2. TWO-VARIABLE REGRESSION ANALYSIS

In order to understand two-variable regression, consider the data given in Table 2.1.
The data in the below table refer to a total **Population** of 42 families with their weekly income (X) and weekly consumption expenditure (Y).

Table 2.1: Weekly family Expenditure (Y), Baht and Income (X), Baht

	X=Weekly family Income, Baht					
	500	600	700	800	900	1000
	360	376	458	610	600	700
	313	475	422	468	531	679
	322	380	498	575	670	730
Y= Weeky	310	382	560	542	630	591
Family Expenditure	390	390	442	588	544	550
	315	425	440	466	565	620
	390	442	-	461	-	695
	400	-	-	-	-	635
Total	2800	2870	2820	3710	3540	5200
Conditional means of Y, $E(Y\|X)$	350	410	470	530	590	650

Notes -

Table 2.2: Conditional Probabilities $p(Y|X_i)$ for the Weekly Family Income (X) and Expenditure (Y)

	X=Weekly family Income, Baht					
	500	600	700	800	900	1000
	1/8	1/7	1/6	1/7	1/6	1/8
	1/8	1/7	1/6	1/7	1/6	1/8
	1/8	1/7	1/6	1/7	1/6	1/8
Y= Weekly	1/8	1/7	1/6	1/7	1/6	1/8
Family Expenditure	1/8	1/7	1/6	1/7	1/6	1/8
	1/8	1/7	1/6	1/7	1/6	1/8
	1/8	1/7	-	1/7	-	1/8
	1/8	-	-	-	-	1/8
Conditional means of Y, $E(Y\|X)$	350	410	470	530	590	650

Notes -

Conditional expected value of weekly consumption expenditure given the income level =X , $E(Y|X)$

Unconditional expected value ,$E(Y)$

Figure 2.1: Conditional Distribution of Expenditure for Various Levels of Income

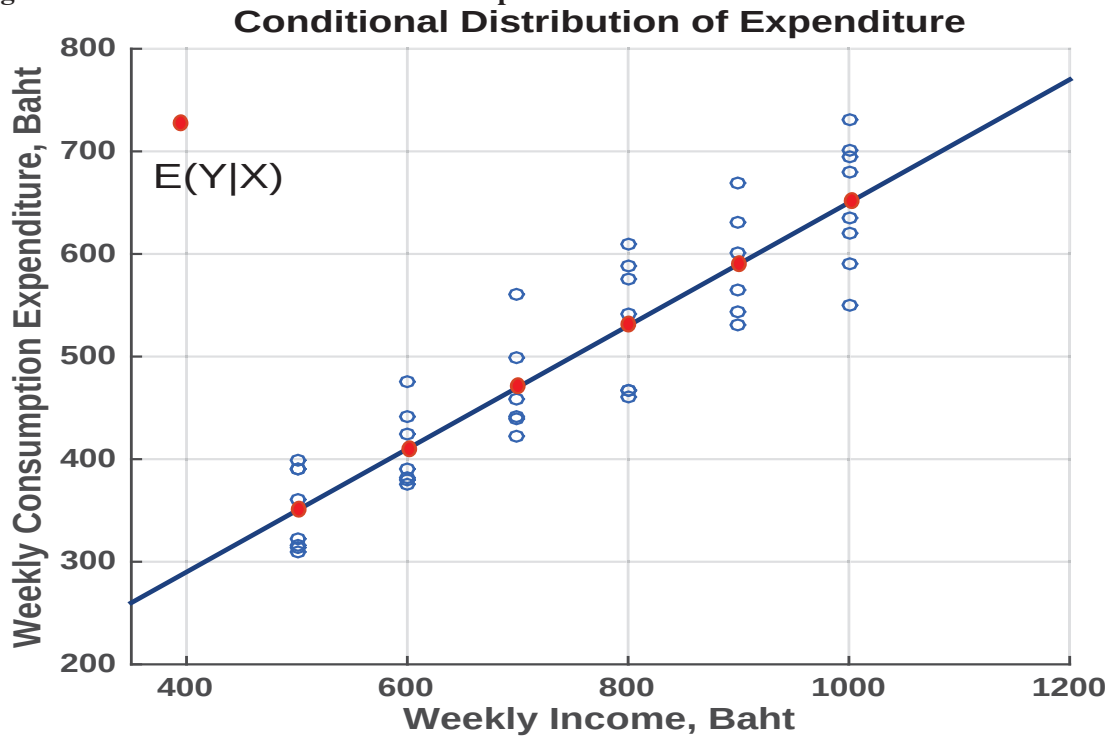
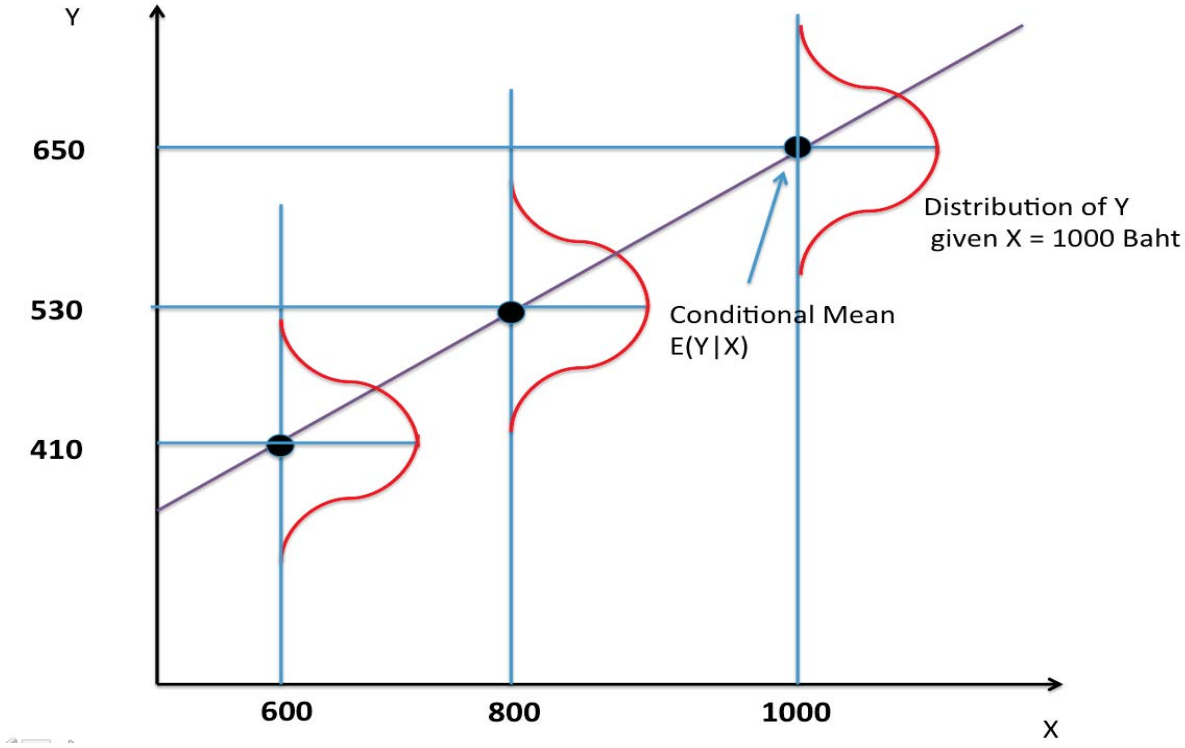


Figure 2.2: Population Regression Line (PRL)



2.1 The Concept of Population Regression Function (PRF)

The population regression function (PRF) can be written as the function of X_i:

2.1.1 What form does the function f(X) assume?

If we assume the PRF $E(Y|X_i)$ is a linear function of X_i, we get

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

2.1.2 What is the meaning of the term LINEAR?

LINEARITY in the variables

LINEARITY in the parameters

2.2 Stochastic Specification of PRF

We can write the **deviation** of an individual Y_i around it expected value as follows:

2.2.1 The roles of the stochastic disturbance term

1. Vagueness of theory

2. Unavailability of data

3. Core variables versus peripheral variables

4. Intrinsic randomness in human behavior

5. Poor proxy variable

6. Principle of parsimony

7.Wrong functional form

2.3 The Sample Regression Function (SRF)

As mentioned, in the real situation, we cannot find out all the population of Y values corresponding to the fixed X's. We only have a sample of Y values corresponding to some fixed X's.

Therefore, our goal in this section is to estimate the population regression line (PRF) on the basis of the **SAMPLE INFORMATION.**

As a result, for the fixed X's as given in table 2.1, we only have a randomly selected sample of Y values. For example, table 2.3 and table 2.4 show a random sample from the population of table 2.1

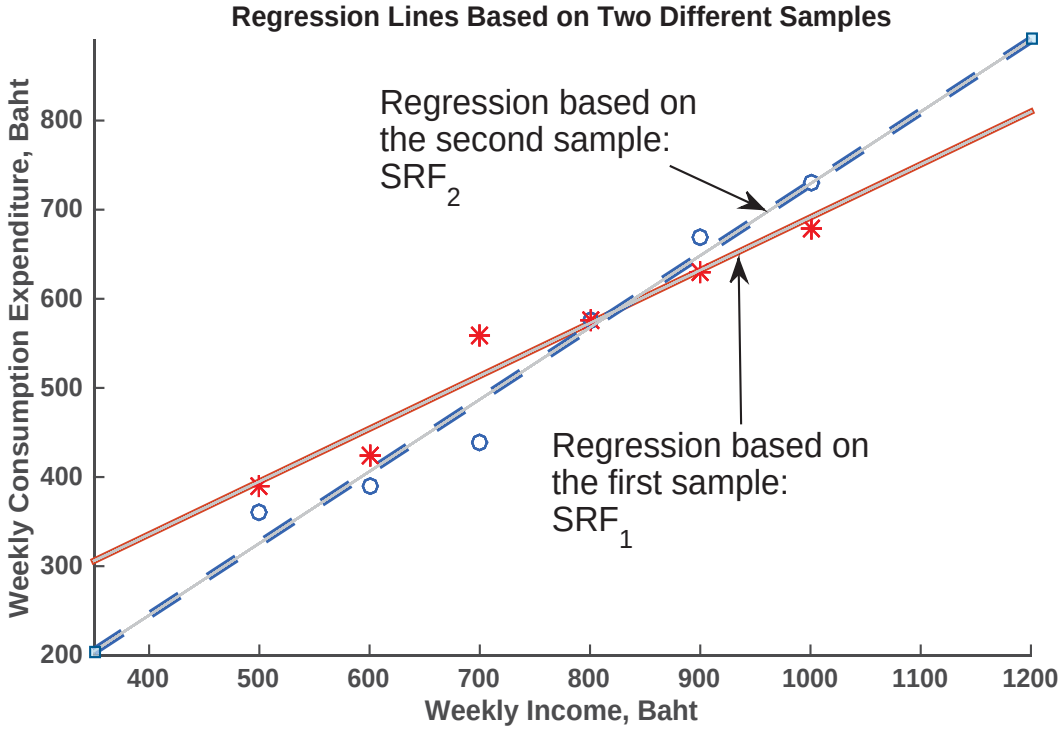
Table 2.3: A Random Sample From the Population

X	Y
500	390
600	425
700	560
800	575
900	630
1000	679

Table 2.4: Another Random Sample From the Population

X	Y
500	360
600	390
700	440
800	575
900	670
1000	730

Figure 2.3: Regression lines based on two different samples



The sample regression function (SRF) can be written as:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

where $\hat{Y}$ is read as "Y-hat"

$\hat{Y}_i$ = estimator of $E(Y|X_i)$

$\hat{\beta}_1$=estimator of β_1

$\hat{\beta}_2$=estimator of β_2

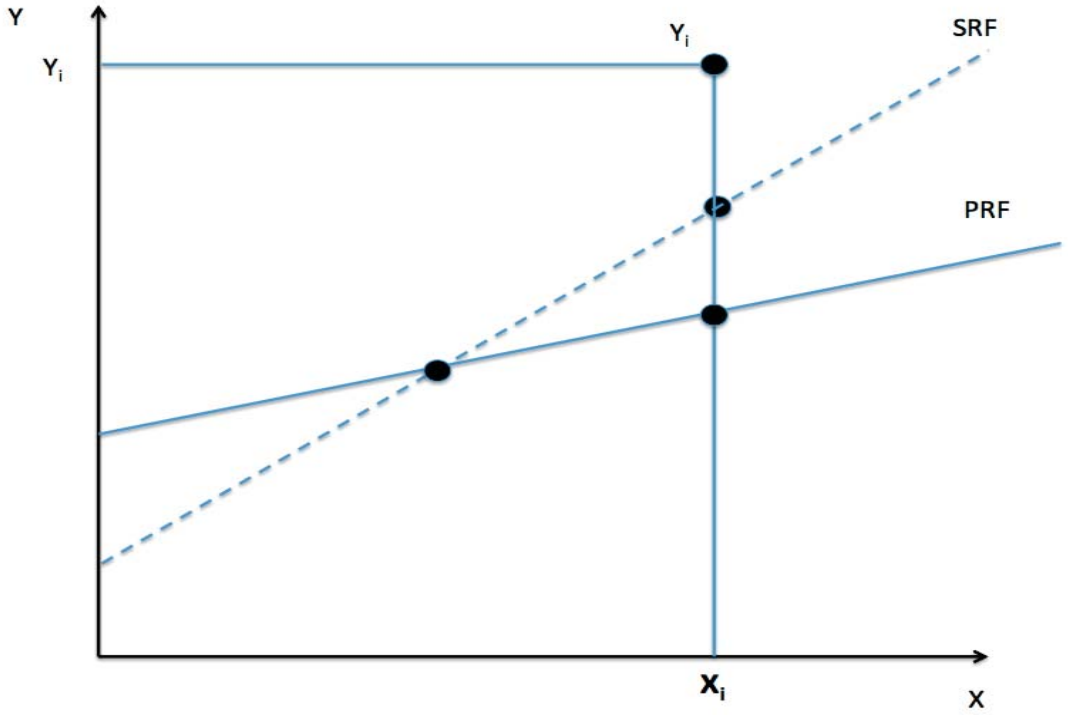
We can express the SRF in its stochastic form as follows:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\mu}_i$$

In sum, our ultimate goal is to estimate
the PRF

on the basis of
the SRF

Figure 2.4: Sample and Population Regression Lines



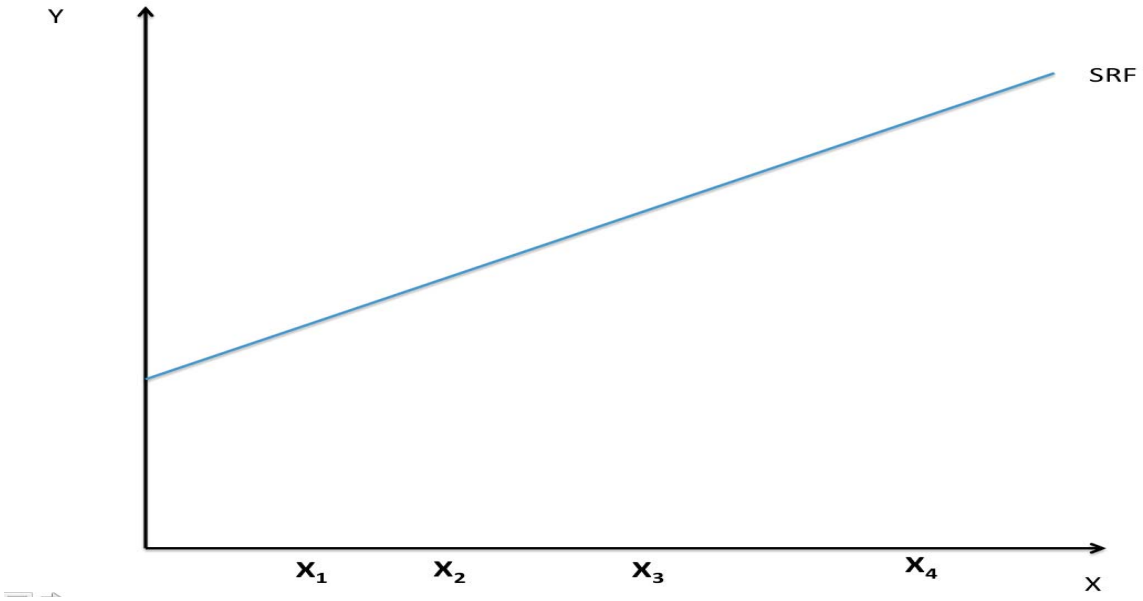
3. REGRESSION:THE PROBLEM OF ESTIMATION

As mentioned in the previous chapter, our main objective is to estimate the population regression function (PRF) based on the basis of the sample regression function (SRF) as accurately as possible.

In this chapter, we are going to discuss the method of estimation:Ordinary Least Squares (OLS)

3.1 The Method of Ordinary Least Squares (OLS)

Figure 3.1: Least-Squares Criterion



3.1.1 The Method to Find Out the Least-Squares Estimators: $\hat{\beta}_1$ and $\hat{\beta}_2$

From the SRF:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

Now, we obtain the **least-squares estimators**:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i^2 - (\sum X_i)^2} \\ &= \bar{Y} - \hat{\beta}_2 \bar{X} \end{aligned} \tag{3.1}$$

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum Xi \sum Yi}{n\sum X_i^2 - (\sum X_i)^2} \tag{3.2}$$

If we define $\bar{X}$ and $\bar{Y}$ to be the sample means of X and Y. Then:

$$\begin{aligned} x_i &= (X_i - \bar{X}) \\ y_i &= (Y_i - \bar{Y}) \end{aligned} \tag{3.3}$$

We can have the alternative expressions for $\hat{\beta}_2$:

$$\begin{aligned} \hat{\beta}_2 &= \frac{\sum x_i y_i}{\sum x_i^2} \\ &= \frac{\sum x_i Y_i}{\sum X_i^2 - n\bar{X}^2} \\ &= \frac{\sum X_i y_i}{\sum X_i^2 - n\bar{X}^2} \end{aligned} \tag{3.4}$$

Show that

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

EXAMPLE

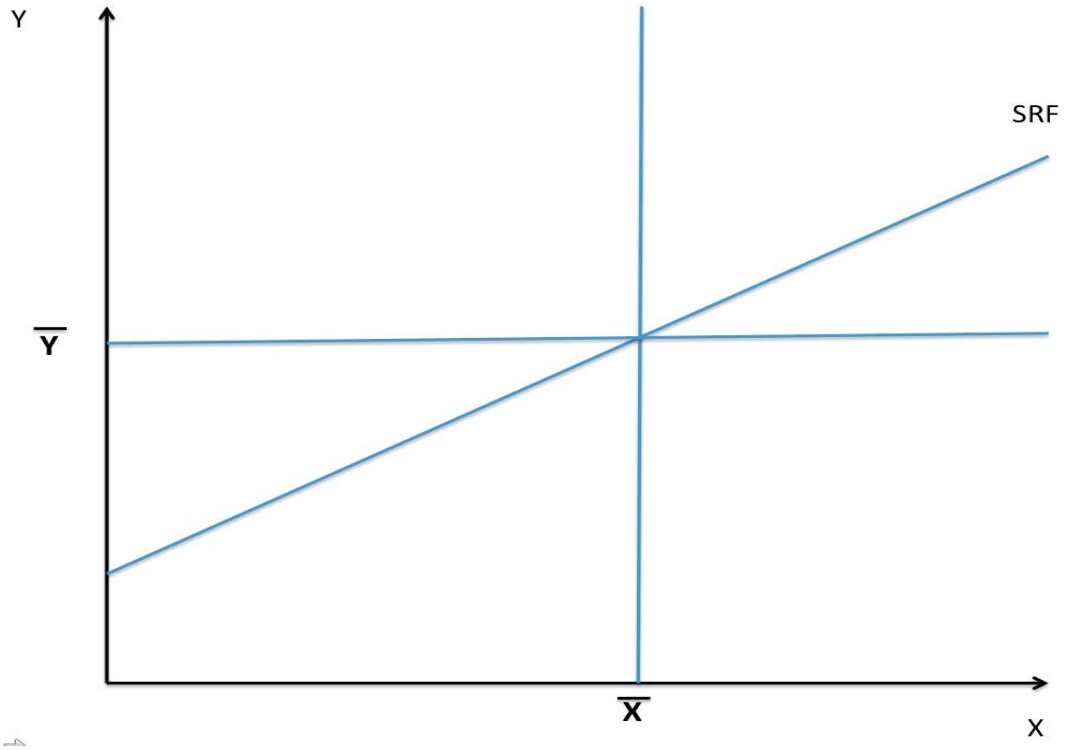
Table 3.1: A Random Sample From the Population

X	Y
500	390
600	425
700	560
800	575
900	630
1000	679

Table 3.2: Raw Data Based on the Sample Data on Table 3.1

	(1) Y_i	(2) X_i	(3) Y_iX_i	(4) X_i^2	(5) $x_i = X_i - \bar{X}$	(6) $y_i = Y_i - \bar{Y}$	(7) x_i^2	(8) x_iy_i	(9) $\hat{Y}_i$	(10) $\hat{u}_i = Y_i - \hat{Y}_i$	(11) $\hat{Y}_i\hat{u}_i$
	390	500	195,000	250,000	−250	−153.17	62,500	38,291.67			
	425	600	255,000	360,000	−150	−118.17	22,500	17,725			
	560	700	392,000	490,000	−50	16.83	2,500	−841.67			
	575	800	460,000	640,000	50	31.83	2,500	1,591.67			
	630	900	567,000	810,000	150	86.83	22,500	13,025			
	679	1,000	679,000	1,000,000	250	135.83	62,500	33,958.33			
Sum	3,259	4,500	2,548,000	3,550,000	0	0	175,000	103,750			
Mean	543.17	750	424,666.67	591,666.670	0	0	29,166.67	17,291.67			

Figure 3.2: Sample Regression Line Based on the Data of Table 3.2



3.1.2 The numerical and statistical properties of OLS estimators

1. The OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are expressed solely in terms of the observable (Sample size) and quantities (i.e X and Y).

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i^2 - (\sum X_i)^2} \\ &= \bar{Y} - \hat{\beta}_2 \bar{X}\end{aligned} \tag{3.5}$$

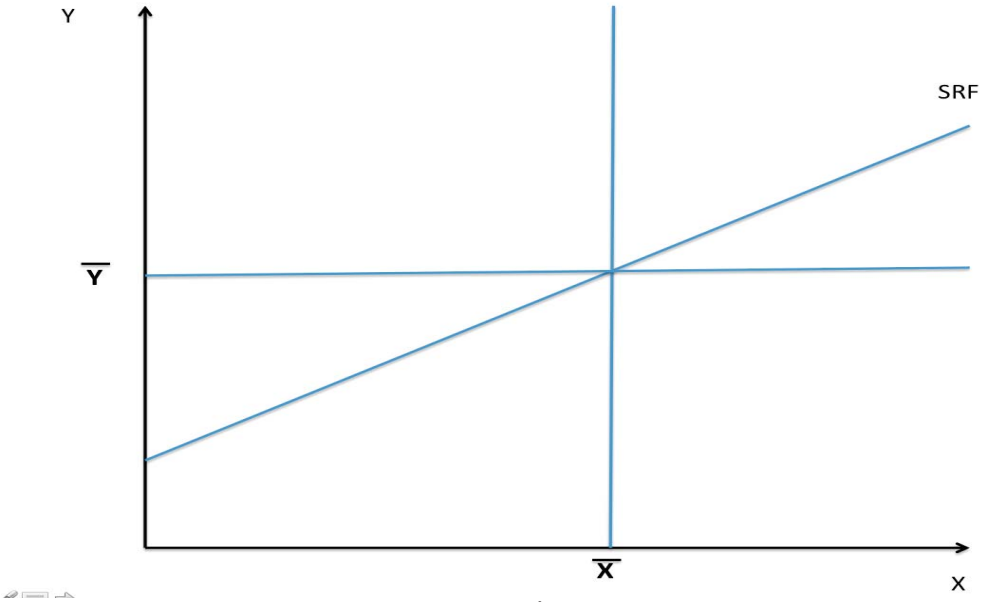
$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum Xi \sum Yi}{n\sum X_i^2 - (\sum X_i)^2} \tag{3.6}$$

2. They are **point estimators**.

3. The regression line has the following properties.

3.1 The sample regression function (SRF) passes through the sample means of Y and X ($\bar{Y}$ and $\bar{X}$).

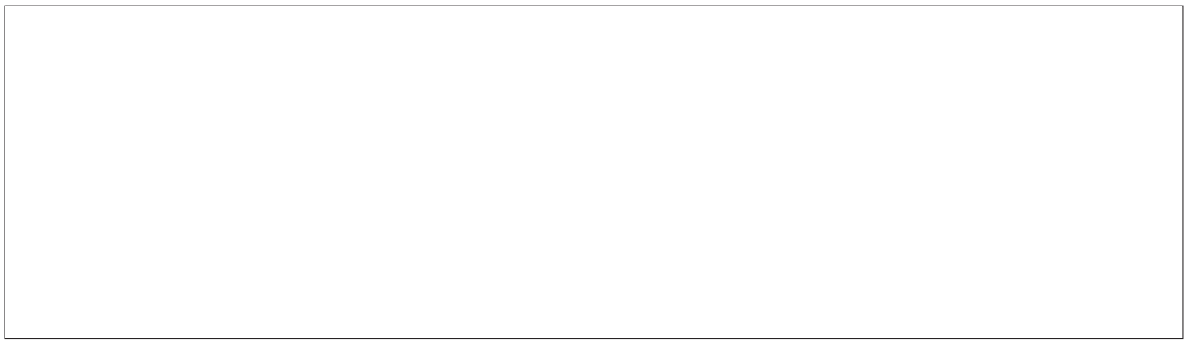
Figure 3.3: The Sample regression Line Passes through the Sample Mean Values of Y and X



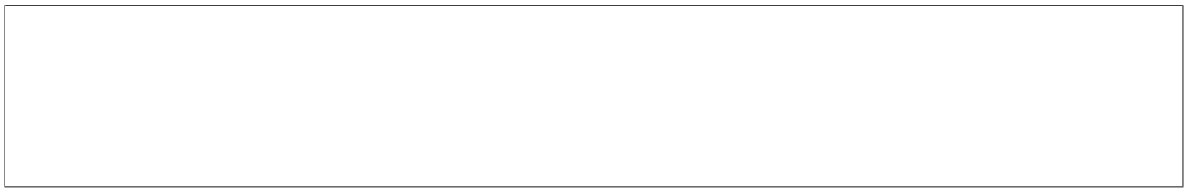
3.2 The mean value of the estimated $Y = \hat{Y}_i$ is equal to the mean value of the actual Y.

3.3. The mean value of the residuals $\hat{u}_i$ is zero.

3.4 The residuals $\hat{u}_i$ are uncorrelated with the predicted Y_i.



3.5 The residuals $\hat{u}_i$ are uncorrelated with X_i.



3.1.3 The Assumptions Underlying the Method of Least Squares

Assumption 1: Linear regression model

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

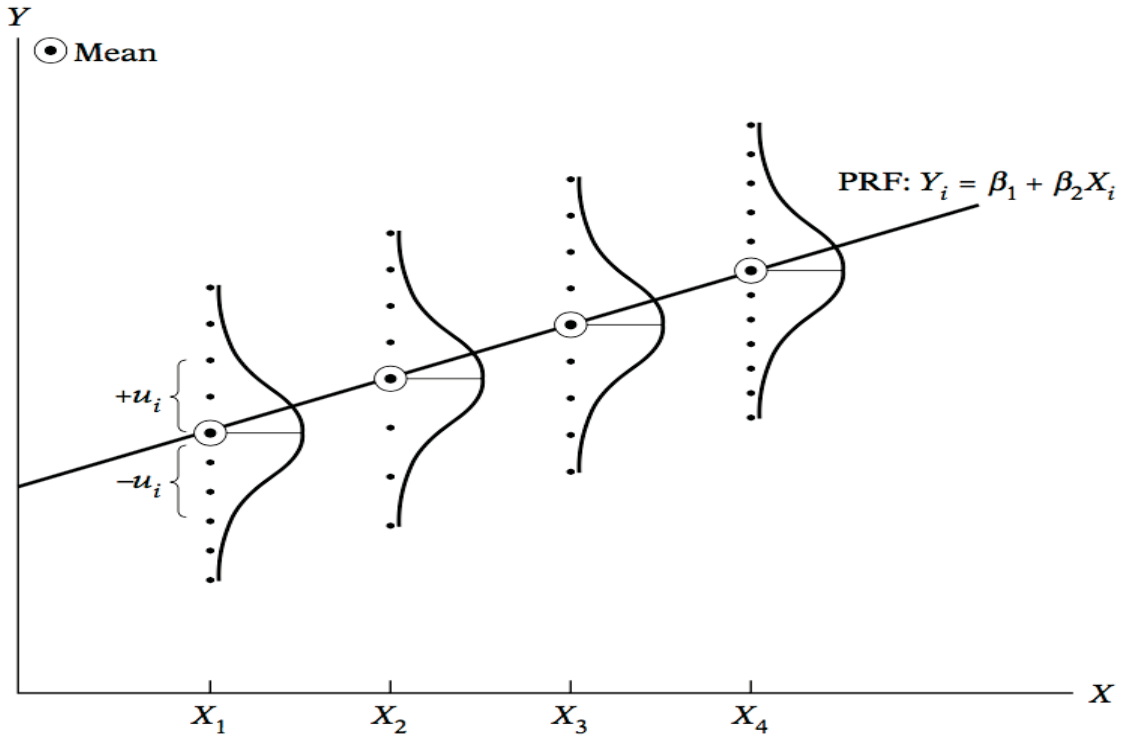
Assumption 2: X values are fixed in repeated sampling

X is assumed to be nonstochastic.

Assumption 3: Zero mean value of disturbance u_i

$$E(u_i|Xi) = 0$$

Figure 3.4: Conditional Distribution of the Disturbances u_i



Assumption 4: Homoscedasticity or Equal Variance of u_i

Figure 3.5: Homoscedasticity

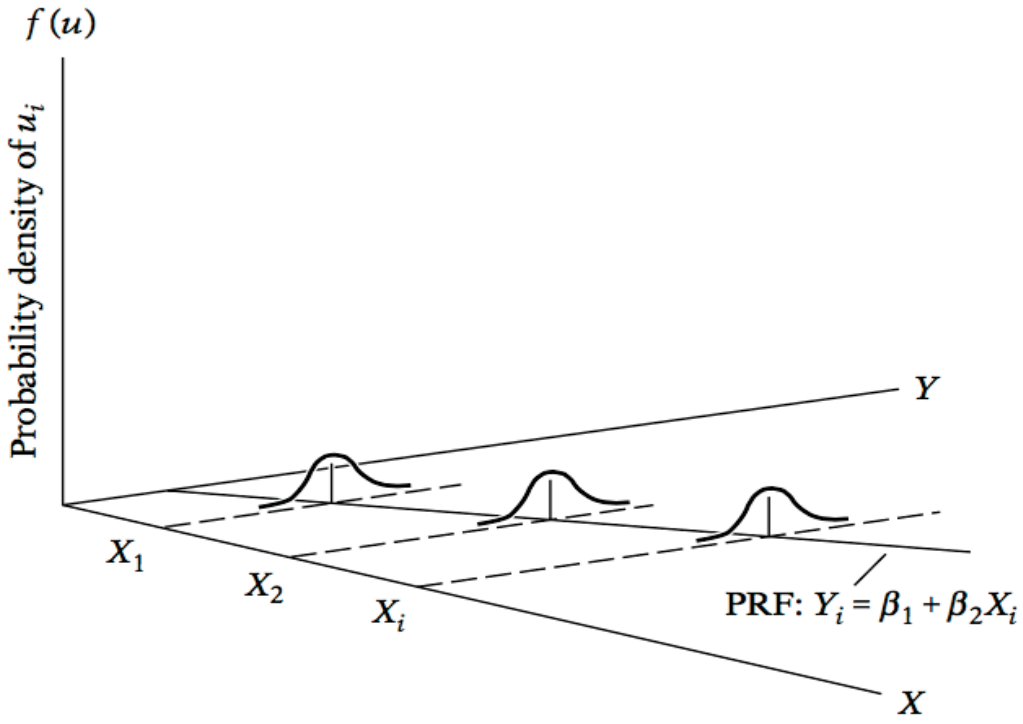
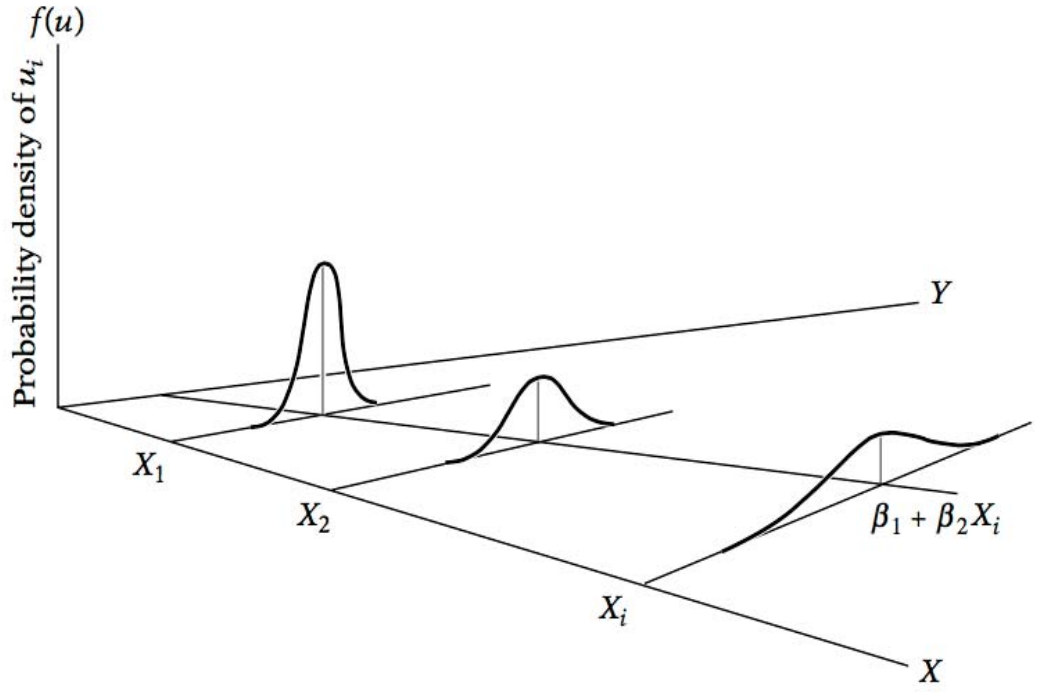


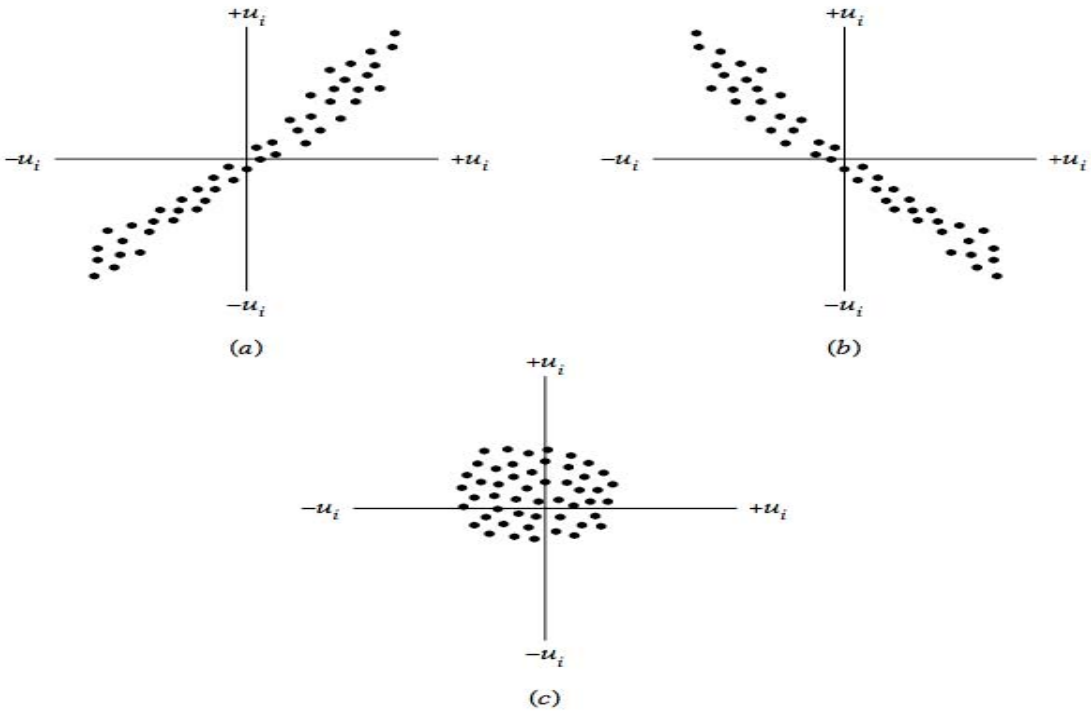
Figure 3.6: Heteroscedasticity



Assumption 5: No Autocorrelation Between the Disturbances

Assumption 6: Zero Covariance Between u_i and X_i

Figure 3.7: Patterns of Correlation Among the disturbances



Assumption 7: The number of observations n must be greater than the number of parameters to be estimated.

Assumption 8:Variability in X values.

Assumption 9: The regression model is correctly specified.

Assumption 10: There is no perfect multicollinearity.

3.1.4 Standard Errors of Least-Squares Estimates

The standard errors of the OLS estimates can be obtained as follows:
We know that

$$\hat{\beta}_2 = \frac{\sum x_i Y_i}{\sum x_i^2} = \sum k_i Yi$$

where

$$k_i = \frac{x_i}{\sum x_i^2}$$

The properties of the weights k_i

1. The k_i are nonstochastic.
2. $\sum k_i = 0$
3. $\sum k_i^2 = \frac{1}{\sum x_i^2}$
4. $\sum k_i x_i = \sum k_i Xi = 1$

Since

$$var(\hat{\beta}_2) = E[\hat{\beta}_2 - E(\hat{\beta}_2)]^2$$

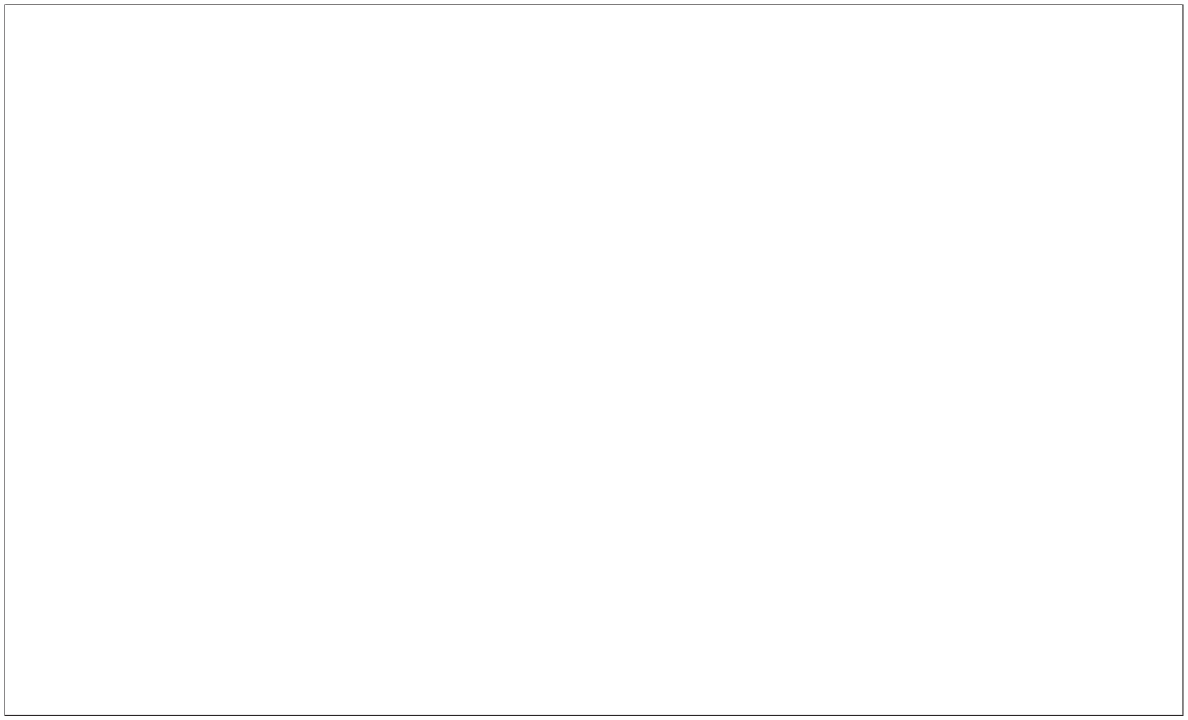
First Step
Find the $E(\hat{\beta}_2)$

Second Step

Using the definition of variance

$$var(\hat{\beta}_2) = E[\hat{\beta}_2 - E(\hat{\beta}_2)]^2$$

The covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$



3.1.5 The Least-Square Estimator of σ^2

In sum, the standard errors of the OLS estimators can be obtained as follow:

$$\begin{aligned} \text{var}(\hat{\beta}_2) &= \frac{\sigma^2}{\sum x_i^2} \\ \text{se}(\hat{\beta}_2) &= \frac{\sigma}{\sqrt{\sum x_i^2}} \end{aligned} \tag{3.7}$$

$$\begin{aligned} \text{var}(\hat{\beta}_1) &= \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2 \\ \text{se}(\hat{\beta}_1) &= \sqrt{\frac{\sum X_i^2}{n\sum x_i^2}}\sigma \end{aligned} \tag{3.8}$$

We can estimate the σ^2 from the data where the formula for the estimated σ^2 is following :

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

where

$$\sum \hat{u_i}^2 = \sum y_i^2 - \hat{\beta_2}^2 \sum x_i^2$$

The alternative expression for computing $\sum \hat{u_i}^2$ is

$$\sum \hat{u_i}^2 = \sum y_i^2 - \frac{(\sum x_i y_i)^2}{\sum x_i^2}$$

The covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$ is:

$$\begin{aligned} \text{cov}(\hat{\beta}_1, \hat{\beta}_2) &= -\bar{X}\,\text{var}(\hat{\beta}_2) \\ &= -\bar{X}\left(\frac{\sigma^2}{\sum x_i^2}\right) \end{aligned} \tag{3.9}$$

3.1.6 Properties of Least-Squares Estimators: The Gauss-Markov Theorem

Given the assumptions of the classical linear regression model, the least-square estimators are satisfied the optimum properties which is known as **"The Gauss- Markov Theorem."** To understand this theorem, we need to know the small-sample properties of an estimator first.

The Small-Sample Properties of An Estimator

1. Unbiasedness

An estimator $\hat{\theta}$ is said to be an unbiased estimator of θ if the expected value of $\hat{\theta}$ is equal to the true θ

$$E(\hat{\theta}) = \theta$$

Therefore, if the expected value of $\hat{\theta}$ is not equal to the true θ, then the estimator is said to be biased. We can calculate the biased as:

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Figure: Biased and Unbiased Estimators

2. Minimum Variance

$\hat{\theta}_1$ is said to be a minimum variance estimator of θ if the variance of $\hat{\theta}_1$ is smaller that or at most equal to the variance of $\hat{\theta}_2$, which is any other estimator of θ

Figure: Minimum Variance

3.Best Unbiased or Efficient Estimator = property 1+ property 2

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of θ and the variance of $\hat{\theta}_1$ is smaller that or at most equal to the variance of $\hat{\theta}_2$, then $\hat{\theta}_1$ is **a minimum-variance unbiased estimator or best unbiased estimator.**

4. Linearity

An estimator $\hat{\theta}$ is said to be a linear estimator of θ if it is a linear function of the sample observations. For example:

$$\bar{X} = \frac{1}{n}\sum X_i = \frac{1}{n}(X_1 + X_2 + ... + X_n)$$

Thus, $\bar{X}$ is a linear estimator because it is a linear function of the X values.

Best Linear Unbiased Estimators : BLUE

The estimator $\hat{\theta}$ is called as the Best Linear Unbiased Estimator **BLUE** if it is satisfied the properties 1,2,4 that is $\hat{\theta}$ is linear, is unbiased, and has the minimum variance in the class of all linear unbiased estimators of θ.

Minimum Mean-Square-Error (MSE) Estimator

The MSE measures dispersion around the true value of the parameter. It is defined as:

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

However, the variance of $\hat{\theta}$ measures the dispersion of the distribution of the distribution of $\hat{\theta}$ around its mean or expected value.

$$\text{var}(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2$$

The relationship between the MSE($\hat{\theta}$) and the var($\hat{\theta}$) is as follows:

An estimator $\hat{\beta}_2$ is said to be a best linear unbiased estimator (BLUE) of β_2 if the following hold:

♣ **It is linear**. It is the linear function of a random variable.

♣ **It is unbiased**. That is $E(\hat{\beta}_2)$ is equal to the true value, β_2

♣ **It has the minimum variance in the class of all such linear unbiased estimators**.

Gauss-Markov Theorem: Given the assumptions of the classical linear regression model, the least-squares estimators, in the class of unbiased linear estimators, have minimum variance, that is , they are BLUE.

3.1.7 A measure of goodness of fit: r^2

In this section, we are going to study the goodness of fit of the fitted regression line to a set of data. Let us consider the following example:

Suppose we were to estimate the family expenditure (Y) based on our information from a random sample (as in Table 3.2).

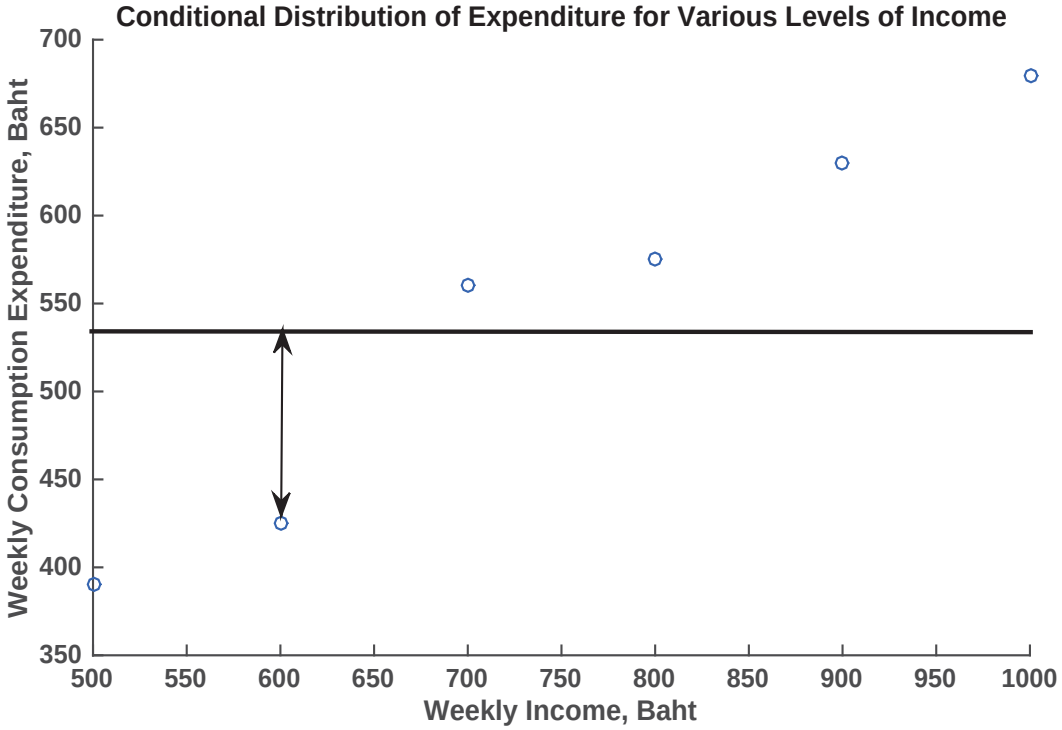
What will happen if we set the estimated Y to be $\bar{Y}$?

Table 3.3: Estimating the expenditure of the household

Family Number (i)	Actual Y_i	Estimate $\hat{Y}_i = \bar{Y}$	**Error in Estimation** $Y_i - \bar{Y}$	**Errors Squared** $(Y_i - \bar{Y})^2$
1	390	543	-153	23460.03
2	425	543	-118	13963.36
3	560	543	17	283.36
4	575	543	32	1013.36
5	630	543	87	7540.03
6	679	543	136	18450.69
Sum	3259	3259	0	64710.83

We can see all this graphically:

Figure 3.8: Graphic Representation



Question: Can we determine the total estimation error for this sample data?

Answer: Yes, we can calculate the total (combined) amount of estimation error for all observations in the sample when **using the mean as the estimate** as following:

$$TSS = \sum (Y_i - \bar{Y})^2$$

It is called the total sum of squares (TSS) which is the total variation of the actual Y values about their sample mean.

Since our objective in estimation is to minimize error (maximize precision), we need to cut down the amount of the estimation error (TSS).

We can achieve this by using information about other variables suspected to be strong predictors (strongly related to) the expenditure of the families.

We now can attempt to estimate the expenditure from the information on the income level of the family, rather than from its own mean.

Table 3.4: Estimating the expenditure of the household with income

Family (i)	Actual Y_i	Income X_i	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$
1	390	500	-250	-153.17	38291.67	62500
2	425	600	-150	-118.17	17725.00	22500
3	560	700	-50	16.83	-841.67	2500
4	575	800	50	31.83	1591.67	2500
5	630	900	150	86.83	13025.00	22500
6	679	1000	250	135.83	33958.33	62500
Sum	3259	4500	0	0	103750	175000

From the table 8, we can calculate the simple regression as following:

Figure 3.9: Breakdown of the variation of Y_i into two components

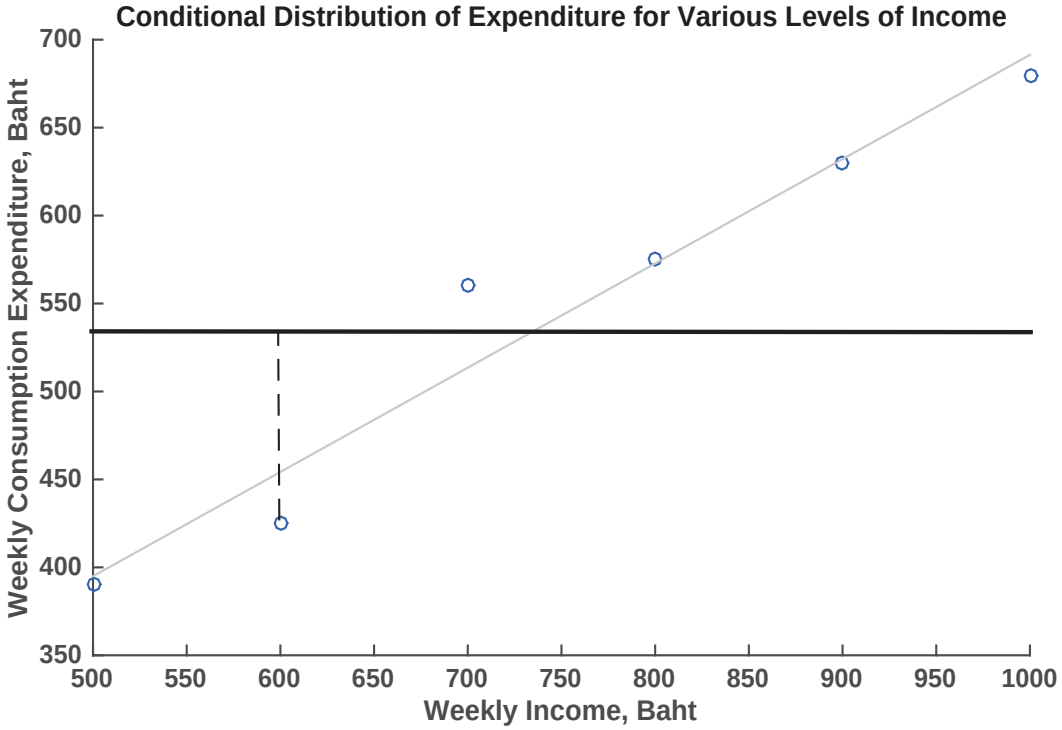


Table 3.5: Estimating the expenditure of the household with income

Family (i)	Actual Y_i	Income X_i	Regression Estimate $\hat{Y}$	Residual $Y-\hat{Y}$	Residual squared $(Y-\hat{Y})^2$
1	390	500	394.95	-4.95	24.53
2	425	600	454.24	-29.24	854.87
3	560	700	513.52	46.48	2160.04
4	575	800	572.81	2.19	4.80
5	630	900	632.10	-2.10	4.39
6	679	1000	691.38	-12.38	153.29
Sum	3259	4500	0	0	3201.90

From the table 9, we can calculate the estimation error we have committed by using the regression line as:

$$RSS = \sum (Y_i - \hat{Y}_i)^2 = \sum \hat{u}_i{}^2$$

where RSS stands for the residual sum of squares. which is the unexplained variation of the Y values about the regression line.

Total Baseline Error using the mean (SS Total) =

New or Remaining Error (SS Error or SS Residual) =

QUESTION: How much of the original estimation error have we explained away (eliminated) by using the regression model (instead of the mean)?

ANS

QUESTION: What % of estimation error have we explained (eliminated by using the regression model?

ANS

QUESTION: What does the remaining% represent?

ANS
Percent of variation (differences) in expenditures that can be accounted for by: (a) all other potential predictors not included in the model, beyond income levels, and (b) unexplainable random/chance variations.

$$r^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

♣ r^2 is a measure of our success regarding accuracy of our estimation effort.

♣ r^2 = % of estimation error that we have been able to explain away by using the regression model, instead of using the mean.

♣ r^2 indicates how much better we can predict Y from information about Xs, rather than from using its own mean.

♣ r^2 = % of differences (variations) in Y values that is explained by (attributable to) differences in X values.

4. Classical Normal Regression Model (CNLRM)

We know that the classical theory of statistical inference consists of:

1. Estimation

We have covered this topic since we were able to estimate the parameters β_1,β_2, and σ^2 by using the method of OLS.

We also proved that these estimators $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\sigma}$ satisfy several desirable statistical properties, such as unbiasedness, minimum variance, and linearity (BLUE property).

However, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\sigma}$ change their values from sample to sample. The following tables show the two different sets of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\sigma}$ depending on the two different sample data.

Table 4.1: Estimating the expenditure of the household with income

Family (i)	Actual Y_i	Income X_i	Regression Estimate $\hat{Y}$	Residual $Y-\hat{Y}$	Residual squared $(Y-\hat{Y})^2$
1	390	500	394.95	-4.95	24.53
2	425	600	454.24	-29.24	854.87
3	560	700	513.52	46.48	2160.04
4	575	800	572.81	2.19	4.80
5	630	900	632.10	-2.10	4.39
6	679	1000	691.38	-12.38	153.29
Sum	3259	4500	0	0	3201.90

If we use this sample data. We can estimate:

$$\hat{\beta}_1 = 98.524$$

$$\hat{\beta}_2 = 0.593$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} = \frac{3201.90}{6-2} = 800.476$$

Table 4.2: Estimating the expenditure of the household with income with another sample data

Family (i)	Actual Y_i	Income X_i	Regression Estimate $\hat{Y}$	Residual $Y-\hat{Y}$	Residual squared $(Y-\hat{Y})^2$
1	360	500	325.71	64.29	4132.65
2	390	600	406.43	18.57	344.90
3	440	700	487.14	72.86	5308.16
4	575	800	567.86	7.14	51.02
5	670	900	648.57	-18.57	344.90
6	730	1000	729.29	-50.29	2528.65
Sum	3165	4500	0	0	12710.29

If we use this sample data. We can estimate:

$$\hat{\beta}_1 = -77.857$$

$$\hat{\beta}_2 = 0.807$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} = \frac{12710.29}{6-2} = 3177.571$$

From the example, you can easily see that these estimators are **RANDOM VARIABLES**. Therefore, we need to learn another part of statistical inference which is called **Hypothesis Testing.**

2. Hypothesis Testing

The main objective is to find out how close of $\hat{\beta}_1$ and $\hat{\beta}_2$ to the true β_1 and the true β_2, respectively. Also, we would like to see how close of $\hat{\sigma}^2$ compared to the true σ^2.

To achieve this goal, we need to know the probability distributions of $\hat{\beta}_1$, $\hat{\beta}_2$,and $\hat{\sigma}^2$.
Consider the estimator of β_2:

$$\hat{\beta}_2 = \sum k_i Y_i$$

We can write the above equation as:

$$\hat{\beta}_2 = \sum k_i(\beta_1 + \beta_2 X_i + u_i)$$

From this equation, the probability distribution of $\hat{\beta}_2$ will depend on the assumption made about the probability distribution of u_i

4.1 The Normality Assumption for u_i

In the classical normal linear regression model (CNLRM), we assume that each u_i is distributed normally :

$$u_i \sim N(0, \sigma^2)$$

where
Mean:

$$E(u_i) = 0$$

Variance:

$$E[u_i - E(u_i)]^2 = E(u_i^2) = \sigma^2$$

$$cov(u_i, u_j) = E\left\{[u_i - E(u_i)][u_j - E(u_j)]\right\} = E(u_i U_j) = 0$$

Therefore,

$$u_i \sim N(0, \sigma^2)$$

Also, u_i and u_j are not only uncorrelated but also independently distributed.

we can then write the above equation as:

$$u_i \sim NID(0, \sigma^2)$$

where NID stands for normally and independently distributed.

4.2 Properties of OLS estimators under the normality assumption

1. They are unbiased.

2. They have minimum variance.

3. By 1+2 properties, they are minimum-variance unbiased, or efficient estimators.

4. $\hat{\beta}_1$ is normally distributed with:

$$\text{Mean:} E(\hat{\beta}_1) = \beta_1$$

$$\text{var}(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n \sum x_i^2} \sigma^2$$

Therefore,

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$$

By the properties of the normal distribution, we can:

5. $\hat{\beta}_2$ is normally distributed with

$$\text{Mean:} E(\hat{\beta}_2) = \beta_2$$

$$\text{var}(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum x_i^2}$$

or more compactly

$$\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\hat{\beta}_2})$$

then we can define the standard normal distribution as

6. $(n-2)(\hat{\sigma}^2/\sigma^2)$ is distributed as the χ^2 (chi-square) distribution with (n-2) df.

7. $(\hat{\beta}_1, \hat{\beta}_2)$ are distributed independently of $\hat{\sigma}^2$

8. $\hat{\beta}_1$ and $\hat{\beta}_2$ have the minimum variance in the entire class of unbiased estimators, whether linear or not.

9. we can find out the probability distribution of Y_i as following:

5. Interval Estimation and Hypothesis Testing

Interval Estimation

5.1 Confidence Intervals for Regression Coefficients β_1 and β_2

In Sum

A $100(1-\alpha)$ percent **confidence interval** for β_2 can be defined as:

$$\hat{\beta}_2 \pm t_{\alpha/2}\text{se}(\hat{\beta}_2)$$

or

$$\Pr[\hat{\beta}_2 - t_{\alpha/2}\text{se}(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2}\text{se}(\hat{\beta}_2)] = 1-\alpha$$

Analogously, we can define $100(1-\alpha)$ percent **confidence interval** for β_1 as:

$$\hat{\beta}_1 \pm t_{\alpha/2}\text{se}(\hat{\beta}_1)$$

or

$$\Pr[\hat{\beta}_1 - t_{\alpha/2}\text{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2}\text{se}(\hat{\beta}_1)] = 1-\alpha$$

Example

Table 5.1: Estimating the expenditure of the household with income

Family (i)	Actual Y_i	Income X_i	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$
1	390	500	-250	-153.17	38291.67	62500
2	425	600	-150	-118.17	17725.00	22500
3	560	700	-50	16.83	-841.67	2500
4	575	800	50	31.83	1591.67	2500
5	630	900	150	86.83	13025.00	22500
6	679	1000	250	135.83	33958.33	62500
Sum	3259	4500	0	0	103750	175000

Table 5.2: Estimating the expenditure of the household with income

Family (i)	Actual Y_i	Income X_i	Regression Estimate $\hat{Y}$	Residual $Y-\hat{Y}$	Residual squared $(Y-\hat{Y})^2$
1	390	500	394.95	-4.95	24.53
2	425	600	454.24	-29.24	854.87
3	560	700	513.52	46.48	2160.04
4	575	800	572.81	2.19	4.80
5	630	900	632.10	-2.10	4.39
6	679	1000	691.38	-12.38	153.29
Sum	3259	4500	0	0	3201.90

Confidence Interval for β_2

Confidence Interval for β_1

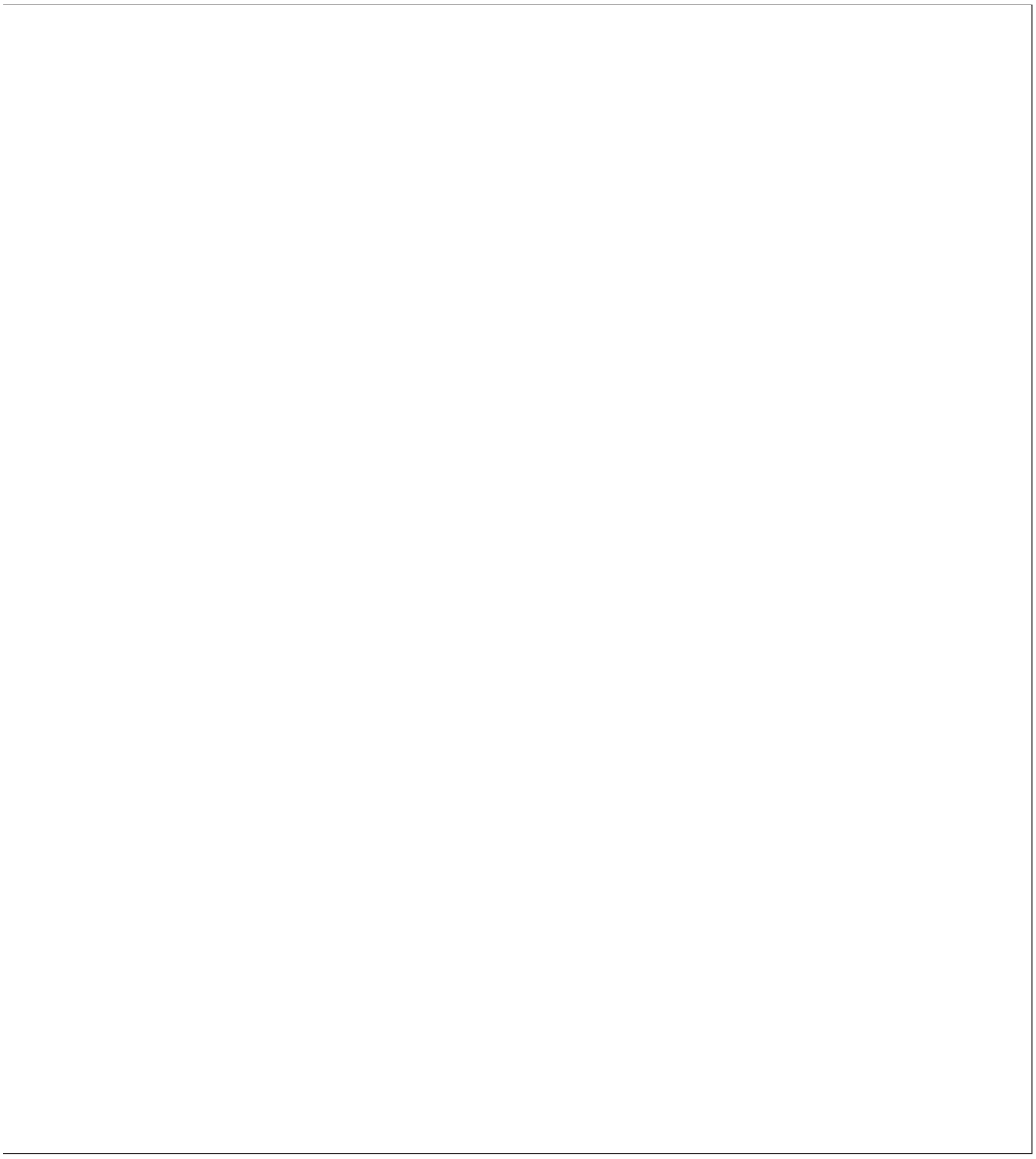
5.2 Confidence Interval for σ^2

5.3 Hypothesis Testing: The Confidence-Interval Approach

Based on our sample data, the estimated marginal propensity to consume (MPC), $\hat{\beta}_2$ is 0.593. Suppose we postulate that

$$H_0 : \beta_2 = 0.6$$

$$H_1 : \beta_2 \neq 0.6$$



5.4 Hypothesis Testing: The Test of Significance Approach

5.4.1 Two-Tail Test

Based on the sample data, the estimated marginal propensity to consume (MPC), $\hat{\beta}_2$ is 0.593. Suppose we postulate that

$$H_0 : \beta_2 = 0.6$$

$$H_1 : \beta_2 \neq 0.6$$

5.4.2 One-Tail Test

Based on the sample data, the estimated marginal propensity to consume (MPC), $\hat{\beta}_2$ is 0.593. Suppose we postulate that

$$H_0 : \beta_2 \leq 0.6$$

$$H_1 : \beta_2 > 0.6$$

We can summarize the decision rules for the t test as follow:

Figure 5.1 The t test of Significance: Decision rules

Type of hypothesis	H_0: the null hypothesis	H_1: the alternative hypothesis	Decision rule: reject H_0 if
Two-tail	$\beta_2 = \beta_2^*$	$\beta_2 \neq \beta_2^*$	$\lvert t \rvert > t_{\alpha/2,df}$
Right-tail	$\beta_2 \leq \beta_2^*$	$\beta_2 > \beta_2^*$	$t > t_{\alpha,df}$
Left-tail	$\beta_2 \geq \beta_2^*$	$\beta_2 < \beta_2^*$	$t < -t_{\alpha,df}$

Notes: β_2^* is the hypothesized numerical value of β_2.
$\lvert t \rvert$ means the absolute value of t.
t_α or $t_{\alpha/2}$ means the critical t value at the α or $\alpha/2$ level of significance.
df: degrees of freedom, $(n-2)$ for the two-variable model, $(n-3)$ for the three-variable model, and so on.
The same procedure holds to test hypotheses about β_1.

5.4.3 Testing the significance of σ^2: The χ^2 test

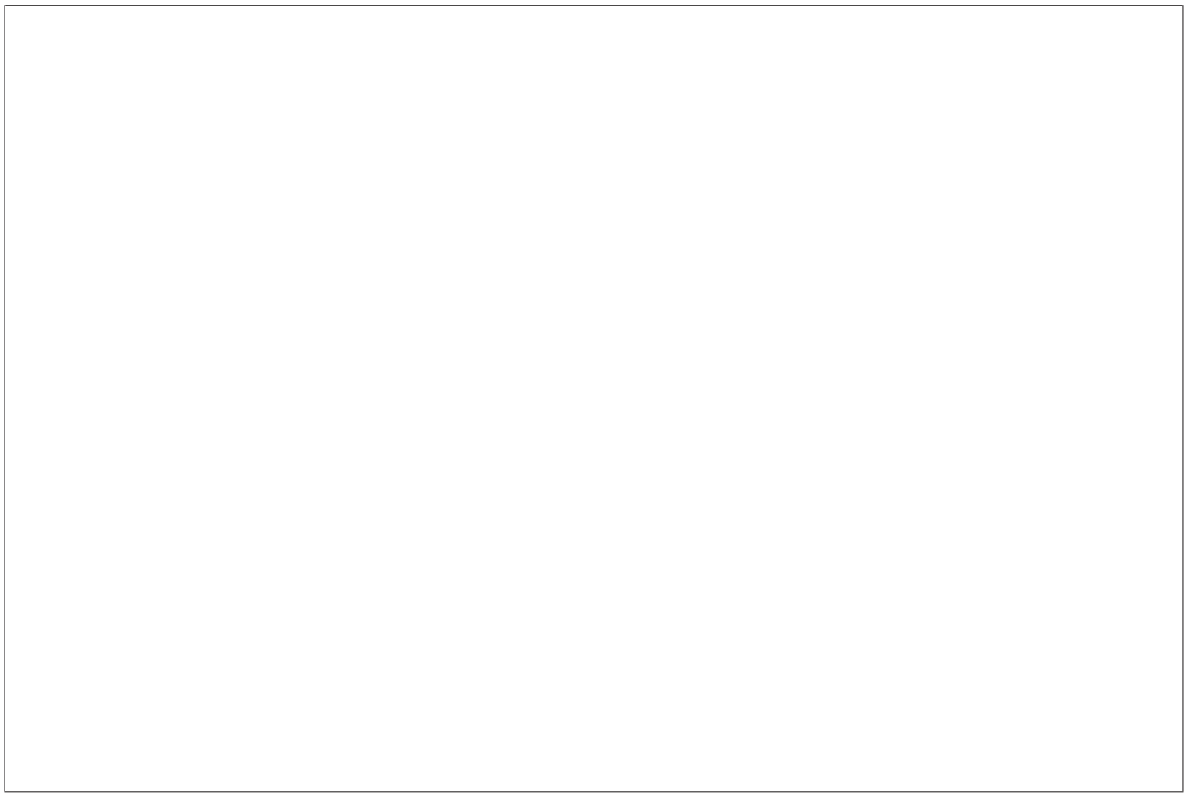


Figure 5.2 The χ^2 Test : Decision rules

H_0: the null hypothesis	H_1: the alternative hypothesis	Critical region: reject H_0 if
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\frac{\text{df}(\hat{\sigma}^2)}{\sigma_0^2} > \chi^2_{\alpha,\text{df}}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\frac{\text{df}(\hat{\sigma}^2)}{\sigma_0^2} < \chi^2_{(1-\alpha),\text{df}}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\frac{\text{df}(\hat{\sigma}^2)}{\sigma_0^2} > \chi^2_{\alpha/2,\text{df}}$ or $< \chi^2_{(1-\alpha/2),\text{df}}$

Note: σ_0^2 is the value of σ^2 under the null hypothesis. The first subscript on χ^2 in the last column is the level of significance, and the second subscript is the degrees of freedom. These are critical chi-square values. Note that df is $(n-2)$ for the two-variable regression model, $(n-3)$ for the three-variable regression model, and so on.

Why do we say "we cannot reject the null hypothesis?" instead of "We accept the null hypothesis"

The Level of Significance: α

Type I error

Type II error

The Exact Level of Significance: The p Value

5.4.1 Regression Analysis and Analysis of Variance

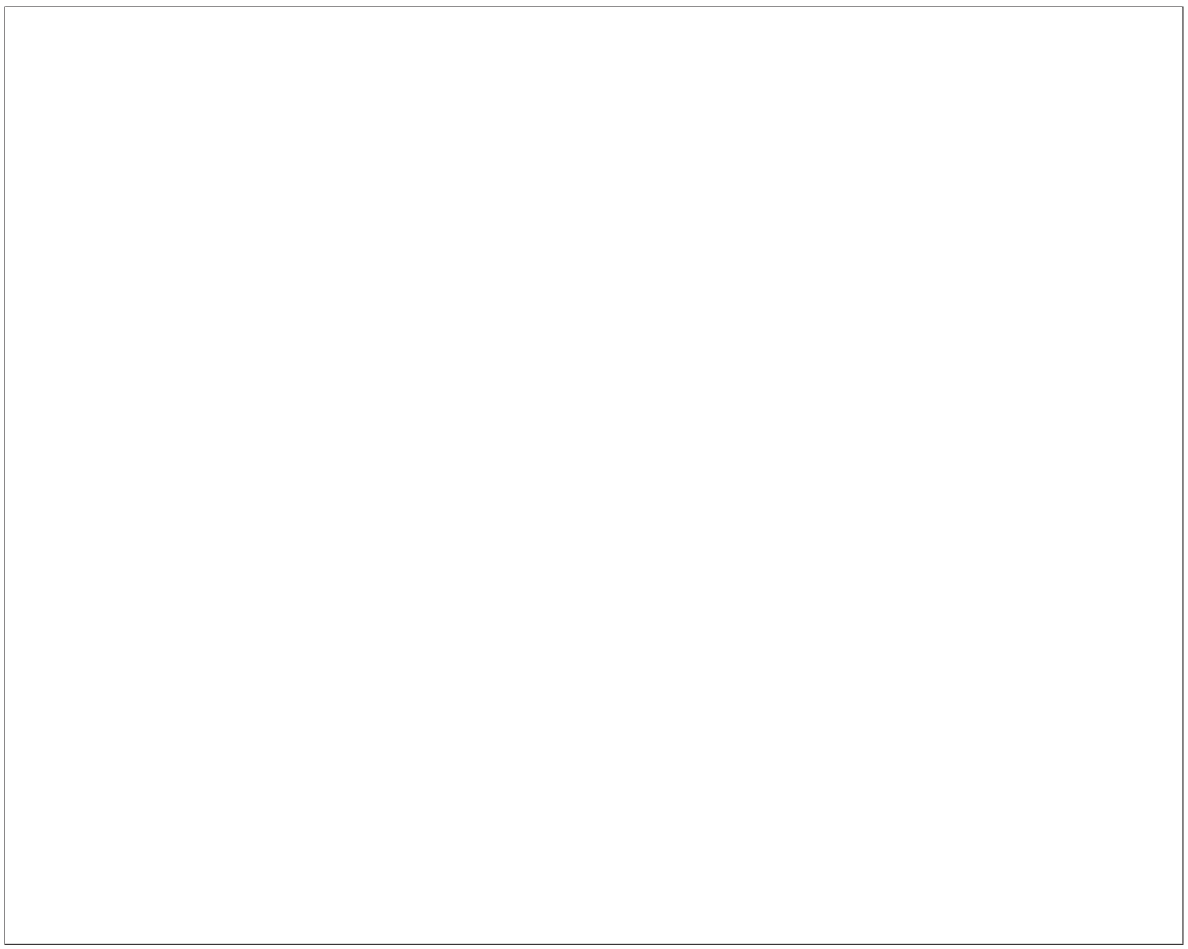


Table 5.3: ANOVA Table for the two-variable regression model

Source of variation	Sum of Square **SS**	**df**	Mean Sum of Square **MSS**
Due to regression (ESS)			
Due to residuals (RSS)			
TSS			

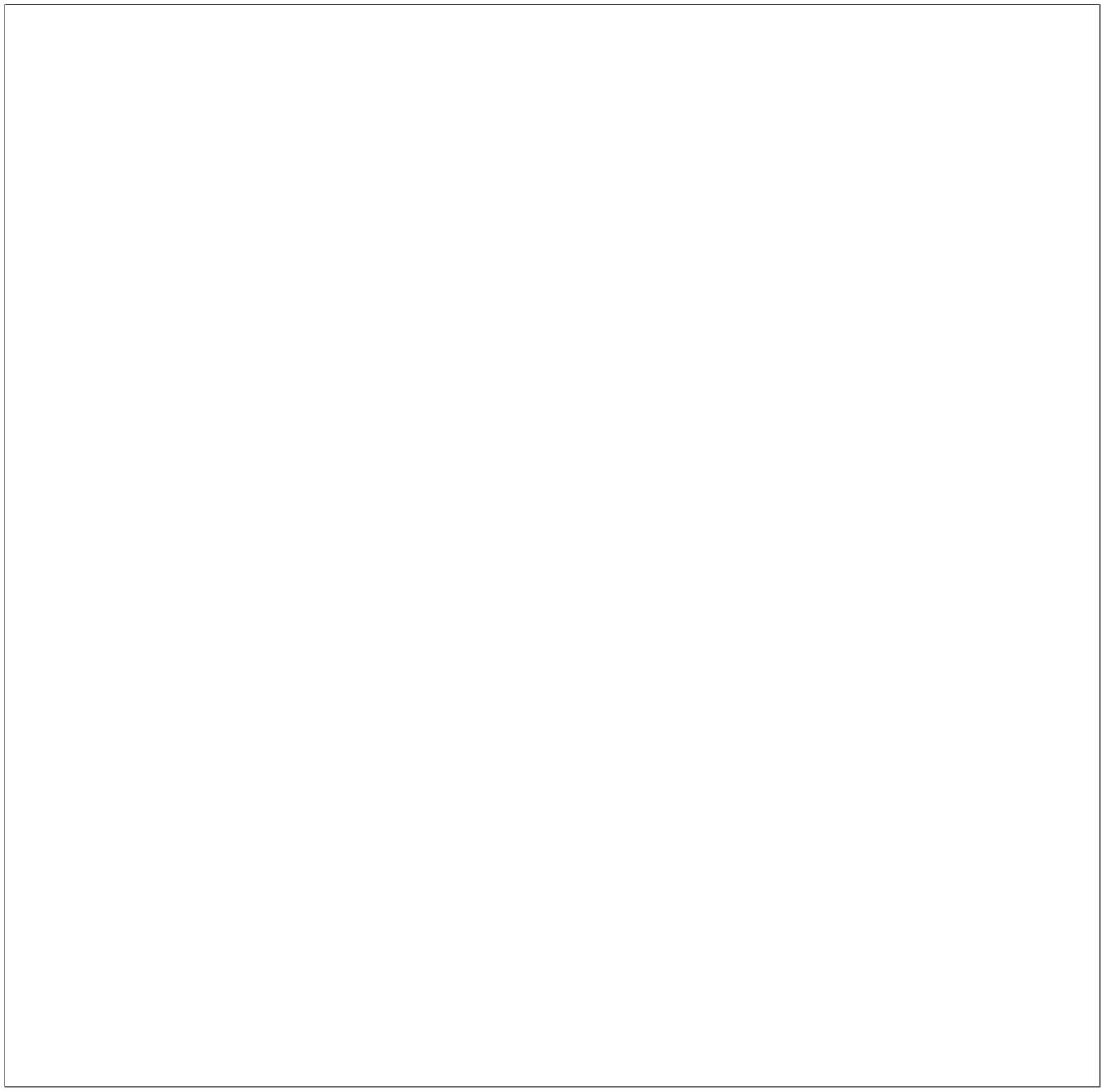


Table 5.4: Estimating the expenditure of the household

Family Number (i)	Actual Y_i	Estimate $\hat{Y}_i = \bar{Y}$	**Error in Estimation** $Y_i - \bar{Y}$	**Errors Squared** $(Y_i - \bar{Y})^2$
1	390	543	-153	23460.03
2	425	543	-118	13963.36
3	560	543	17	283.36
4	575	543	32	1013.36
5	630	543	87	7540.03
6	679	543	136	18450.69
Sum	3259	3259	0	64710.83

Table 5.5: Estimating the expenditure of the household with income

Family (i)	Actual Y_i	Income X_i	Regression Estimate $\hat{Y}$	Residual $Y - \hat{Y}$	Residual squared $(Y - \hat{Y})^2$
1	390	500	394.95	-4.95	24.53
2	425	600	454.24	-29.24	854.87
3	560	700	513.52	46.48	2160.04
4	575	800	572.81	2.19	4.80
5	630	900	632.10	-2.10	4.39
6	679	1000	691.38	-12.38	153.29
Sum	3259	4500	0	0	3201.90

Table 5.6: ANOVA Table: Estimating the expenditure of the household with income

Source of variation	Sum of Square **SS**	**df**	Mean Sum of Square **MSS**
Due to regression (ESS)			
Due to residuals (RSS)			
TSS			

(After MID-TERM)

5.5 The problem of prediction

Based on our sample data, we have the following sample regression:

We can use the above regression to "Predict" or "Forecast" the future consumption expenditure Y corresponding to some given level of income X

There are two kinds of predictions which are:

[1] **Mean prediction** We will predict the conditional mean value of Y corresponding to a chosen X (i.e X_0)

[2] **Individual Prediction** We will predict an individual Y value corresponding to (i.e X_0)

Mean Prediction

We know that

$$\hat{Y_0} \sim N(E(\hat{Y_0}), \text{var}(\hat{Y_0}))$$

where

$$E(\hat{Y_0}) = \beta_1 + \beta_2 X_0$$

and

$$\text{var}(\hat{Y_0}) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right]$$

If we replace the unknown σ^2 by the unbiased estimator $\hat{\sigma}^2$ we can get

$$t = \frac{\hat{Y}_0 - (\beta_1 + \beta_2 X_0)}{se(\hat{Y}_0)}$$

which is the t distribution with n-2 df.

Therefore, we can derive the confidence interval for the true $E(Y_0|X_0)$ as following:

$$Pr[\hat{\beta}_1 + \hat{\beta}_2 X_0 - t_{\frac{\alpha}{2}} se(\hat{Y}_0) \leq \beta_1 + \beta_2 X_0 \leq \hat{\beta}_1 + \hat{\beta}_2 X_0 + t_{\frac{\alpha}{2}} se(\hat{Y}_0)] = 1 - \alpha$$

Example

Individual Prediction

We can prediction an individual Y value, Y_0 corresponding to a given X value (X_0) :

but the variance in this case is:

$$\text{var}(Y_0 - \hat{Y}_0) = E[Y_0 - \hat{Y}_0]^2 = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right]$$

We can show that Y_0 follows the normal distribution:

$$Y_0 \sim N(\hat{Y}_0, \text{var}(Y_0 - \hat{Y}_0))$$

Therefore, we can construct the confidence interval for the Y_0 as well.

From our example:

6. Extensions of The Two-Variable Linear Regression Mode

6.1 Functional Form of regression Models

We will consider the following models:

[1] The log-linear model

[2] Semilog models

[3] Reciprocal models

[4] The logarithmic reciprocal model

The Log-linear Model

The Semilog Models

The Reciprocal Models

The Logarithmic Reciprocal Models

6.2 Regression Through the Origin

In this section, we consider the case that the two-variable PRF assumes the following form:

$$Y_i = \beta_2 X_i + u_i$$

This model is called **the regression through the origin** where the intercept term $\hat{\beta}_1$ is absent from the model.

Example

Since it is the linear regression model, we can apply the Ordinary Least Square (OLS) to estimate the formula for $\hat{\beta}_2$

Let us first write the sample regression function (SRF) as:

$$Y_i = \hat{\beta}_2 X_i + \hat{u}_i$$

We would like to minimize

$$\sum \hat{u_i}^2 = \sum (Y_i - \hat{\beta}_2 X_i)^2$$

therefore,

$$\hat{\beta}_2 = \frac{\sum X_i Y_i}{\sum X_i^2}$$

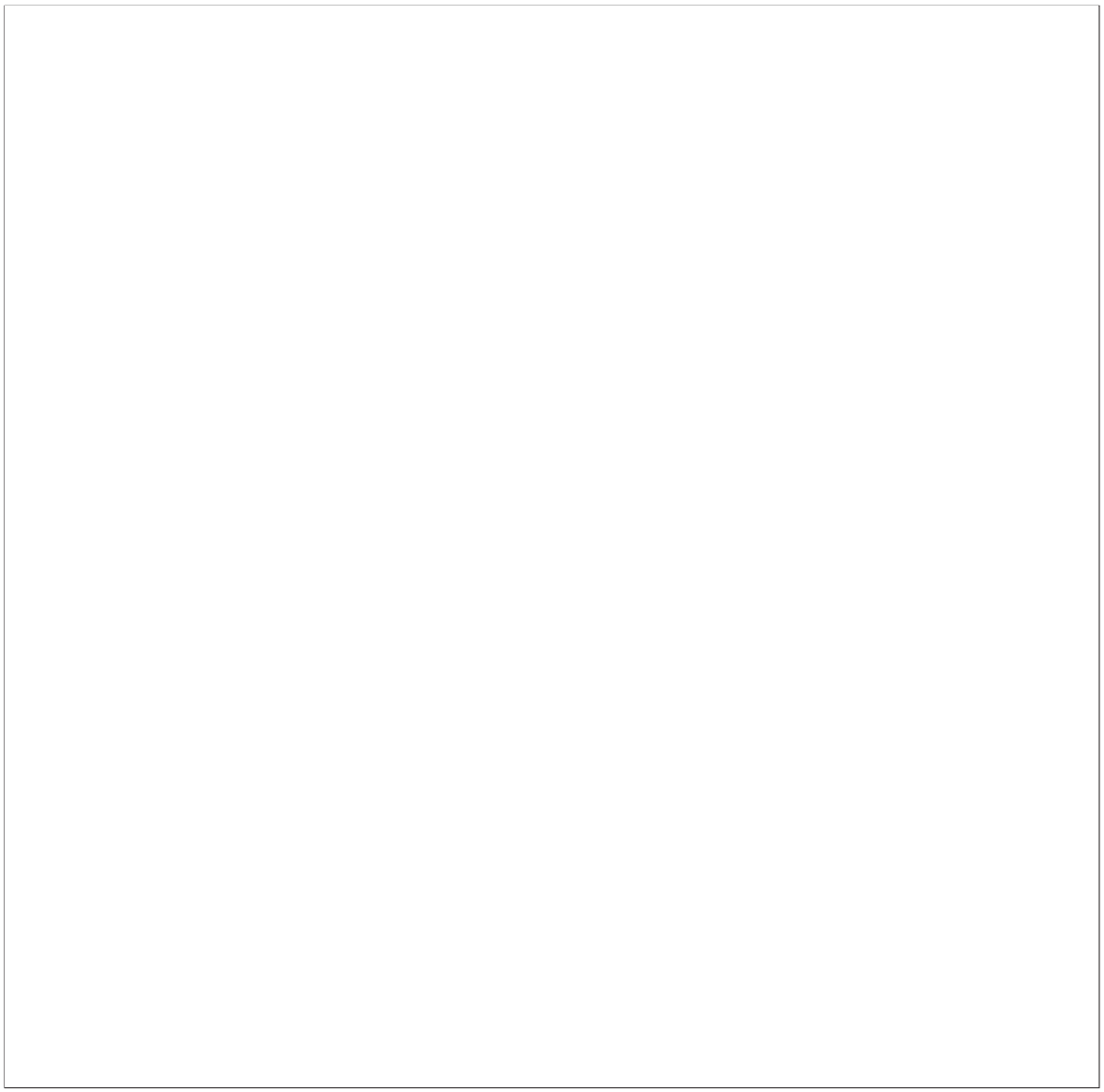
Now we can find out the variance of $\hat{\beta}_2$

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum X_i^2}$$

$$\hat{\sigma^2} = \frac{\sum \hat{u_i}^2}{n-1}$$

It should be noted that we get the condition $\sum \hat{u}_i X_i = 0$ from the normal equation. However, with the regression through the origin model, we cannot get the condition $\sum \hat{u}_i = 0$.

For the zero-intercept model, r^2 can be negative, whereas for the conventional model it cannot be negative.



Since the conventional r^2 is not appropriate for the regressions that do not contain the intercept, we therefore compute what is known as the **raw** r^2 instead:

$$\text{raw } r^2 = \frac{(\sum X_i Y_i)^2}{\sum X_i^2 \sum Y_i^2}$$

This raw r^2 has its value between 0 and 1, but we cannot directly compare its value to the conventional r^2 value. For this reason, some researchers do not report the r^2 value for zero intercept regression models.

6.2.1 Scaling and Units of Measurements

Consider our old example given in table 18 which refer to weekly family expenditure (Y) and Income (x), in baht.

Table 6.1: Weekly family Expenditure (Y), Baht and Income (X), (Unit:Baht)

X	Y
500	360
600	390
700	440
800	575
900	670
1000	730

By using the OLS estimation, we get the following results:

Now, we are interested in changing the units of our data. For example, we would prefer to express our sample data in the unit of 1000 baht. By using the new unit of X and Y, we can report our data in 1000 baht as in the following table.

Table 6.2: Weekly family Expenditure (Y), Baht and Income (X), (Unit: 1000 Baht)

X	**Y**
0.5	0.360
0.6	0.390
0.7	0.440
0.8	0.575
0.9	0.670
1	0.730

With the new unit, we would like to answer these two questions:
1. Do the units in which the regressand (Y) and regressor/s (X) are measured make any difference in the regression results?
2. If so, what is the sensible course to follow in choosing units of measurement for regression analysis?

To answer these questions, let:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

where Y is the weekly family expenditure and X is the income, in baht.

Now, let w_1 and w_2 are constants, called the **Scale factors**. For example, in our data, if we need to use the unit of 1000 baht instead, we can directly multiple the original data in table 18 with the scale factors equal to 0.001. In other words, $w_1 = w_2 = \frac{1}{1000} = 0.001$.

Define

$$Y_i^* = w_1 Y_i$$

$$X_i^* = w_1 X_i$$

Now consider the regression using Y_i^* and X_i^* variables:

$$Y_i^* = \hat{\beta}_1^* + \hat{\beta}_2^* X_i^* + \hat{u}_i{}^*$$

$\hat{u_i^*} = ?$

Our target is to find out the relationship between the following pairs:

1. $\hat{\beta}_1$ and $\hat{\beta}_1^*$

2. $\hat{\beta}_2$ and $\hat{\beta}_2^*$

3.var($\hat{\beta}_1$) and var($\hat{\beta}_1^*$)

4.var($\hat{\beta}_2$) and var($\hat{\beta}_2^*$)

5.$\hat{\sigma}^2$ and $\hat{\sigma}^{*2}$

6. r_{xy}^2 and $r_{x^*y^*}^2$

1. $\hat{\beta}_1$ and $\hat{\beta}_1^*$

2. $\hat{\beta}_2$ and $\hat{\beta}_2^*$

3.var($\hat{\beta}_1$) and var($\hat{\beta}_1^*$)

4.var($\hat{\beta}_2$) and var($\hat{\beta}_2^*$)

5.$\hat{\sigma}^2$ and $\hat{\sigma}^{*2}$

6. r^2_{xy} and $r^2_{x^*y^*}$

7. Multiple Regression Analysis: The Problem of Analysis

Three-Variable Model: Notation and Assumptions

Let us consider the following three-variable PRF as:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

where

Y_i is the dependent variable (regressand)

X_{2i} and X_{3i} are the regressors or the explanatory variables

u_i is the stochastic disturbance term

Remark: the subscript i is denoted the observation i from our sample data.

In case our data are time series, the subscript t will denote the t observation.

β_1 means the average value of Y when X_2 and X_3 are set equal to zero

β_2 and β_3 are called the partial regression coefficients.

We will talk about the meaning of β_1 and β_2 shortly after knowing the assumptions of the classical linear regression model (CLRM)

Under the CLRM, we assume:

1. Zero mean value of u_i

2. No serial correlation

3. Homoscedasticity

4. Zero covariance between u_i and each X variable, or

5. No specification bias or

The model is correctly specified.

6. No exact collinearity between the X variables or

By the above assumptions, we can find out the conditional expectation of Y_i:

The meaning of partial coefficients:

β_2

β_3

7.1 OLS Estimation of the Partial Regression Coefficients

In order to find the OLS estimators, we need to write down the sample regression function (SRF) corresponding to the PRF:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i$$

From the FOC, we then get the normal equations:

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3$$
$$\sum Y_i X_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i}$$
$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum X_{3i}^2$$

We therefore get:

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$$
$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$
$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

Variance and Standard Errors of OLS Estimators

$$var(\hat{\beta}_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3i}^2 + \bar{X}_3{}^2 \sum x_{2i}^2 - 2\bar{X}_2 \bar{X}_3 \sum x_{2i} x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} x_{3i})^2} \right] * \sigma^2$$
$$se(\hat{\beta}_1) = +\sqrt{var(\hat{\beta}_1)}$$

$$var(\hat{\beta}_2) = \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} * \sigma^2$$
$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$$
$$se(\hat{\beta}_2) = +\sqrt{var(\hat{\beta}_2)}$$

$$var(\hat{\beta}_3) = \frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} * \sigma^2$$
$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$$
$$se(\hat{\beta}_3) = +\sqrt{var(\hat{\beta}_2)}$$

$$cov(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1-r_{23}^2)\sqrt{\sum x_{2i}^2}\sqrt{\sum x_{3i}^2}}$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-3}$$

7.2 Properties of OLS Estimators

Properties of OLS Estimators (Cont:)

Properties of OLS Estimators (Cont:)

The Multiple Coefficient of Determination R^2 and the Multiple Coefficient of Correlation R

In this section, we will study how to measure the proportion of the variation in Y explained by the variables X_2 and X_3 jointly. This is the same concept of r^2 that we have learned before.

The quantity that gives this information is known as the **the multiple coefficient of determination** and is denoted by R^2.

To derive R^2, we firstly write down the following equation:

$$\begin{aligned} Y_i &= \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{u}_i \\ &= \hat{Y}_i + \hat{u}_i \end{aligned} \tag{7.1}$$

where $\hat{Y}_i$ is the estimated value of Y_i from the fitted regression line and is an estimator of true $E(Y_i|X_{2i}, X_{3i})$.

7.1 may be written as

$$\begin{aligned} y_i &= \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \\ &= \hat{y}_i + \hat{u}_i \end{aligned} \tag{7.2}$$

Squaring 7.2 on both sides and summing over the sample values, we obtain

$$\begin{aligned} \sum y_i^2 &= \sum \hat{y}_i{}^2 + \hat{u}_i{}^2 + 2\sum \hat{y}_i \hat{u}_i \\ &= \sum \hat{y}_i{}^2 + \sum \hat{u}_i{}^2 \end{aligned} \tag{7.3}$$

$$\begin{aligned} R^2 &= \frac{ESS}{TSS} \\ &= \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2} \end{aligned} \tag{7.4}$$

The three-or-more-variable analogue of r is the coefficient of multiple correlation, denoted by R, and it is a measure of the degree of association between Y and all the explanatory variables jointly. Although r can be positive or negative, R is always taken to be positive.

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} \left(\frac{1}{1 - R_j^2} \right)$$

7.2.1 R^2 and the Adjusted R^2

It should be noted that he R^2 is a nondecreasing function of the number of explanatory variables. Thus, when the number of regressors increases, R^2 almost invariably increases and never decreases. **In other words, an additional X variable will not decrease R^2!**

To explain this fact, let we write down the definition of R^2 again:

$$\begin{aligned} R^2 &= \frac{ESS}{TSS} \\ &= 1 - \frac{RSS}{TSS} \\ &= 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2} \end{aligned} \tag{7.5}$$

Therefore, in comparing two regression models **with the same dependent variable but differing number of X variables**, one should be very wary of choosing the model with the highest R^2.

In light of comparing two R^2 terms, we have to take into account the number of X variables present in the model. To achieve this goal , we can consider the alternative coefficient of determination, which is as follows:

k= the number of parameters in the model including the intercept term.
n= the number of observations in the sample data.

The above equation is known as **the adjusted** R^2, denoted by $\bar{R}^2$. The term adjusted means adjusted for the df associated with the sums of squares entering into 7.5.

We can rewrite the the adjusted R^2 as:

We can also get the equation which shows the relationship between $\bar{R}^2$ and R^2:

Besides R^2 and $\bar{R}^2$ as goodness of fit measures, other criteria are often used to judge the adequacy of a regression model. Two of these are **Akaike's Information criterion and Amemiya's Prediction criteria**, which are used to select between competing models. We will discuss these criteria in greater detail later.

8. Multiple Regression Analysis: The Problem of Inference

In this chapter, we will extend the ideas of interval estimation and hypothesis testing developed there to models involving three or more variables.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

We have already known that if our objective is to do interval estimation and hypothesis testing, we need to assume that the u_i follow the normal distribution with zero mean and constant variance σ^2

With the normality assumption and the CLRM assumptions, we know that:

[1] The OLS estimations of partial regression coefficients are best linear unbiased estimators (BLUE).

[2] The estimators $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ are normally distributed with means equal to true β_1,β_2, and β_3 and variances are following:

$$var(\hat{\beta}_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3i}^2 + \bar{X}_3{}^2 \sum x_{2i}^2 - 2\bar{X}_2\bar{X}_3 \sum x_{2i}x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i}x_{3i})^2}\right] * \sigma^2$$

$$se(\hat{\beta}_1) = +\sqrt{var(\hat{\beta}_1)}$$

$$var(\hat{\beta}_2) = \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2} * \sigma^2$$

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2(1 - r_{23}^2)}$$

$$se(\hat{\beta}_2) = +\sqrt{var(\hat{\beta}_2)}$$

$$var(\hat{\beta}_3) = \frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2} * \sigma^2$$

$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2(1 - r_{23}^2)}$$

$$se(\hat{\beta}_3) = +\sqrt{var(\hat{\beta}_3)}$$

Moreover, $\frac{(n-3)\hat{\sigma^2}}{\sigma^2}$ follows the χ^2 distribution with n-3 df. We can also show that, if we replace the true σ^2 by its unbiased estimator $\hat{\sigma^2}$ in the computation of the standard errors, we then get

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)}$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)}$$

$$t = \frac{\hat{\beta}_3 - \beta_3}{se(\hat{\beta}_3)}$$

follows the t distribution with n-3 df.

Example Consider the following regression:

$$\begin{array}{rcl} \widehat{\log(\text{salary})} & = & 4.32 + 0.280\ \log(\text{sales}) + 0.0174\ \text{ROE} + 0.00024\ \text{ROS} \\ se & = & (0.32)\ (0.035)\qquad\quad (0.0041)\qquad\quad (0.00054) \end{array} \tag{8.1}$$

$$R^2 = 0.283$$

where
salary =salary of CEO
sales=annual firm sales
ROE=return on equity in percent
ROS =return on firm's stock

Interpreter the partial regression coefficients

Questions What about the statistical significance of the observed results?

For the coefficient of log(sales) of 0.280, Is this coefficient statistically significant different from zero?

For the coefficient of ROE of 0.0174, Is this coefficient statistically significant different from zero?

For the coefficient of ROS of 0.00024, Is this coefficient statistically significant different from zero?

Are these three coefficients statistically significant?
To answer these questions, we have to learn the kinds of hypothesis testing.

8.1 Hypothesis Testing About Individual Regression Coefficients

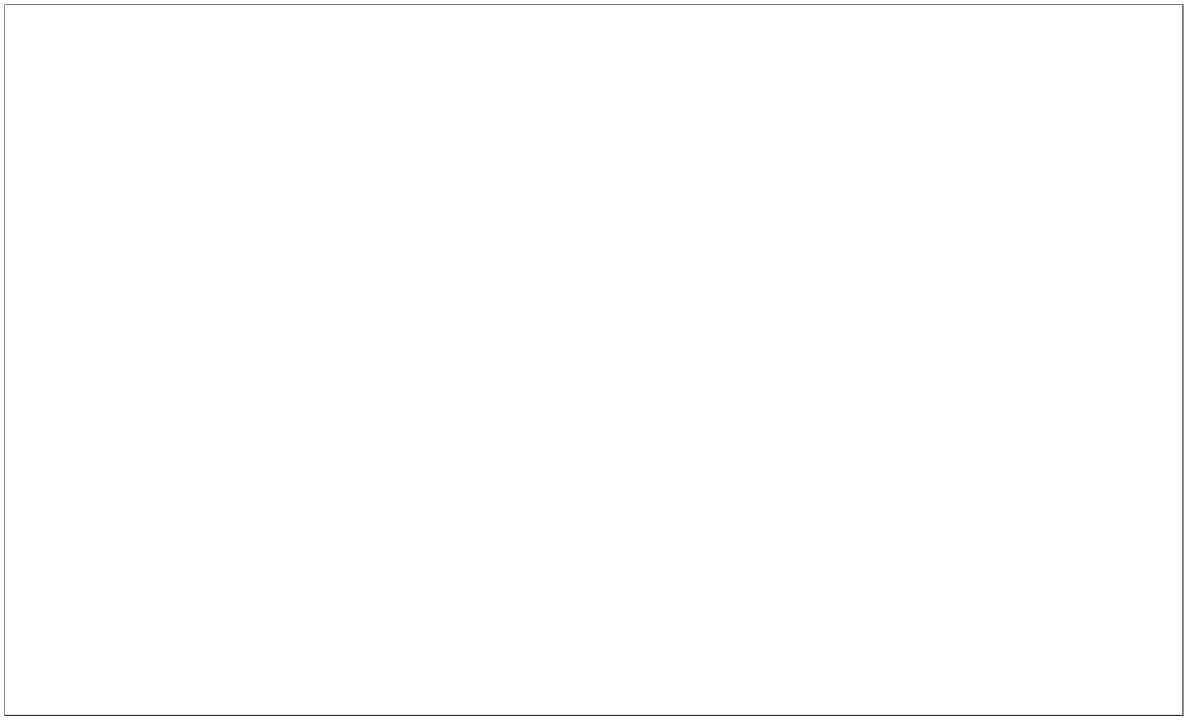
We can use the t-test to test a hypothesis about any individual partial regression coefficient.

8.1.1 Two-tail test:

Let us postulate that

$H_0: \ \beta_2 = 0$

$H_1: \ \beta_2 \neq 0$



8.1.2 One-tail test:

Let us postulate that

$H_0: \quad \beta_2 \leq 0$

$H_1: \quad \beta_2 > 0$

8.2 Testing The Overall Significance of the Sample Regression

In the previous section, we test the significance of the estimated partial regression coefficients individually, that is under the separate hypothesis that each true population partial regression coefficient was zero. But now we are interested in testing β_2 , β_3 and β_4 are jointly or simultaneously equal to zero. In other words, we would like to test the following hypothesis:

$$H_0 \;\; \beta_2 = \beta_3 = \beta_4 = 0$$

In order to reach this goal, we have to learn the following test.

The Analysis of Variance Approach to Testing the Overall Significance of an Observed Multiple Regression: The F-Test

The joint hypothesis can be tested by the **Analysis of Variance (ANOVA)** which can be demonstrated as follows:

Table 8.1: ANOVA Table for the three-variable regression model

Source of variation	Sum of Square **SS**	**df**	Mean Sum of Square **MSS**
Due to regression (ESS)			
Due to residuals (RSS)			
TSS			

Decision Rule Given the k- variable regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$$

To test the hypothesis

$$H_0 : \quad \beta_2 = \beta_3 = ... = \beta_k = 0$$

(i.e ., all slope coefficients are simultaneously zero) versus

H_1 Not all slope coefficients are simultaneously zero

If $F > F_\alpha(k-1, n-k)$, we reject H_o ; otherwise we cannot reject it, where $F_\alpha(k-1, n-k)$ is the critical F value at the α level of significance and (k-1) numerator df and (n-k) denominator df.

An important Relationship between R^2 and F

Table 8.2: ANOVA Table in Terms of R^2

Source of variation	Sum of Square **SS**	**df**	Mean Sum of Square **MSS**
Due to regression (ESS)			
Due to residuals (RSS)			
TSS			

Decision Rule Testing the overall significance of a regression in terms of R^2

Given the k- variable regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$$

To test the hypothesis

$$H_0 : \quad \beta_2 = \beta_3 = ... = \beta_k = 0$$

(i.e ., all slope coefficients are simultaneously zero) versus

H_1 Not all slope coefficients are simultaneously zero

Compute

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

If $F > F_\alpha(k-1, n-k)$, we reject H_o ; otherwise we cannot reject it, where $F_\alpha(k-1, n-k)$ is the critical F value at the α level of significance and (k-1) numerator df and (n-k) denominator df.

8.3 The "Incremental" or "Marginal" Contribution of an Explanatory Variable

Let consider the following regression:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + u_i$$

Having run the above regression, let us suppose we decide to add the additional variable, X_{3i}, to the model and obtain the multiple regression as follow:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Comparing between these two regressions, we might need to answer the below questions:

[1]. What are the marginal, or incremental, contribution of X_{3i}, knowing that X_{2i} is already in the model and that it is significantly related to Y_i.

[2]. Is the incremental contribution of X_{3i} statistically significant?

[3]. What is the criterion for adding variables to the model?

By contribution we mean whether the additional of the variable, X_{3i}, to the model increases ESS (and hence R^2) "significantly" in relation to the RSS. This contribution is called **the incremental, or marginal** contribution of an additional variable.

To assess the incremental contribution of X_3 after allowing for the contribution of X_2, we form

$$\begin{aligned} F &= \frac{Q_2/\text{df}}{Q_4/\text{df}} \\ &= \frac{(ESS_{new} - ESS_{old})/\text{number of new regressors}}{RSS_{new}/df(\text{=n-number of parameters in the new model})} \end{aligned} \tag{8.2}$$

Under the normality assumption of u_i and CLRM assumptions, this F value follows the F distribution with 1 and n-number of parameters in the new model.

Table 8.3: ANOVA Table To Assess Incremental Contribution of A Variable(s)

Source of variation	Sum of Square **SS**	**df**	Mean Sum of Square **MSS**
ESS due to X_2 alone	$Q_1 = \hat{\alpha_2}^2 \sum x_2^2$	1	$\frac{Q_1}{1}$
ESS due to the addition of X_3	$Q_2 = Q_3 - Q_1$	1	$\frac{Q_2}{1}$
ESS due to both X_2,X_3	$Q_3 = \hat{\beta}_2 \sum x_{2i} y_i + \hat{\beta}_3 \sum x_{3i} y_i$	2	$\frac{Q_3}{2}$
RSS	$Q_4 = Q_5 - Q_3$	n-3	$\frac{Q_4}{n-3}$
TSS	$Q_5 = \sum y_i^2$	n-1	

As usual method, we can re write 8.2 in term of R^2 only. Thus the F ratio of 8.2 is equivalent to the following F ratio:

$$\begin{aligned} F &= \frac{R^2_{new} - R^2_{old}/\text{df}}{(1 - R^2_{new})/\text{df}} \\ &= \frac{(R^2_{new} - R^2_{old})/\text{number of new regressors}}{1 - R^2_{new}/df(\text{=n-number of parameters in the new model})} \end{aligned} \tag{8.3}$$

This F ratio follows the F distribution with 1 and n-number of parameters in the new model.

Example

Consider the child mortality example. We considered the behavior of child mortality (CM) in relation to per capita GNP (PGNP). There we found that PGNP has a negative impact on CM, as one would expect. Now let us bring in female literacy as measured by the female literacy rate (FLR). A priori, we expect that FLR too will have a negative impact on CM. Our sample consists of 64 countries.

In model 1, we regressed child mortality (CM) on per capita GNP (PGNP) and female literacy rate (FLR).

Model 1:

$$\begin{aligned} \widehat{CM_i} &= 263.6416 - 0.0056PGNP_i - 2.2316FLR_i \\ se &= (11.5932) \quad (0.0019) \qquad (0.2099) \quad R^2 = 0.7077 \end{aligned} \tag{8.4}$$

Now we extend this model to model 2 by including total fertility rate (TFR):

Model 2:

$$\begin{aligned} \widehat{CM_i} &= 168.3067 - 0.00555GNP_i - 1.7680FLR_i + 12.8686TFR_i \\ se &= (32.8916) \quad (0.0018) \qquad (0.2480) \qquad (?) \quad R^2 = 0.7474 \end{aligned} \tag{8.5}$$

Questions

1.How would you choose between models 1 and 2? Which statistical test would you use to answer this question? Show the necessary calculations.
2. We have not given the standard error of the coefficient of TFR. Can you find it out? (Hint: Recall the relationship between the t and F distributions.

8.4 Testing the Equality of Two Regression Coefficients

Suppose we have the following model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + ... + \beta_k X_{ki} + u_i$$

We would like to test the hypotheses:

$$H_0 : \ \beta_3 = \beta_4 \ \text{ or } \ (\beta_3 - \beta_4) = 0$$
$$H_1 : \ \beta_3 \neq \beta_4 \ \text{ or } \ (\beta_3 - \beta_4) \neq 0$$

Under the classical assumptions, it can be shown that:

$$t = \frac{(\hat{\beta}_3 - \hat{\beta}_4) - (\beta_3 - \beta_4)}{se(\hat{\beta}_3 - \hat{\beta}_4)}$$

where the t follows the t distribution with (n-k) df because the above equation is a k-variable model, where k is the total number of parameters estimated, including the constant term.

The $se(\hat{\beta}_3 - \hat{\beta}_4)$ is calculated from the following formula:

$$se(\hat{\beta}_3 - \hat{\beta}_4) = \sqrt{var(\hat{\beta}_3) + var\hat{\beta}_4 - 2cov(\hat{\beta}_3, \hat{\beta}_4)}$$

Example

among other things, you were asked to consider the following demand function for chicken:

$$\begin{aligned} \widehat{\ln Y_t} &= 2.0328 + 0.4515 \ln X_{2t} - 0.3772 \ln X_{3t} \\ se &= (0.1162) \quad (0.0247) \qquad (0.0635) \quad R^2 = 0.9801 \end{aligned} \tag{8.6}$$

where Y = per capita consumption of chicken, lb, X_2 = real disposable per capita income, \$, X_3 = real retail price of chicken per lb.

Question

For the above demand function, how would you test the hypothesis that the income elasticity is equal in value but opposite in sign to the price elasticity of demand? Show the necessary calculations. [Note: $\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.00142$. and the sample data = 23 observations]

8.5 Restricted Least Squares: Testing Linear Equality Restriction

In economic theories, the coefficients in a regression model need to satisfy some linear equality restrictions, For example, in microeconomics, consider the Cobb-Douglas production function:

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{u_i}$$

where Y=output, X_2 = labor input, and X_3=capital input. We can transform the above equation to be the log form as:

$$\ln Y_i = \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i$$

where $\beta_0 = \ln \beta_1$

Now, if there are the constant returns to scale, economic theory would suggest that

$$\beta_2 + \beta_3 = 1$$

which is an example of a linear equality restriction.
In order to test the above linear equality restriction, we can follow two approaches which are:

[1]. The t-test approach

[2]. The F-test approach: Restricted Least Squares.

First Approach: The t-Test
A test of the hypothesis or restriction can be conducted by the t-test:

$$t = \frac{(\hat{\beta}_2 + \hat{\beta}_3) - (\beta_2 + \beta_3)}{se(\hat{\beta}_2 + \hat{\beta}_3)}$$

where the t follows the t distribution with (n-k) df for a k-variable model, where k is the total number of parameters estimated, including the constant term. In this case, df=n-3.

The $se(\hat{\beta}_2 + \hat{\beta}_3)$ is calculated from the following formula:

$$se(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{var(\hat{\beta}_2) + var\hat{\beta}_3 + 2cov(\hat{\beta}_2, \hat{\beta}_3)}$$

Example
Consider the Cobb-Douglas production function to the Mexican economy (1955-1974: n=20):

$$\begin{aligned} \widehat{\ln GDP_t} &= -1.6524 + 0.3397 \ln Labor_t + 0.8460 \ln Capital_t \\ t &= (-2.7259) \quad (1.8295) \qquad\quad (9.0625) \quad R^2 = 0.9951 \;\; RSS_{UR} = 0.0136 \end{aligned} \tag{8.7}$$

where GDP = Real GDP, Millions of 1960 pesos, *Labor* = Employment, Thousands of People, *Capital* = Fixed Capital, Millions of 1960 pesos.

Question
As you can see, the output/labor elasticity is about 0.34 and the output/capital elasticity is about 0.85. If we add these coefficients, we obtain 1.19, suggesting that perhaps the Mexican economy during the stated time period was experiencing increasing returns to scale. However, we do not know if 1.19 is statistically different from 1.

Therefore, we have to test this linear equality restriction.

8.6 The F-Test Approach: Restricted Least Squares

From the Cobb-Douglas production function:

$$\ln Y_i = \beta_0 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i \tag{8.8}$$

if there are the constant returns to scale, economic theory would suggest that

$$\beta_2 + \beta_3 = 1$$

We can rewrite it as:

$$\beta_2 = 1 - \beta_3$$

or

$$\beta_3 = 1 - \beta_2$$

Using either of these equalities, we can eliminate one of the β coefficients. Therefore, we can rewrite the Cobb-Douglas production function as:

$$\ln (Y_i/X_{2i}) = \beta_0 + \beta_3 \ln (X_{3i}/X_{2i}) + u_i \tag{8.9}$$

where $\frac{Y_i}{X_{2i}}$=output/labor ratio
$\frac{X_{3i}}{X_{2i}}$= capital labor ratio.

It should be noted that:
8.8 is known as **unrestricted Least Squares (URLS)**
8.9 is known as **restricted Least Squares (RLS)**

We can compare the unrestricted and restricted least-squares regressions by applying the F-test as follows:

$\sum \hat{U}_{UR}^2$=RSS of the unrestricted regression 8.8

$\sum \hat{U}_{R}^2$ = RSS of the restricted regression 8.9

m=number of linear restrictions (in this example, we have 1 restriction)

k= number of parameters in the unrestricted regression

n= number of observations

Then, we have

$$\begin{aligned} F &= \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/(n-k)} \\ &= \frac{(\sum \hat{U}_R^2 - \sum \hat{U}_{UR}^2)/m}{\sum \hat{U}_{UR}^2/(n-k)} \end{aligned} \tag{8.10}$$

follows the F-distribution with m, (n-k) df.

We can also rewrite the F-test in terms of R^2 as follows:

$$F = \frac{R_{UR}^2 - R_R^2/\text{m}}{(1-R_{UR}^2)/\text{n-k}} \tag{8.11}$$

Example
Consider the Cobb-Douglas production function to the Mexican economy(1955-1974: n=20):

$$\begin{aligned} \widehat{\ln GDP_t} &= -1.6524 + 0.3397 \ln Labor_t + 0.8460 \ln Capital_t \\ t &= (-2.7259) \quad (1.8295) \qquad (9.0625) \quad R^2 = 0.9951 \;\; RSS_{UR} = 0.0136 \end{aligned} \tag{8.12}$$

where GDP = Real GDP, Millions of 1960 pesos, *Labor* = Employment, Thousands of People, *Capital* = Fixed Capital, Millions of 1960 pesos.

The restriction of constant return to scale, which gives the following regression:

$$\begin{aligned} \widehat{\ln (GDP/Labor)_t} &= -0.4947 + 1.0153 \ln (Capital/Labor)_t \\ t &= (-4.0612) \quad (28.1056) \quad R_R^2 = 0.9777 \;\; RSS_R = 0.0166 \end{aligned} \tag{8.13}$$

8.7 Testing for Structural or Parameter Stability of Regression Models: The Chow Test

Sometime when we estimate the regression model, it may happen that there is a **Structural Change** in the relationship between the regressand Y and the regressors X's, especially the model involving time series data. The structural change may be due to the external forces (i.e the financial crisis of 2007-2008) or due to policy changes (such as the switch from a fixed exchange rate system to a flexible exchange rate system in 1997).

The question is **"How do we figure out that there is a structural change in our sample data?"**

To answer this question, consider the following example.

Based on the sample data, we found out that in 1982 the United State suffers its worst peacetime regression. This event might disturb the relationship between savings and DPI.

To see this effect, we can divide our sample data into two time periods: 1970-1981 (Pre-1982 crisis) and 1982-1995 (Post-1982 crisis).

Therefore we have three possible regressions:

Time period 1970-1981: $Y_t = \beta_1 + \beta_2 X_t + u_1 t$ where $n_1 = 12$

Time period 1982-1995: $Y_t = \gamma_1 + \gamma_2 X_t + u_2 t$ where $n_2 = 14$

Time period 1970-1995: $Y_t = \alpha_1 + \alpha_2 X_t + u_t$ where $n = n_1 + n_2 = 26$

For our sample data, we can get the following results:

Time period 1970-1981:

$$\begin{aligned} \hat{Y}_t &= 1.0161 + 0.0803 X_t \\ t &= (0.00873) \quad (9.6015) \end{aligned} \tag{8.14}$$

$R^2 = 0.9021 \quad RSS_1 = 1785.032 \quad df = 10$

Time period 1982-1995:

$$\begin{aligned} \hat{Y}_t &= 153.4947 + 0.0148 X_t \\ t &= (4.6922) \quad (1.7707) \end{aligned} \tag{8.15}$$

$R^2 = 0.2971 \quad RSS_2 = 10,005.22 \quad df = 12$

Time period 1970-1995:

$$\begin{aligned} \hat{Y}_t &= 62.4226 + 0.0376 X_t \\ t &= (4.8917) \quad (8.8937) \end{aligned} \tag{8.16}$$

$R^2 = 0.7672 \quad RSS_3 = 23,248.30 \quad df = 24$

We can apply **the Chow test** to investigate the structural changes that may be caused by differences in the intercept or the slope coefficient or both.

The chow test assumes that:

[1] $u_{1t} \sim N(0, \sigma^2)$ and $u_{2t} \sim N(0, \sigma^2)$

[2] The two error terms u_{1t} and u_{2t} are independently distributed.

Chow Test

$$H_0: \text{ There is no structural change in the model}$$

$$H_1: \text{ There is structural change in the model}$$

Then, we need to construct the F-ratio:

$$F = \frac{(RSS_R - RSS_{UR})/k}{RSS_{UR}/(n_1 + n_2 - 2k)} \tag{8.17}$$

where the F ratio follows the F distribution with k and $(n_1 + n_2 - 2k)$ df in the numerator and denominator, respectively.

We do not reject the null hypothesis of parameter stability (i.e no structural change) if the computed F value does not exceed the critical value F value obtained from the F table.