Regression analysis

Pravat Uprety
Assistant Professor
Central Department of Statistics
Tribhuvan University

Example (p=4)

Sales volume (Y)	Price (X1)	No of stores (X2)	Level of quality (X3)	No of advertisement (X4)

What is Econometrics

- Literally speaking, the word “measurement in economics.”

- The application of statistical and mathematical methods to the analysis of economic data, with the purpose of giving empirical content to economic theories and verifying them.

- Econometrics is based upon the development of statistical methods for estimating economic relationships, testing economic theories, and evaluating and implementing government and business policy.

- The most common application of econometrics is the forecasting of macroeconomic and business variables such as GDP, Private consumption, ROA, ROE etc.

Methodology of econometrics

- The statement of economic/business theory of formulation of hypothesis
- Specification of the econometric model to test the theory or hypothesis
- Estimation of parameters of the specified model
- Verification or statistical inference
- Forecasting and policy formulation

Hypothesis

- Hypothesis is that aspect of economic theory which is to be tested for empirical validity.
- Example: To test the Keynesian consumption theory : if we frame the statement ‘Consumption is a function of income’ it represents a hypothesis.

Model specification

- The model is an algebraic representation of a real world process
- At the stage of model specification, we decide on the precise form of functional relationship between consumption and income
- $Y_i = \beta_0 + \beta_1 X_i$ …………………… (1)

Where Y = Consumption and X = Income. The subscript i refers to the case of a particular individual (i = 1,2, …., n)

However, the reality is that the relationship between consumption and income is not exact i.e. persons with same income level are found to have different levels of consumptions so that we write the model 1 as

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ ……………..(2)

Here ε_i is called the stochastic term or disturbance term or error term.

Equation (2) represents an econometric specification of the consumption – income relationship as against mathematical specification of such relationship provided by (1)

Estimation

- Our objective here is to obtain estimates or numerical values of the unknown parameters of the model (2) by using any one of the estimation technique. Some popular estimation techniques are
- Ordinary Least Squares (OLS)
- Maximum Likelihood Estimates (MLE)
- Method of Moment

Necessary Assumptions for estimation

i) the mean or expected value of disturbance term ε is zero.

$E(\varepsilon_i) = 0$ for all i.

ii) The disturbances have uniform variance which is known as the assumption of homoskedasticity.

$\text{Var}(\varepsilon_i) = \sigma^2$ constant for all i. The violation of this assumption creates an econometric problem called heteroskedasticity.

iii) The disturbances are uncorrelated which is known as the assumption of serial independence or non autocorrelation.

i.e. Cov (ε_i , ε_j) = 0 for $i \neq j$

The violation of this assumption creates the problem of serial correlation or autocorrelation.

iv) ε is normally distributed.

This assumption is necessary for conducting statistical tests of significance of the parameters estimated

v) X is a non-stochastic variable with fixed values in repeated samples

BLUE Properties
(Best linear unbiased estimates)

- Unbiased ness
- Linearity
- Best ness (minimum variance)
- Consistency

Data for econometric analysis

- Cross sectional data
- Time series data
- Panel data

Terminology and notation

Dependent variable	Independent variable
Explained variable	Explanatory variable
Predictand	Predictor
Regressand	Regressor
Response	Stimulus
Endogenous	Exogeneous
Outcome	Covariate
Controlled variable	Control variable

The simple linear regression model

- Studying the relationship between two variables only (one dependent and one independent) is called the simple regression model. It is written as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{(for population)}$$

$$Y_i = b_0 + b_1 X_i + e_i \quad \text{(for sample)}$$

Example 1: Ceo salary and return on equity

- Let y be annual salary (salary) in thousands of dollars and x be the average return on equity (roe) for the CEO’s firm (Return on equity is defined in terms of net income as a percentage of common equity).
- Using the data, the OLS regression line relating salary to roe is

= 963.191 + 18.501 x $\hat{y}$

$$\hat{y}$$

Meaning of y-intercept and slope (regression coefficient)

- If the return on equity is zero then the predicted salary is the intercept (963.191), which equals $963191 since salary is measured in thousand.
- If the return on equity increases by one unit (percentage), then salary is predicted to change by about 18.5 or $18,500

Goodness of fit

- To measure how well the explanatory or independent variable (x) explains the dependent variable (y), coefficient of determination (R^2) is used.
- It is defined as

$$R^2 = SSE/TSS \quad = 1 – SSR/TSS$$

Where

SSE = explained sum of squares (explained variation)

SSR = Residuals sum of squares (Unexplained variation)

SST = Total sum of squares (total variation)

The value of R^2 always lies between 0 and 1.

Example: ceo salary and return on equity

$$\hat{y} = 963.191 + 18.501\,x$$

$$n = 209 \qquad R^2 = 0.0132$$

The firm’s return on equity explains only about 1.3% of the variation in salaries.

Log transformation

- Generally, log transformation is used to obtain a constant elasticity model.
- A constant elasticity model is
- log (salary) = β_0 + β_1 log(sales) + ε

Where sales is annual firm sales, measured in million of dollars.

β_1 is the elasticity of salary with respect to sales.

Example

GDP (million) Y	Export (million) X

Ln(GDP) ln (Y)	Ln (Export) ln (X)

- Estimating this equation by OLS gives

- log = 4.882 + 0.257 log x

$\hat{y}$ n= 209, R^2 = 0.211

the coefficient of log x is the estimated elasticity of salary with respect to sales. It implies that 1% increase in firm sales increases CEO salary by about 0.257%.

log = 4.882 + 0.231 x

= 844.882 + 18.28 log x

$\hat{y}$

$\hat{y}$

Multiple regression analysis

- Studying the relationship between one dependent and two or more than two independent (explanatory) variables
- The general multiple linear regression model for population can be written as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \qquad + \beta_p x_p + \varepsilon$$

Where β_0 = intercept

β_1 = is the parameter associated with x_1

β_2 = is the parameter associated with x_2 and so on

For sample data it is written as

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \qquad + b_p x_p + e$$

assumption

- Zero mean of ε_i $E(\varepsilon_i\ ) = 0$ for each i.
- Homoskedastcity var $(\varepsilon_{\,i}\,) = \sigma^2$ constant
- Non autocorrelation Cov $(\varepsilon_i\ , \varepsilon_j\,) = 0$ where $\varepsilon_i \neq \varepsilon_j$
- Normality: ε_i is normally distributed
- Non stochastic Xs, the values of the X-variables are same in repeated samples
- Zero covariance between ε_i and X variables.

$$\text{Cov}\ (\varepsilon_i\ , X_{1i}\,) = \text{Cov}\ (\varepsilon_i\ , X_{2i}\,) = 0$$

- No exact linear relationship exists between the X variables, i.e. Xs are not correlated (no multicollinearity)

Model Specification and Assumption (in vector and matrix form)

- The general population regression model involving the dependent variable Y_i and the independent (explanatory) variables X_{1i}, X_{2i},…………X_{ki} is specified as

Y_i	X1i	X2i		Xki
Y1	X11	X21		Xk1
Y2				
Yn				

The multiple regression equation is written as

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots\ldots\ldots\ldots\ldots\ldots + \beta_k X_{ki} + \epsilon_i$

This equation gives the following set of simultaneous equations:

$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \ldots\ldots\ldots\ldots\ldots\ldots + \beta_k X_{k1} + \epsilon_1$

$Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \ldots\ldots\ldots\ldots\ldots\ldots + \beta_k X_{k2} + \epsilon_2$

……………………………………………………………

……………………………………………………………

$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \ldots\ldots\ldots\ldots\ldots\ldots + \beta_k X_{kn} + \epsilon_n$

The system of equation can be written in the matrix form as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 \;\; X_{11} X_{21} \ldots\ldots\; X_{k1} \\ 1 \;\; X_{12} X_{22} \ldots\ldots\; X_{k2} \\ \\ \\ 1 \;\; X_{1n} X_{2n} \ldots\ldots\; X_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

That is

$\underline{Y} = X\underline{\beta} + \underline{\in}$ ------------------------(1)

Where $\underline{Y}$ is an (n X 1) vector of observations on dependent variable.

X is an [n X (k+1)] matrix of n observations on k variables X_{1i}, X_{2i},............X_{ki}, and the first column of 1 represents the intercept term.

$\underline{\beta}$ is a [(k+1) X 1] vector of parameters to be estimated and

$\underline{\in}$ is an (n X1) vector of disturbances.

Assumptions

1) Zero mean of $\underline{\in}$

That is

$$E\,(\underline{\in}) = E\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ . \\ . \\ E(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ 0 \end{bmatrix}$$

2) Constant variance of $\underline{\in}$

$\mathrm{Var}\,(\underline{\in}) = E\,(\underline{\in}\ \underline{\in}^t)$

$$= E\left[\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix}(\varepsilon_1\ \ \varepsilon_2.......\varepsilon_n)\right]$$

$$= E\begin{bmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 \ldots\ldots & \varepsilon_1\varepsilon_n \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 \ldots\ldots & \varepsilon_2\varepsilon_n \\ \\ \\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 \ldots\ldots & \varepsilon_n^2 \end{bmatrix} = \begin{bmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) \ldots\ldots & E(\varepsilon_1\varepsilon_n) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) \ldots\ldots & E(\varepsilon_2\varepsilon_n) \\ \\ \\ E(\varepsilon_n\varepsilon_1) & E(\varepsilon_n\varepsilon_2) \ldots\ldots & E(\varepsilon_n^2) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 \ldots\ldots\ldots 0 \\ 0 & \sigma^2 \ldots\ldots\ldots 0 \\ \\ 0 & 0 \ldots\ldots\ldots \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 \ldots\ldots\ldots 0 \\ 0 & 1 \ldots\ldots\ldots 0 \\ \\ 0 & 0 \ldots\ldots\ldots 1 \end{bmatrix}$$

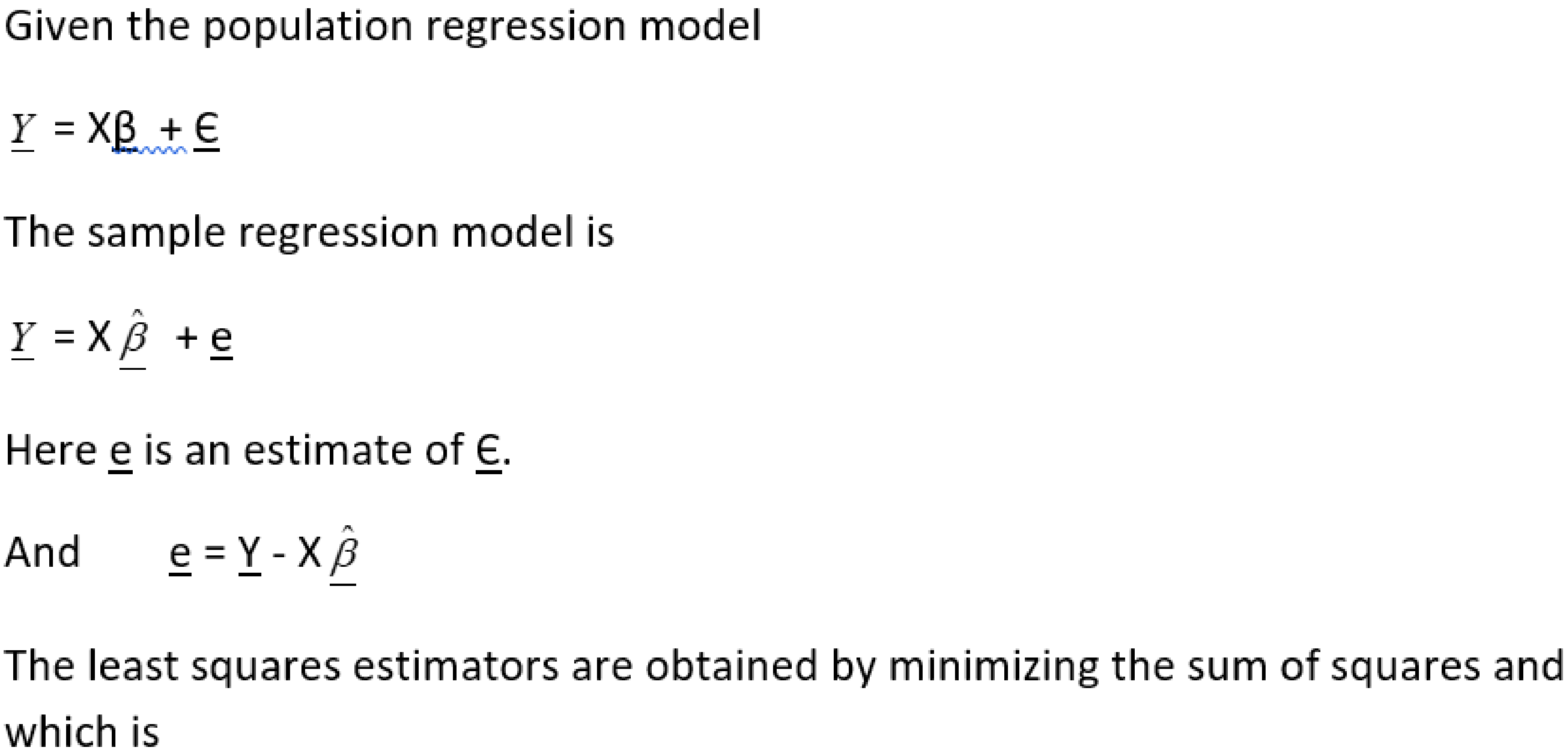
Var ($\underline{\in}$) = σ^2 I [Where I is (nXn) identity matrix]

3. Non stochastic Xs : This implies that all explanatory variables are non stochastic and hence, independent of the Єs.

4. Linear independence of Xs: This means that the explanatory variables do not form a linearly dependent set. In other word, rank of X matrix denoted by ρ(X) must be equal to number of explanatory variables in the model, which is k (no multicollinearity).

5. The vector $\underline{Є}$ has a multivariate normal distribution.

OLS estimation (derive OLS for multiple regression model)



Given the population regression model

$$\underline{Y} = X\underline{\beta} + \underline{\in}$$

The sample regression model is

$$\underline{Y} = X\underline{\hat{\beta}} + \underline{e}$$

Here $\underline{e}$ is an estimate of $\underline{\in}$.

And $\qquad \underline{e} = \underline{Y} - X\underline{\hat{\beta}}$

The least squares estimators are obtained by minimizing the sum of squares and which is

$$\sum e_i^2 = e_1^2 + e_2^2 + e_3^2 + \dots\dots\dots\dots + e_n^2$$

$$= (e_1\ e_2 \dots\dots\dots\dots e_n) \begin{pmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{pmatrix}$$

$$= \underline{e}^t\underline{e}$$

$$= (\underline{Y} - X\underline{\hat{\beta}})^t(\underline{Y} - X\underline{\hat{\beta}})$$

$$= (\underline{Y}^t - \underline{\hat{\beta}}^t X^t)((\underline{Y} - X\underline{\hat{\beta}})$$

$$= \underline{Y}^t\underline{Y} - \underline{Y}^t X\underline{\hat{\beta}} - \underline{\hat{\beta}}^t X^t\underline{Y} + \underline{\hat{\beta}}^t X^t X\underline{\hat{\beta}}$$

$$= \underline{Y}^t\underline{Y} - (\underline{\hat{\beta}}^t X^t\underline{Y})^t - \underline{\hat{\beta}}^t X^t\underline{Y} + \underline{\hat{\beta}}^t X^t X\underline{\hat{\beta}}$$

$$= \underline{Y}^t\underline{Y} - (\underline{\hat{\beta}}^t X^t\underline{Y}) - \underline{\hat{\beta}}^t X^t\underline{Y} + \underline{\hat{\beta}}^t X^t X\underline{\hat{\beta}}$$ (transpose of scalar = scalar)

$$= \underline{Y}^t\underline{Y} - 2\underline{\hat{\beta}}^t X^t\underline{Y} + \underline{\hat{\beta}}^t X^t X\underline{\hat{\beta}}$$

For least squares

$$\frac{\partial \Sigma e_i^2}{\partial \underline{\hat{\beta}}} = 0$$

$$\frac{\partial(\underline{Y}^t \underline{Y} - 2\,\underline{\hat{\beta}}^t X^t \underline{Y} + \underline{\hat{\beta}}^t X^t X \underline{\hat{\beta}})}{\partial \underline{\hat{\beta}}} = 0$$

$$-2\, X^tY + 2\, X^t X \underline{\hat{\beta}} = 0$$

$$(X^t X)\underline{\hat{\beta}} = X^tY \text{ -------------------------(2)}$$

The equations contained in 2 are called OLS normal equations in the context of the general linear model.

Therefore

$$\underline{\hat{\beta}} = (X^t X)^{-1} (X^tY) \text{ -----------------(3)}$$

The vector contains estimators for all unknown parameters.

Software output and Interpretation

Format of anova table

Source	Degrees of freedom	Sum of squares	Mean sum of squares	F-value	P-value
Regression	K	SSR	MSR=SSR/p	F= MSR/MSE	
Residual or Error	n-K-1	SSE	MSE=SSE/n-p-1		
Total	n-1	TSS or SST			

- Where, n= sample size

K = number of independent variables
SSR = explained sum of squares (explained variation)
SSE = Residuals sum of squares (Unexplained variation)
SST = Total sum of squares (total variation)
The ANOVA table is used to test the overall goodness of fit or testing of all regression coefficients simultaneously

By using following anova table obtained from 30 observations

Source	SS	Df	MSS	F
Regression	500	4	?	?
Error	?	?	?	
Total	700	?		

COMPLETE THE GIVEN ANOVA TABLE

OBTAIN COEFFICIENT OF DETERMINATION AND STANDARD ERROR

By using following anova table obtained from 30 observations

Source	SS	Df	MSS	F
Regression	500	4 =p	125	15.625
Error	200	25 = n-p-1	8	
Total	700	29 = n-1		

R^2 = 500/700 = 0.71

S_{YX} = SQR (MSE) = SQR OF 8 =

Format of coefficient table

Predictor	bi (Unstandardized regression coeff)	Sbi (Unstandardized standard error)	t-stat t = bi/Sbi	P-value
Constant	b_0	Sb_0	b_0 /Sb_0	
X_1	b_1	Sb_1	b_1 /Sb_1	
X_2	b_2	Sb_2	b_2 /Sb_2	
.	.		.	
.	.		.	
X_p	b_p	Sb_p	b_p /Sb_p	

Where, b_i = regression coefficient of X_i

Sb_i = Standard error of regression coefficient

The coefficient table is used to test the individual impact of each X_i on Y.

From ANOVA table and coefficient table

- We can compute
- i) Multiple coefficient of determination

$$R^2 = SSR/TSS \quad = 1 - SSE/TSS$$

It measures the proportion of variation in dependent variable that is explained by all explanatory variables.

Suppose $R^2 = 0.856$, $p = 5$

85.6% of variation in dependent variable is explained by 5 independent/explanatory variables.

Adjusted R^2

- Adj R^2 = 1 – {(1-R^2) } $\frac{(n-1)}{(n-p-1)}$

- It measures the proportion of variation in dependent variable that is explained by all independent variables after adjusting for given degrees of freedom.

- It is used to select the model.

2. Standard error of estimate (Syx) [ANOVA table]

i.e. Syx = $\sqrt{MSE}$

It measures the average variation of observed values of dependent variable around its fitted equation.

Suppose, Syx = 6.89

i.e. the average variation of observed values of dependent variable around its fitted equation is 6.89

3. We can develop the estimating equation and prediction of dependent variable (coefficent table)

- The estimating equation is written as
- $\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + \qquad + b_p x_p$

4. Confidence interval estimate for the population slope or regression coefficient (β_i)

$$b_i \pm t_{n-p-1,\, \alpha}\ Sb_i$$

5. Testing significance of individual regression coefficient (t test-coefficent table)

- Null hypothesis (H_0): $\beta_i = 0$
- Alternative hypothesis (H_1): $\beta_i \neq 0$

From software we get the t –value and corresponding p-value (in coefficient table)

Decision: Cal t = |t| (Critical Value)

tab t = $t_{n-p-1, \alpha}$

If cal t ≤ tab t do not reject H_o

cal t > tab t reject H_o

If p-value ≥ α (level of significance)

we do not reject null hypothesis

If p-value < α (level of significance)

we reject null hypothesis

6. Testing the overall significance of regression (F test – ANOVA table)

- Null hypothesis (H_0): $\beta_1 = \beta_2 = \ldots\ldots\ldots = \beta_p = 0$
- Alternative hypothesis (H_1) : not all βs are simultaneously zero.

Or at least one $\beta i \neq 0$

From software we get the F-value and corresponding p-value.

From table = Tabulated value = $F_{p,\, n-p-1}$ at α %

Decision: If cal F ≤ tab F do not reject

If cal F > tab F reject

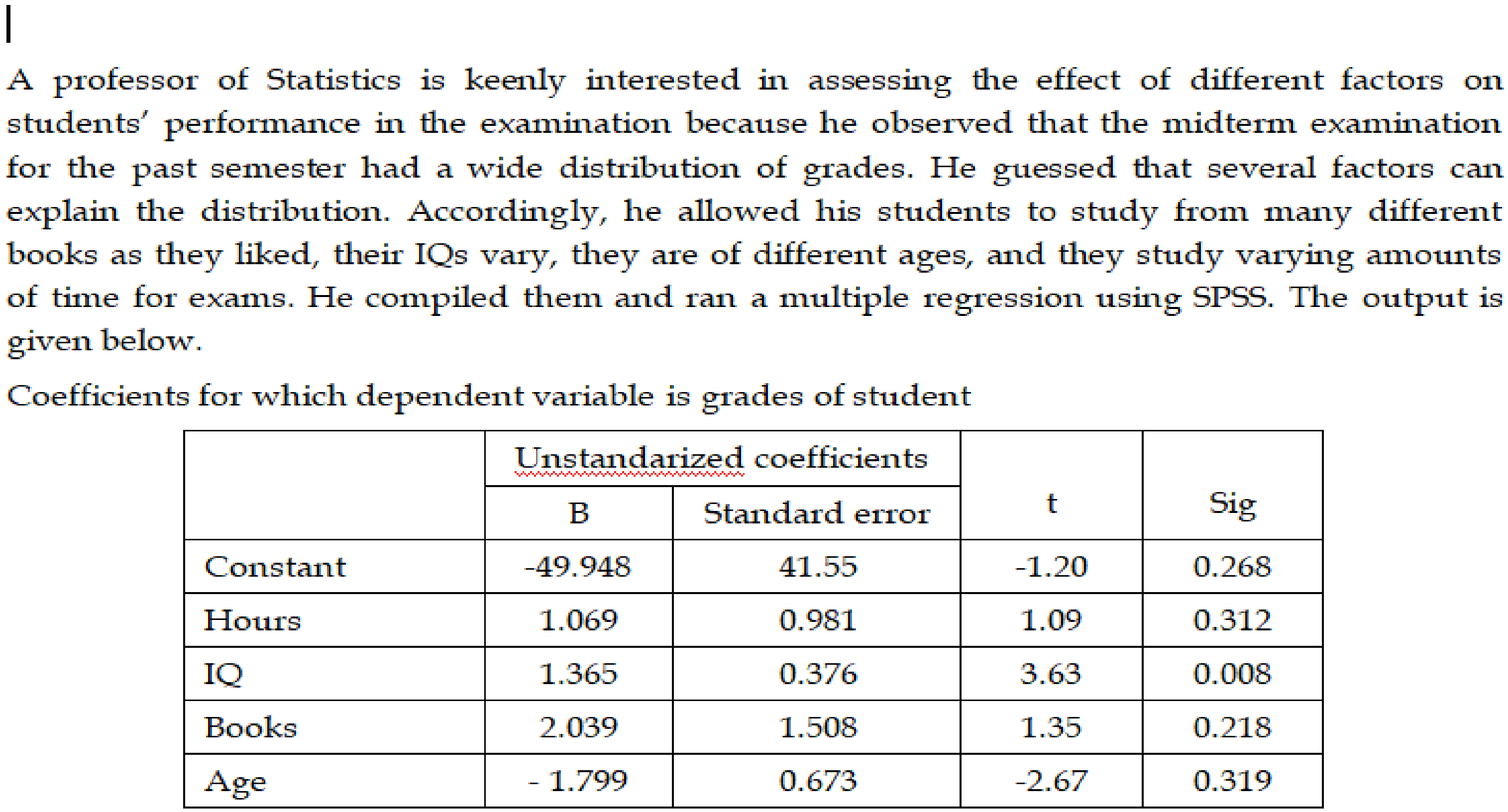
Or

If p-value ≥ α (level of significance)

we do not reject null hypothesis

If p-value < α (level of significance)

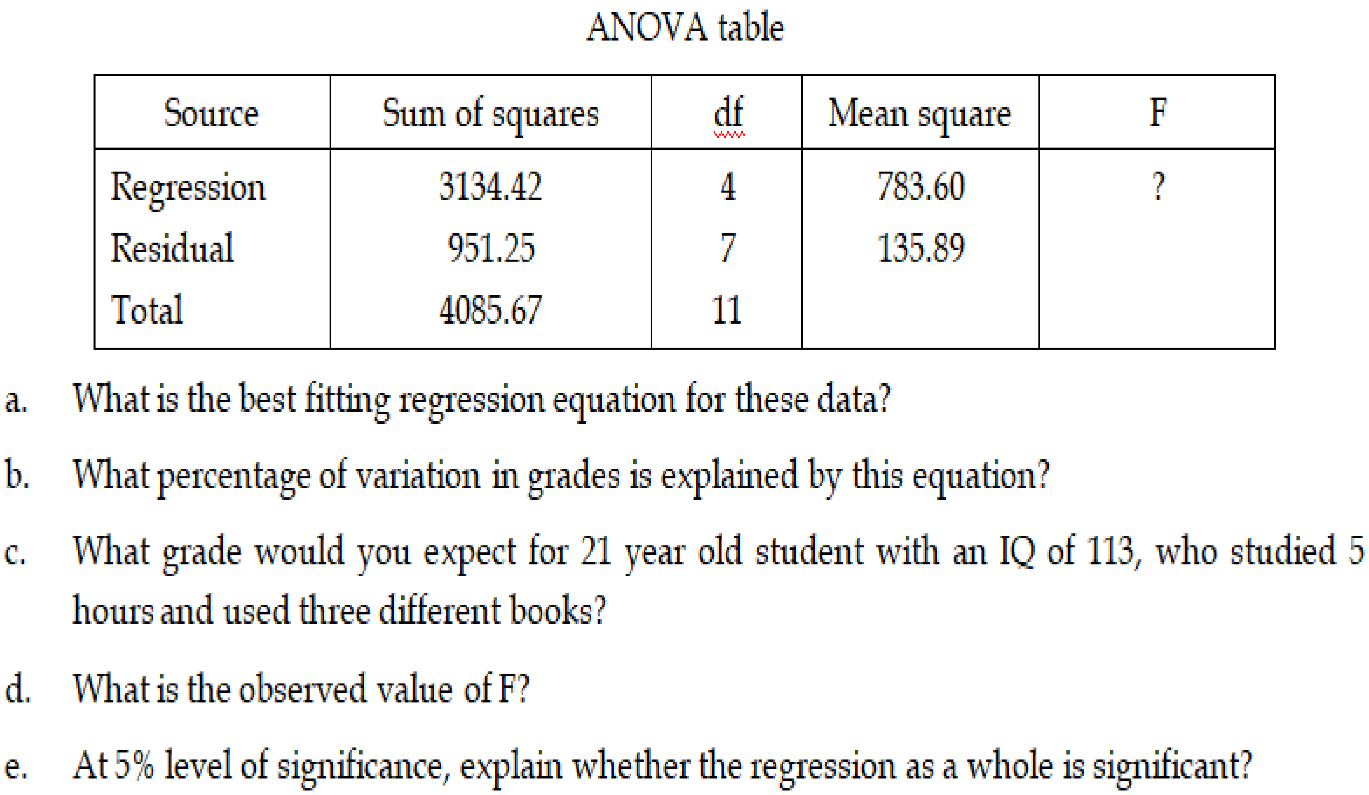
we reject null hypothesis



A professor of Statistics is keenly interested in assessing the effect of different factors on students' performance in the examination because he observed that the midterm examination for the past semester had a wide distribution of grades. He guessed that several factors can explain the distribution. Accordingly, he allowed his students to study from many different books as they liked, their IQs vary, they are of different ages, and they study varying amounts of time for exams. He compiled them and ran a multiple regression using SPSS. The output is given below.

Coefficients for which dependent variable is grades of student

	Unstandarized coefficients		t	Sig
	B	Standard error		
Constant	-49.948	41.55	-1.20	0.268
Hours	1.069	0.981	1.09	0.312
IQ	1.365	0.376	3.63	0.008
Books	2.039	1.508	1.35	0.218
Age	- 1.799	0.673	-2.67	0.319



ANOVA table

Source	Sum of squares	df	Mean square	F
Regression	3134.42	4	783.60	?
Residual	951.25	7	135.89	
Total	4085.67	11		

a. What is the best fitting regression equation for these data?

b. What percentage of variation in grades is explained by this equation?

c. What grade would you expect for 21 year old student with an IQ of 113, who studied 5 hours and used three different books?

d. What is the observed value of F?

e. At 5% level of significance, explain whether the regression as a whole is significant?

Solution

a. We have the response variable(Y) is the grades obtained by the students. The independent variables are hours(X_1), IQ(X_2) books(X_3) and age(X_4). The estimated multiple regression equation for 4 independent variables is:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

Substituting the values of regression coefficient for each predictor variables from the generated table, the best fitting regression equation for theses data will be:

$$\hat{Y} = -49.948 + 1.069X_1 + 1.365X_2 + 2.039X_3 - 1.799X_4$$

Where X1, X2, X3 and X4 represents hours, IQ, books and age of students respectively.

b. We have coefficient of multiple determination $R^2 = 76.7\%$. This shows that 76.7% of the total variation in grades(Y) is explained by this multiple regression equation.

c. The expected value of grades(Y) for 21 year old(i.e. $X_4 = 21$) student with IQ of 113(i.e. $X_2 = 113$) who studied 5 hours(i.e. $X_1 = 5$) and used 3 different books(i.e. $X_3 = 3$) will be:

$$\hat{Y} = -49.948 + 1.069(5) + 1.365(113) + 2.039(3) - 1.799(21) = 77.98$$

d. The observed value of F

$$F = \frac{MSR}{MSE} = \frac{783.60}{135.89} = 5.77$$

e. Null hypothesis, $H_0: \beta_1 = \ldots\ldots\ldots = \beta_4 = 0$ i.e. there is no linear relationship between the dependent variable and independent variables

Alternative hypothesis, H_1: At least one $\beta_j \neq 0$, for j = 1, 2, 3, 4 i.e. there is linear relationship between the dependent variable and at least one of the independent variables.

We have from (d) that F = 5.77. Table value of F at 5 % level of significance with (4, 7) = 4.12. Calculated value of F is greater than table value of F at 5% level of significance with (4, 7) degrees of freedom i.e. 5.77 > 4.12. We reject the null hypothesis and concluded that the regression coefficient as a whole is significant.

1. A manager selects a representative sample of 24 monthly customer bills taken from several recent heating seasons. The manager considers kilowatt hours per month (Y) as a linear function of square feet heated space (X1), an index of roof insulation quality (X2), presence/absence of insulated windows (X3), mean temperature (X4), and heat pump/electric forced air (X5). A SPSS output is as follows:

	Unstandardized Coefficients		t	p-value
	bi	Sbi		
(Constant)	6356.17	838.701	?	
X1	0.56038	0.15811	?	0.000
X2	-31.2077	8.95905	?	0.025
X3	-327.503	149.169	?	0.001
X4	-113.895	16.2604	?	0.000
X5	-621.458	147.828	?	0.000

ANOVA

Source	Sum of Squares	df	Mean Square	F	P-value
Regression	?	?	?	?	0.000
Residual	2166000	?	?		
Total	14370000	23			

i) Complete above Coefficient table and ANOVA table.

ii) Test the significance of the estimated regression coefficient of X_3 at the 5% significance level.

iii) Construct 99% confidence interval estimate for the regression coefficient of square feet heated space.

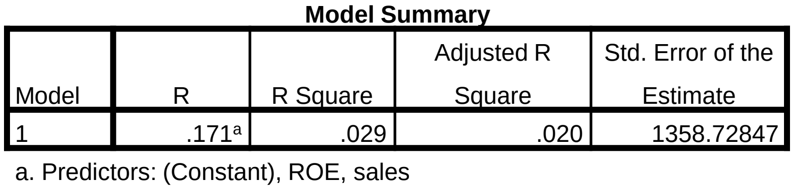
iv) Compute the standard error of the estimate and interpret its meaning.

v) Compute the R^2 and adjusted R^2 then interpret its meaning.

vi) Given that X1=1295, X2= 18, X3= 5, X4=3, X5=1 predict the average Kilowatt hours per month.

vii) Set up the null and alternative hypothesis, carry out F-test and interpret your result.

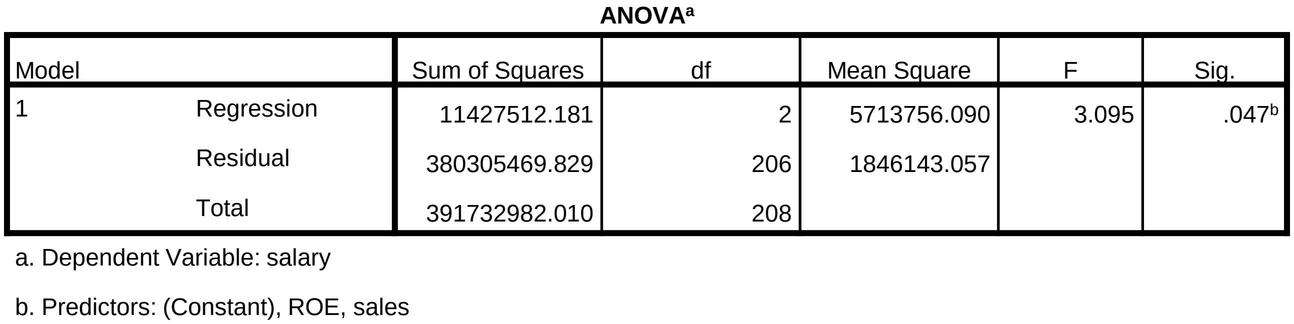
Example: CEO SALARY ANOVA table and coefficient table



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.171[a]	.029	.020	1358.72847

a. Predictors: (Constant), ROE, sales

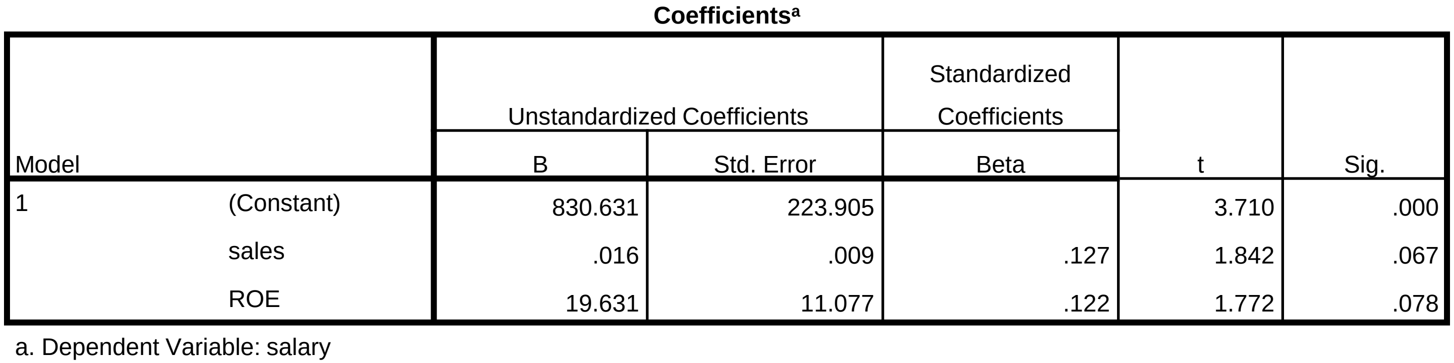


ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11427512.181	2	5713756.090	3.095	.047[b]
	Residual	380305469.829	206	1846143.057		
	Total	391732982.010	208			

a. Dependent Variable: salary

b. Predictors: (Constant), ROE, sales



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	830.631	223.905		3.710	.000
	sales	.016	.009	.127	1.842	.067
	ROE	19.631	11.077	.122	1.772	.078

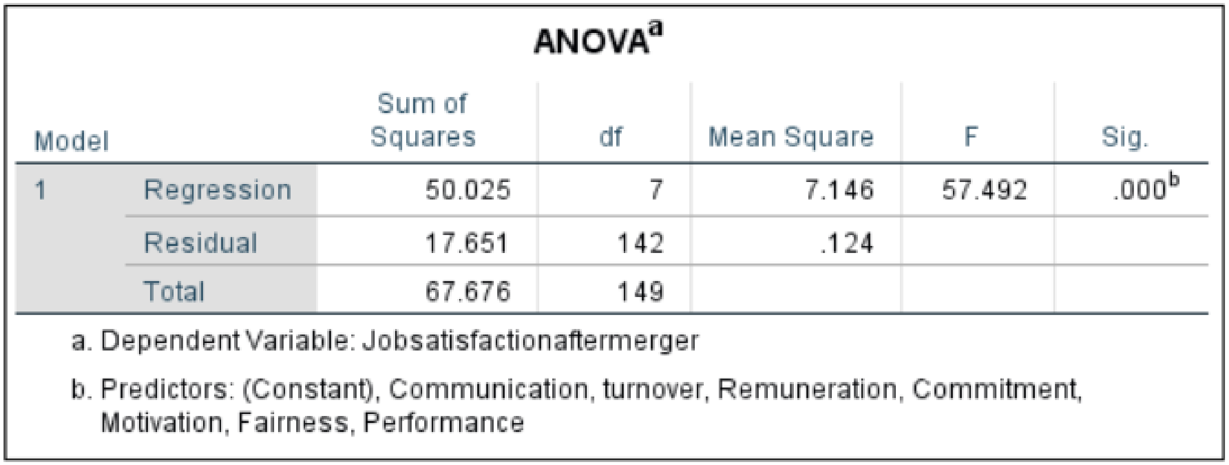
a. Dependent Variable: salary

Example ceo salary

$$\hat{y} = 830.63 + 0.163\ x_1 + 19.63\ x_2$$

$n = 209$, $R^2 = 0.029$

Example 2:

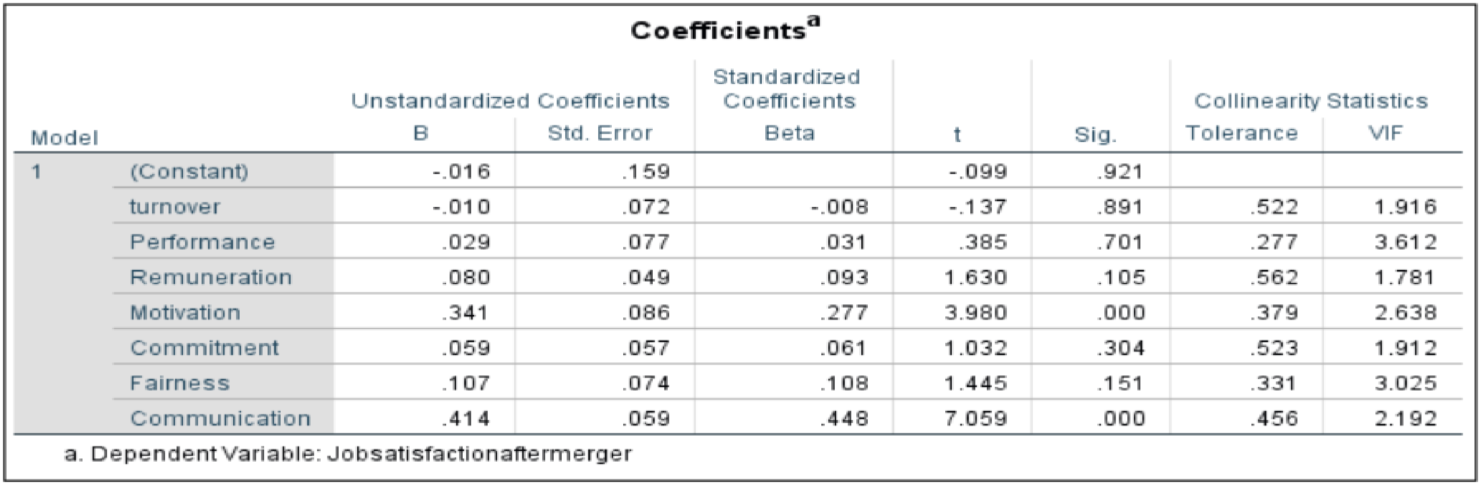


ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	50.025	7	7.146	57.492	.000[b]
	Residual	17.651	142	.124		
	Total	67.676	149			

a. Dependent Variable: Jobsatisfactionaftermerger

b. Predictors: (Constant), Communication, turnover, Remuneration, Commitment, Motivation, Fairness, Performance



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-.016	.159		-.099	.921		
	turnover	-.010	.072	-.008	-.137	.891	.522	1.916
	Performance	.029	.077	.031	.385	.701	.277	3.612
	Remuneration	.080	.049	.093	1.630	.105	.562	1.781
	Motivation	.341	.086	.277	3.980	.000	.379	2.638
	Commitment	.059	.057	.061	1.032	.304	.523	1.912
	Fairness	.107	.074	.108	1.445	.151	.331	3.025
	Communication	.414	.059	.448	7.059	.000	.456	2.192

a. Dependent Variable: Jobsatisfactionaftermerger

Dummy variable

Qualitative factors often come in the form of binary information:

a person is female or male,

private bank or public bank,

a person does or does not own a personal computer,

a person does or does not own a car.

In all of these examples, the relevant information can be captured by defining a binary variable or a zero – one variable. In econometrics, binary variables are most commonly called dummy variables.

THANK YOU