spaced levels with equal numbers of observations per level

1.5 Matrix Approach to Simple Linear Regression Analysis

1.5.1 Matrices

- A matrix is a rectangular array of elements in rows and columns
- The dimension of the matrix is given by $r \times c$ where r is the number of rows and c is the number of columns in the matrix
- To identify the elements of a matrix $\mathbf{A}$, let a_{ij} represent the element in the ith row and the jth column
- A matrix with r rows and c columns can be represented in full as

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1c} \\ a_{21} & a_{22} & \dots & a_{2j} & \dots & a_{2c} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ij} & \dots & a_{ic} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \dots & a_{rj} & \dots & a_{rc} \end{bmatrix}$$

  or in abbreviated form as

$$\mathbf{A} = [a_{ij}] \quad i = 1\dots, r; j = 1, \dots, c$$

  or simply by a boldface symbol, such as $\mathbf{A}$

- A matrix is said to be square if the number of rows equals the number of columns
- A matrix containing only one column in called a column vector, or simply a vector; likewise, a matrix containing only one row is called a row vector
- The transpose of a matrix $\mathbf{A}$ is another matrix, denoted by $mtvTA$, that is obtained by interchanging corresponding columns and rows of the matrix $\mathbf{A}$

$$\underbrace{\mathbf{A}}_{r\times c} = \begin{bmatrix} a_{11} & \dots & a_{1c} \\ \vdots & \ddots & \vdots \\ a_{r1} & \dots & a_{rc} \end{bmatrix} = [a_{ij}] \quad i = 1, \dots, r; j = 1, \dots, c$$

$$\underbrace{\text{tr}\{A\}}_{c\times r} = \begin{bmatrix} a_{11} & \dots & a_{r1} \\ \vdots & \ddots & \vdots \\ a_{1c} & \dots & a_{rc} \end{bmatrix} = [a_{ji}] \quad j = 1, \dots, c; i = 1, \dots r$$

  The element in the ith row and the jth column in $\mathbf{A}$ is found int he jth row and ith column in tr$\{A\}$
- Two matrices $\mathbf{A}$ and $\mathbf{B}$ are said to be equal if they have the same dimension and if all corresponding elements are equal; conversely, if two matrices are equal, their corresponding elements are equal

1.5.2 Matrix Addition and Subtraction

- Adding and subtracting two matrices require that they have the same dimension
- The sum, or difference, of two matrices is another matrix whose elements each consist of the sum, of difference, of the corresponding elements of the two matrices
- In general, if

$$\underbrace{\mathbf{A}}_{r\times c}[a_{ij}] \quad \underbrace{\mathbf{B}}_{r\times c} = [b_{ij}] \quad i = 1, \dots, r; j = 1, \dots, c$$

  then

$$\underbrace{\mathbf{A} + \mathbf{B}}_{r\times c} = [a_{ij} + b_{ij}] \qquad \underbrace{\mathbf{A} - \mathbf{B}}_{r\times c} = [a_{ij} - b_{ij}]$$

1.5.3 Matrix Multiplication

- A scalar is an ordinary number of a symbol representing a number
- In multiplication of a matrix by a scalar, every element of the matrix is multiplied by the scalar
- If every element of a matrix has a common factor, this factor can be taken outside the matrix and treated as a scalar
- In general, if $\mathbf{A} = [a_{ij}]$ and k is a scalar, then

$$k\mathbf{A} = \mathbf{A}k = [ka_{ij}]$$

- The matrix multiplication of two matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{AB}$ and is defined only when the number of columns in $\mathbf{A}$ equals the number of rows in $\mathbf{B}$ so that there will be corresponding terms in the cross products
- When obtaining the product $\mathbf{AB}$, $\mathbf{A}$ is said to be postmultiplied by $\mathbf{B}$, or $\mathbf{B}$ is premultiplied by $\mathbf{A}$
- In general, if $\mathbf{A}$ has dimension $r \times c$ and $\mathbf{B}$ has dimension $c \times s$, the product $\mathbf{AB}$ is a matrix of dimension $r \times s$ whose element in the ith row and jth column is

$$\sum_{k=1}^{c} a_{ik} b_{kj}$$

so that

$$\underbrace{\mathbf{AB}}_{r \times s} = \left[\sum_{k=1}^{c} a_{ik} b_{kj} \right] \quad i = 1, \ldots, r; j = 1, \ldots, s$$

1.5.4 Special Types of Matrices

- If $\mathbf{A} = \text{tr}\{A\}$, then $\mathbf{A}$ is said to be symmetric; a symmetric matrix necessarily is square
- A diagonal matrix is a square matrix whose off-diagonal elements are all zeros

$$\begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix}$$

- The identity matrix or unit matrix is denoted by $\mathbf{I}$ and is a diagonal matrix whose elements on the main diagonal are all 1s

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- Premultiplying or postmultiplying any $r \times r$ matrix $\mathbf{A}$ by the $r \times r$ identify matrix $\mathbf{I}$ leaves $\mathbf{A}$ unchanged

$$\mathbf{IA} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{AI}$$

- In general, for any $r \times r$ matrix $\mathbf{A}$,

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}$$

The identity matrix can be inserted or dropped from a matrix expression whenever it is convenient to do so

- A scalar matrix is a diagonal matrix whose main-diagonal elements are the same, and can be represented as $k\mathbf{I}$, where k is the scalar

$$\begin{bmatrix} k & 0 & 0 \\ 0 & k & 0 \\ 0 & 0 & k \end{bmatrix} = k \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = k\mathbf{I}$$

- Multiplying an $r \times r$ matrix $\mathbf{A}$ by the $rtimesr$ scalar matrix $k\mathbf{I}$ is equivalent to multiplying $\mathbf{A}$ by the scalar k
- A column vector with all elements 1 will be denoted by $\mathbf{1}$

$$\underbrace{\mathbf{1}}_{r\times 1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

- A square matrix with all elements 1 will be denoted by $\mathbf{J}$

$$\underbrace{\mathbf{J}}_{r\times r} = \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}$$

- Note that for an $n \times 1$ vector $\mathbf{1}$,

$$\underbrace{\text{tr}\{1\}\mathbf{1}}_{1\times 1} = \begin{bmatrix} 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = [n] = n$$

and

$$\underbrace{\mathbf{1}\,\text{tr}\{1\}}_{n\times n} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} 1 & \dots & 1 \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix} = \underbrace{\mathbf{J}}_{n\times n}$$

- A zero vector is a vector containing only zeros
- The zero column vector will be denoted as $\mathbf{0}$

$$\underbrace{\mathbf{0}}_{r\times 1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

1.5.5 Linear Dependence and Rank of Matrix

- When c scalars $k_1, \dots, k_c$, not all zero, can be found such that

$$k_1\mathbf{C}_1 + k_2\mathbf{C}_2 + \cdots + k_c\mathbf{C}_c = \mathbf{0}$$

where $\mathbf{0}$ denotes the zero column vector and $\mathbf{C}_i$ are column vectors of a matrix, the $c$column vectors are linearly dependent; if the only set of scalars for which the equality holds is $k_1 = 0, \dots, k_c = 0$, the set of c column vectors is linearly independent

- The rank of a matrix is defined to be the maximum number of linearly independent columns in the matrix
- The rank of a matrix is unique and can equivalently be defined as the maximum number of linearly independent rows
- The rank of an $r \times c$ matrix cannot exceed $\min(r, c)$, the minimum of the two values r and c
- When a matrix is the product of two matrices, its rank cannot exceed the smaller of the two ranks for the matrices being multiplied

1.5.6 Inverse of a Matrix

- The inverse of a matrix $\mathbf{A}$, denoted by $\mathbf{A}^{-1}$, is another matrix such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

  where $\mathbf{I}$ is the identity matrix

- Note that the inverse of a matrix is defined only for square matrices; even so, many square matrices do not have inverses; if a square matrix does have an inverse, the inverse is unique

- The inverse of a diagonal matrix is a diagonal matrix consisting simply of the reciprocals of the elements on the diagonal

- An inverse of a square $r \times r$ matrix exists if the rank of the matrix is r; such a matrix is said to be nonsingular or of full rank

- An $r \times r$ matrix with rank less than r is said to be singular and not of full rank, and does not have an inverse

- The inverse of an $r \times r$ matrix of full rank also has rank r

- If

$$\underbrace{\mathbf{A}}_{2\times 2} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

  then

$$\underbrace{\mathbf{A}^{-1}}_{2\times 2} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \begin{bmatrix} \frac{d}{D} & -\frac{b}{D} \\ -\frac{c}{D} & \frac{a}{D} \end{bmatrix}$$

  where

$$D = ad - bc$$

  is called the determinant of the matrix $\mathbf{A}$; if $\mathbf{A}$ were singular, its determinant would equal zero and no inverse of $\mathbf{A}$ would exist

- If

$$\underbrace{\mathbf{B}}_{3\times 3} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix}$$

  then

$$\underbrace{\mathbf{B}^{-1}}_{3\times 3} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix}^{-1} = \begin{bmatrix} A & B & C \\ D & E & F \\ G & H & K \end{bmatrix}$$

  where

$$\begin{aligned} A &= (ek - fh)/Z \quad B = -(bk - ch)/Z \quad C = (bf - ce)/Z \\ D &= -(dk - fg)/Z \quad E = (ak - cg)/Z \quad F = -(af - cd)/Z \\ G &= (dh - eg)/Z \quad H = -(ah - bg)/Z \quad K = (ae - bd)/Z \end{aligned}$$

  and

$$Z = a(ek - fh) - b(dk - fg) + c(dh - eg)$$

  is called the determinant of the matrix $\mathbf{B}$

- Computing the inverse of matrices with higher number of rows/columns require a large amount of computation
- In matrix algebra, if there is an equation $\mathbf{AY} = \mathbf{C}$, both sides can be premultiplied by $\mathbf{A}^{-1}$, assuming $\mathbf{A}$ has an inverse, to obtain a solution for $\mathbf{Y}$

$$\mathbf{Y} = \mathbf{A}^{-1}\mathbf{C}$$

1.5.7 Some Basic Results for Matrices

- $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
- $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
- $\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB}$
- $k(\mathbf{A} + \mathbf{B}) = k\mathbf{A} + k\mathbf{B}$
- $(\mathbf{IA})' = \mathbf{A}$
- $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$
- $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$
- $(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- $(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$
- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$

1.5.8 Random Vectors and Matrices

- A random vector or a random matrix contains elements that are random variables
- The expected value of a random vector is a vector whose elements are the expected values of the random variables that are the elements of the random vector; for a random vector $\mathbf{Y}$, the expectation is

$$\underbrace{\mathbf{E}[\mathbf{Y}]}_{n\times 1} = [E[Y_i]] \quad i = 1, \ldots, n$$

- The expectation of a random matrix is a matrix whose elements are the expected values of the corresponding variables in the original matrix; for a random matrix $\mathbf{Y}$ with dimension $n \times p$, the expectation is

$$\underbrace{E[\mathbf{Y}]}_{n\times p} = [E[Y_{ij}]] \quad i = 1, \ldots, n; j = 1, \ldots, p$$

- The variances and covariances of a random vector $\mathbf{Y}$ are assembled in a variance-covariance matrix of $\mathbf{Y}$, denoted by $\sigma^2[\mathbf{Y}]$; the variance-covariance matrix for an $n \times 1$ random vector $\mathbf{Y}$ is

$$\underbrace{\sigma^2[\mathbf{Y}]}_{n \times n} = \begin{bmatrix} \sigma^2[Y_1] & \sigma[Y_1, Y_2] & \dots & \sigma[Y_1, Y_n] \\ \sigma[Y_2, Y_1] & \sigma^2[Y_2] & \dots & \sigma[Y_2, Y_n] \\ \vdots & \vdots & \ddots & \vdots \\ \sigma[Y_n, Y_1] & \sigma[Y_n, Y_2] & \dots & \sigma^2[Y_n] \end{bmatrix}$$

  Note that $\sigma^2[\mathbf{Y}]$ is a symmetric matrix since $\sigma[Y_i, Y_j] = \sigma[Y_j, Y_i]$ for all $i \neq j$

- The derivation of the variance-covariance matrix arises from

$$\sigma^2[\mathbf{Y}] = E[[\mathbf{Y} - E[\mathbf{Y}]][\mathbf{Y} - E[\mathbf{Y}]]']$$

- Suppose $\mathbf{W}$ is a random vector, $\mathbf{A}$ is a constant matrix (a matrix whose elements are fixed) and $\mathbf{Y}$ is a random vector, then if $\mathbf{W} = \mathbf{AY}$, then

$$\begin{aligned} E[\mathbf{A}] &= \mathbf{A} \\ E[\mathbf{W}] &= E[\mathbf{AY}] = \mathbf{A}E[\mathbf{Y}] \\ \sigma^2[\mathbf{W}] &= \sigma^2[\mathbf{AY}] = \mathbf{A}\sigma^2[\mathbf{Y}]\mathbf{A}' \end{aligned}$$

  where $\sigma^2[\mathbf{Y}]$ is the variance-covariance matrix of $\mathbf{Y}$

- The density function for the multivariate normal distribution is best given in matrix form; Let the observations vector $\mathbf{Y}$ contain observation on each of the p Y variables as defined as usual:

$$\underbrace{\mathbf{Y}}_{p \times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix}$$

  The mean vector $E[\mathbf{Y}]$, denoted by $\boldsymbol{\mu}$, contains the expected values for each of the p Y variables

$$\underbrace{\boldsymbol{\mu}}_{p \times 1} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

  The variance-covariance matrix $\sigma^2[\mathbf{Y}]$, denoted by $\boldsymbol{\Sigma}$, contains the variances and covariances of the p Y variables

$$\underbrace{\boldsymbol{\Sigma}}_{p \times p} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{bmatrix}$$

  where σ_1^2 denotes the variance of Y_1, σ_{12} denotes the covariance of Y_1 and Y_2, and so forth

- The density function of the multivariate normal distribution can be stated as follows:

$$f(\mathbf{Y}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{Y} - {}^{-})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - {}^{-})\right]$$

  where $|\sum|$ is the determinant of the variance-covariance matrix $\boldsymbol{\Sigma}$

- Note that when there are $p = 2$ variables, the multivariate normal density function simplifies to the bivariate normal density function
- The multivariate normal density function has properties that correspond to the ones described for the bivariate normal distribution, such as, if $Y_1, \ldots, Y_p$ are jointly normally distributed (i.e., they follow the multivariate normal distribution), the marginal probability distribution of each variable Y_k is normal, with mean μ_k and standard deviation σ_k

1.5.9 Simple Linear Regression Model in Matrix Terms

- The normal error regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \ldots, n$$

which simplifies to

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_n + \varepsilon_n \end{aligned}$$

- To develop simple linear regression in matrix terms, define the following:

$$\underbrace{\mathbf{Y}}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \underbrace{\mathbf{X}}_{n\times 2} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \quad \underbrace{\boldsymbol{\beta}}_{2\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \underbrace{\boldsymbol{\varepsilon}}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Then the normal error regression model can be rewritten as

$$\underbrace{\mathbf{Y}}_{n\times 1} = \underbrace{\mathbf{X}}_{n\times 2}\underbrace{\boldsymbol{\beta}}_{2\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{n\times 1}$$

This is derived from

$$\begin{aligned} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} &= \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\ &= \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\ &= \begin{bmatrix} \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 X_n + \varepsilon_n \end{bmatrix} \end{aligned}$$

- $\mathbf{XB}$ is the vector of the expected values of the Y_i observations since $E[Y_i] = \beta_0 + \beta_1 X_i$

$$\underbrace{E[\mathbf{Y}]}_{n\times 1} = \underbrace{\mathbf{XB}}_{n\times 1}$$

- The $\mathbf{X}$ matrix may be considered to contained a column vector consisting of 1s and another column vector consisting of the predictor variable observations X_i
- The condition $E[\varepsilon_i] = 0$ in matrix terms is

$$\underbrace{E[\boldsymbol{\varepsilon}]}_{n\times 1} = \underbrace{\mathbf{0}}_{n\times 1}$$

since

$$\begin{bmatrix} E[\varepsilon_1] \\ E[\varepsilon_2] \\ \vdots \\ E[\varepsilon_n] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

- The condition that the error terms have constant variance σ^2 and that all covariances $\sigma[\varepsilon_i, \varepsilon_j]$ for $i \neq j$ are zero (since the ε_i are independent) is expressed in matrix terms through the variance-covariance matrix of the error terms

$$\underbrace{\boldsymbol{\sigma}^{\mathbf{2}}[\boldsymbol{\varepsilon}]}_{n\times n} = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

which can be expressed as

$$\underbrace{\boldsymbol{\sigma}^{\mathbf{2}}[\boldsymbol{\varepsilon}]}_{n\times n} = \underbrace{\sigma^2\mathbf{I}}_{n\times n}$$

- The normal error regression model in matrix terms is

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}$ is a vector of independent normal random variables with $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $\boldsymbol{\sigma}^{\mathbf{2}}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}$

1.5.10 Least Squares Estimation of Regression Parameters

- The normal equations

$$nb_0 + b_1 \sum X_I = \sum Y_i$$
$$b_0 \sum X_I + b_1 \sum X_i^2 = \sum X_i Y_i$$

in matrix terms are

$$\underbrace{\mathbf{X}'\mathbf{X}}_{2\times 2}\underbrace{\mathbf{b}}_{2\times 1} = \underbrace{\mathbf{X}'\mathbf{Y}}_{2\times 1}$$

where $\mathbf{b}$ is the vector of the least squares regression coefficients

- This is derived as follows:

$$\underbrace{\mathbf{X'X}}_{2\times2} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

$$\underbrace{\mathbf{X'Y}}_{2\times1} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

$$\begin{aligned} \mathbf{X'Xb} &= \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \\ &= \begin{bmatrix} nb_0 + b_1 \sum X_i \\ b_0 \sum X_i + b_1 \sum X_i^2 \end{bmatrix} \\ &= \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix} \\ &= \mathbf{X'Y} \end{aligned}$$

- To obtain the estimated regression coefficients from the normal equations by matrix methods, premultiply both sides by the inverse of $\mathbf{X'X}$ to get the following

$$\underbrace{\mathbf{b}}_{2\times1} = \underbrace{(\mathbf{X'X})^{-1}}_{2\times2} \underbrace{\mathbf{X'Y}}_{2\times1}$$

  The estimators b_0 and b_1 in $\mathbf{b}$ are the same as given before

- A comparison of the normal equations and $\mathbf{X'X}$ shows that whenever the columns of $\mathbf{X'X}$ are linearly dependent, the normal equations will be linearly dependent also; no unique solutions can then be obtained for b_0 and b_1; fortunately, in most regression applications, the columns of $\mathbf{X'X}$ are linearly independent, leading to unique solutions for b_0 and b_1

1.5.11 Fitted Values and Residuals

- Let the vector of the fitted values $\hat{Y}_i$ be denoted by $\hat{\mathbf{Y}}_i$

$$\underbrace{\hat{\mathbf{Y}}}_{n\times1} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$$

  In matrix notation, this is

$$\underbrace{\hat{\mathbf{Y}}}_{n\times1} = \underbrace{\mathbf{X}}_{n\times2} \underbrace{\mathbf{b}}_{2\times1}$$

  because

$$\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} b_0 + b_1 X_1 \\ b_0 + b_1 X_2 \\ \vdots \\ b_0 + b_1 X_n \end{bmatrix}$$

- The matrix result for $\hat{\mathbf{Y}}$ can be expressed as

$$\underbrace{\hat{\mathbf{Y}}}_{n\times 1} = \underbrace{\mathbf{H}}_{n\times n}\underbrace{\mathbf{Y}}_{n\times 1}$$

where

$$\underbrace{\mathbf{H}}_{n\times n} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

so that

$$\hat{\mathbf{Y}} = \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{H}}\mathbf{Y}$$

The square $n \times n$ matrix $\mathbf{H}$ is called the hat matrix

- The matrix $\mathbf{H}$ is symmetric and has following property (called idempotency):

$$\mathbf{HH} = \mathbf{H}$$

- A matrix $\mathbf{M}$ is said to be idempotent if $\mathbf{MM} = \mathbf{M}$
- Let the vector of the residuals $e_i = Y_i - \hat{Y}_i$ be denoted by $\mathbf{e}$:

$$\underbrace{\mathbf{e}}_{n\times 1} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

In matrix notation, this becomes

$$\underbrace{\mathbf{e}}_{n\times 1} = \underbrace{\mathbf{Y}}_{n\times 1} - \underbrace{\hat{\mathbf{Y}}}_{n\times 1} = \underbrace{\mathbf{Y}}_{n\times 1} - \underbrace{\mathbf{Xb}}_{n\times 1}$$

- The residuals e_i can be expressed as linear combinations of the response variable observations Y_i in matrix form as follows:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

which can be simplified to

$$\underbrace{\mathbf{e}}_{n\times 1} = (\underbrace{\mathbf{I}}_{n\times n} - \underbrace{\mathbf{H}}_{n\times n})\underbrace{\mathbf{Y}}_{n\times 1}$$

where $\mathbf{H}$ is the hat matrix and the matrix $\mathbf{I} - \mathbf{H}$ is also symmetric and idempotent

- The variance-covariance matrix of the vector of residuals $\mathbf{e}$ involves the matrix $\mathbf{I} - \mathbf{H}$ as follows:

$$\underbrace{\boldsymbol{\sigma}^{\mathbf{2}}[\boldsymbol{e}]}_{n\times n} = \sigma^2(\mathbf{I} - \mathbf{H})$$

and is estimated by

$$\underbrace{\mathbf{s}^{\mathbf{2}}[\mathbf{e}]}_{n\times n} = MSE(\mathbf{I} - \mathbf{H})$$

1.5.12 Analysis of Variance Results

- The total sum of squares SSTO, in matrix notation, is derived to be

$$SSTO = \mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}$$

  where $\mathbf{J}$ is a matrix of 1s

- The sum of square errors SSE, in matrix notation, is derived to be

$$SSE = \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b})$$

  which is equivalent to

$$SSE = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}'$$

- The sum of square due to regression SSR, in matrix notation, is derived to be

$$SSR = \mathbf{b}'\mathbf{X}'\mathbf{Y}' - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}$$

- In general, a quadratic form is defined as

$$\underbrace{\mathbf{Y}'\mathbf{A}\mathbf{Y}}_{1\times 1} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}Y_iY_j \quad \text{where } a_{ij} = a_{ji}$$

  $\mathbf{A}$ is a symmetric $n \times n$ matrix and is called the matrix of the quadratic form

- The ANOVA sum of squares SSTO, SSE and SSR are all quadratic forms, as can be seen by reexpressing $\mathbf{b}'\mathbf{X}'$

$$\mathbf{b}'\mathbf{X}' = (\mathbf{X}\mathbf{b})' = \hat{\mathbf{Y}}' = (\mathbf{H}\mathbf{Y})' = \mathbf{Y}'\mathbf{H}$$

  This result then leads to the following

$$\begin{aligned} SSTO &= \mathbf{Y}'\left[\mathbf{I} - \left(\frac{1}{n}\right)\mathbf{J}\right]\mathbf{Y} \\ SSE &= \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} \\ SSR &= \mathbf{Y}'\left[\mathbf{H} - \left(\frac{1}{n}\right)\mathbf{J}\right]\mathbf{Y} \end{aligned}$$

  Each of these sum of squares can be seen to be of the form $\mathbf{Y}'\mathbf{A}\mathbf{Y}$, where the three $\mathbf{A}$ matrices are:

$$\begin{aligned} \mathbf{I} - \left(\frac{1}{n}\right)\mathbf{J} \\ \mathbf{I} - \mathbf{H} \\ \mathbf{H} - \left(\frac{1}{n}\right)\mathbf{J} \end{aligned}$$

  Since each of these $\mathbf{A}$ matrices is symmetric, SSTO, SSE and SSR are quadratic forms, with the matrices of the quadratic forms give above

- Quadratic forms play an important role in statistics because all sums of squares in the analysis of variance for linear statistical models can be expressed as quadratic forms

1.5.13 Inferences in Regression Analysis

- The variance-covariance matrix of $\mathbf{b}$, in matrix notation, is

$$\underbrace{\boldsymbol{\sigma^2[b]}}_{2\times 2} = \sigma^2(\mathbf{X'X})^{-1}$$

  which can be shown to be

$$\underbrace{\boldsymbol{\sigma^2[b]}}_{2\times 2} = \begin{bmatrix} \frac{\sigma^2}{n} + \frac{\sigma^2\overline{X}^2}{\sum(X_i-\overline{X})^2} & -\frac{\overline{X}\sigma^2}{\sum(X_i-\overline{X})^2} \\ -\frac{\overline{X}\sigma^2}{\sum(X_i-\overline{X})^2} & \frac{\sigma^2}{\sum(X_i-\overline{X})^2} \end{bmatrix}$$

- When MSE is substituted for σ^2 in the variance-covariance matrix, the estimated variance-covariance matrix of $\mathbf{b}$, or $\boldsymbol{s^2[b]}$, is obtained

$$\underbrace{\boldsymbol{s^2[b]}}_{2\times 2} = MSE(\mathbf{X'X})^{-1} = \begin{bmatrix} \frac{MSE}{n} + \frac{MSE\overline{X}^2}{\sum(X_i-\overline{X})^2} & -\frac{\overline{X}MSE}{\sum(X_i-\overline{X})^2} \\ -\frac{\overline{X}MSE}{\sum(X_i-\overline{X})^2} & \frac{MSE}{\sum(X_i-\overline{X})^2} \end{bmatrix}$$

- To estimate the mean response at X_h, let

$$\underbrace{\mathbf{X}_h}_{2\times 1} = \begin{bmatrix} 1 \\ X_h \end{bmatrix} \quad \text{or} \quad \underbrace{\mathbf{X}_h'}_{1\times 2} = \begin{bmatrix} 1 & X_h \end{bmatrix}$$

  Then the fitted value in matrix notation is

$$\hat{\mathbf{Y}}_h = \mathbf{X}_h'\mathbf{b}$$

  since

$$\mathbf{X}_h'\mathbf{b} = \begin{bmatrix} 1 & X_h \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} b_0 + b_1X_h \end{bmatrix} = \begin{bmatrix} \hat{Y}_h \end{bmatrix} = \hat{Y}_h$$

  Note that $\mathbf{X}_h'\mathbf{b}$ is a 1×1 matrix and thus the final result can be written as a scalar

- The variance of $\hat{Y}_h$, in matrix notation, is

$$\sigma^2[\hat{Y}_h] = \sigma^2\mathbf{X}_h'(\mathbf{X'X})^{-1}\mathbf{X}_h$$

  and can be reexpressed using the variance-covariance matrix of the estimated regression coefficients as follows

$$\sigma^2[\hat{Y}_h] = \mathbf{X}_h'\boldsymbol{\sigma^2[b]}\mathbf{X}_h$$

- The estimated variance of $\hat{Y}_h$, in matrix notation, is

$$s^2[\hat{Y}_h] = MSE(\mathbf{X}_h'(\mathbf{X'X})^{-1}\mathbf{X}_h)$$

- The estimated variance $s^2[\text{pred}]$, in matrix notation, is

$$s^2[\text{pred}] = MSE(1 + \mathbf{X}_h'(\mathbf{X'X})^{-1}\mathbf{X}_h)$$