

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 ($i = 1, 2, \dots, n$)

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

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \text{(2)}$$

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.

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

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**.

$$\text{i.e. Cov } (\varepsilon_i, \varepsilon_j) = 0 \text{ for } i \neq j$$

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

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

\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 \$963,191 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

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

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

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

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

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
- $\log(\text{salary}) = \beta_0 + \beta_1 \log(\text{sales}) + \varepsilon$

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

$\ln(\text{GDP})$ $\ln(Y)$	$\ln(\text{Export})$ $\ln(X)$

- Estimating this equation by OLS gives

- $\log \hat{y} = 4.882 + 0.257 \log x$
 $n = 209, R^2 = 0.211$

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%.

$$\log \hat{y} = 4.882 + 0.231 x$$

$$= 844.882 + 18.28 \log x$$

\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

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

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

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

assumption

- Zero mean of ε_i $E(\varepsilon_i) = 0$ for each i .
- Homoskedasticity $\text{var}(\varepsilon_i) = \sigma^2$ constant
- Non autocorrelation $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ 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 $X_{1i}, X_{2i}, \dots, X_{ki}$ is specified as

Y_i	X_{1i}	X_{2i}		X_{ki}
Y_1	X_{11}	X_{21}		X_{k1}
Y_2				
Y_n				

The multiple regression equation is written as

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

This equation gives the following set of simultaneous equations:

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

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

.....

.....

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

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

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

That is

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

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

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

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

$\underline{\epsilon}$ is an $(n \times 1)$ vector of disturbances.

Assumptions

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

That is

$$E(\underline{\epsilon}) = E \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix} = \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ . \\ . \\ E(\epsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ 0 \end{bmatrix}$$

|

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

$$\text{Var}(\underline{\epsilon}) = E(\underline{\epsilon} \underline{\epsilon}^t)$$

$$= E \left[\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix} (\epsilon_1 \ \epsilon_2 \dots \epsilon_n) \right]$$

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

$$= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Var ($\underline{\varepsilon}$) = $\sigma^2 I$ [Where I is ($n \times n$) identity matrix]

3. Non stochastic X s : This implies that all explanatory variables are non stochastic and hence, independent of the ϵ s.
4. Linear independence of X s: This means that the explanatory variables do not form a linearly dependent set. In other word, rank of X matrix denoted by $\rho(X)$ must be equal to number of explanatory variables in the model, which is k (no multicollinearity).
5. The vector $\underline{\epsilon}$ has a multivariate normal distribution.

OLS estimation (derive OLS for multiple regression model)

Given the population regression model

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

The sample regression model is

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

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

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

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

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

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

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

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

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

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

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

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

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

For least squares

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

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

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

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

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

Therefore

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

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		

$$R^2 = 500/700 = 0.71$$
$$S_{YX} = \text{SQR (MSE)} = \text{SQR OF } 8 =$$

Format of coefficient table

Predictor	b_i (Unstandardized regression coeff)	Sb_i (Unstandardized standard error)	t-stat $t = b_i/Sb_i$	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

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

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 R²

- $$\text{Adj } R^2 = 1 - \left\{ (1 - R^2) \frac{(n-1)}{(n-p-1)} \right\}$$
- 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 (S_{yx}) [ANOVA table]

i.e. $S_{yx} = \sqrt{MSE}$

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

Suppose, $S_{yx} = 6.89$

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 (coefficient table)

- The estimating equation is written as

- $\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$

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

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

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

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

If cal t \leq tab t do not reject H_0

cal t $>$ tab t reject H_0

If p-value $\geq \alpha$ (level of significance)

we do not reject null hypothesis

If p-value $< \alpha$ (level of significance)

we reject null hypothesis

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

- Null hypothesis (H_0): $\beta_1 = \beta_2 = \dots = \beta_p = 0$
- Alternative hypothesis (H_1) : not all β s 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 $\text{cal } F \leq \text{tab } F$ do not reject

If $\text{cal } F > \text{tab } F$ reject

Or

If $\text{p-value} \geq \alpha$ (level of significance)

we do not reject null hypothesis

If $\text{p-value} < \alpha$ (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

	<u>Unstandardized coefficients</u>		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		

- What is the best fitting regression equation for these data?
- What percentage of variation in grades is explained by this equation?
- What grade would you expect for 21 year old student with an IQ of 113, who studied 5 hours and used three different books?
- What is the observed value of F?
- 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 these data will be:

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

Where X_1 , X_2 , X_3 and X_4 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, $H_0: \beta_1 = \dots = \beta_4 = 0$ 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 $X_1=1295$, $X_2=18$, $X_3=5$, $X_4=3$, $X_5=1$ 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			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	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

BLUE Properties and Multicollinearity

Pravat Uprety

BLUE Properties

(Best linear unbiased estimates)

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

BLUE properties

1. The estimators are linear, that is, they are linear functions of the dependent variable Y . Linear estimators are easy to understand and deal with compared to nonlinear estimators.
2. The estimators are unbiased, that is, in repeated applications of the method, on average, the estimators are equal to their true values.
3. In the class of linear unbiased estimators, OLS estimators have minimum variance. As a result, the true parameter values can be estimated with least possible uncertainty; an unbiased estimator with the least variance is called an efficient estimator.

Unbiased ness

Unbiased ness

Given the general linear model

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

And we know that

$$\underline{\hat{\beta}} = (\underline{X}^t \underline{X})^{-1} (\underline{X}^t \underline{Y})$$

Now substituting $\underline{Y} = X\underline{\beta} + \underline{\epsilon}$ in the equation

we get,

$$\begin{aligned}\underline{\hat{\beta}} &= (\underline{X}^t \underline{X})^{-1} \underline{X}^t (X\underline{\beta} + \underline{\epsilon}) \\ &= (\underline{X}^t \underline{X})^{-1} (\underline{X}^t \underline{X}) \underline{\beta} + (\underline{X}^t \underline{X})^{-1} \underline{X}^t \underline{\epsilon} \\ &= \underline{\beta} + (\underline{X}^t \underline{X})^{-1} \underline{X}^t \underline{\epsilon}\end{aligned}$$

Taking expectation on both sides

$$E(\underline{\hat{\beta}}) = E[\underline{\beta} + (\underline{X}^t \underline{X})^{-1} \underline{X}^t \underline{\epsilon}]$$

$$= E(\underline{\beta}) + (\underline{X}^t \underline{X})^{-1} \underline{X}^t E(\underline{\epsilon})$$

$$= \underline{\beta} + (\underline{X}^t \underline{X})^{-1} \underline{X}^t (0)$$

$$E(\underline{\hat{\beta}}) = \underline{\beta}$$

This proves that $\underline{\hat{\beta}}$ is an unbiased estimator of $\underline{\beta}$.

Linearity

We know that

$$\underline{\hat{\beta}} = (\underline{X^t X})^{-1} (\underline{X^t Y})$$

And it is clear that $\hat{\beta}$'s have a linear relation with Y with the weights being functions of X data, which are non stochastic.

Bestness

We know that

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

$$\text{Or, } \hat{\underline{\beta}} - \underline{\beta} = (\underline{X}^t \underline{X})^{-1} \underline{X}^t \underline{\epsilon}$$

The variance of $\hat{\underline{\beta}}$ is

$$\begin{aligned} \text{Var}(\hat{\underline{\beta}}) &= E[(\hat{\underline{\beta}} - \underline{\beta})(\hat{\underline{\beta}} - \underline{\beta})^t] \\ &= E[(\underline{X}^t \underline{X})^{-1} \underline{X}^t \underline{\epsilon} \{(\underline{X}^t \underline{X})^{-1} \underline{X}^t \underline{\epsilon}\}^t] \\ &= E[(\underline{X}^t \underline{X})^{-1} \underline{X}^t \underline{\epsilon} \underline{\epsilon}^t \underline{X} (\underline{X}^t \underline{X})^{-1}] \\ &= (\underline{X}^t \underline{X})^{-1} \underline{X}^t E(\underline{\epsilon} \underline{\epsilon}^t) \underline{X} (\underline{X}^t \underline{X})^{-1} \\ &= (\underline{X}^t \underline{X})^{-1} \underline{X}^t \sigma^2 \underline{X} (\underline{X}^t \underline{X})^{-1} \\ &= \sigma^2 (\underline{X}^t \underline{X})^{-1} (\underline{X}^t \underline{X}) (\underline{X}^t \underline{X})^{-1} \\ &= \sigma^2 (\underline{X}^t \underline{X})^{-1} \text{ -----(2)} \end{aligned}$$

Example: ceo salary and return on equity

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

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

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
- $\log(\text{salary}) = \beta_0 + \beta_1 \log(\text{sales}) + \varepsilon$

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

β_1 is the elasticity of salary with respect to sales.

##Command in R

- `lm (y~x, data = name of data frame)`

##Open CEOSAL1 data

```
attach(CEOSAL1)
```

```
dim(CEOSAL1)
```

```
head(CEOSAL1)
```

##Using dplyr

```
library(dplyr)
```

```
CEOSAL1 %>% summarise(avg_sal=mean(salary), avg_roe=mean(roe))
```

##Summary statistics using R

```
mean(salary)
```

```
mean(roe)
```

```
cov(salary, roe)
```

```
var(salary)
```

```
var(roe)
```

##Manual Calculation in R

```
(slope= cov(roe,salary)/var(roe))
```

```
(yintercept = mean(salary)-slope*mean(roe))
```

Regression analysis

- ##Direct regression
- `lm(salary~roe)`
- `reg_ceo<-lm(salary~roe)`
- `summary(reg_ceo)`
- `salhat<-fitted(reg_ceo)`
- `uhat<-resid(reg_ceo)`

- `cbind(CEOSAL1, salhat,uhat)`
- `BIC(reg_ceo)`
- `detach(CEOSAL1)`

Log linear model

- For the estimation of logarithmic or semi logarithmic models, the `lm` formula can be directly used
- `lm(log(salary)~log(sales), data =CEOSAL1)`

Common used technique

```
reg_ceo<-lm(salary~roe, data=CEOSAL1)
```

```
summary(reg_ceo)
```

```
max(CEOSAL1$salary)
```

```
plot(density(CEOSAL1$salary))
```

##Scatter diagram

```
plot(CEOSAL1$roe, CEOSAL1$salary)
```

```
plot(CEOSAL1$roe, CEOSAL1$salary, ylim=c(0,4000))
```

```
abline(reg_ceo) [to draw the regression line in plot]
```

##Log linear model

```
ln_sal<-lm(log(salary)~log(sales), data =CEOSAL1)
```

```
summary(ln_sal)
```

Matrix approach

```
##determine sample size and no. of regressors:
```

```
n<-nrow(CEOSAL1); k<-1
```

```
n
```

```
#Extract y
```

```
y<- CEOSAL1$salary
```

```
x <- (cbind(1, CEOSAL1$roe))
```

```
dim(x)
```

```
head(x)
```

```
##Parameter estimate (matrix approach)
```

```
(bhat<-solve(t(x)%*%x) %*% t(x) %*%y)
```

```
#Residual
```

```
uhat<-y-x%*%bhat
```

```
Uhat
```

```
mean(uhat)
```

```
var(uhat)
```

Packages

- ##to get better presentation we can install jtools package
- `install.packages("jtools")`
- `library(jtools)`

- `install.packages("huxtable")`
- `library(huxtable)`

- `install.packages("car")`
- `library(car)`

- `summary(reg_ceo)`
- `summ(reg_ceo)`

Example: Wage data and dummy in independent variable

Wage: Hourly wage in dollars, which is the dependent variable.

The explanatory variables, or regressors, are as follows:

Female: Gender, coded 1 for female, 0 for male

Nonwhite: Race, coded 1 for nonwhite workers, 0 for white workers

Union: Union status, coded 1 if in a union job, 0 otherwise

Education: Education (in years)

##Opening file

- `wage <- read_excel("wage.xls")`
- `library(dplyr)`
- `head(wage)`

#Regression

- `reg_wage<- lm(wage~female+nonwhite+union+education+exper,
data=wage)`
- `reg_wage`
- `summary (reg_wage)`

##to obtain confidence interval estimate

- `confint(reg_wage)`

Running regression

- `reg_wage<- lm(wage~female+nonwhite+union+education+exper, data=wage)`
- `reg_wage1<- lm(wage~female+nonwhite+union+education, data=wage)`
- `summ(reg_wage1)`

##To compare two or more models

- `export_summs(reg_wage, reg_wage1)`
- `summ(reg_wage1, scale = TRUE, vifs = TRUE, part.corr = TRUE, confint = TRUE, pvals = FALSE)`

Result using summary command

Residuals:

Min	1Q	Median	3Q	Max
-20.781	-3.760	-1.044	2.418	50.414

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-7.18334	1.01579	-7.072	2.51e-12	***
female	-3.07488	0.36462	-8.433	< 2e-16	***
nonwhite	-1.56531	0.50919	-3.074	0.00216	**
union	1.09598	0.50608	2.166	0.03052	*
education	1.37030	0.06590	20.792	< 2e-16	***
exper	0.16661	0.01605	10.382	< 2e-16	***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Result using summ command after installing jtools (summ(reg_wage))

$F(5, 1283) = 122.61, p = 0.00$

$R^2 = 0.32$

$Adj. R^2 = 0.32$

Standard errors: OLS

	Est.	S.E.	t val.	p
(Intercept)	-7.18	1.02	-7.07	0.00
female	-3.07	0.36	-8.43	0.00
nonwhite	-1.57	0.51	-3.07	0.00
union	1.10	0.51	2.17	0.03
education	1.37	0.07	20.79	0.00
exper	0.17	0.02	10.38	0.00

Measuring goodness of fit in multiple regression analysis

- The goodness of fit of the estimated model is understood in terms of value of R^2 (coefficient of multiple determination) and R^2 statistic provides a measure of proportion of total variation in the dependent variable that is explained by the independent/explanatory variables in the model.

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{ESS}{TSS}$$

Adj R^2 penalizes R^2 for adding additional explanatory variables in the model so that Adj R^2 values of models having different number of explanatory variables become comparable.

However, Adj R^2 is not only statistic (criteria) used by the researchers to select the best fit model from a set of alternative models.

other model selection criteria that

- Akaike Information Criteria (AIC)

$$AIC = -2 * \log(L) + 2 * k$$

- Schwarz Bayesian Criteria (SBC) (BIC)

$$BIC = -2 * \log(L) + k * \ln(n)$$

- Hannan-Quinn Criteria (HQC)

$$HQ = -2 * \log(L) + 2 * k * \ln(\ln(n))$$

While selecting a model based on these criteria, we select the one that reports minimum values for all these criteria (statistics) compared to an alternative model.

Model Selection

- ####Model Selection criteria
- `anova(reg_wage1, reg_wage)`
- ####By using AIC
- `AIC(reg_wage1, reg_wage)`
- `BIC(reg_wage1, reg_wage)`

- `library(car)`
- `outlierTest(reg_wage)`

Usefull Function/Command

Function	Action
<code>summary()</code>	Displays detailed results for the fitted model
<code>coefficients()</code>	Lists the model parameters (intercept and slopes) for the fitted model
<code>confint()</code>	Provides confidence intervals for the model parameters (95% by default)
<code>fitted()</code>	Lists the predicted values in a fitted model
<code>residuals()</code>	Lists the residual values in a fitted model
<code>anova()</code>	Generates an ANOVA table for a fitted model, or an ANOVA table comparing two or more fitted models
<code>vcov()</code>	Lists the covariance matrix for model parameters
<code>AIC()</code>	Prints Akaike's Information Criterion
<code>plot()</code>	Generates diagnostic plots for evaluating the fit of a model
<code>predict()</code>	Uses a fitted model to predict response values for a new dataset

Useful functions for regression diagnostics (car package)

Function	Purpose
<code>qqPlot()</code>	Quantile comparisons plot
<code>durbinWatsonTest()</code>	Durbin–Watson test for autocorrelated errors
<code>crPlots()</code>	Component plus residual plots
<code>ncvTest()</code>	Score test for nonconstant error variance
<code>spreadLevelPlot()</code>	Spread-level plots
<code>outlierTest()</code>	Bonferroni outlier test
<code>avPlots()</code>	Added variable plots
<code>influencePlot()</code>	Regression influence plots
<code>scatterplot()</code>	Enhanced scatter plots
<code>scatterplotMatrix()</code>	Enhanced scatter plot matrixes
<code>vif()</code>	Variance inflation factors

Multicollinearity

- Multicollinearity refers to a situation where there are high inter correlations among the explanatory variables of a multiple regression model. Multicollinearity problem arises only in the context of multiple regressions, it is considered as a problem because when the explanatory variables are highly correlated, most of their variation is common so that there is little variation unique to each variable.
- In empirical econometrics, while estimating multiple regression model, quite often we obtain unsatisfactory results in the sense that a good number of the estimated coefficients are found to be statistically insignificant. This happens when variances and hence standard errors of the estimated coefficients are large. This is possible when there is little variation in explanatory variables or high inter correlations among the explanatory variables or both.

Overall test (F test) -----Reject

Individual test (t test) -----Do not reject

(problem of multicollinearity)

Perfect correlation between two explanatory variables ----- (-1 or +1)

We can not run the regression

Y	X1	X2	X3	X4	X5
		Highly correlated			

Sources of Multicollinearity

- Multicollinearity may arise for several reasons
 - 1) Multicollinearity may arise because of faulty data collection method. For example, sampling over a limited range of values of explanatory variables creates this problem.
 - 2) Multicollinearity problem arises when the explanatory variables share a common time trend. This is the case in time series regressions. For instance, in regression of GDP on money supply and prices, the two explanatory variables (money supply and prices) are likely to be highly correlated because when money supply rises in an economy, price level also rises.
 - 3) When lagged values of the same variable are included as explanatory variables, we are likely to have multicollinearity problem. For example, in a time series regression of area under cultivation for a crop (say wheat) on its current and past prices, we may have multicollinearity problem because prices of the crop at different time points are generally correlated.
 - 4) There are many cross section regressions where we face high inter correlations among the explanatory variables.
 - For example, in a regression of consumption expenditure of persons on their income and education levels, we typically have strong correlation between income and education variables as the persons reporting higher incomes are usually found to be more educated.

Sources

5. Multicollinearity arises for faulty specification of the model. For example, adding polynomial terms (X and X^2) when X ranges are small will create this problem.
6. Multicollinearity arises in the over determined model where number of explanatory variables (k) is greater than number of observations (n).
7. In a situation of dummy variable trap we have multicollinearity problem. In this situation, the model includes an intercept term and the number of dummies is equal to number of categories.

Consequences of Multicollinearity

- No multicollinearity
- Perfect multicollinearity
 - under perfect multicollinearity it is not possible to compute the values of OLS estimates
- Imperfect multicollinearity

Detection of multicollinearity

- 1) Correlation matrix
- 2) F statistic and t statistic
- 3) Klen's rule of thumb
- 4) Variance inflation factor (VIF)
- 5) Tolerance method (TOL)
 - $VIF = 1$ (no multicollinearity)
 - $1 < VIF < 5$ (less multicollinearity)
 - $5 \leq VIF \leq 10$ (Moderate multicollinearity)
 - $VIF > 10$ (High multicollinearity)

Correlation matrix

- Since multicollinearity is caused by inter correlation among the explanatory variables, some idea about this problem may be obtained by computing simple or zero order correlations between the explanatory variables. To understand these correlations, the researchers usually obtain the correlation matrix by using different software.
- However, it is to be remembered that using simple correlation coefficient to understand presence of multicollinearity is a valid procedure when the model has two explanatory variables. When the model includes more than two explanatory variables instead of simple correlations, we should consider the partial correlations which examine the influence of one variable upon another after eliminating the effects of all other variables. The statistical significance of partial correlation coefficient may also be tested by applying the t test procedure, but formula for computation of t is different here.

-

F statistic and t statistic

- F-statistic and t-statistic
- In the data having multicollinearity problem we observe that its presence generates the high variances for the OLS estimates, thereby providing low t-ratios and hence statistically **insignificant regression** results that is there seems contradictory result in F-test and t-test.

Klein's Rule of thumb:

Klein (1962) suggested a rule of thumb according to which multicollinearity would be regarded as a problem if $R_Y^2 < R_K^2$, where R_Y^2 is the squared multiple correlation coefficient between the dependent variable Y_i and explanatory variables $X_{1i}, X_{2i}, \dots, X_{ki}$, and R_K^2 is squared multiple correlation coefficient between K^{th} explanatory variable and other explanatory variables.

Variance Inflation Factor

|
When the multicollinearity is present, the variance of the estimated coefficient of K^{th} explanatory variable is measured by

$$\text{Var}(\hat{\beta}_k) = \frac{\sigma_\varepsilon^2}{\sum x_{ki}^2 (1 - R_k^2)}$$

Under the ideal situation when there is no multicollinearity, $R_k^2 = 0$, so that,

$$\text{Var}(\hat{\beta}_k) = \frac{\sigma_\varepsilon^2}{\sum x_{ki}^2}$$

The VIF compares these two situations by taking a ratio of the two variances.

Decision criteria by using VIF

Thus,

$$\text{VIF}(\hat{\beta}_k) = \frac{\sigma_\varepsilon^2}{\sum x_{ki}^2 (1 - R_k^2)} / \frac{\sigma_\varepsilon^2}{\sum x_{ki}^2} = \frac{1}{(1 - R_k^2)}$$

VIF = 1 (no multicollinearity)

$1 < \text{VIF} < \underline{5}$ (less multicollinearity)

$5 \leq \text{VIF} \leq 10$ (Moderate multicollinearity)

$\text{VIF} > 10$ (High multicollinearity)

Tolerance method (TOL) = $1/\text{VIF}$ $= 1 - R_k^2$

Remedial Technique

- 1) By increasing sample size
- 2) Transformation of variables
- 3) Using extraneous estimates
- 4) Dropping the variables
 - We have to drop the variable having highest VIF (>10) at a time. Again run the regression of remaining variables check whether there is VIF >10 or not and repeat the same process until we get VIF < 10 .

QQ test for normality

- `qqPlot(reg_wage)`
- `qqPlot(reg_wage, id.method="identify", simulate=TRUE, main="Q-Q Plot")`
- `##To check the multicollinearity`
- `vif(reg_wage)`
- `residualPlot(reg_wage)`
- `##To check the autocorrelation`
- `durbinWatsonTest(reg_wage)`

Usefull Function/Command

Function	Action
<code>summary()</code>	Displays detailed results for the fitted model
<code>coefficients()</code>	Lists the model parameters (intercept and slopes) for the fitted model
<code>confint()</code>	Provides confidence intervals for the model parameters (95% by default)
<code>fitted()</code>	Lists the predicted values in a fitted model
<code>residuals()</code>	Lists the residual values in a fitted model
<code>anova()</code>	Generates an ANOVA table for a fitted model, or an ANOVA table comparing two or more fitted models
<code>vcov()</code>	Lists the covariance matrix for model parameters
<code>AIC()</code>	Prints Akaike's Information Criterion
<code>plot()</code>	Generates diagnostic plots for evaluating the fit of a model
<code>predict()</code>	Uses a fitted model to predict response values for a new dataset

Useful functions for regression diagnostics (car package)

Function	Purpose
<code>qqPlot()</code>	Quantile comparisons plot
<code>durbinWatsonTest()</code>	Durbin–Watson test for autocorrelated errors
<code>crPlots()</code>	Component plus residual plots
<code>ncvTest()</code>	Score test for nonconstant error variance
<code>spreadLevelPlot()</code>	Spread-level plots
<code>outlierTest()</code>	Bonferroni outlier test
<code>avPlots()</code>	Added variable plots
<code>influencePlot()</code>	Regression influence plots
<code>scatterplot()</code>	Enhanced scatter plots
<code>scatterplotMatrix()</code>	Enhanced scatter plot matrixes
<code>vif()</code>	Variance inflation factors

Autocorrelation

Pravat Uprety

Central Department of Statistics

Tribhuvan University

Autocorrelation

- One of the assumptions of the Classical Linear Regression Model (CLRM) is that the disturbance term of the model is independent.

$$\text{Cov}(\epsilon_t, \epsilon_s) = E(\epsilon_t, \epsilon_s) = 0 \text{ for } t \neq s.$$

This feature of regression disturbance is known as **serial independence or non autocorrelation**. It implies that the value of disturbance term in one period is not correlated with its value in another period.

In time series the disturbance term at period t may be related with the disturbance term at $t-1, t-2, \dots$ and $t+1, t+2, \dots$ and so on. In that case $\text{cov}(\epsilon_t, \epsilon_s) \neq 0$ for $t \neq s$ and we say that the disturbances are autocorrelated.

Specification of Autocorrelation Relationship

$$\epsilon_t = \rho \epsilon_{t-1} + u_t \text{ ----- (1)}$$

$$E(u_t) = 0$$

$$\text{Var}(u_t) = E(u_t^2) = \sigma_u^2 \text{ for all } t.$$

u_t is normally distributed

$$E(u_t u_{t-1}) = 0$$

Equation 1 is known as first order autoregression scheme and it is denoted by AR (1).

By successive substitution for $\epsilon_t, \epsilon_{t-1}, \dots$ in (1) we get,

$$\begin{aligned}
 \epsilon_t &= \rho(\rho \epsilon_{t-2} + u_{t-1}) + u_t \\
 &= \rho^2 \epsilon_{t-2} + \rho u_{t-1} + u_t \\
 &= \rho^2 (\rho \epsilon_{t-3} + u_{t-2}) + \rho u_{t-1} + u_t \\
 &= \rho^3 \epsilon_{t-3} + \rho^2 u_{t-2} + \rho u_{t-1} + u_t \\
 &= u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \rho^3 u_{t-3} + \dots \text{-----}(2)
 \end{aligned}$$

This shows that under the first-order autoregressive scheme, the effect of past disturbances wears off gradually as $|\rho| < 1$.

- ρ represents the correlation coefficient between ϵ_t and ϵ_{t-1} (first order autocorrelation), ρ^2 is the correlation coefficient between ϵ_t and ϵ_{t-2} (second order autocorrelation) and ρ^s is the correlation coefficient between ϵ_t and ϵ_{t-s} (S order autocorrelation).
- When $\rho = 0$ there is no autocorrelation because ϵ_t becomes u_t and which does not suffer from autocorrelation.
- The strength of autocorrelation becomes high as ρ approaches to unity (+1).
- If ρ approaches to -1, the strength of autocorrelation becomes high again.

Mean

$$E(\epsilon_t) = 0$$

Variance of ϵ_t

$$\text{Var}(\epsilon_t) = \frac{\sigma_u^2}{1 - \rho^2} = \sigma_\epsilon^2$$

Covariance of ϵ_t and ϵ_{t-1}

$$\text{Cov}(\epsilon_t, \epsilon_{t-1}) = \rho \sigma_\epsilon^2$$

Consequences of autocorrelation are

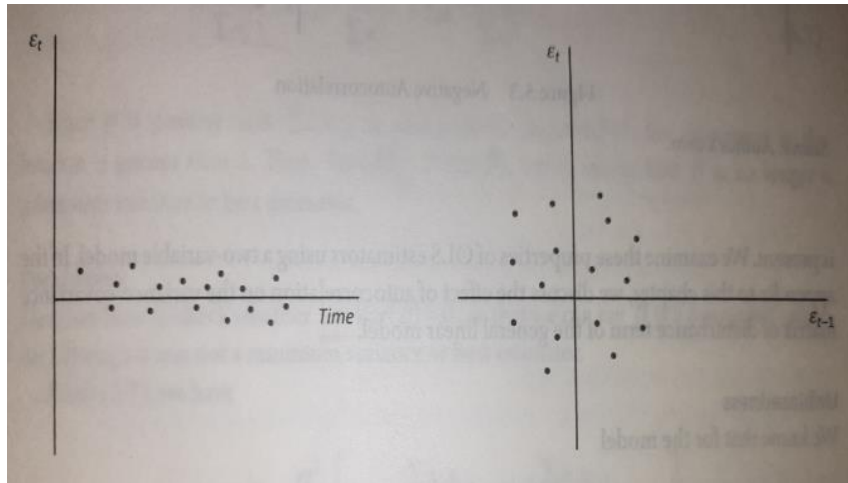
- 1) The OLS estimators are still unbiased and consistent.
- 2) The OLS estimators are no longer minimum variance or best estimators. Hence, they are not efficient and BLUE.
- 3) If we disregard the problem of autocorrelation and believe that all assumptions are valid, following problems will arise
 - The estimated variance of disturbance term will be under estimate of its true variance.
 - The standard error of the estimated slope coefficient will be much smaller if it is computed usual OLS formula. Which will provide spurious (wrong) impression about the statistical significance.
 - The usual t and F test will become invalid.

Detection

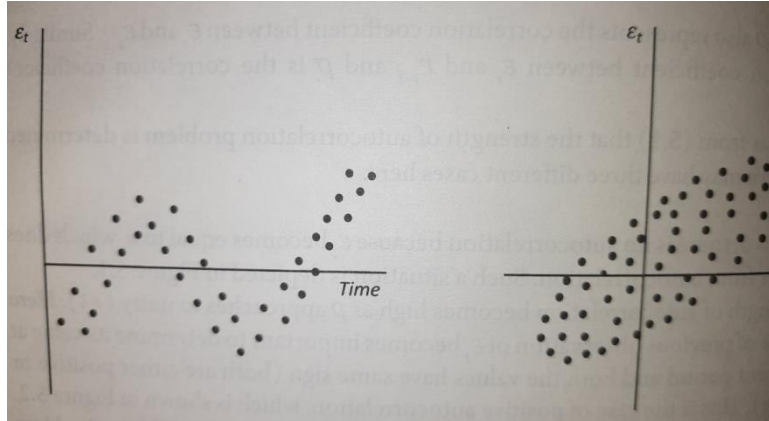
- **Graphical Method (Residual Plot)**

- Graph of ϵ_t against time
- Graph of ϵ_t against ϵ_{t-1}

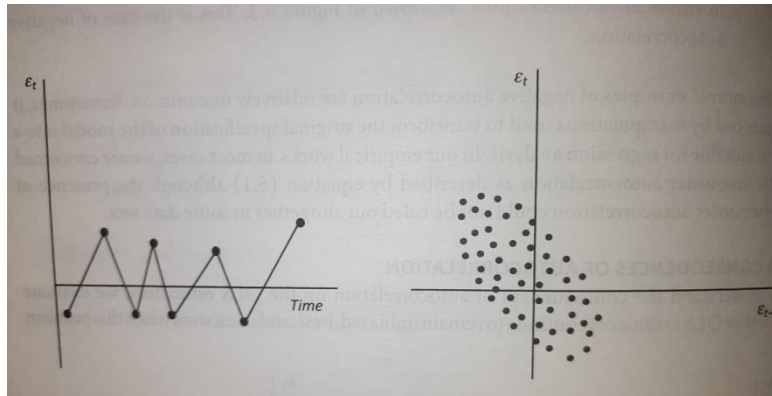
1. When $\rho = 0$ there is no autocorrelation because ϵ_t becomes u_t and which does not suffer from autocorrelation.



2) The strength of autocorrelation becomes high as ρ approaches to unity (+1). Here the value of previous observation of ϵ_t becomes important to determine its value at the current period and both the values have same sign (both are either positive or negative). This is the case of +ve autocorrelation.



3) If ρ approaches to -1, the strength of autocorrelation becomes high again. Here also the past values of the disturbance term become important in determining its value in the current period but the signs of the disturbance term switch in consecutive observations. This is the case of -ve autocorrelation.



Durbin Watson Test (D-W test)

This is the simplest and most widely used test for autocorrelation. It is based on following assumptions:

- a) The regression model includes a constant or intercept term.
- b) We are examining presence of first order autocorrelation.
- c) The regression model does not include a lagged dependent variable as an explanatory variable.

Now, to understand how the test is performed, consider the model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + \epsilon_t$$

$$\text{Where } \epsilon_t = \rho \epsilon_{t-1} + u_t \quad |\rho| < 1$$

Null hypothesis (H_0): $\rho = 0$

$$\text{Or } H_0 : \rho \geq 0$$

$$\text{Or } H_0 : \rho \leq 0$$

Alternative hypothesis (H_1): $\rho \neq 0$

$$\text{Or } H_1 : \rho < 0$$

$$\text{Or } H_1 : \rho > 0$$

The Durbin-Watson (D-W) statistic is

$$d = \frac{\sum_{t=2}^n (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^n \hat{\varepsilon}_t^2} = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} = \frac{\sum_{t=2}^n e_t^2 + \sum_{t=2}^n e_{t-1}^2 - 2 \sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2}$$

For large sample

$$d = \frac{2 \sum_{t=1}^n e_t^2 - 2 \sum_{t=1}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2} = 2 \left[1 - \frac{\sum_{t=1}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2} \right] = 2(1-\rho)$$

Decision Rule

From D-W table

We can get the value of d_L and d_U for given value of n , k and α .

- If $0 < d < d_L$ reject H_0 (+ve autocorrelation)
- If $d_L \leq d \leq d_U$ no conclusion
- If $d_U < d < 4 - d_U$ do not reject H_0 (no significant autocorrelation)
- If $4 - d_U \leq d \leq 4 - d_L$ no conclusion
- If $4 - d_L < d \leq 4$ reject H_0 (-ve autocorrelation)

Limitations

- 1) It can not be used for testing higher order autocorrelation
- 2) This test is biased towards non-rejection of null hypothesis when a lagged dependent variable is included in the model as an explanatory variable.
- 3) It becomes inapplicable if the model does not contain intercept term.
- 4) Sometime, the test produces no conclusion result.
- 5) This test is not robust for small sample.

Example

Suppose that the residuals for a set of data collected over 12 consecutive time periods were as follows:

Time Period	1	2	3	4	5	6	7	8	9	10	11	12
Residual	+5	+6	+3	-4	-4	+2	0	-5	-3	+4	-2	-2

Compute the Durbin-Watson statistic. At the 0.05 level of significance, is there evidence of positive autocorrelation among the residuals?

Breusch-Godfrey Lagrange Multiplier Test (LM test)

Given the problems of Durbin-Watson test, it is always good to use some other test for autocorrelation. Although there are many alternative here, the most widely used test is the Breusch- Godfrey Lagrange Multiplier test (BG test) developed by [Breusch \(1978\)](#) and [Godfrey \(1978\)](#). This test can pick up higher order autocorrelation and can be performed using many econometrics software packages, including [Eviews/Stata/R/python](#). It is also a more powerful test as it is not biased towards non rejection of the null hypothesis when lagged dependent variables is included in the model as an explanatory variable.

To understand the BG test procedure, consider the model,

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + \epsilon_t \quad \text{-----(1)}$$

Where,

$$\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + \dots + \rho_p \epsilon_{t-p} + u_t \quad \text{-----(2)}$$

Null hypothesis (H_0): $\rho_1 = \rho_2 = \rho_3 = \dots = \rho_p = 0$ (no auto correlation)

Alternative hypothesis (H_1): at least one ρ is not zero (autocorrelation)

Steps

- 1) Estimate equation (1) by OLS and estimate $\hat{\varepsilon}_t = e_t$
- 2) Run the auxiliary regression of e_t on $X_{1t}, X_{2t}, \dots, X_{kt}, e_{t-1}, e_{t-2}, \dots, e_{t-p}$.
- 3) To test the validity of above null hypothesis compute LM statistic as follows
$$LM = (n-p) R^2$$

Where R^2 is coefficient of determination of auxiliary equation

n is no of observations

p is order of autocorrelation

This LM statistic follows a chi square distribution with degrees of freedom P .

- 4) If this $\chi^2 \leq \chi^2_{p, \alpha}$ (tabulated)
We do not reject H_0 (no autocorrelation)
If this $\chi^2 > \chi^2_{p, \alpha}$ (tabulated)
We reject H_0 (autocorrelation)

Remedial Measure

- When the value of ρ is known (Transformation)

Consider the model,

$$Y_t = \alpha + \beta X_t + \epsilon_t \text{ -----(1)}$$

Where ϵ_t is assumed to follow first order autocorrelation, so that

$$\epsilon_t = \rho \epsilon_{t-1} + u_t \text{ -----(2)}$$

Now lagging equation (1) by one period and multiplying it by ρ .

$$\rho Y_{t-1} = \rho \alpha + \rho \beta X_{t-1} + \rho \epsilon_{t-1} \text{ -----(3)}$$

Now, subtracting equation (3) from equation (1)

$$Y_t - \rho Y_{t-1} = (1-\rho) \alpha + \beta (X_t - \rho X_{t-1}) + (\epsilon_t - \rho \epsilon_{t-1}) \text{ -----(4)}$$

$$\text{Therefore } Y_t^* = \alpha^* + \beta X_t^* + u_t \quad [\text{since } u_t = \epsilon_t - \rho \epsilon_{t-1}]$$

- It is clear that we lost one observation with this transformation. In order to avoid this loss of observation, it is suggested that Y_1 and X_1 should be transformed for first observation as follows:

$$Y_1^* = Y_1 \sqrt{1 - \rho^2} \quad X_1^* = X_1 \sqrt{1 - \rho^2}$$

The transformation that generated Y_t^* , X_t^* and α^* is known as quasi- differencing or generalized differencing.

When the value of ρ is unknown

Cochrane – Orcutt (1949) Iterative Procedure

Under this model, we estimate the following equation by OLS method

$$Y_t = \alpha + \beta X_t + \epsilon_t$$

Then we obtain $\hat{\epsilon}_t = e_t$

And which is used to calculate $\hat{\rho}$ as

$$\hat{\rho} = \frac{\sum e_t e_{t-1}}{\sum e_t^2}$$

Then,

$$Y_t - \hat{\rho} Y_{t-1} = \alpha^* + \beta (X_t - \hat{\rho} X_{t-1}) + (\epsilon_t - \hat{\rho} \epsilon_{t-1})$$

Thus a two step estimation is involved here

Suppose, α' and β' are Cochrane-Orcutt estimation of α and β obtaining new set of residuals

$$\epsilon'_t = Y_t - \alpha' - \beta' X_t$$

Also estimating new ρ say $\hat{\rho}$ as

$$\hat{\rho} = \frac{\sum e'_t e'_{t-1}}{\sum e'^2_t}$$

Again use $\hat{\rho}$ to construct the model,

$$Y_t - \hat{\rho} Y_{t-1} = \alpha^* + \beta (X_t - \hat{\rho} X_{t-1}) + (\epsilon_t - \hat{\rho} \epsilon_{t-1})$$

And this procedure is continued until estimated values of $\hat{\alpha}$ and $\hat{\beta}$ converge.

Note: Taking difference (in time series)

Command in R

Data File: hprice1

```
library(car)
```

```
reg_hprice<-lm(price~lotsize+sqrft+bdrms, data=hprice1)
```

```
summary(reg_hprice)
```

```
summ(reg_hprice)
```

Residual Plot

```
residualPlot(reg_hprice)
```

Durbin Watson Test

```
durbinWatsonTest(name of fitted model)
```

```
durbinWatsonTest(reg_hprice)
```

Breusch-Godfrey test

- ```
bgtest(reg_hprice)
```



# Heteroskedasticity

Pravat Uprety  
Central Department of Statistics  
Tribhuvan University

# Heteroskedasticity

For the sample two variable model

$$Y_i = \alpha + \beta X_i + \epsilon_i \text{ -----(1)}$$

We assumed that the variances of disturbance term  $\epsilon_i$  is constant for all observations

$$\text{i.e. } \text{var}(\epsilon_i) = E(\epsilon_i^2) = \sigma^2 \text{ (constant for all } i) \text{ -----(2)}$$

- this feature of disturbances term of the regression model is known as **homoskedasticity**.

However, it is quite common in regression analysis to have cases where the variance of disturbance term becomes variable rather than remaining constant.

- In this situation the disturbance is said to be **heteroskedasticity**.
- i.e.  $\text{var}(\epsilon_i) = E(\epsilon_i^2) = \sigma_i^2 \text{ -----(3)}$
- which means that the variance of disturbance term can change for every different observation in the sample  $i = 1, 2, \dots, n$ .

# Sources

- 1) When we are dealing with micro-economic or cross sectional data, we are very likely to have a heteroskedasticity problem.
- 2) Presence of outliers in data may cause heteroskedasticity.
- 3) Heteroskedasticity may arise if some relevant variables have been mistakenly omitted.
- 4) Inclusion of explanatory variables in the model whose distributions are skewed.
- 5) Heteroskedasticity may also arise due to incorrect data transformation and incorrect functional form.

# Consequences

## Unbiased ness

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

This shows that  $\hat{\beta}$  remains unbiased when the disturbance term of the model  $\epsilon_i$  is heteroskedasticity.

## Bestness

$$\text{Thus } \text{Var}(\hat{\beta})|_{\text{heteroskedasticity}} > \text{Var}(\hat{\beta})|_{\text{homoskedasticity}}$$

So, there is no longer a minimum variance and hence not best estimator.

$\hat{\beta}$  is unbiased but not the best.

## Consistence

:  $\hat{\beta}$  is consistent when the disturbance term is heteroskedastic.

# Consequences

- The OLS estimators continue to remain unbiased and consistent under heteroskedasticity.
- Heteroskedasticity increases the variances of the distributions of estimator of  $B$  thereby turning the OLS estimators inefficient (not best

Heteroskedasticity also affects the variance of OLS and their standard error. In fact, the presence of heteroskedasticity, in general causes the OLS method to underestimate the variances and hence standard error of the estimators. As a consequence, we have higher than expected values of  $t$  and  $F$  statistic. As the OLS estimators are unbiased under heteroskedasticity, the forecasts generated on the basis of the estimated model will also be unbiased.

# Detection Techniques

## Graphical method

We may graphically examine the presence of heteroskedasticity by plotting the squared residuals ( $\epsilon_i^2$ ) against the explanatory variable ( $X_i$ ) to which it is suspected the disturbance variance is related. Since  $\epsilon_i^2$  is unknown, its proxy measure  $\epsilon_i^2 = e_i^2$  is used.



# Breusch-Pagan-Godfrey Test

Breusch-Pagan Godfrey (1978) developed a Lagrange Multiplier (LM) test to examine the presence of heteroskedasticity in data

Considering the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \epsilon_i \quad \text{-----}(1)$$

And suppose that

$$\text{Var}(\epsilon_i) = \sigma_i^2 = f(\gamma_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + \dots + \gamma_r Z_{ri})$$

This implies that  $\text{Var}(\epsilon_i)$  is the function of non-stochastic  $Z$ s. Here  $Z$ s represent a set of variables that we think determine the variance of the disturbance term  $\epsilon_i$ . Usually, the explanatory variables of (1) are used for  $Z$ s.

The steps involved in the Breusch-Pagan Godfrey test are the following.

Step 1: Estimate model (1) by OLS method and obtain the estimated residuals  $\hat{e}_i = \hat{\varepsilon}_i$ .

Step 2: Run the auxiliary regression

$$e_i^2 = \gamma_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + \dots + \gamma_r Z_{ri} + v_i \quad (2)$$

Step 3: Construct the null and alternative hypothesis

$H_0: \gamma_1 = \gamma_2 = \dots = \gamma_r = 0$  (Homoskedasticity)

$H_1$ : At least one of the  $\gamma_s$  is non zero (Heteroskedasticity)

Step 4: Compute  $LM = nR^2$  where  $n$  = number of observations used to estimate auxiliary regression model (2) and  $R^2$  is the coefficient of determination of this regression.

(Note that LM-statistic follows a  $\chi^2$  distribution with degrees of freedom  $r$ )

Step 5: If  $LM = \chi^2 \leq \chi_r^2$  at  $\alpha\%$

Then we do not reject  $H_0$

If  $LM = \chi^2 > \chi_r^2$  at  $\alpha\%$

Then we reject  $H_0$

And we conclude that there is significant evidence of heteroskedasticity in data.

[If we have corresponding p-value then we can use the p-value approach]

# Park Test

Considering the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \epsilon_i \quad \text{-----(1)}$$

|

Step 1: Estimate model (1) by OLS method and obtain the estimated residuals  $\hat{e}_i = \hat{\epsilon}_i$ .

Step 2: Run the auxiliary regression


$$\ln(e_i^2) = \alpha_0 + \alpha_1 \ln Z_{1i} + \alpha_2 \ln Z_{2i} + \dots + \alpha_r \ln Z_{ri} + v_i \quad \text{-----(2)}$$

Step 3: Construct the null and alternative hypothesis

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_r = 0$  (Homoskedasticity)

$H_1$ : At least one of the  $\alpha_s$  is non zero (Heteroskedasticity)

Step 4: Compute  $LM = nR^2$  where  $n$  = number of observations used to estimate auxiliary regression model (2) and  $R^2$  is the coefficient of determination of this regression.  
(Note that LM-statistic follows a  $\chi^2$  distribution with degrees of freedom  $r$ )

 Step 5: If  $LM = \chi^2 \leq \chi_r^2$  at  $\alpha\%$   
Then we do not reject  $H_0$

If  $LM = \chi^2 > \chi_r^2$  at  $\alpha\%$   
Then we reject  $H_0$

And we conclude that there is significant evidence of heteroskedasticity in data.  
[If we have corresponding p-value then we can use the p-value approach]

# Goldfeld -Quandt Test

- Goldfeld and Quandt (1965) proposed a test of heteroskedasticity that may be applied when **one of the explanatory variables is suspected to be the heteroskedasticity culprit**.
- The basic idea behind this test is that if the variances of the disturbances are the same across all observations (i.e., homoskedasticity), then the variance of one part of the sample should be the same as the variance of another part of the sample. Under this test, we start by assuming that  $\sigma_i^2$  is proportional to the size of  $X_i$ . It is also assumed that the disturbance term of the model ( $\epsilon_i$ ) is normally distributed and satisfies other regression assumptions.
- The hypothesis are
  - Null hypothesis ( $H_0$ ):  $\text{Var}(\epsilon_i | X_i) = \sigma^2$ , a constant (Homoskedasticity)
  - Alternative hypothesis ( $H_1$ ):  $\text{Var}(\epsilon_i | X_i) = \sigma_i^2$ , a variable (Heteroskedasticity)

# Steps

- Step 1: Identify the variable to which the variance of disturbance term is suspected to be related.
- Step 2: Sort the raw data in ascending order (starting with lowest and going to be highest) of the values of  $X_i$ .
- Step 3: Cut out some central observations ( $c$ ), breaking data set in two distinct sets-first one with low values of  $X_i$ . Note that there is no clear rule about how many observations to be cut out. Goldfeld and Quandt suggested the value of  $c$  as 8 if the sample size is about 30.
- Step 4: Fit separate regression by OLS to the first and last  $(n-c)/2$  observations.
- Step 5: Compute residual sum squares for the two regressions, which are denoted by  $RSS_1$  and  $RSS_2$ , respectively.

Step 6: Compute the F-ratio of two RSSs as

$$F = \frac{RSS_2 / df_2}{RSS_1 / df_1}$$

Where  $df_1 = df_2 = (n-c)/2 - k = (n-c-2k)/2$   
(here k is no of parameters to be estimated)

Step 7: Compute critical value of F with  $df_1$ ,  $df_2$  and level of significance ( $\alpha$ ).

If cal F > Critical value of F

We reject  $H_0$

Then we conclude that there is heteroskedasticity in data

# Limitations

- 1) The determination of an appropriate value of  $c$  sometimes becomes difficult.
- 2) In a model that involves a large number of explanatory variables, there may be difficulty in identifying the X-variable with which to sort the data.
- 3) It does not consider cases where heteroskedasticity is caused by more than one explanatory variable.



# Remedial Techniques

Log Transformation

Weighted least squares method

# Generalized Least Squares (GLS)

- Consider the model
- $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$  -----(1)
- And suppose that there is heteroskedasticity so that  $\text{Var}(\epsilon_i) = \sigma_i^2$ .
- If we know specifically the form of heteroskedasticity we may utilize such information to suitably transform our model to overcome the problem of heteroskedasticity and obtain efficient estimates for unknown parameters of our model.

# Case a: Suppose heteroskedasticity is of the form $\sigma_i^2 = \sigma^2 Z_i^2$

Now dividing the model (1) by  $Z_i$ , so that the transformed model is

$$\frac{Y_i}{Z_i} = \beta_1 \frac{1}{Z_i} + \beta_2 \frac{X_{2i}}{Z_i} + \beta_3 \frac{X_{3i}}{Z_i} + \frac{\varepsilon_i}{Z_i}$$

$$Y_i^* = \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + \beta_3 X_{3i}^* + v_i \text{ -----(2)}$$

Then there is no heteroskedasticity in the transformed model (2) because

$$\text{Var}(v_i) = E(v_i^2) = E\left(\frac{\varepsilon_i}{Z_i}\right)^2 = \frac{E(\varepsilon_i^2)}{Z_i^2} = \frac{\sigma_i^2}{Z_i^2} = \frac{\sigma^2 Z_i^2}{Z_i^2} = \sigma^2 \text{ (Constant)}$$

Case b: Heterokedasticity is of the form  $\sigma_i^2 = \sigma^2 X_{1i}$

In this case we transform model (1) by dividing through  $\sqrt{X_{1i}}$

$$\frac{Y_i}{\sqrt{X_{1i}}} = \beta_1 \frac{1}{\sqrt{X_{1i}}} + \beta_2 \frac{X_{2i}}{\sqrt{X_{1i}}} + \beta_3 \frac{X_{3i}}{\sqrt{X_{1i}}} + \frac{\varepsilon_i}{\sqrt{X_{1i}}}$$

$$Y_i^* = \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + \beta_3 X_{3i}^* + v_i \text{ -----(3)}$$

|

For the transformed model, the variance of the disturbance term is again constant

$$\text{Var}(v_i) = E(v_i^2) = E\left(\frac{\varepsilon_i}{\sqrt{X_{1i}}}\right)^2 = \frac{E(\varepsilon_i^2)}{X_{1i}} = \frac{\sigma_i^2}{X_{1i}} = \frac{\sigma^2 X_{1i}}{X_{1i}} = \sigma^2 \text{ (Constant)}$$

Thus the transformed model (3) is free from the problem of heteroskedasticity.

Case c Hetersokedasticity is of the form  $\sigma^2 X_{1i}^2$

Here the transformed model is

$$\frac{Y_i}{X_{1i}} = \beta_1 \frac{1}{X_{1i}} + \beta_2 \frac{X_{2i}}{X_{1i}} + \beta_3 \frac{X_{3i}}{X_{1i}} + \frac{\varepsilon_i}{X_{1i}}$$

$$Y_i^* = \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + \beta_3 X_{3i}^* + v_i \text{ -----(3)}$$

Then there is no heteroskedasticity in the transformed model (3) because

$$\text{Var}(v_i) = E(v_i^2) = E\left(\frac{\varepsilon_i}{X_{1i}}\right)^2 = \frac{E(\varepsilon_i)^2}{X_{1i}^2} = \frac{\sigma^2 X_{1i}^2}{X_{1i}^2} = \sigma^2 \text{ (Constant)}$$

The above procedures of estimating the transformed model to overcome the heteroskedasticity problem are known as generalized least squares.

## Weighted Least Squares (WLS)

The above cases are also referred to as cases of weighted least squares (WLS) method. This is because we may view  $\frac{1}{Z_i}$  in (1),  $\frac{1}{\sqrt{X_{1i}}}$  in (2) and  $\frac{1}{X_{1i}}$  in (3) as weight ( $w_i$ )

Then above models can be expressed as

$$(w_i Y_i) = \beta_1 w_i + \beta_2 (w_i X_{2i}) + \beta_3 (w_i X_{3i}) + (\epsilon_i w_i) \text{ -----(4)}$$

# Detection Techniques (in R)

## Normal Q-Q plot checks

```
qqPlot(reg_hprice)
```

## Breuch Pagan Test

- library(lmtest)
- bptest(reg)

```
bptest(reg_hprice)
```

## White test

```
bptest(reg_hprice, ~fitted(reg_hprice)+I(fitted(reg_hprice)^2))
```

# Log transformation

- ###Taking log

```
reg_lnhprice<-lm(log(price)~log(lotsize)+log(sqrft)+log(bdrms),
data=hprice1)
```

## BP test

```
bptest(reg_lnhprice)
```

## White test

```
bptest(reg_lnhprice, ~fitted(reg_lnhprice)+I(fitted(reg_lnhprice)^2))
```



# Another Example: SMOKE.dta (Wooldridge)

- `install.packages("wooldridge")`
- `library(wooldridge)`
- `data(smoke, package="wooldridge")`  
`head(smoke)`
- `data(smoke, package = 'Wooldridge')`
- Or open SMOKE.dta file
- `dim(SMOKE)`
- `head(SMOKE)`

```
reg_smoke<-lm(cigs~log(income)+log(cigpric)+educ+age+l(age^2)+restaurn, data=SMOKE)
summary(reg_smoke)
bptest(reg_smoke)
```

- `logu2<-log(resid(reg_smoke)^2)`
- `varreg<-lm(logu2~log(income)+log(cigpric)+educ+age+l(age^2)+restaurn,  
data=SMOKE)`
- `##Weight`
- `w<-1/exp(fitted(varreg))`
- `wls<-  
lm(cigs~log(income)+log(cigpric)+educ+age+l(age^2)+restaurn,weight=w,  
data=SMOKE)`
- `summary(wls)`
- `bptest(wls)`

# Logistic Regression

Pravat Uprety

Central Department of Statistics

Tribhuvan University

## Models with dummy dependent variable

The model in which the dependent variable is a binary variable is also called binary choice model. In such a model, the dependent variable actually involves only two choices indicating presence or absence of an attribute or quality.

There are three important approaches to dealing with dummy dependent variable models or binary choice models and they are:

Linear Probability Model (LPM)

Logit Model

- Probit Model

## Linear Probability Model

To know what determines car ownership status of the families in a locality and supposing that car ownership status of the families is determined by their family incomes. For this, following points need to be noted

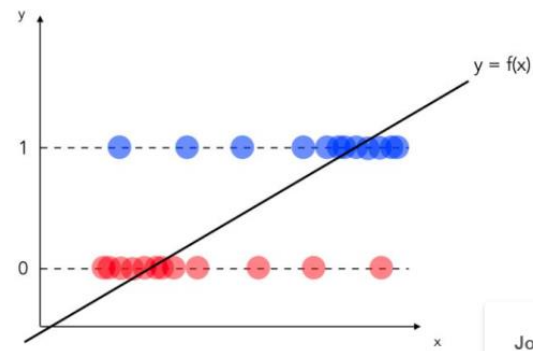
- i) Car status is a qualitative variable.
- ii) However, car ownership status may be quantified by using a binary or dummy variable that is assigned value 1 if the family owns a car and value 0 if the family does not own a car.
- iii) Family income is a usual quantitative variable

The model

$$Y_i = \alpha + \beta X_i + \epsilon_i \quad \text{-----(1)}$$

- Where,  $X_i$  = Family income

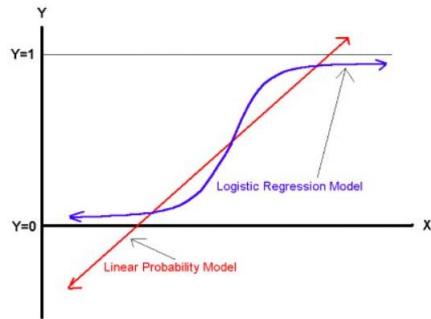
$Y_i = 1$  if the family owns a car  
= 0 otherwise



J0

## Logit Model

- The linear probability model (LPM) reveals that when the dependent variable is a binary variable, a non linear specification of the model appears more appropriate. Specifically it seems appropriate to fit some kind of S – shaped (or **sigmoid curve**) to the observed data points as following graph.



The sigmoid curve has following features:

- 1) It resembles as elongated S.
- 2) The tails of the sigmoid curve level off before reaching  $p=0$  or  $p=1$  so that the problem of impossible values of estimated probability is avoided.
- 3) More importantly, the sigmoid curve, resemble the cumulative distribution function (CDF) of a random variables. So we can choose a suitable CDF to represent the sigmoid curve to capture 0-1 representation for the dependent variable.

# Logistic Regression

- Binary logistic regression is a form of regression which is used when the dependent variable is a true or forced dichotomy and the independent variables are of any type.
- Logistic regression can be used to predict a categorical dependent variable on the basis of continuous and/or categorical independent variables; to determine the effect size of the independent variables on the dependent variable; to rank the relative importance of independent variables; to assess interaction effects; and to understand the impact of covariate control variables. The impact of predictor variables is usually explained in terms of odds ratios, which is the key effect size measure for logistic regression.

- Logistic regression applies maximum likelihood estimation after transforming the dependent into a logit variable.
- A logit is the natural log of the odds of the dependent equaling a certain value or not (usually 1 in binary logistic models, or the highest value in multinomial models).
- Logistic regression estimates the odds of a certain event (value) occurring. This means that logistic regression calculates changes in the log odds of the dependent, not changes in the dependent itself as does OLS regression.



## Specification of logit model

$$P_i = P(Y_i=1) = F(z_i) = \frac{1}{1+e^{-z_i}}$$

Where  $P_i$  is the probability of  $Y_i = 1$

$F(z_i)$  is the cdf of the cumulative logistic function

$Z_i = \alpha + \beta X_i$ , is a predictor variable

$e = 2.71828$

$$1 - P_i = 1 - \frac{1}{1 + e^{-z_i}} = \frac{1 + e^{-z_i} - 1}{1 + e^{-z_i}} = \frac{e^{-z_i}}{1 + e^{-z_i}} = \frac{1}{\frac{1 + e^{-z_i}}{e^{-z_i}}} = \frac{1}{1 + e^{z_i}}$$

$$\frac{P_i}{1 - P_i} = \frac{1}{1 + e^{-z_i}} / \frac{1}{1 + e^{z_i}} = \frac{1 + e^{z_i}}{1 + e^{-z_i}} = \frac{1 + e^{z_i}}{1 + \frac{1}{e^{z_i}}} = \frac{1 + e^{z_i}}{\frac{e^{z_i} + 1}{e^{z_i}}} = e^{z_i}$$

$$\text{So } \ln\left(\frac{P_i}{1 - P_i}\right) = z_i = \alpha + \beta X_i \text{ -----(2)}$$

Hence,  $\frac{P_i}{1-P_i}$  is called odds ratio in favour of the event occurring and  $\ln\left(\frac{P_i}{1-P_i}\right)$  is the log odds ratio (also called **logit of p**)

In model (2) we considered only one explanatory variable. But, if necessary, more explanatory variables can be added by supposing that  $Z_i$  is a linear function of a set of predictor variables.

$$Z_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_3 X_{3i} \dots\dots\dots + \beta_k X_{ki}$$

$$\log \left[ \frac{P(Y = 1)}{1 - P(Y = 1)} \right] = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_3 X_{3i} \dots\dots\dots + \beta_k X_{ki}$$

The logistic regression model is a special case of a GLM. The random component for the (success, failure) outcomes has a binomial distribution. The link function of  $\pi = P(Y = 1)$  is the *logit* function,  $\log[\pi/(1 - \pi)]$ , symbolized by “logit( $\pi$ ).” Logistic regression models are often called *logit models*. Whereas  $P(Y = 1)$  is restricted to the 0 to 1 range, the logit can be any real number. The real numbers are also the potential range for linear predictors. This model therefore does not have the structural limitation that the linear probability model has.

### Estimation of logit model

It is not possible to estimate the logit (2) by OLS method for two reasons

This is a non linear model

$\ln\left(\frac{P_i}{1 - P_i}\right)$  is not a familiar quantity.

# Logistic regression has many analogies to OLS regression:

logit coefficients correspond to b coefficients in the logistic regression equation;

the standardized logit coefficients correspond to beta weights;

and a pseudo R<sup>2</sup> statistic is available to summarize the overall strength of the model.

Unlike OLS regression, however, logistic regression does not assume linearity of relationship between the raw values of the independent variables and raw values of the dependent; does not require normally distributed variables; does not assume homoscedasticity; and in general has less stringent requirements.

# Assumptions of logistic regression

**Results:** Models the natural log of the odds (logits) of success probability

**Associated Dependent Variable:** Dichotomous

**Probability Distribution:** Binomial

**Link:** Logit

**Estimator:** Maximum likelihood

**Assumptions:** 1. Non-autocorrelation.  $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$  for  $i \neq j$

2. Linear relationship between the predictors and the log odds (logit) of the dependent variable.

3. Absence of high partial multicollinearity.

4. Large samples.

5. Adequate expected cell frequencies.

# Measuring goodness of fit

Effron's  $R^2$

Effron revised the  $R^2$  formula that is used in the context of OLS regression in such a way that the binary feature of dependent variable is taken

$$\text{Effron's } R^2 = 1 - \frac{n}{n_1 n_2} \sum (Y - \hat{Y})^2$$

# McFadden's Pseudo $R^2$

McFadden's Pseudo  $R^2$  is a very popular measure of goodness of fit in the context of binary dependent variable models. It is also known as the log-likelihood ratio index (LRI). It is called Pseudo  $R^2$  because although there is no measure equivalent to  $R^2$  in the maximum likelihood method, it tends to provide an interpretation similar to  $R^2$  statistic.

The Pseudo  $R^2$  formula is obtained by comparing the value of log likelihood of initial regression model ( $\ln L$ ) with the value of log likelihood that would have been obtained with only the intercept term in the regression model ( $\ln L_0$ ).

It is defined as

$$\text{Pseudo } R^2 = 1 - \frac{\ln L}{\ln L_0}$$



## Examining the overall significance of regression

- For the purpose of examining the overall significance of logit/probit models, we may use the likelihood ratio statistic (LR-statistic). Two important features of LR statistic are: (i) it uses variations in likelihood as a basis for test, and (ii) it is useful to assess the explanatory power of the model, i.e., to understand overall significance of regression.

$$LR = 2 \ln \frac{L}{L_0} = 2 (\ln L - \ln L_0)$$

It has been shown that  $LR \sim \chi^2$  with degrees of freedom  $K$  (number of explanatory variables). We reject null hypothesis if  $\chi^2 > \text{Critical (Tab) value of } \chi^2$ .

**Odds:** An odds is a ratio formed by the probability that an event occurs divided by the probability that the event does not occur. In binary logistic regression, the odds is usually the probability of getting a “1” divided by the probability of getting a “0”.

**Odds ratio:** An odds ratio is the ratio of two odds, such as the ratio of the odds for men and the odds for women. Odds ratios are the main effect size measure for logistic regression, reflecting in this case what difference gender makes as a predictor of some dependent variable. An odds ratio of 1.0 (which is 1:1 odds) indicates the variable has no effect. The further from 1.0 in either direction, the greater the effect.

**Log odds:** The log odds is the coefficient predicted by logistic regression and is called the “logit”. It is the natural log of the odds of the dependent variable equaling some value (ex., 1 rather than 0 in binary logistic regression). The log odds thus equals the natural log of the probability of the event occurring divided by the probability of the event not occurring:  $\ln(\text{odds}(\text{event})) = \ln(\text{prob}(\text{event})/\text{prob}(\text{nonevent}))$

Logit: The “logit function” is the function used in logistic regression to transform the dependent variable prior to attempting to predict it. Specifically, the logit function in logistic regression is the log odds, explained above. The “logit” is the predicted value of the dependent variable. “Logit coefficients” are the b coefficients in the logistic equation used to arrive at the predicted value.

Parameter estimates: These are the logistic (logit or b) regression coefficients for the independent variables and the constant in a logistic regression equation, much like the b coefficients in OLS regression. Synonyms for parameter estimates are unstandardized logistic regression coefficients, logit coefficients, log odds-ratios, and effect coefficients. Parameter estimates are on the right-hand side of the logistic regression equation and logits are on the left-hand side. The logistic regression equation itself is:

$$z = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k$$

Where z is the log odds of the dependent variable =  $\ln(\text{odds}(\text{event}))$ . The "z" is the “logit”, also called the log odds.

- $\text{Exp}(b)$  = the odds ratio for an independent variable = the natural log base  $e$  raised to the power of  $b$ . The odds ratio of an independent variable is the factor by which the independent variable increases or (if negative) decreases the log odds of the dependent variable. The term “odds ratio” usually refers to odds ratios for independent variables.

# TEST FOR INDIVIDUAL PREDICTORS

- In logistic regression there is an analogous statistics known as the Wald statistics. Wald statistics tell us whether the beta coefficient for that predictor is significantly different from zero. If the coefficient is significantly different from zero then we can say that the predictor is making a significant contribution to the prediction of the outcome (Y).

The Wald statistic is defined as

$$W = \frac{\hat{\beta}_i^2}{\{S.E(\hat{\beta}_i)\}^2}$$

# Summarizing Predictive Power: Classification Tables

A *classification table* cross-classifies the binary outcome  $y$  with a prediction of whether  $y = 0$  or  $1$ . The prediction for observation  $i$  is  $\hat{y} = 1$  when its estimated probability  $\hat{\pi}_i > \pi_0$  and  $\hat{y} = 0$  when  $\hat{\pi}_i \leq \pi_0$ , for some cutoff  $\pi_0$ . One possibility is to take  $\pi_0 = 0.50$ . However, if a low (high) proportion of observations have  $y = 1$ , the model fit may never (always) have  $\hat{\pi}_i > 0.50$ , in which case one never (always) predicts  $\hat{y} = 1$ . Another possibility takes  $\pi_0$  as the sample proportion of 1 outcomes, which is  $\hat{\pi}_i$  for the model containing only an intercept term.

# Two useful summaries of predictive power are

$$\text{Sensitivity} = P(\hat{y} = 1 \mid y = 1)$$

$$\text{Specificity} = P(\hat{y} = 0 \mid y = 0)$$

$$\begin{aligned} P(\text{correct classif.}) &= P(y = 1 \text{ and } \hat{y} = 1) + P(y = 0 \text{ and } \hat{y} = 0) \\ &= P(\hat{y} = 1 \mid y = 1)P(y = 1) + P(\hat{y} = 0 \mid y = 0)P(y = 0) \\ &= \text{sensitivity}[P(y = 1)] + \text{specificity}[1 - P(y = 1)], \end{aligned}$$

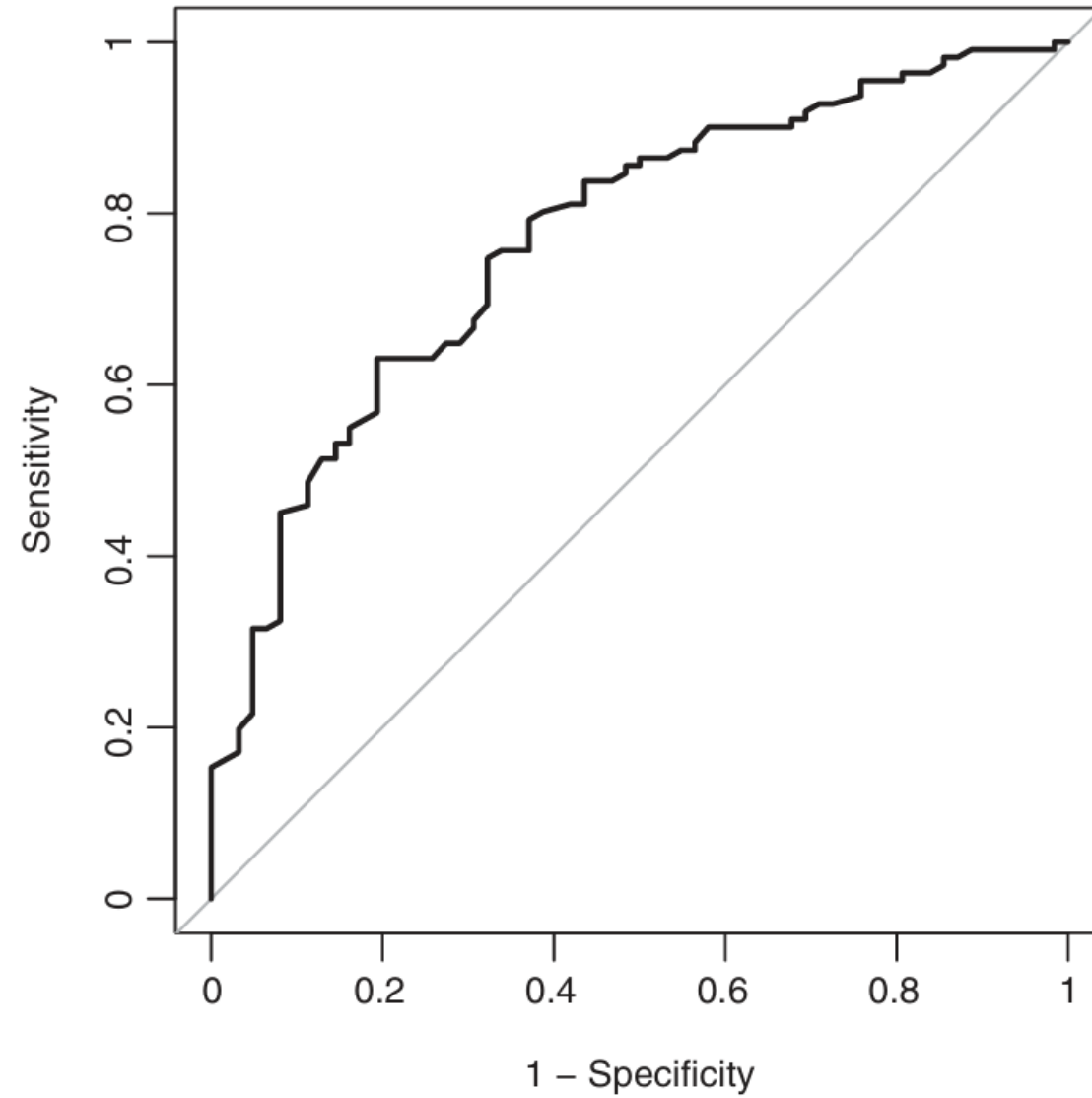
# Summarizing Predictive Power: ROC Curves

A *receiver operating characteristic* (ROC) curve is a plot that shows the sensitivity and the specificity of the predictions for all the possible cutoffs  $\pi_0$ . This curve is more informative than a classification table, because it summarizes predictive power for all possible  $\pi_0$ .

The ROC curve plots sensitivity on the vertical axis versus  $(1 - \text{specificity})$  on the horizontal axis. When  $\pi_0$  gets near 0, almost all predictions are  $\hat{y} = 1$ ; then, sensitivity is near 1, specificity is near 0, and the point for  $(1 - \text{specificity}, \text{sensitivity})$  has coordinates near  $(1, 1)$ . When  $\pi_0$  gets near 1, almost all predictions are  $\hat{y} = 0$ ; then, sensitivity is near 0, specificity is near 1, and the point for  $(1 - \text{specificity}, \text{sensitivity})$  has coordinates near  $(0, 0)$ . The ROC curve usually has a concave or nearly concave shape connecting the points  $(0, 0)$  and  $(1, 1)$ .



For a particular value of specificity, better predictive power corresponds to higher sensitivity. Therefore, the better the predictive power, the higher is the ROC curve. Because of this, the area under the curve provides a single value that summarizes predictive power. The greater the area, the better the predictive power. This measure of predictive power is called the *concordance index*. Consider all pairs of observations  $(i, j)$  such that  $y_i = 1$  and  $y_j = 0$ . The concordance index estimates the probability that the predictions and the outcomes are *concordant*, which means that the observation with the larger  $y$  also has the larger  $\hat{\pi}$ . A concordance value of 0.50 means predictions were no better than random guessing.



# HOSMER- LEMESHOW TEST

- The goodness of fit (the degree of closeness of the model-predicted value to the corresponding observed value) is useful for applying to the regression model. The Hosmer-Lemeshow goodness of fit statistics (Hosmer & Lemeshow, 2000) proposes a Pearson's statistic to compare the observed and fitted counts for the partition.
- In general, the Hosmer-Lemeshow goodness of fit test divides subject into deciles based on predicted probabilities and computes a chi square from observed and expected frequencies.

# Example and interpretation

| Step | Chi-square | df | Sig. |
|------|------------|----|------|
| 1    | 1.228      | 3  | .746 |

We observed from the example, that the Hosmer-Lemeshow chi-square statistic with 3 degree of freedom is 1.228 with p value 0.746. The large p value signifies that there is no difference between the observed and predicted values, implying that the model fits the data at an acceptable level.

# Nagelkerke R<sup>2</sup>

- Coefficient of determination ( $R^2$ ) is the proportion of the variation in the dependent variable that can be explained by predictors in the model. While coefficient of determination ( $R^2$ ) is a very valuable measure of how well linear regression model fits, it is less useful in logistic regression.
- As logistic regression is not a linear model, we can't calculate  $R^2$  directly as for linear regression model. However, test like -2log likelihood gives the pseudo  $R^2$  (Nagelkerke  $R^2$ ) statistics which are based on comparing the likelihood of the current model to the null model (without any predictors). A large pseudo  $R^2$  indicate that more of the variation is explained by the model, from a minimum of 0 to maximum of 1. It should also be noted that pseudo  $R^2$  values tend to be very low for logistic regression model, much lower than for linear regression model. This is because we are trying to predict the outcome where as the model only given us the probability of outcomes.

# Logistic regression (command in R)

- `##Open MROZ.dta data file`
- `head(MROZ)`

## Linear Probability Model

- `linprob<-lm(inlf~nwifeinc+educ+exper+l(exper^2)+age+kidslt6+kidsge6, data=MROZ)`
- `summary(linprob)`
- `###for prediction.`
- `pred<-list(nwifeinc=c(100,0),educ=c(5,15), exper=c(0,10),age=c(20,52),kidslt6=c(2,0), kidsge6=c(0,0))`
- `predict(linprob,pred)`

# Logistic model

- **### Logistic regression**

- `glm(formula=inlf~nwifeinc+educ+exper+l(exper^2)+age+kidslt6+kidsg  
e6,family=binomial(link=logit), data=MROZ)`

- `logit_femp<-  
glm(formula=inlf~nwifeinc+educ+exper+l(exper^2)+age+kidslt6+kidsg  
e6,family=binomial(link=logit), data=MROZ)`

- `summary(logit_femp)`

- **# Logit model odds ratios**

- `exp(logit_femp$coefficients)`

###Log likelihood

- logLik(logit\_femp)

###McFadden's pseudo R2

$1 - \text{logit\_femp}\$deviance / \text{logit\_femp}\$\text{null.deviance}$

library(lmtest)

lrtest(logit\_femp)



# Data File: Smoking

- Run the logistic regression taking smoker as a dependent variable and age, educ, income and pcigs79 as independent variables.

# Data file: Smoking

|             | Estimate   | Std. Error | z      | value    | Pr(> z ) |  |
|-------------|------------|------------|--------|----------|----------|--|
| (Intercept) | 2.745e+00  | 8.292e-01  | 3.311  | 0.000931 | ***      |  |
| age         | -2.085e-02 | 3.739e-03  | -5.577 | 2.44e-08 | ***      |  |
| educ        | -9.097e-02 | 2.067e-02  | -4.402 | 1.07e-05 | ***      |  |
| income      | 4.720e-06  | 7.170e-06  | 0.658  | 0.510356 |          |  |
| pcigs79     | -2.232e-02 | 1.247e-02  | -1.789 | 0.073538 | .        |  |
| ---         |            |            |        |          |          |  |

# Meaning

- The variables age and education are highly statistically significant and have the expected signs.
- As age increases, the value of the logit decreases, that is, as people age, they are less likely to smoke.
- More educated people are less likely to smoke, perhaps due to the ill effects of smoking.
- The price of cigarettes has the expected negative sign and is significant at about the 10% level. The higher the price of cigarettes, the lower is the probability of smoking.
- Income has no statistically visible impact on smoking, perhaps because expenditure on cigarettes may be a small proportion of family income.

# Meaning of coefficient

- holding other variables constant, if, for example, education increases by one year, the average logit value goes down by 0.09 , that is, the log of odds in favor of smoking goes down by about 0.09.
- Other coefficients are interpreted similarly.

# Stepwise Regression

Pravat Uprety

# Stepwise Regression

- One option in regression analysis is to bring all possible independent variables into the model in one step. This is what we have done previously.
- We use the term full regression to describe this approach. Another option for developing a regression model is called stepwise regression. Stepwise regression, as the name implies, develops the least squares regression equation in steps, either through forward selection, backward elimination, or standard stepwise regression

# Forward Selection

- The forward selection procedure begins (Step 1) by selecting a single independent variable from all those available. The independent variable selected at Step 1 is the variable that is most highly correlated with the dependent variable. A t-test is used to determine if this variable explains a significant amount of the variation in the dependent variable. At Step 1, if the variable is statistically significant, it is selected to be part of the final model used to predict the dependent variable. If it is not significant, the process is terminated. If no variables are found to be significant, the researcher will have to search for different independent variables than the ones already tested.

- In the next step (Step 2), a second independent variable is selected based on its ability to explain the remaining unexplained variation in the dependent variable. Recall that the coefficient of determination  $R^2$  measures the proportion of variation explained by all of the independent variables in the model. Thus, after we select the first variable (say,  $x_1$ ),  $R^2$  indicates the percentage of variation this variable explains. The forward selection routine then computes all possible two-variable regression models, with  $x_1$  included, and determines the  $R^2$  for each model. The coefficient of partial determination at Step 2 is the proportion of the as yet unexplained variation (after  $x_1$  is in the model) that the additional variable explains. The independent variable that adds the most to  $R^2$ , given the variable(s) already in the model, is the one we select. Then, we conduct a t-test to determine if the newly added variable is significant. This process continues until either we have entered all available independent variables or the remaining independent variables do not add appreciably to  $R^2$ . For the forward selection procedure, the model begins with no variables. We enter variables one at a time, and after a variable is entered, it cannot be removed.



# Backward Elimination

- Backward elimination is the reverse of the forward selection procedure. In the backward elimination procedure, all variables are forced into the model to begin the process. Then we remove the variables one insignificant variable at a time until no more insignificant variables are found. Once we have removed a variable from the model, it cannot be re-entered

# Standard Stepwise Regression

- The standard stepwise procedure (sometimes referred to as forward stepwise regression—not to be confused with forward selection) combines attributes of both backward elimination and forward selection. The standard stepwise method serves one more important function. If two or more independent variables are correlated, a variable selected in an early step may become insignificant when other variables are added at later steps. The standard stepwise procedure will drop this insignificant variable from the model. Standard stepwise regression also offers a means of observing multicollinearity problems, because we can see how the regression model changes as each new variable is added to it.

# Mallows' Cp

- **Mallows' Cp** is a metric that is used to pick the best [regression model](#) among several different models.
- It is calculated as:
- $C_p = \text{RSS}_p / S^2 - n + 2(P+1)$

- **Where**

$\text{RSS}_p$ : The residual sum of squares for a model with  $p$  predictor variables

$S^2$ : The residual mean square for the model (estimated by MSE)

$n$ : The sample size

$P$ : The number of predictor variables

Mallows'  $C_p$  is used when we have several potential predictor variables that we'd like to use in a regression model and we'd like to identify the best model that uses a subset of these predictor variables.

- We can identify the “best” regression model by identifying the model with the lowest  $C_p$  value that is less than  $P+1$ , where  $P$  is the number of predictor variables in the model.