QUESTION ONE: 30 MARKS

i. **Briefly discuss the components of time series data (8mks)**

a) Trend- measure the average change in the variable per unit time
b) Seasonal variation- periodic variation that occurs with some degree of regulations within a year or shorter
c) Cyclical variation- recurring up and down movement which are extended over a long period
d) Irregular variation- random fluctuation which happens due to errors

ii. **Explain the main problem of using non stationary time series data and elaborate how to transform non stationary time series (4mks)**

1. **Spurious Relationships**: In regression analysis, non-stationary data can produce misleading results, indicating relationships that do not actually exist.
2. **Unreliable Predictions**: Forecasts based on non-stationary data can be highly inaccurate since the underlying patterns and trends change over time.
3. **Inconsistent Statistical Measures**: Mean, variance, and other measures are not constant, complicating the analysis and interpretation of the data.

1. **Differencing:** This involves subtracting the previous observation from the current observation. First-order differencing can often remove trends and seasonality. For higher-order trends, second-order differencing (differencing the differenced series) may be necessary.

$$Y_t' = Y_t - Y_{t-1}$$

For second-order differencing:

$$Y_t'' = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$$

2. **Log Transformation:** Applying a logarithm to the series can stabilize the variance and transform a multiplicative relationship into an additive one, making it easier to handle trends and seasonal effects.

$$Y_t' = \log(Y_t)$$

3. **Detrending:** This involves removing the trend component from the series. It can be done through regression or by subtracting a moving average from the series.

$$Y_t' = Y_t - \text{Trend Component}$$

4. **Seasonal Adjustment:** For data with strong seasonal patterns, seasonal adjustment techniques such as seasonal differencing (subtracting the value from the same season in the previous year) can be used.

$$Y_t' = Y_t - Y_{t-s}$$

QUESTION TWO: 20 MARKS

I. **Discuss the difference between logit and probit models (5mks)**

1. Distribution of the Error Terms

- **Logit Model:** Assumes that the error terms follow a logistic distribution. The logistic distribution has slightly heavier tails than the normal distribution, which means it gives more weight to extreme values.
- **Probit Model:** Assumes that the error terms follow a standard normal distribution (Gaussian distribution). This assumption results in a different shape for the cumulative distribution function (CDF) compared to the logistic distribution.

2. Link Function

- **Logit Model:** Uses the logistic function as the link function. The logistic function is given by:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}}$$

This function maps any real-valued number to the (0, 1) interval, suitable for modeling probabilities.

- **Probit Model:** Uses the cumulative distribution function of the standard normal distribution as the link function. The probit function is given by:

$$P(Y = 1|X) = \Phi(\beta_0 + \beta_1 X_1+$$

iii. **Outline how you would test for autocorrelation using Durbin Watson Test (3mks)**

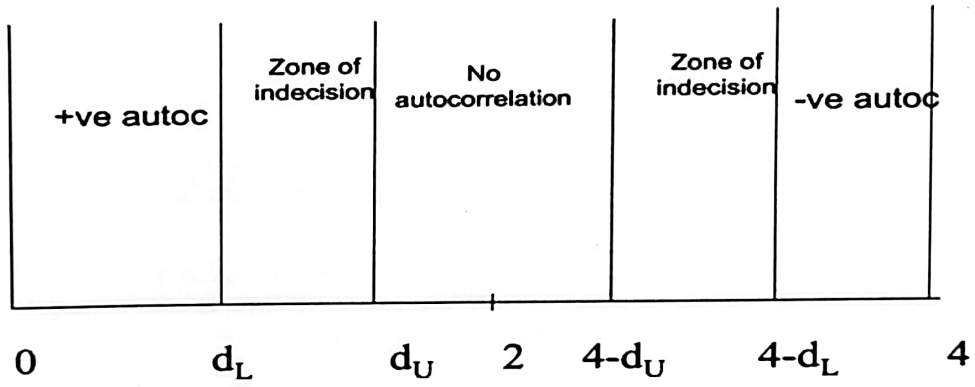
The Durbin-Watson d statistic is defined as:

$$d = \frac{\sum_{t=2}^{t=n} (e_t - e_{t-1})^2}{\sum_{t=1}^{t=n} e_t^2}$$

Step 1: Estimate the model by OLS and obtain the residuals

Step 2: Calculate the DW statistic

Step 3: Construct the table with the calculated DW statistic and the d_U, d_L, $4-d_U$ and $4-d_L$ critical values.



iv. **Illustrate clearly how you would estimate OLS parameters from the following linear regression model (10mks)**

$Y=\beta 0+\beta 1X+u$

THE LAST PAGE

v. **Briefly discuss any two assumptions of multiple linear regression (5mks)**

a) There is a linear relationship between dependent variable and independent variables.
b) There is no correlation between two or more explanatory variables
c) The variance of the error term is constant in regardless of the value of Y
d) The probability distribution error term is normal

QUESTION TWO

You estimated a regression model and obtained the following results (SE denotes standard error)

$$Return = 1200 - 300inflation + 500lending\ rate$$

$$SE\beta_0 = 100 \quad SE\beta_1 = 150 \quad SE\beta_2 = 1000$$

a. State the hypothesis

Test the null hypothesis, H_0 implies that inflation and lending rate has no significant effect on the amount of profit

$$H_0: \beta_1 = 0, H_0: \beta_2 = 0$$

Vs

alternative hypothesis H_A implies that inflation and lending rate has significant effect on the amount of profit

$$H_A: \beta_1 \neq 0, H_A: \beta_2 \neq 0$$

b. Show how you calculated the test statistics

T- test is given as;

For inflation $t = \frac{b_1 - \beta_1}{SE\beta_1}$

where, $b_1 = -300$ $\beta_1 = 0$ and $SE\beta_1$=150

$t = \frac{-300-0}{150} = \frac{-300}{150} = -2.0$

For lending rate $t = \frac{b_2 - \beta_2}{SE\beta_2}$

where, $b_2 = -300$ $\beta_2 = 0$ and $SE\beta_2$=150

$t = \frac{500-0}{1000} = \frac{500}{1000} = 0.5$

c. **State the decision rule you used**

Reject null hypothesis if;

$$t = \frac{b_1 - \beta_1}{SE\beta_1} \geq t_{n-k,\frac{\alpha}{2}}$$

Where $t_{n-k,\frac{\alpha}{2}}$ *is the critical value*

Do not Reject null hypothesis if;

$$t = \frac{b_1 - \beta_1}{SE\beta_1} \leq t_{n-k,\frac{\alpha}{2}}$$

Where $t_{n-k,\frac{\alpha}{2}}$ *is the critical value*

d. **What would you conclude from the results of the test**

Suppose coefficient of inflation and lending rate is statistically significant at 5% significant level (the critical value for the t-test 5% significant level is 1.96) ≥ 1.96

Recall that inflation t = -2.0 and lending rate t = 0.5 ≤ 1.96

Then it is clear that, $t = \frac{b_1 - \beta_1}{SE\beta_1} \leq t_{n-k,\frac{\alpha}{2}}$

Thus, we fail to reject null hypothesis, and conclude that there is no statistically significant different from zero

QUESTION THREE: 20 MARKS

i. Using an example explain the dummy trap (10mks)

ii. Briefly explain the causes of autocorrelation (10mks)

1. **Inertia** - Macroeconomics data often exhibit business cycles.
2. **Model Specification Error- eg. Exclusion of a variable**
   - ➢ True model: $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t$
   - ➢ Estimated model: $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + v_t$
3. **Cobweb Phenomenon**
   - ➢ In agricultural market, the supply reacts to price with a lag of one time period because supply decisions take time to implement. This is known as the cobweb phenomenon.
4. **Data Manipulation**
   - data 'massaging' can lead to patterns in error term. eg by taking a moving average of observations, the errors will no longer be independent of one another.

QUESTION FOUR: 20 MARKS

Discuss reasons for lags (6mks)

Informational reasons – informational assymetric will causes people not to react freely due to the changing factors

Psychological reasons – due to inertia people cannot make correct reasons due to changing factor

Institutional reasons – people cannot react freely to the changing factor due to the contractual obligations

Technological reasons-

It occurs when two dummy variables relating to same aspect are included. arises when many dummy variables are included describing a given number of groups

$$Y_i=\beta_1+\beta_2X_{2i}+\beta_3D_i+u_i$$

Suppose we code D_2 *is* (*female and male*) with 1 and 0 otherwise respectively

D = 1 for male

0 for female

$$Y_i=\beta_1+\beta_2X_{2i}+\beta_3D_i+u_i$$

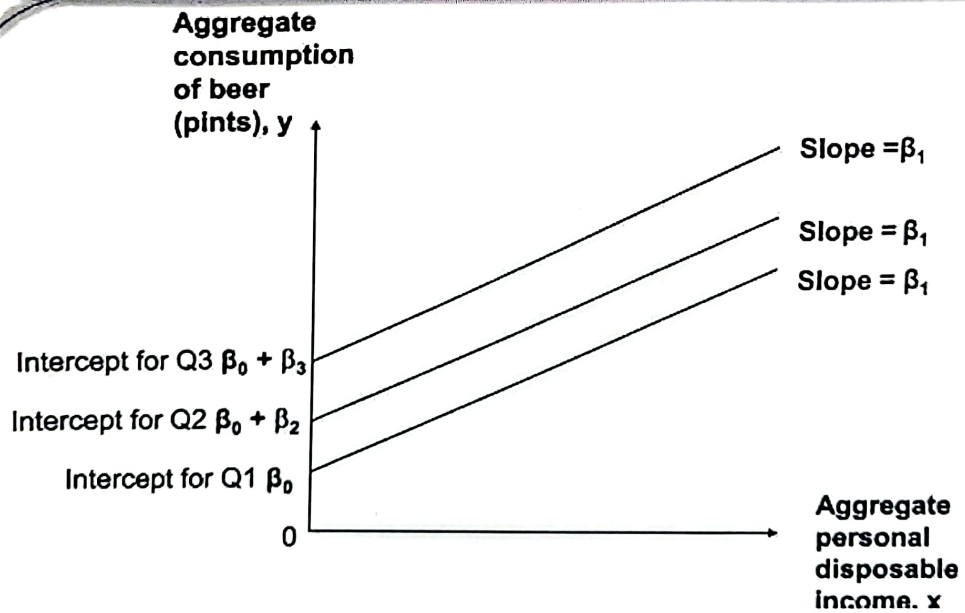
Now we have two cases

$D_i=0$ $\quad Y_i=\beta_1+\beta_2X_{2i}+\beta_3(0)+u_i$

$$Y_i=\beta_1+\beta_2X_{2i}+u_i$$

$D_i=1$ $\quad Y_i=\beta_1+\beta_2X_{2i}+\beta_3(1)+u_i$

$$Y_i=(\beta_1+\beta_3)+\beta_2X_{2i}+u_i$$



Given the following system of equation depicting demand and supply

$$Q = \alpha_1 P + \alpha_2 X + \epsilon_d$$

$$Q = \beta_1 P + \epsilon_s$$

a. **Identify and explain the endogenous and exogenous variables? (4 marks)**

We have two endogenous variable Q, P and one exogenous variable X

Thus, the endogenous variables Q, P are determined by exogenous variable X in the model and by the error terms

b. **Present the reduction form of the structured system of equation (10 marks)**

We need to solve for endogenous variable P and Q since they are determined inside the model,

Now to solve foe P and Q

$$Q_d = Q_S$$

$$\beta_1 P + \epsilon_s = \alpha_1 P + \alpha_2 X + \epsilon_d$$

$$\beta_1 P - \alpha_1 P = \alpha_2 X + \epsilon_d - \epsilon_s$$

$$P(\beta_1 - \alpha_1) = \alpha_2 X + \epsilon_d - \epsilon_s$$

$$\frac{P(\beta_1 - \alpha_1)}{(\beta_1 - \alpha_1)} = \frac{\alpha_2}{(\beta_1 - \alpha_1)} X + \frac{\epsilon_d - \epsilon_s}{(\beta_1 - \alpha_1)}$$

$$P = \frac{\alpha_2}{(\beta_1 - \alpha_1)} X + \frac{\epsilon_d - \epsilon_d}{(\beta_1 - \alpha_1)}$$

$$P = \pi_1 X + V_1$$

Now we need to solve for Q

$$Q = \beta_1 P + \epsilon_d$$

$$\text{But } P = \frac{\alpha_2}{(\beta_1 - \alpha_1)} X + \frac{\epsilon_d - \epsilon_s}{(\beta_1 - \alpha_1)}$$

$$Q = \beta_1 \left[\frac{\alpha_2}{(\beta_1 - \alpha_1)} X + \frac{\epsilon_d - \epsilon_s}{(\beta_1 - \alpha_1)}\right] + \epsilon_d$$

$$Q = \left[\frac{\beta_1 \alpha_2}{(\beta_1 - \alpha_1)} X + \frac{\beta_1 \epsilon_d - \beta_1 \epsilon_s}{(\beta_1 - \alpha_1)}\right] + \epsilon_s$$

$$Q = \pi_2 X + V_2$$

Thus, the parameters $\pi_2 and \pi_2$ *are* called reduced form parameters and $V_2 and V_2$ are called reduced form errors.

REG No.

Write on both sides of the paper

Question

Do not write in either margin

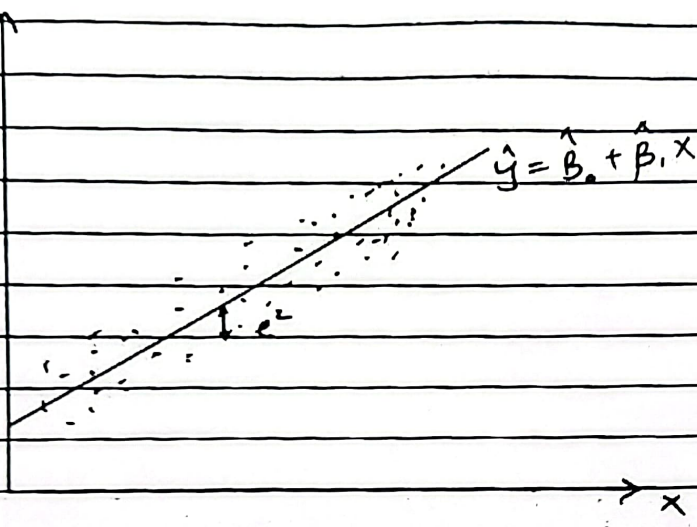
Assignment 2. C01/0489/2020

Derive the OLS estimation for a simple linear Regression model (10marks).

Solution:

Let simple linear Regression model be;

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



$\hat{\beta}_0$ and $\hat{\beta}_1$ is estimated using OLS

$$e_i = y_i - \hat{y}_i \quad , \quad e_i^2 = (y_i - \hat{y}_i)^2 \quad , \quad \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

Where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ → Estimate.

Now,

$$L = \text{Min} \sum_{i=1} e_i^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Now we can consider the two Partial Derivative, That is First order condition (F.O.C)

$$\frac{\partial L}{\partial \beta_0} = 0 \quad , \quad \frac{\partial L}{\partial \beta_1} = 0.$$

FOR $\hat{\beta}_0$

$$\text{Thus } \frac{\partial L}{\partial \beta_0} = -2 \sum (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i) = 0$$

Divide both sides by -2

$$\sum (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i) = 0$$

$$\sum y_i - \sum \hat{\beta}_0 - \hat{\beta}_1 \sum x_i = 0$$

Recall that $\frac{1}{n} \sum x_i = \bar{x}$, $\frac{1}{n} \sum y_i = \bar{y}$

Also $\sum x_i = n\bar{x}$, $\sum y_i = n\bar{y}$

REG No.

Write on both sides of the paper

Question

~~$\bar{x}\bar{y} \quad \bar{y}n$~~

$$\sum x_i y_i - \bar{y}\sum x_i + \hat{B}_1 \bar{x}\sum x_i - \hat{B}_1 \sum x_i^2 = 0$$

$$\sum x_i y_i - n\bar{x}\bar{y} + n\hat{B}\bar{x}^2 - \hat{B}_1 \sum x_i^2 = 0$$

$$n\hat{B}_1 \bar{x}^2 - \hat{B}_1 \sum x_i^2 = -\sum x_i y_i + n\bar{x}\bar{y}$$

$$\hat{B}_1 \left(n\bar{x}^2 / \sum x_i^2\right) = n\bar{x}\bar{y} - \sum x_i y_i$$

$$n\bar{x}^2 / \sum x_i^2 \qquad n\bar{x}^2 - \sum x_i^2$$

$$\hat{B}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$$\sum y_i - \sum \hat{\beta}_0 - \hat{\beta}_1 \sum x_i = 0$$

$$n\bar{y} - n\hat{\beta}_0 - \hat{\beta}_1 n\bar{x} = 0$$

$$n(\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}) = 0$$ Divide both sides by n

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0$$ make $\hat{\beta}_0$ the subject of the equation

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

For $\hat{\beta}_1$

Recall that

$$\frac{\partial L}{\partial x_i} = -2 \sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$ Divide both sides by -2

$$\sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Recall that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$\sum_{i=1}^{n} x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} (x_i y_i - x_i \bar{y} + \hat{\beta}_1 \bar{x} x_i - \hat{\beta}_1 x_i^2) = 0$$

Recall that

$$\frac{1}{n}\sum x_i = \bar{x}, \quad \frac{1}{n}\sum y_i = \bar{y},$$

$$\sum x_i = n\bar{x} \qquad \sum y_i = n\bar{y}$$

Also

$$\sum x_i y_i = n\bar{x}\bar{y}$$

REG No.

Write on both sides of the paper

Question

Do not write in either margin

~~$\hat{\beta}_0 \bar{x} \bar{y} - \hat{\beta}_1 n$~~

$$\sum x_i y_i - \bar{y} \sum x_i + \hat{\beta}_1 \bar{x} \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$$

$$\sum x_i y_i - n\bar{x}\bar{y} + n\hat{\beta}\bar{x}^2 - \hat{\beta}_1 \sum x_i^2 = 0$$

$$n\hat{\beta}_1 \bar{x}^2 - \hat{\beta}_1 \sum x_i^2 = -\sum x_i y_i + n\bar{x}\bar{y}$$

$$\hat{\beta}_1 (n\bar{x}^2 - \sum x_i^2) = n\bar{x}\bar{y} - \sum x_i y_i$$

$$\frac{n\bar{x}^2 - \sum x_i^2}{n\bar{x}^2 - \sum x_i^2} \qquad \frac{n\bar{x}^2 - \sum x_i^2}{n\bar{x}^2 - \sum x_i^2}$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

b) **Why does the classical linear model assume that there is no multicollinearity among the explanatory variables? (6mks)**

the reasoning is this: If multicollinearity is perfect in the sense of $\alpha 1+\alpha 2+\cdots+\alpha k=0$, the regression coefficients of the X variables are indeterminate and their standard errors are infinite.

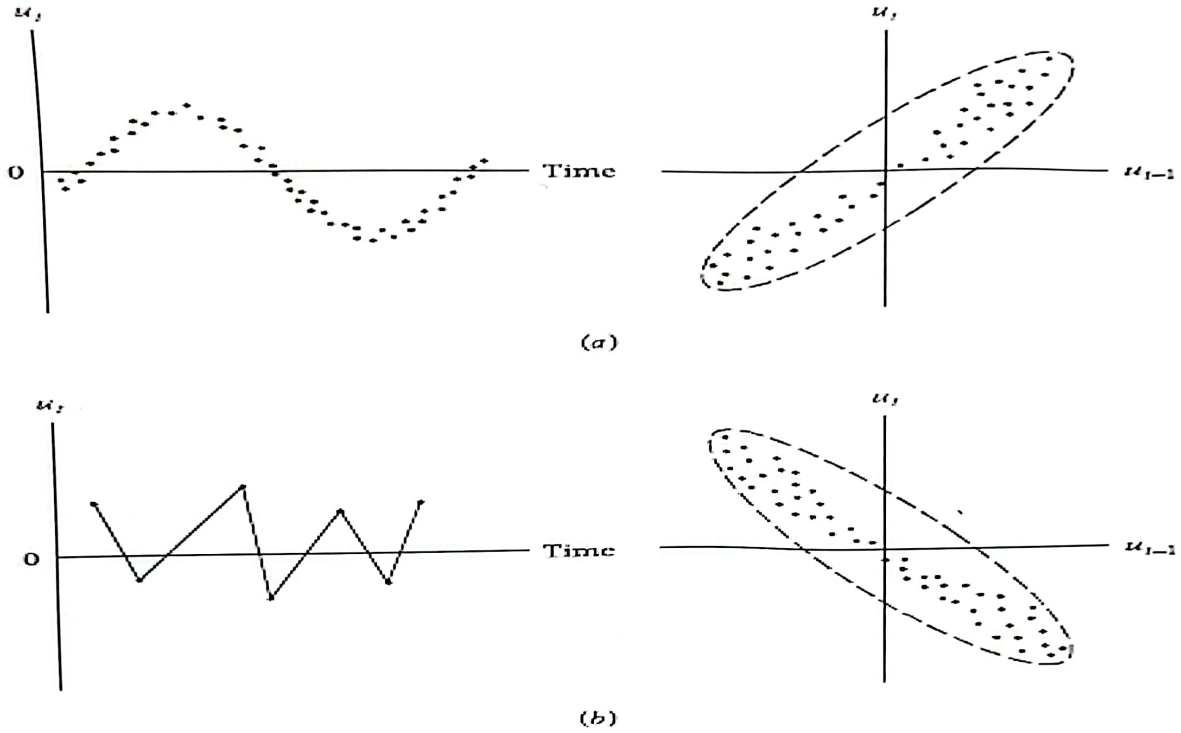
If multicollinearity is less than perfect, as $\alpha 1+\alpha 2+\cdots+\alpha k+\varepsilon i=0$,

the regression coefficients, although determinate, possess large standard errors (in relation to the coefficients themselves), which means the coefficients cannot be estimated with great precision or accuracy.

c) **Briefly explain three reasons for heteroscedasticity (6mks)**

i. presences of outliers- extreme values that can cause the variance to increase in one part of the data
ii. a structural shift in data- the variance changes at different level of the explanatory variable data often change due to population.
iii. Incorrect model specification – omitting relevant variables that influence the variance of error term

d) **Using graphs, illustrate the difference between positive and negative autocorrelation (6mks)**



(a) Positive and (b) negative autocorrelation.

- If positive autocorrelation exists
  - Residuals will follow a sine wave-type
  - Negative residuals tend to be followed by negative residuals while positive residuals tend to be followed by positive residuals
  - Any jaggedness due to random white noise
- If negative autocorrelation exists
  - then negative numbers are followed immediately by positive numbers in almost all cases
  - Any jaggedness due to white noise

e) Explain the difference between a stochastic and a deterministic relationship (4mks)

If the error terms is included in the model, the model is called stochastic, while if the error term is excluded in the model we refer it as deterministic model.

$Y = \beta_0 + \beta_1 X_1 + \varepsilon_i$ is stochastic

$Y = \beta_0 + \beta_1 X_1 ..$ is deterministic

f) **Explain the difference between linear in the variables and linear in the parameters (2mks)**

The model is linear in variable when the relationship between the independent variables and dependent variables is linear

The model is linear in parameter if the relationship between the dependent variable and the parameter is linear.

QUESTION TWO

Omondi sought to investigate determinants of performance among students at Nairobi University. He estimated the following regression model.

$$Pi = \beta 0 + \beta 1Ti + \beta 2Ai + \varepsilon i$$

$$\beta 0 = 0.6, \beta 1 = 0.1, \beta 2 = -0.2, SE(\beta 0) = 1.1, SE(\beta 1) = 0.01$$

$$(SE(\beta 2) = 0.13, R^2 = 0.76$$

Where: T is time spent reading, A is class attendance and P is performance.SE

denotes standard error.

Use the above results to answer the following:

i. **State both null and alternative hypotheses for the above model (6Mks)**

$$Pi = 0.6 + 0.1Ti - 0.2Ai + \varepsilon i$$

Test the null hypothesis, H_0 implies that time spend reading and class attendance has no significant effect on the amount of performance

$$H_0: \beta_1 = 0, H_0: \beta_2 = 0$$

Vs

alternative hypothesis H_A implies that time spend reading and class has significant effect on the amount of performance

$$H_A: \beta_1 \neq 0, H_A: \beta_2 \neq 0$$

ii. What does εi represent (2mks)

an error term represents the variation of dependent variables that is not explained by the independent variable,

it means that the variation of performance is not explained by the time spend reading and class attendance,

ii. Interpret R square (4mks)

R^2 implies that % variation of dependent variable can be explained by the movement of the independent variables.

$R^2 = 0.76$ it means that 63% variation of performance can be explained by the time spend reading and class attendance

Since $R^2 = 76\%$ is $\geq 50\%$ hence it's a good fit model

iv. Interpret the coefficients (8Mks)

$$Pi = 0.6 + 0.1Ti - 0.2Ai + \varepsilon i$$

1. Performance is expected to increase by 0.1 for every additional one unit of time spend reading.
2. Performance is expected to decrease by 0.2 for every additional one unit of class attendance.
3. Performance is expected to remain constant if there is zero time spend reading such that other factors remain constant.
4. Performance is expected to remain constant if there is zero class attendances such that other factors remains constants.

QUESTION THREE

a) Explain the rationale for using the adjusted R square (4mks)

It helps to identifies the percentage of variance in the target field that is explained by the input or output

Adjusted R^2 is the modified version of R^2 which is better measure of goodness fit in the regression model

b) Explain the practical consequences of high multicollinearity (6mks)

i. Estimate will remain unbiased- Even if an equation has significant multicollinearity, the estimates of the ßs still will be centered around the true population ßs if the first six Classical Assumptions are met for a correctly specified equation.

ii. The variance and standard error of the estimates will increase- This is the principal consequence of multicollinearity. Since two or more of the explanatory variables are significantly related, it becomes difficult to precisely identify the separate effects of the multicollinear variables

iii. Estimate will become very sensitive to change in specification- The addition or deletion of an explanatory variable or of a few observations will often cause major changes in the values of the ßs when significant multicollinearity exists

iv. The overall fit of the equation and the estimate of the coefficient of non-multicollinearity will largely be unaffected- Even though the individual t-scores are often quite low in a multicollinear equation, the overall fit of the equation, as measured by overline R ^ 2 . will not fall much, if at all, in the face of significant multicollinearity

v. The computed t-test will fall-

c) **Consider a two-variable model $Yi = \beta 1 + \beta 1 Xi + \varepsilon i$. Suppose that the heteroscedastic variances are known. Explain the method of generalized least squares (10mks)**

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \textbf{(11.3.1)}$$

which for ease of algebraic manipulation we write as

$$Y_i = \beta_1 X_{0i} + \beta_2 X_i + u_i \qquad \textbf{(11.3.2)}$$

where $X_{0i} = 1$ for each i. The reader can see that these two formulations are identical.

Now assume that the heteroscedastic variances σ_i^2 are *known*. Divide (11.3.2) through by σ_i to obtain

$$\frac{Y_i}{\sigma_i} = \beta_1 \left(\frac{X_{0i}}{\sigma_i}\right) + \beta_2 \left(\frac{X_i}{\sigma_i}\right) + \left(\frac{u_i}{\sigma_i}\right) \qquad \textbf{(11.3.3)}$$

which for ease of exposition we write as

$$Y_i^* = \beta_1^* X_{0i}^* + \beta_2^* X_i^* + u_i^* \qquad \textbf{(11.3.4)}$$

where the starred, or transformed, variables are the original variables divided by (the known) σ_i. We use the notation β_1^* and β_2^*, the parameters of the transformed model, to distinguish them from the usual OLS parameters β_1 and β_2.

What is the purpose of transforming the original model? To see this, notice the following feature of the transformed error term u_i^*:

$$\begin{aligned} \text{var}(u_i^*) = E(u_i^*)^2 &= E\left(\frac{u_i}{\sigma_i}\right)^2 \\ &= \frac{1}{\sigma_i^2} E(u_i^2) \qquad \text{since } \sigma_i^2 \text{ is known} \\ &= \frac{1}{\sigma_i^2}(\sigma_i^2) \qquad \text{since } E(u_i^2) = \sigma_i^2 \\ &= 1 \end{aligned} \qquad \textbf{(11.3.5)}$$

QUESTION FOUR

a) Explain how you can informally detect heteroscedasticity (4mks)

Graphical Method

- Plot residuals against time where residuals are estimates of disturbance term; Can highlight violations; Look for nonrandom patterns
- We can also plot ordinary residuals against lagged ordinary residuals
- If positive autocorrelation exists
  - Residuals will follow a sine wave-type
  - Negative residuals tend to be followed by negative residuals while positive residuals tend to be followed by positive residuals
  - Any jaggedness due to random white noise
- If negative autocorrelation exists
  - then negative numbers are followed immediately by positive numbers in almost all cases
  - Any jaggedness due to white noise
  -

b) What is the remedial measure for autocorrelation when the coefficient of first order correlation (ρ) is known? Briefly explain. (10mks)

Consider $Yi = \beta 1 + \beta 1 Xi + \varepsilon i$

Let us multiply equation (12.13) on both the sides by ρ. We obtain:

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho u_{t-1} \quad \ldots (12.14)$$

Let us now subtract equation (12.14) from equation (12.11) to obtain:

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + v_t \quad \ldots (12.15)$$

Note that we have used v_t for the new disturbance term above. Let us now denote:

$$Y_t^* = (Y_t - \rho Y_{t-1})$$

$$X_t^* = (X_t - \rho X_{t-1})$$

$$\beta_1^* = \beta_1(1-\rho)$$

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1}$$

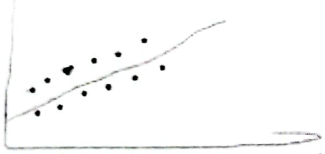
The transformed model will be

$$Y_t^* = \beta_1^* + \beta_2 X_t^* + v_t \qquad \ldots (12.16)$$

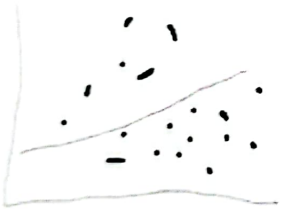
Now, the transformed variables Yt^* and Xt^* will have the desirable BLUE property. The estimators obtained by applying the OLS method to (12.16) are called the Generalized Least Squares (GLS) estimators. The transformation as suggested above is known as the Cochrane-Orcutt transformation procedure.

c) Distinguish between perfect and imperfect multicollinearity (6mks)

Perfect multicollinearity it means that independent variable are perfectly correlated or have positive linear relationship with more than two explanatory variables



imperfect multicollinearity it means that there is correlation between more than two explanatory variables but have no positive linear relationship



QUESTION FIVE

a) Briefly explain the remedial measure for heteroscedasticity when the error variance is known. (10mks)

The GLS procedure is the same as the WLS where we have weights, w_t, adjusting our variables.

Define $w_t = 1/\sigma_t$, and rewrite the original model as:

$$w_t Y_t = \beta_1 w_t + \beta_2 X_{2t} w_t + \beta_3 X_{3t} w_t + \ldots + \beta_k X_{kt} w_t + u_t w_t$$

Where if we define as $w_t Y_{t-1} = Y^*_t$ and $X_{it} w_t = X^*_{it}$

we get

$$Y^*_t = \beta^*_1 + \beta^*_2 X^*_{2t} + \beta^*_3 X^*_{3t} + \ldots + \beta^*_k X^*_{kt} + u^*_t$$

b) Explain the Durbin Watson test (4Mks)

The Durbin-Watson d statistic is defined as:

$$d = \frac{\sum_{t=2}^{t=n} (e_t - e_{t-1})^2}{\sum_{t=1}^{t=n} e_t^2}$$

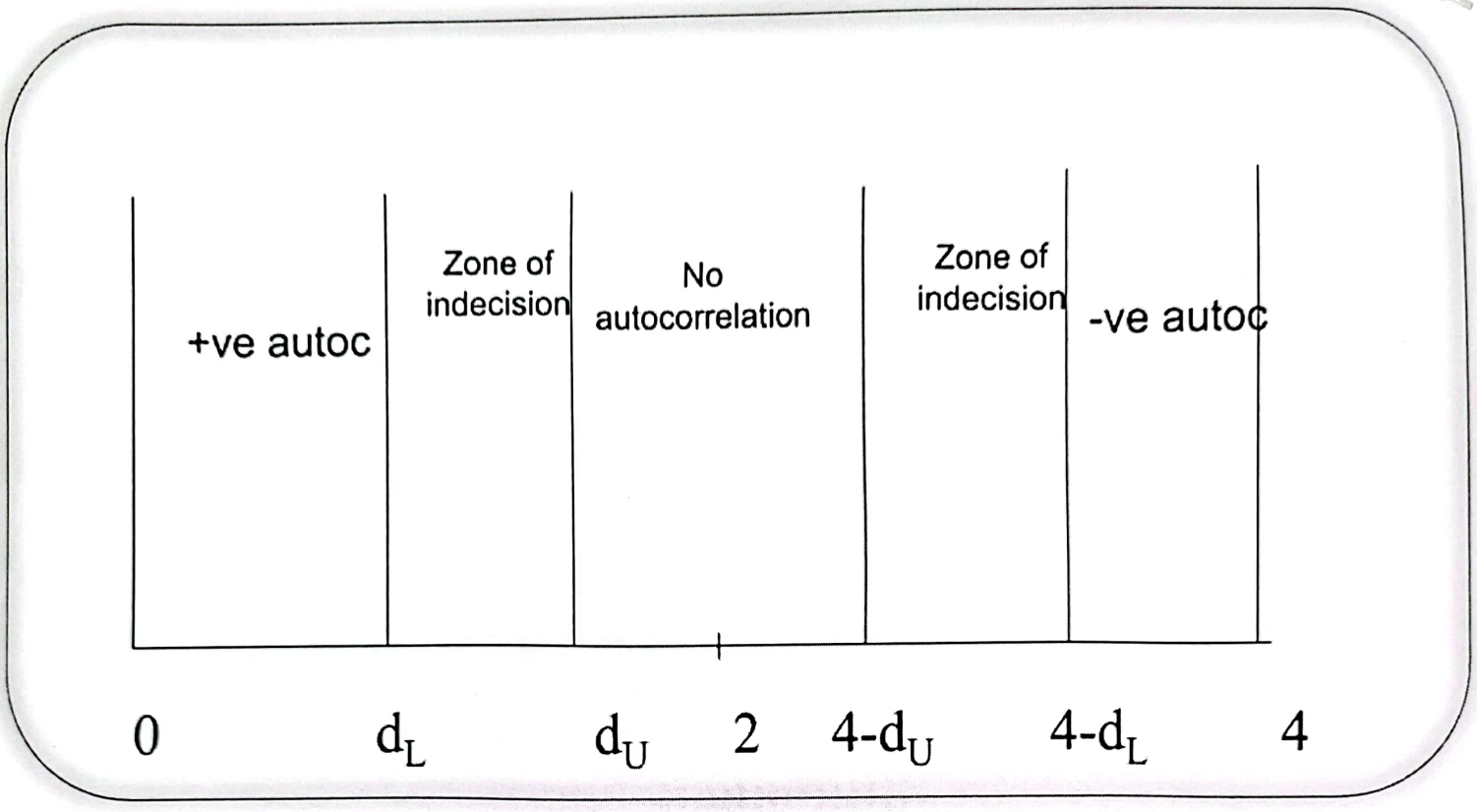
The following assumptions should be satisfied:

1. The regression model includes a constant
2. Autocorrelation is assumed to be of first-order only
3. The equation does not include a lagged dependent variable as an explanatory variable

Step 1: Estimate the model by OLS and obtain the residuals

Step 2: Calculate the DW statistic

Step 3: Construct the table with the calculated DW statistic and the d_U, d_L, $4-d_U$ and $4-d_L$ critical values.



c) Give reasons why we introduce a stochastic disturbance term in a regression equation (6mks)

(i) The error may arise due to the omission of relevant variables in the model. We know that each economic variable is affected by so many economic variables at the same time. If any one of the relevant variables is left out and cannot be included in the model, errors may arise. In such a case, an error is bound to occur in the model.

(ii) Errors may also arise due to the non-availability of data. The econometric study is frequently based upon the assumption that we have large samples of accurate data. But unfortunately, we will not find an example in which reliable and representative data are available. Thus, in the absence of accurate data, the sampling error arises in the model.

(iii) Errors may also arise due to the misspecification of the functional form between economic variables. Sometimes, we assume that the relationship is linear between variables, but it may be non-linear. In such a case, the forecast is bound to be incorrect. This type of error arises mainly due to the misunderstanding of the investigator.