BUS 173
Applied Statistics

Lecture 10-14

Prepared By:
Mohammad Kamrul Arefin
Lecturer, School of Business, North South University

Measures of Relationships Between Variables

✓Covariance: Covariance is a measure of the linear relationship between two variables. A positive value indicates a direct or increasing linear relationship and a negative value indicates a decreasing linear relationship.

A sample covariance is

$$Cov(x,y) = S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{1}{n-1}[\sum xy - \frac{\sum x \sum y}{n}]$$

✓**Correlation**: If the change in one variable effects a change in the other variable, the variables are said to be correlated.

✓If the increase (decrease) in one variable results in the corresponding increase in the other i.e. if the changes are in the same direction, the variables are positively correlated. e.g. Height and weight of a group of people.

✓If the increase (decrease) in one variable results in the corresponding decrease (increase) in the others, i.e. if the changes are in the opposite direction, the variables are negatively correlated. e.g. Volume and pressure of perfect gas.

$$\text{Correlation Coefficient } (x, y) = r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$$= \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\{\sum x_i^2 - \frac{(\sum x_i)^2}{n}\}\{\sum y_i^2 - \frac{(\sum y_i)^2}{n}\}}}$$

✓**Correlation**:

✓The correlation coefficient ranges from -1 to +1.

✓When r =0 there is no linear relationship between x and y but not necessarily a lack of relationship.

✓The closer "r" is to +1, represents strong positive relationship.

✓The closer "r" is to -1, represents strong negative relationship.

✓Correlation indicates whether there is any relation between the variables and correlation coefficient measures the extent of relationship between them.

✓**Regression**: Regression measures the probable movement of one variable in term of the other. Therefore regression is used for prediction or forecasting purpose.

✓Suppose the movement of the variable Y is dependent on the movement of X variable. Hence Y is dependent variable and X is independent variable. Let the regression line of Y on X be

$$\hat{y} = \beta_1 + \beta_2 x + u_i \text{ where } \beta_1 = \text{intercept; } \beta_2 = \text{slope}$$

$$\beta_2 = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{S_{xy}}{S_{xx}} \qquad \beta_1 = \overline{Y} - b\overline{X}$$

Example

The table shows the math achievement test scores for a random sample of n = 10 college freshmen, along with their final calculus grades.

Student	1	2	3	4	5	6	7	8	9	10
Math test, x	39	43	21	64	57	47	28	75	34	52
Calculus grade, y	65	78	52	82	92	89	73	98	56	75

Use your calculator to find the sums and sums of squares.

$$\sum x = 460 \qquad \sum y = 760$$

$$\sum x^2 = 23634 \quad \sum y^2 = 59816$$

$$\sum xy = 36854$$

$$\bar{x} = 46 \qquad \bar{y} = 76$$

Example

$$S_{xx} = 23634 - \frac{(460)^2}{10} = 2474$$

$$S_{yy} = 59816 - \frac{(760)^2}{10} = 2056$$

$$S_{xy} = 36854 - \frac{(460)(760)}{10} = 1894$$

$$b = \frac{1894}{2474} = .76556 \quad \text{and} \quad a = 76 - .76556(46) = 40.78$$

$$\text{Bestfitting line:} \quad \hat{y} = 40.78 + .77x$$

Goodness of fit:

- The overall goodness of fit of the regression is measured by the coefficient of determination, r^2.
- It explains what proportion of variation in the dependent variable is explained by the explanatory variable.
- $0 \le r^2 \le 1$: the closer it is to 1, the better is the fit.
- e.g. if $r^2 = 0.92$, it means that 92% of variation in Y is explained by X.
- In the case of multivariate regression, the coefficient of determination is denoted by R^2.

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} = \frac{s_{xy}^2}{s_{xx}s_{yy}},$$

$$\mathrm{y_i} = \text{actual value, } \hat{\mathrm{y}}_\mathrm{i} = \text{predicted value}$$

The Analysis of Variance

TSS= Total Sum of Squares = $\sum (y_i - \bar{y})^2 = S_{yy}$ =SSR+SSE

SSR= Sum of Squares of Regression= $\sum (\hat{y}_i - \bar{y})^2 = \frac{s_{xy}^2}{s_{xx}}$

=Variation in y explained by regression

SSE= Sum of Squares of Error $= \sum \hat{u}_i^2 = \sum (y_i - \hat{y}_i)^2,$

$= S_{yy} - \frac{s_{xy}^2}{s_{xx}}$ = unexplained variation in y

y_i = actual value, $\hat{y}_i$ = predicted value

The Analysis of Variance

We calculate

$$\text{SSR} = \frac{(S_{xy})^2}{S_{xx}} = \frac{1894^2}{2474}$$

$$= 1449.9741$$

$$\text{SSE= Total SS- SSR}$$

$$= S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

$$= 2056 - 1449.9741$$

$$= 606.0259$$

The ANOVA Table

Total df = **n -1**

Regression df = **K=1**

Error df = **n –k – 1 = n - 2**

Mean Squares

MSR = SSR/(1)

MSE = SSE/(n-2)

Source	df	SS	MS	F
Regression	K=1	SSR	SSR/(1)	MSR/MSE 1/(n-2) df
Error	(n -k-1)=n-2	SSE	SSE/(n-2)	
Total	n -1	Total SS		

The Calculus Problem

$$SSR = \frac{(S_{xy})^2}{S_{xx}} = \frac{1894^2}{2474} = 1449.9741$$

$$SSE = \text{Total SS} - SSR = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

$$= 2056 - 1449.9741 = 606.0259$$

Source	df	SS	MS	F
Regression	1	1449.9741	1449.9741	19.14
Error	8	606.0259	75.7532	
Total	9	2056.0000		

Estimation and Prediction

To estimate the average value of y when $x = x_0$:

$$\hat{y} \pm t_{\alpha/2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

To predict a particular value of y when $x = x_0$:

$$\hat{y} \pm t_{\alpha/2}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

The Calculus Problem

- Estimate the average calculus grade for students whose achievement score is 50 with a 95% confidence interval.

Calculate $\hat{y} = 40.78424 + .76556(50) = 79.06$

$$\hat{y} \pm 2.306\sqrt{75.7532\left(\frac{1}{10} + \frac{(50-46)^2}{2474}\right)}$$

79.06 ± 6.55 or 72.51 to 85.61.

The Calculus Problem

- Estimate the calculus grade for **a particular student** whose achievement score is 50 with a 95% confidence interval.

$$\text{Calculate } \hat{y} = 40.78424 + .76556(50) = 79.06$$

$$\hat{y} \pm 2.306\sqrt{75.7532\left(1 + \frac{1}{10} + \frac{(50-46)^2}{2474}\right)}$$

$$79.06 \pm 21.11 \quad \text{or } 57.95 \text{ to } 100.17.$$

Notice how much wider this interval is!

Minitab Output

Confidence and prediction intervals when $x = 50$

```
Predicted Values for New Observations
New Obs    Fit      SE Fit    95.0% CI           95.0% PI
1         79.06       2.84   (72.51, 85.61)  (57.95,100.17)

Values of Predictors for New Observations
New Obs            x
1               50.0
```

Two variable regression: Interval estimation and Hypothesis Testing

To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_2 is zero. Two tests are commonly used. Both require an estimate of σ^2, the variance of u_i in the regression model.

Suppose that an OLS regression of consumption (Y_i) against a constant and income (X_i):
$Y_i = \beta_1 + \beta_2 X_{2i} + u_i$ yields the sample regres. line:

$$\hat{Y}_i = 24.45 + 0.5091 X_i$$

24.25, 0.5091 are single (point) estimates of the unknown β_1, β_2, respectively.

Two variable regression: Interval estimation and Hypothesis Testing

The simple linear regression model is

$$\hat{y} = \beta_1 + \beta_2 x + u_i \text{ where } \beta_1 = \text{intercept; } \beta_2 = \text{slope}$$

If *x* and *y* are linearly related, we must have $\beta_2 \neq 0$. The purpose of the *t* test is to see whether we can conclude that $\beta_2 \neq 0$. We will use the sample data to test the following hypotheses about the parameter $\beta_2 \neq 0$.

$$H_0 : \beta_2 = 0$$
$$H_A : \beta_2 \neq 0$$

If H_0 is rejected, we will conclude that $\beta_2 \neq 0$ and that a statistically significant relationship exists between the two variables. However, if H_0 cannot be rejected, we will have insufficient evidence to conclude that a significant relationship exists.

Two variable regression: Interval estimation and Hypothesis Testing

$$T_{statistic,\ \text{(n-k-1)d.f}} = \frac{\hat{\beta}_2 - \beta_{2_0}}{\sigma_{\hat{\beta}_2}} = \frac{\hat{\beta}_2 - \beta_{2_0}}{\sqrt{\frac{MSE}{S_{xx}}}}$$

K= no. of explanatory variables in the model

Confidence Interval for β_2

$$= \hat{\beta}_2 \pm t_{\alpha/2}(\sigma_{\hat{\beta}_2}) = \hat{\beta}_2 \pm t_{\alpha/2}\left(\sqrt{\frac{MSE}{S_{xx}}}\right)$$

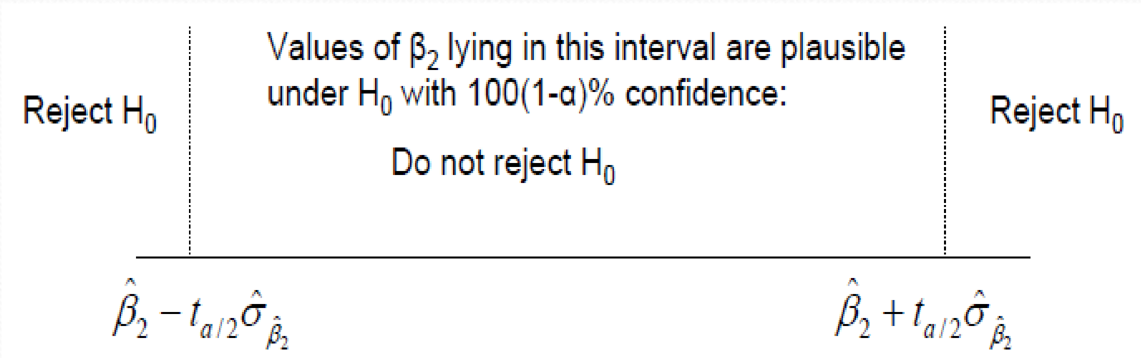
Exercise: Confidence Interval Calculation

If $\hat{\beta}_2$=0.5091, $\sigma_{\hat{\beta}_2}$ =0.0357, degrees of freedom=8, α=5%

Confidence Interval for β_2

$$= \hat{\beta}_2 \pm t_{\alpha/2}(\sigma_{\hat{\beta}_2}) = 0.5091 \pm 2.306x0.0357 = 0.4268 \text{ to } 0.5914$$

A null hypothesis that is commonly tested is H_0: β_2=0 , i.e. that the slope coefficient is zero, indicating no relationship between X and Y.



95% Confidence Interval for $\beta_2 = 0.4268$ to 0.5914

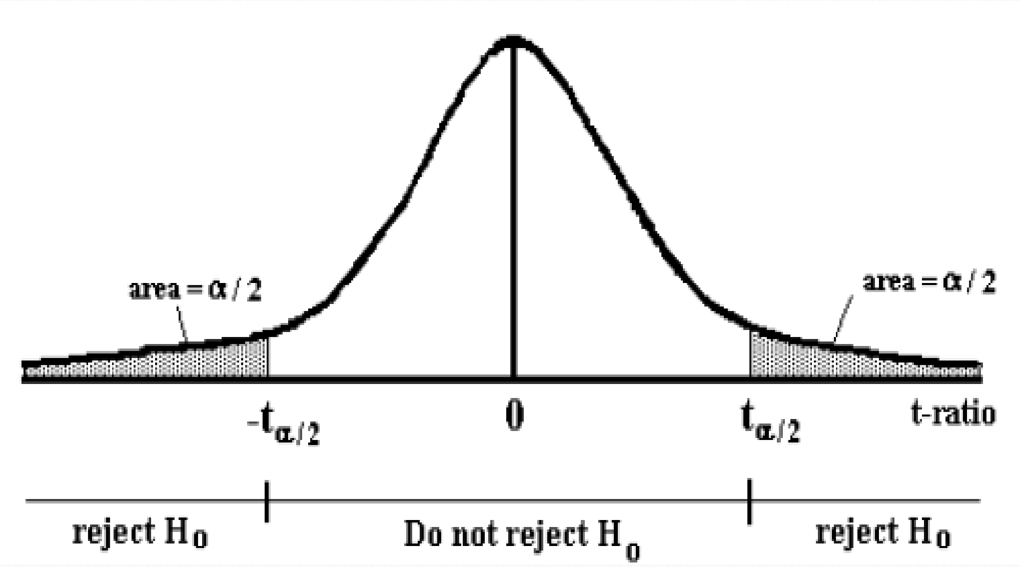
Example: If H_0: $\beta_2=0.3$ against H_1: $\beta_2 \neq 0.3$

We can reject null with 95% confidence, since 0.3 (i.e. β_2 under the null) lies outside the 95% confidence interval.

t-test decision rules:

Type of hypothesis	H_0	H_1	Reject H_0 if
Two-tail	$\beta_2 = \beta_2^*$	$\beta_2 \neq \beta_2^*$	$\lvert t \rvert > t_{(n-k),\alpha/2}$
Right-tail	$\beta_2 \leq \beta_2^*$	$\beta_2 > \beta_2^*$	$t > t_{(n-k),\alpha}$
Left-tail	$\beta_2 \geq \beta_2^*$	$\beta_2 < \beta_2^*$	$t < -t_{(n-k),\alpha}$

Two-tailed Test



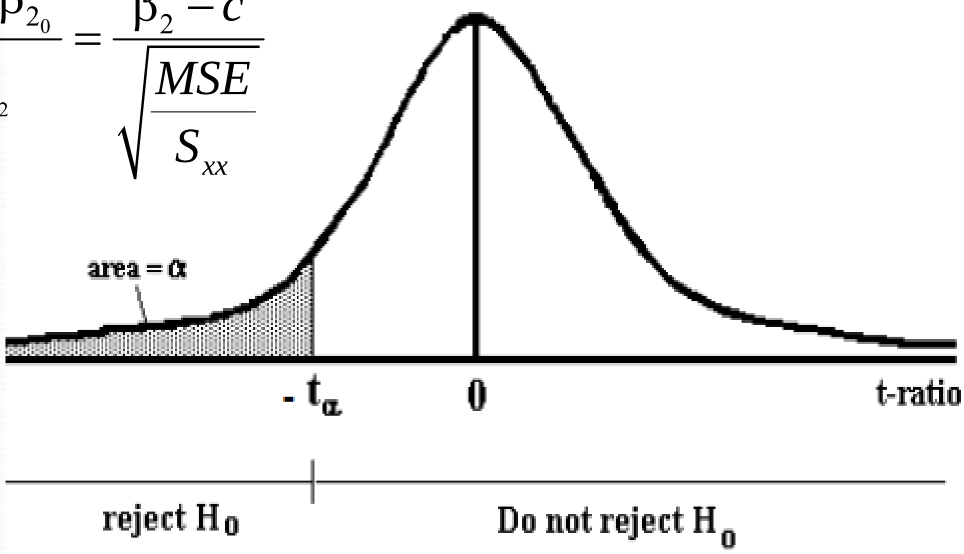
Reject the null hypothesis only if the estimated t-statistic falls in one of the rejection regions (critical regions) depicted above as shaded area

One-tailed (left) Test

$$T_{statistic,\ \text{(n-k-1)d.f}} = \frac{\hat{\beta}_2 - \beta_{2_0}}{\sigma_{\hat{\beta}_2}} = \frac{\hat{\beta}_2 - c}{\sqrt{\dfrac{MSE}{S_{xx}}}}$$

$$H_0 = \beta_2 \geq c$$

$$H_A = \beta_2 < c$$



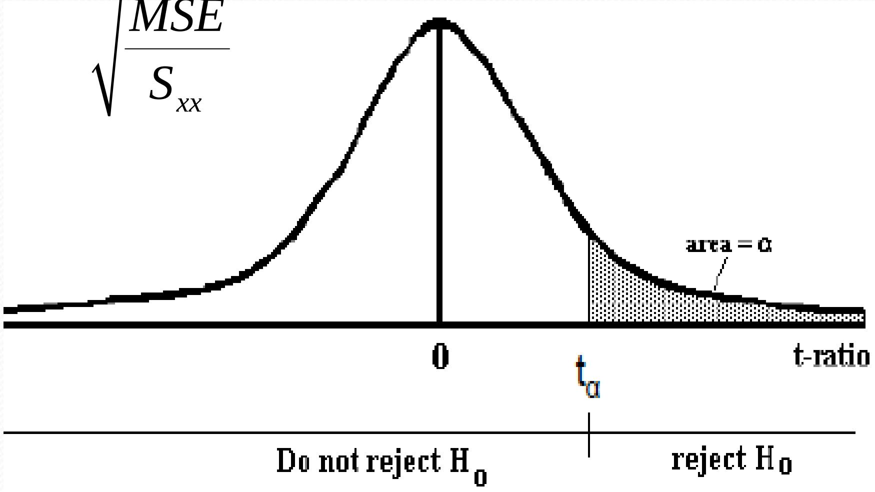
Reject the null hypothesis only if the estimated t-statistic falls in the rejection region (critical region) depicted above as shaded area

One-tailed (Right) Test

$$T_{statistic,\ \text{(n-k-1)d.f}} = \frac{\hat{\beta}_2 - \beta_{2_0}}{\sigma_{\hat{\beta}_2}} = \frac{\hat{\beta}_2 - c}{\sqrt{\dfrac{MSE}{S_{xx}}}}$$

$$H_0 = \beta_2 \leq c$$

$$H_A = \beta_2 > c$$



Reject the null hypothesis only if the estimated t-statistic falls in the rejection region (critical region) depicted above as shaded area

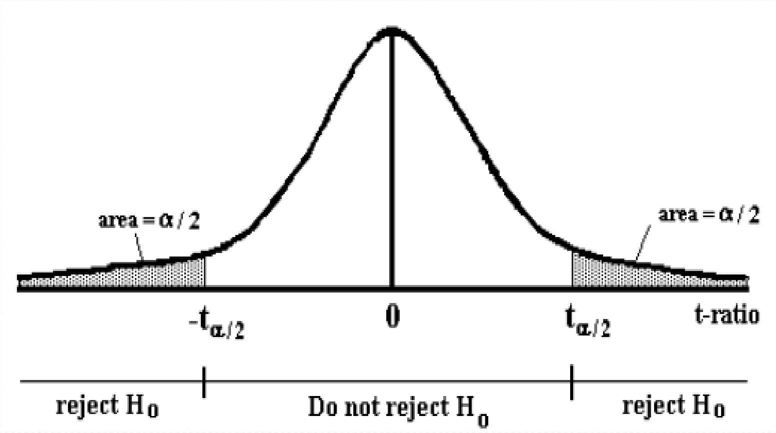
Example:

If $\hat{\beta}_2$=0.5091, $\sigma_{\hat{\beta}_2}$=0.0357, degrees of freedom=8, α=5%

$H_0 : \beta_2 = 0.3$; $H_A : \beta_2 \neq 0.3$

$$T_{statistic,\ (\text{n-k-1})\text{d.f}} = \frac{\hat{\beta}_2 - \beta_{2_0}}{\sigma_{\hat{\beta}_2}}$$

$$= \frac{0.5091 - 0.3}{0.0357} = 5.86$$



Critical value, $\text{t}_{(n-k-1)d.f\,;\,\alpha/2} = \text{t}_{8d.f\,;\,0.025} = 2.306$

Since $|t| = |5.86| = 5.86 > \text{t}_{8d.f\,;\,0.025} = 2.306$

Reject H_0

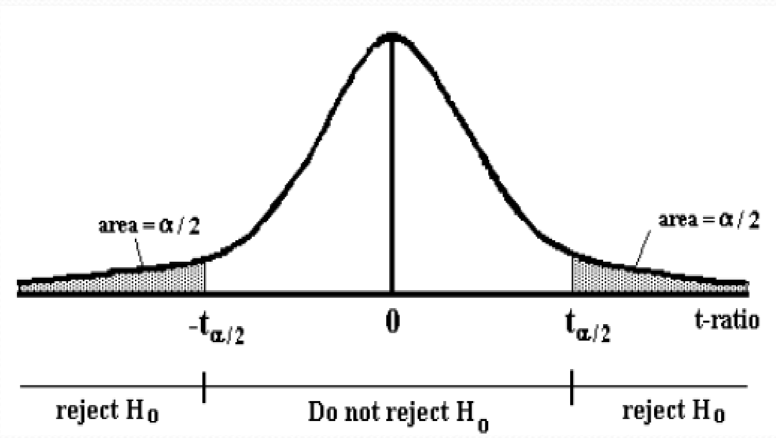
Example:

If $\hat{\beta}_2$=0.5091, $\sigma_{\hat{\beta}_2}$=0.0357, degrees of freedom=8, α=5%

$H_0 : \beta_2 = 0.3$; $H_A : \beta_2 \neq 0.3$

$$T_{statistic,\ (\text{n-k-1})\text{d.f}} = \frac{\hat{\beta}_2 - \beta_{2_0}}{\sigma_{\hat{\beta}_2}}$$

$$= \frac{0.5091 - 0.3}{0.0357} = 5.86$$



Critical value, $\text{t}_{(n-k-1)d.f\,;\,\alpha/2} = \text{t}_{8d.f\,;\,0.025} = 2.306$

Since $|t| = |5.86| = 5.86 > \text{t}_{8d.f\,;\,0.025} = 2.306$

Reject H_0

The Calculus Problem

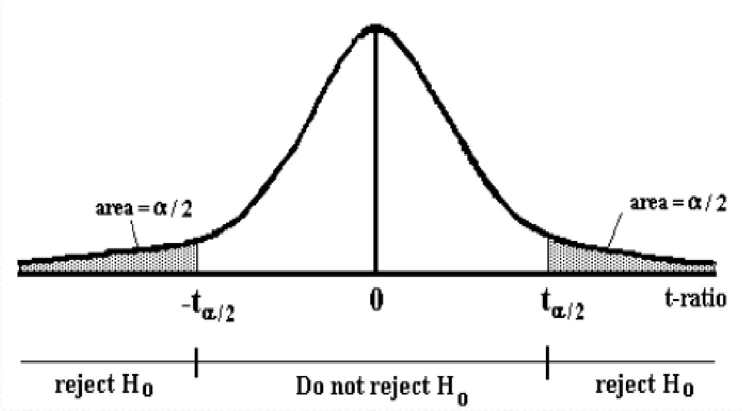
Is there a significant relationship between the calculus grades and the test scores at the 5% level of significance?

If $\hat{\beta}_2$=0.77, $\sigma_{\hat{\beta}_2} = \sqrt{\dfrac{MSE}{S_{xx}}} = \sqrt{\dfrac{75.7532}{2474}} = 0.1749$, degrees of freedom=8, α=5%

$$H_0 : \beta_2 = 0 \; ; \; H_A : \beta_2 \neq 0$$

$$T_{statistic,\ (n\text{-}k\text{-}1)d.f} = \frac{\hat{\beta}_2 - \beta_{2_0}}{\sigma_{\hat{\beta}_2}}$$

$$= \frac{0.77 - 0}{0.1749} = 4.40$$



$$\text{Critical value, } t_{(n-k-1)d.f\,;\ \alpha/2} = t_{8d.f\,;\ 0.025} = 2.306$$

$$\text{Since } |t| = |4.40| = 4.40 > t_{8d.f\,;\ 0.025} = 2.306; \text{ Reject } H_0$$

This evidence is sufficient to conclude that a significant relationship exists between student population and quarterly sales.

Example

The table shows student population and quarterly sales data for 10 armand's pizza parlours.

Pizza Parlours	1	2	3	4	5	6	7	8	9	10
St. Population (,000) (x)	2	6	8	8	12	16	20	20	22	26
Quarterly Sales, (,000) (y)	58	105	88	118	117	137	157	169	149	202

Use your calculator to find the sums and sums of squares.

$$\sum x = 140; \sum x^2 = 2528; \sum xy = 21040$$

$$\sum y = 1300; \sum y^2 = 184730; \bar{x} = 14; \bar{y} = 130$$

$$\text{Standard deviation of X } (s_x) = 7.944$$

$$\text{Standard deviation of Y } (s_y) = 41.806$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 2528 - \frac{(140)^2}{10} = 568;$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 184730 - \frac{(1300)^2}{10} = 15730;$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 21040 - \frac{140x1300}{10} = 2840;$$

$$\hat{\beta}_2 = b = \frac{S_{xy}}{S_{xx}} = \frac{2840}{568} = 5, \ \hat{\beta}_1 = \overline{y} - \beta_2 \overline{x} = 130 - 5x14 = 60$$

The regression model is: $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x = 60 + 5x$

Sum of squares of regression (SSR) = $\frac{S_{xy}^2}{S_{xx}} = \frac{2840^2}{568} = 14200$

Total Sum of Squares (TSS) = $S_{yy} = 15730$

$\text{TSS} = \text{SSR} + \text{SSE}$

Sum of Squares of Error (SSE) = $\text{TSS} - \text{SSR} = 15730 - 14200 = 1530$

The ANOVA Table

Total df = **n -1**

Regression df = **K=1**

Error df = **n –k – 1 = n - 2**

Mean Squares

MSR = SSR/(1)

MSE = SSE/(n-2)

Source	df	SS	MS	F
Reg.	K=1	SSR=14,200	SSR/(1)=14,200	MSR/MSE =14,200/191.25=74.248 1/(n-2) df=1/8 df
Error	$(n -k-1)=8$	SSE= 1530	SSE/(8)=191.25	
Total	n -1=9	TSS=15,730		

$$\text{Sum of squares of regression (SSR)} = \frac{S_{xy}^2}{S_{xx}} = \frac{2840^2}{568} = 14200$$

$$\text{Total Sum of Squares (TSS)} = S_{yy} = 15730$$

$$r^2 = \frac{SSR}{TSS} x100\% = \frac{14200}{15730} x100\% = 90.27\%$$

90.27% of the variability in sales can be explained by the linear relationship between the size of the student population and sales.

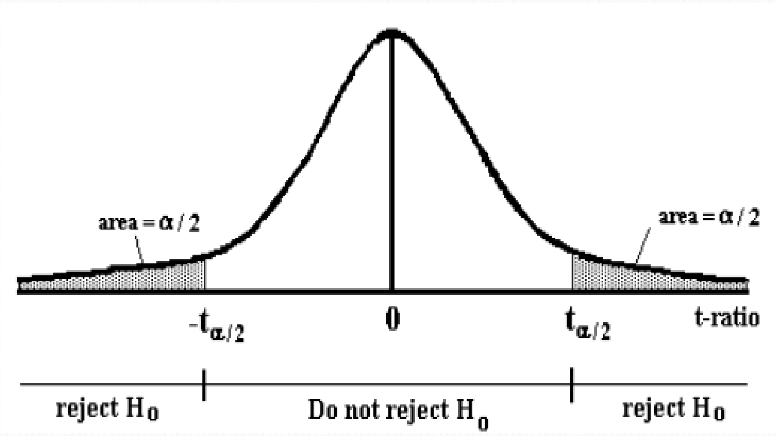
Example:

$$\text{If } \hat{\beta}_2 = 5,\ \sigma_{\hat{\beta}_2} = \sqrt{\frac{MSE}{S_{xx}}} = \sqrt{\frac{191.25}{568}} = 0.58,\ \text{degrees of freedom}=8,\ \alpha=5\%$$

$$H_0 : \beta_2 = 0\ ;\ H_A : \beta_2 \neq 0$$

$$T_{statistic,\ \text{(n-k-1)d.f}} = \frac{\hat{\beta}_2 - \beta_{2_0}}{\sigma_{\hat{\beta}_2}}$$

$$= \frac{5-0}{0.58} = 8.62$$



$$\text{Critical value, } t_{(n-k-1)d.f;\ \alpha/2} = t_{8d.f;\ 0.025} = 2.306$$

$$\text{Since } |t| = |8.62| = 8.62 > t_{8d.f;\ 0.025} = 2.306;\ \text{Reject } H_0$$

This evidence is sufficient to conclude that a significant relationship exists between student population and quarterly sales.

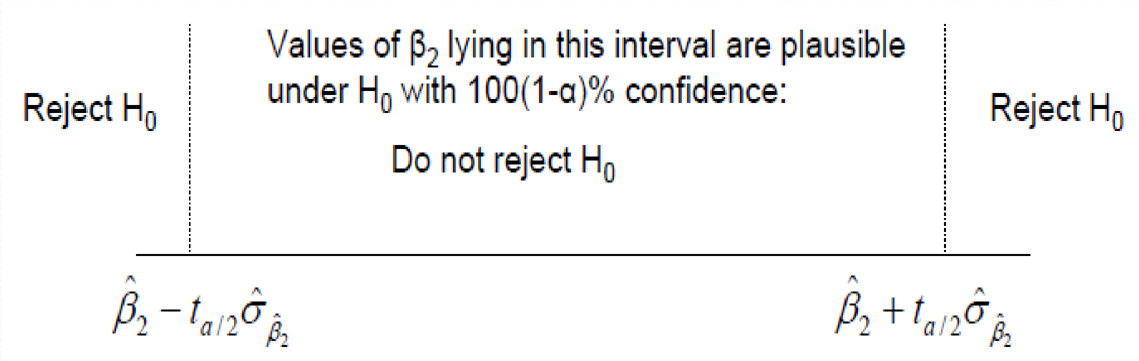
Exercise: Confidence Interval Calculation

If $\hat{\beta}_2$=5, $\sigma_{\hat{\beta}_2}$ =0.58, degrees of freedom=8, α=5%

Confidence Interval for β_2

$$=\hat{\beta}_2 \pm t_{\alpha/2}(\sigma_{\hat{\beta}_2}) = 5 \pm 2.306x0.58 = 3.66 \text{ to } 6.33$$

A null hypothesis that is commonly tested is H_0: β_2=0 , i.e. that the slope coefficient is zero, indicating no relationship between X and Y.



95% Confidence Interval for $\beta_2 = 3.66$ to 6.33

Example: If H_0: β_2=0 against H_1: $\beta_2 \neq 0$

We can reject null with 95% confidence, since 0 (i.e. β_2 under the null) lies outside the 95% confidence interval. We can conclude that a significant statistical relationship exists between the size of the student population and quarterly sales.

F-test:

An *F* test, based on the *F* probability distribution, can also be used to test for significance in regression. With only one independent variable, the *F* test will provide the same conclusion as the *t* test; that is, if the *t* test indicates $\beta \neq 0$ and hence a significant relationship, the *F* test will also indicate a significant relationship. But with more than one independent variable, only the *F* test can be used to test for an overall significant relationship.

95% Confidence Interval for $\beta_2 = 3.66$ to 6.33

Example: If H_0: β_2=0 against H_1: $\beta_2 \neq 0$

F-statistic= MSR/MSE =14,200/191.25=74.248; 1/(n-2) df=1/8 df

If F-statistic> F_α [k/(n-k-1) df: Reject the Null H_0

Here, α =5% = 0.05; $F_{0.5}$, 1/8 df =5.32

i.e. F-statistic=74.248 > $F_{0.5}$, 1/8 df =5.32

Reject the null. Therefore from the F-test also we can conclude that there is a statistically significant relationship exists between size of the student population and quarterly sales or the regression model is statistically significant at 5% level.

FIGURE 14.10 MINITAB OUTPUT FOR THE ARMAND'S PIZZA PARLORS PROBLEM

```
The regression equation is
Sales = 60.0 + 5.00 Pop
```

Estimated regression equation

Predictor	Coef	SE Coef	T	p
Constant	60.000	9.226	6.50	0.000
Pop	5.0000	0.5803	8.62	0.000

S = 13.8293   R-sq = 90.3%   R-sq(adj) = 89.1%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	14200	14200	74.25	0.000
Residual Error	8	1530	191		
Total	9	15730			

ANOVA table

Multiple Regression Model:

- The two-variable regression model is often inadequate in practise: e.g. consumption is affected not only by income but also by wealth.
- The two-variable model needs to be extended by adding more explanatory variables. The simplest multiple regression model is the three-variable regression (with two explanatory variables X_2, X_3):

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

  Y: dependent variable, u: stochastic error term
- β1: intercept term, it shows the average effect on Y of all the variables excluded from the model (or the average value of Y when: X2i = X3i = 0).
- β2 , β3: partial regression (or slope) coefficients

Multiple Regression Model:

- β2 = ∂Y/ ∂X2: measures the change in the mean value of Y per unit change of X2, holding X3 constant (i.e. the 'direct' effect of ΔX2 on E(Y), net of any influence that X3 may have).
- β3 = ∂Y/ ∂X3: measures the change in the mean value of Y per unit change of X3, holding X2 constant (i.e. the 'direct' effect of ΔX3 on E(Y), net of any influence that X2 may have).

  Goodness of fit: the multiple coefficient of determination R^2 and the adjusted R^2
- The overall goodness of fit of the regression is measured by the R^2.
- It explains what proportion of variation in the dependent variable (Y) is explained by the explanatory variables (X_2 and X_3) *jointly*: there is little point in trying to allocate the R^2 value to its constituent regressors.
- $0 \leq R^2 \leq 1$, If $R^2=1$, the fitted regression line explains 100% of the variation in Y.
- If $R^2=0$, the model doesn't explain any of the variation in Y.
- Typically, R^2 lies between these extreme values.
- The fit of the model is said to be 'better' the closer is R^2 to 1.

Goodness of fit: the multiple coefficient of determination R^2 and the adjusted R^2

- An important property of R^2 is that it is a non-decreasing function of the number of explanatory variables present in the model.
- As the number of regressors increases, R^2 invariably increases and never decreases (i.e. an additional X variable will never decrease R^2). This is due to the fact that as the number of X variables increases, the SSE is likely to decrease.

- **Implication**: when comparing two regression models with the same dependent variable (Y) but differing number of X variables, one should be very suspicious of choosing the model with the highest R^2.
- To compare two R^2 terms, we must take into account the number of variables present in the model. This can be done by using the adjusted R^2:

$$\text{Adjusted R}^2 = 1-(1-\text{R}^2)\frac{n-1}{n-k-1}$$

- where k is the number of parameters in the model (including the intercept term), n is the sample size.

- The term adjusted means adjusted for the degrees of freedom associated with the sum of squares terms.

- For k>1: Adjusted $R^2 < R^2$, implying that as the number of X variables increases the adjusted R^2 increases by less than the un-adjasted R^2.

- The adjusted R^2 can be negative, although R^2 is necessarily non-negative.

- It is good practice to use the adjusted R^2 instead of the R^2 because R^2 tends to give an overly optimistic picture of the fit of the regression, particularly when the number of explanatory variables is not very small compared with the number of observations.

- When comparing two models on the basis of R^2 (whether adjusted or not) the sample size and the dependent variable must be the same.

- Sometimes researchers choose among alternative models solely on the basis of maximizing the adjusted R^2. This can be dangerous , since it's not unusual to obtain a very high adjusted R^2 but find that some of the regression coefficients are either statistically insignificant or have signs that are contrary to a priori expectations (e.g. negative income coefficient in consumption model!).

- Also, sometimes we can obtain low adjusted R^2 without meaning that the model is necessarily bad (e.g. in stock returns regressions adjusted R^2 can be less than 0.1).

Example: Child Mortality Regression Model

We use OLS to regress child mortality (C) on per capita GNP (PGNP) and the female literacy rate (FLR) for a sample of 64 countries (n=64).

$C_i = \beta_1 + \beta_2 PGNP_i + \beta_3 FLR_i + u_i \quad (k=3)$

The results are:

$$\hat{C}_i = 263.6416 - 0.056 PGNP_i - 2.2316 FLR_i$$
$$se = \quad (11.5932) \quad (0.0019) \qquad (0.2099)$$

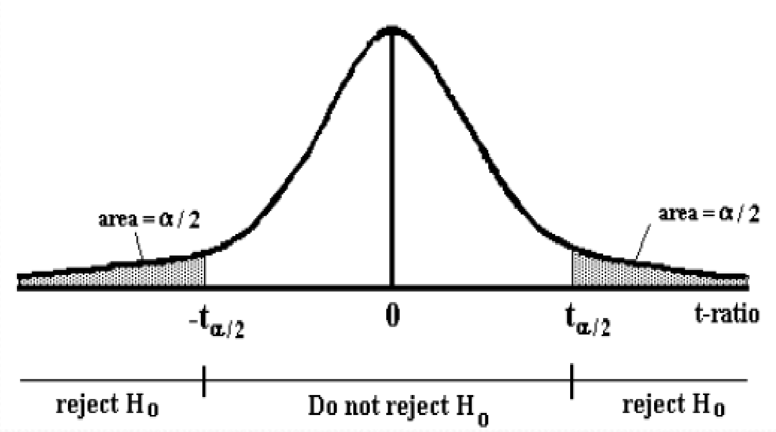
Use the t-test of significance to test at α=5%:

4

Example:

If $\hat{\beta}_2$=-0.056, $\sigma_{\hat{\beta}_2}$ =0.0019, degrees of freedom=n-k-1=64-2-1-61, α=5%

$H_0 : \beta_2 = 0$; $H_A : \beta_2 \neq 0$



$$T_{statistic,\ \text{(n-k-1)d.f}} = \frac{\hat{\beta}_2 - \beta_{2_0}}{\sigma_{\hat{\beta}_2}}$$

$$= \frac{-0.056 - 0}{0.0019} = -29.47$$

$$\text{Critical value, } t_{(n-k-1)d.f\,;\ \alpha/2} = t_{61d.f\,;\ 0.025} = 2$$

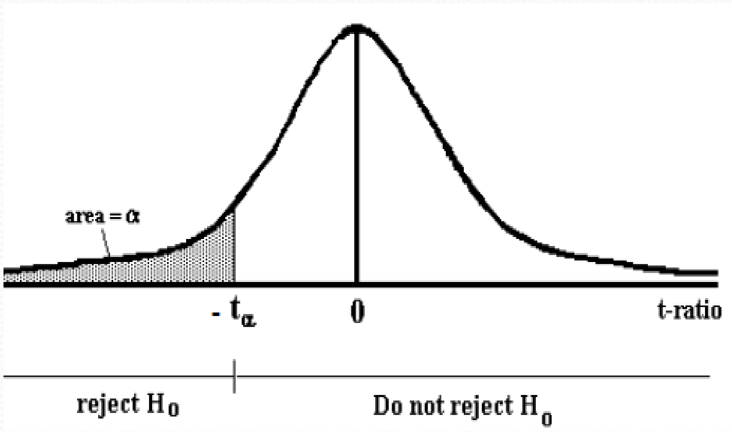
$$\text{Since } |t| = |-29.47| = 29.47 > t_{61d.f\,;\ 0.025} = 2;\ \text{Reject } H_0$$

β_2 is statistically significant, i.e. statistically different from zero: holding FLR constant, PGNP has a significant (negative) impact on child mortality.

Example:

If $\hat{\beta}_3$=-2.2316, $\sigma_{\hat{\beta}_3}$ =0.2099, degrees of freedom=n-k-1=64-2-1=61, α=5%

$H_0 : \beta_3 \geq 0$; $H_A : \beta_3 < 0$



$$T_{statistic,\ \text{(n-k-1)d.f}} = \frac{\hat{\beta}_3 - \beta_{3_0}}{\sigma_{\hat{\beta}_3}}$$

$$= \frac{-2.2316 - 0}{0.2099} = -10.631$$

$$\text{Critical value, } t_{(n-k-1)d.f\,;\ \alpha} = t_{61d.f\,;\ 0.05} = 1.67$$

$$\text{Since t=-10.631<-}t_{61d.f\,;\ 0.05} = -1.67;\ \text{Reject } H_0$$

implying that β_3< 0: holding PGNP constant, FLR has a significant negative impact on child mortality.

Note that most econometric packages (including Microfit) automatically report the t-statistic (T-ratio) for the null hypothesis: $\beta_j = 0$ (j=1..k).

Ordinary Least Squares Estimation

****

Dependent variable is DLS

132 observations used for estimation from 1873 to 2004

****

Regressor	Coefficient	Standard Error	T-Ratio[Prob]
CON	0.030335	0.016211	1.8713[.064]
DLS(-1)	0.049932	0.090173	0.55374[.581]
DLP	0.44044	0.25890	1.7012[.091]

****

The probabilities indicate that the null hypothesis of statistical insignificance cannot be rejected for all coefficients at the 1% and 5% level of significance. The null can be rejected at the 10% level of significance for the CON and DLP coefs.

7

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:

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

we must compute the test statistic:

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

where: $\hat{\sigma}_{(\hat{\beta}_3-\hat{\beta}_4)} = \sqrt{\hat{\sigma}^2_{\hat{\beta}_3} + \hat{\sigma}^2_{\hat{\beta}_4} - 2Cov(\hat{\beta}_3, \hat{\beta}_4)}$

- If $|t| > t_{(n-k),\alpha/2}$, reject H_0 at the α% level.

Example: Cubic Cost Function

- Use OLS to regress Total Cost (Y) on Output (X), Output Squared (X^2) and Output Cubed (X^3) using a sample of 10 obs (n=10). Results are:

$$\hat{Y}_i = 141.7667 + 63.477X_i - 12.9615X_i^{\,2} + 0.9396X_i^{\,3} \qquad (k=4)$$

$$se = \quad (6.3753) \quad (4.7786) \quad (0.9857) \qquad (0.0591)$$

$$Cov(\hat{\beta}_3, \hat{\beta}_4) = -0.0576$$

9

$H_0 : \beta_3 = \beta_4$ (equal coefficient for X^2 and X^3 term)

$H_1 : \beta_3 \neq \beta_4$

$$t = \frac{(\hat{\beta}_3 - \hat{\beta}_4) - (\beta_3 - \beta_4)}{\hat{\sigma}_{(\hat{\beta}_3-\hat{\beta}_4)}} = \frac{(-12.9615 - 0.9396) - 0}{\sqrt{(0.9857)^2 + (0.0591)^2 - 2(-0.0576)}}$$

$$= -13.313$$

- Since $|t| = 13.313 > t_{(10-4),0.05/2} = 2.447$ reject H_0 at the 5% level of significance.

3. Joint Hypothesis testing: F-test

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 joint parameter significance, form:

$H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$ (i.e. all the slope coefficients are simultaneously equal to zero).

H_1 : Not all slope coefficients are simultaneously zero

$$\text{F-statistic} = \frac{MSR}{MSE} = \frac{\frac{SSR}{K}}{\frac{SSE}{n-k-1}} = \frac{\frac{R^2}{K}}{\frac{1-R^2}{n-k-1}}$$

If F-statistic> F_α [k/(n-k-1) df: Reject the Null H_0 at α % level of significance

We should point out that it is possible to reject (not reject), via the *t*-test, the hypothesis that a particular slope coefficient is zero and yet not reject (reject) via the *F*-test the joint hypothesis that all slope coefficients are zero.

In other words: testing a series of single (individual) hypotheses is *not* equivalent to testing those hypotheses jointly.

Dependent variable is DLS

132 observations used for estimation from 1873 to 2004

Regressor	Coefficient	Standard Error	T-Ratio[Prob]
CON	0.030335	0.016211	1.8713[.064]
DLS(-1)	0.049932	0.090173	0.55374[.581]
DLP	0.44044	0.25890	1.7012[.091]
	F-stat.	F(2, 129)	1.9764[.143]

The value of the F-statistic is 1.9764

If F-statistic=1.9764> F_{α} [k/(n-k-1 df)= $F_{0.05}$ [2/(129 df)=3.04 : Reject the Null H_0 at 5 % level of significance

We can see that F-statistic=1.9764< $F_{0.05}$ [2/(129 df)=3.04

Thus at 5 % level of significance we fail to reject the H_0: $\beta_2 = \beta_3 = 0$

Regression versus Causation

• A statistical relationship in itself (however strong) can never establish causal connection. To attribute causality, one must appeal to a priori or theoretical considerations.

• For example, there is no statistical reason to assume that rainfall doesn't depend on crop yield. The fact we treat crop yield as the Y variable and rainfall as the X variable is due to common sense: we cannot control rainfall by varying crop yield.

Differences between regression and correlation:

- In correlation analysis we treat any (two) variables symmetrically; there is no distinction between dependent and explanatory variables.

-In regression analysis we treat the dependent variable (Y) as stochastic or random and the independent (X) as fixed or non-stochastic.

VI. Terminology and Notation

Dependent variable (Y)	**Explanatory variable (X)**
Explained variable	Independent variable
Predictand	Predictor
Regressand	Regressor
Response	Stimulus
Endogenous	Exogenous
Outcome	Covariate
Controlled variable	Control variable
Left-hand side variable	Right-hand side variable

- **Random (stochastic) variable**: a variable that can take on any set of values with a given probability.

8

Two Variable Regression Model: Estimation

The Method of Ordinary Least Squares (OLS)

• The OLS method is used to estimate the non-directly observable population regression function (PRF) on the basis of the sample regression function.

• The OLS method is extensively used in regression analysis because it is intuitively appealing and mathematically simpler than alternative estimation methods (e.g. maximum likelihood estimation).

• The OLS was developed first by C.F. Gauss in 1821.

Least Squares Criterion: specify the SRF (by choosing values for the estimators) so that it is as close as possible to the actual PRF, by minimising the sum of squared residuals (RSS):

$$\min_{\beta_1 \beta_2} \sum_{i=1}^{n} \hat{u}_i^2 = \min_{\beta_1 \beta_2} \text{SSE}$$

where, $\hat{\mathrm{u}}_i = y_i - \hat{y}_i$ is the residual or error term

Assumptions underlying the OLS method

The Gaussian or Classical Linear Regression Model (CLRM) makes 10 assumptions:

1. **Linearity**: the regression model is linear in the parameters
2. **X non-stochastic**: Values taken by the regressor X are considered fixed in repeated samples.
3. **Zero mean value of error term**: Given the value of X, the conditional mean of the random error term ui is zero: $E(u_i|X_i) = 0$.
4. **Homoskedasticity**: Given the value of X, the variance of u_i is the same for all observations: $Var(u_i|X_i)=\sigma^2$
5. **No autocorrelation (or no serial correlation)**: Given any two values of X: X_i and X_j ($i \neq j$), the correlation between any two error terms u_i and u_j is zero: $Cov(u_i , u_j \mid X_i , X_j)=0$; Thus: we don't worry about the other influences that might act on Y as a result of possible intercorrelations among the u's.

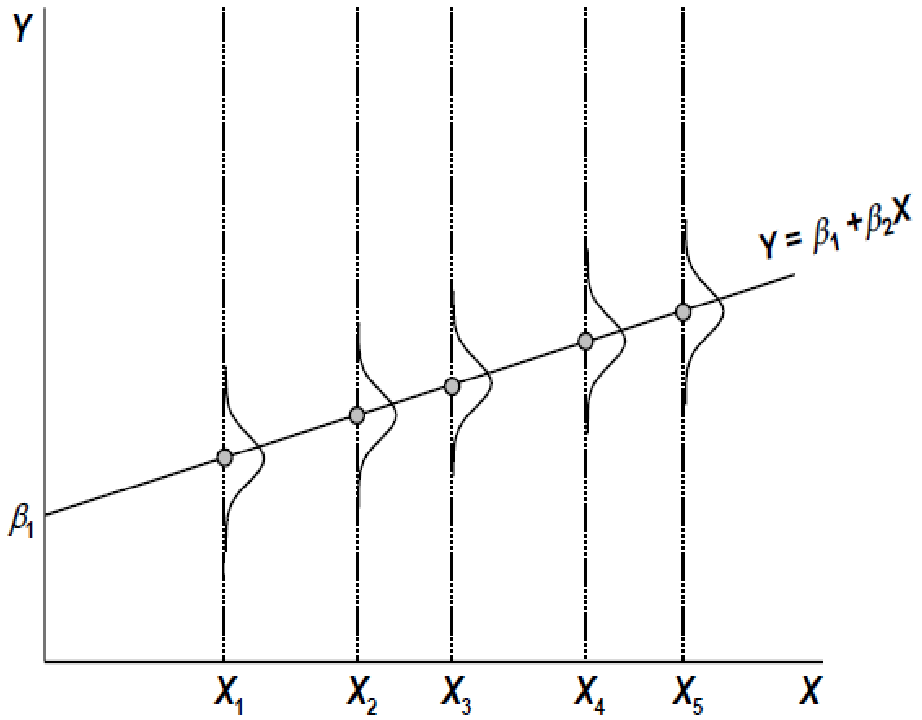
6. **Zero covariance between ui and Xi** : $Cov(u_i , X_i)=0$; Thus: the error term and the explanatory variable are uncorrelated, so that it possible to assess their individual effects on Y.

7. The number of observations (n) must be greater than the number of explanatory variables. Thus: if $Yi = \beta_1 + \beta 2X_i + u_i$, we need at least two observations on Y and X to estimate the 2 unknowns β_1 , β_2 .

8. Variability in X values: the X values in a given sample must not all be the same.

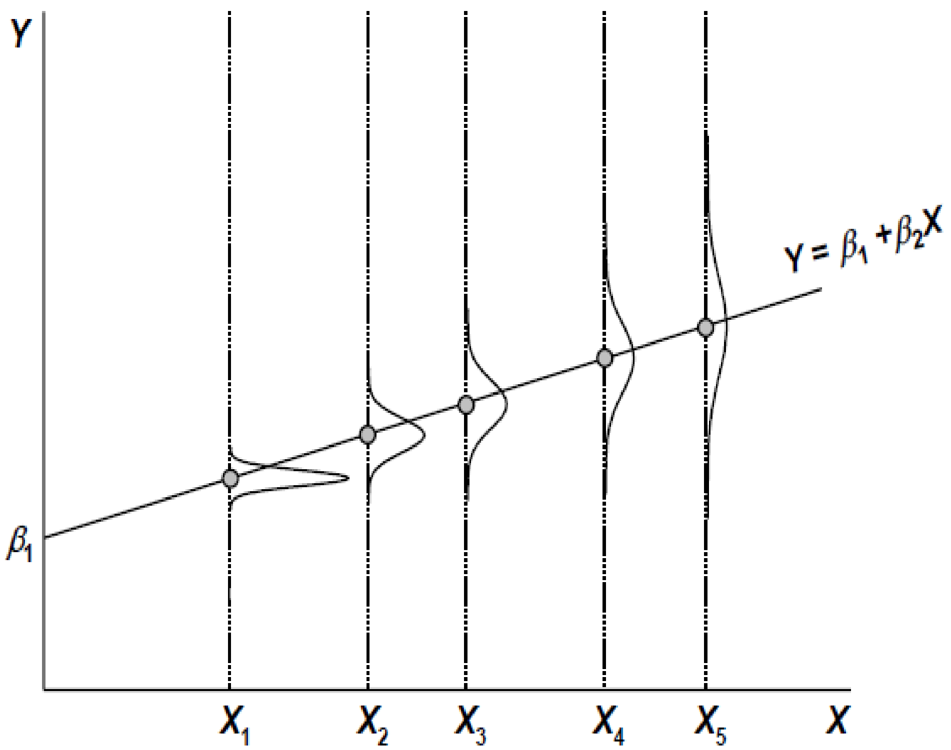
9. The regression model must be correctly specified.

10. There is no multicollinearity: there is no perfect (exact) linear relationship among the X's. This assumption applies to the case where there is more than one explanatory variable, e.g. : $Yi = \beta 1 + \beta 2X2i + \beta 3X3i + ui$



Homoskedasticity implies that the variation around the regression line is the same across the X values; it neither increases or decreases as X varies: $Var(u_i|X_i) = \sigma^2$

12



In contrast, *heteroskedasticity* is the situation when the variance is no longer constant but varies with X: $Var(u_i|X_i) = \sigma_i^2$