**Q. N0. 1 Part A**

**1, Pooled OLS** is employed when you select a different sample for each year/month/period of the panel data. ... If you are using the same sample along all periods, than your results are correct by now and Fixed or Random effects models are recommended.

**2, Fixed effects model** is a [statistical model](https://en.wikipedia.org/wiki/Statistical_model) in which the model [parameters](https://en.wikipedia.org/wiki/Parameter) are fixed or non-random quantities. This is in contrast to [random effects models](https://en.wikipedia.org/wiki/Random_effects_model) and [mixed models](https://en.wikipedia.org/wiki/Mixed_model) in which all or some of the model parameters are random variables.

**3, Random effects model**also called a variance components model, is a statistical model where the model parameters are random variables.

**Part B**

**Dummy Variable,** A dummy variable (aka, an indicator variable) is a numeric variable that represents categorical data, such as gender, race, political affiliation, etc.

Technically, dummy variables are dichotomous, [quantitative](https://stattrek.com/statistics/dictionary.aspx?definition=quantitative-variable) variables. Their range of values is small; they can take on only two quantitative values. As a practical matter, regression results are easiest to interpret when dummy variables are limited to two specific values, 1 or 0. Typically, 1 represents the presence of a qualitative attribute, and 0 represents the absence.

Whenever you have a regression model with dummy variables, you can always see how the variables are being used to represent multiple subgroup equations by following the two steps described above: create separate equations for each subgroup by substituting the dummy values.

**Q.No. 2 A**

**Heteroscedasticity** is a problem because ordinary least squares (OLS) regression assumes that all residuals are drawn from a population that has a constant variance (homoscedasticity). To satisfy the regression assumptions and be able to trust the results, the residuals should have a constant variance.

In statistics, heteroskedasticity (or heteroscedasticity) happens when the standard deviations of a predicted variable, monitored over different values of an independent variable or as related to prior time periods, are non-constant. ... Heteroskedasticity often arises in two forms: conditional and unconditional.

To check for heteroscedasticity, you need to assess the residuals by fitted value plots specifically. Typically, the telltale pattern for heteroscedasticity is that as the fitted values increases, the variance of the residuals also increases.

Heteroscedasticity means unequal scatter. In [regression analysis](https://statisticsbyjim.com/glossary/regression-analysis/), we talk about heteroscedasticity in the context of the [residuals](https://statisticsbyjim.com/glossary/residuals/) or error term. Specifically, heteroscedasticity is a systematic change in the spread of the residuals over the range of measured values. Heteroscedasticity is a problem because [ordinary least squares](https://statisticsbyjim.com/glossary/ordinary-least-squares/) ([OLS](https://statisticsbyjim.com/glossary/ordinary-least-squares/)) [regression](https://statisticsbyjim.com/glossary/regression-analysis/) assumes that all residuals are drawn from a [population](https://statisticsbyjim.com/glossary/population/) that has a constant variance (homoscedasticity).

To satisfy the regression assumptions and be able to trust the results, the residuals should have a constant variance. In this blog post, I show you how to identify heteroscedasticity, explain what produces it, the problems it causes, and work through an example to show you several solutions.

**Part 2 B**

**Panel** data can model both the common and individual behaviors of groups. Panel data contains more information, more variability, and more efficiency than pure time series data or cross-sectional data. Panel data can detect and measure statistical effects that pure time series or cross-sectional data can't.

Panel data has the advantage that it is possible to observe the before- and after-effects of receiving the treatment of the same individual as well as providing the possibility of isolating the effects of treat- ment from other factors affecting the outcome

Pooled data occur when we have a “time series of cross sections,” but the observations in each cross section do not necessarily refer to the same unit. Panel data refers to samples of the same cross-sectional units observed at multiple points in time.

Cross-sectional data refers to data collected by observing many subjects (such as individuals, firms or countries/regions) at the same point of time, or without regard to differences in time. ... Panel analysis uses panel data to examine changes in variables over time and differences in variables between subjects.

**Q.No.3**

**1, A stochastic process** is a system which evolves in time while undergoing chance fluctuations. We can describe such a system by defining a family of **random** variables, {X t }, where X t measures, at time t, the aspect of the system which is of interest.

A **stochastic process** is defined as a collection of random variables X={Xt:t∈T} defined on a common probability space, taking values in a common set S (the state space), and indexed by a set T, often either N or [0, ∞) and thought of as time (discrete or continuous respectively) (Oliver, 2009).

* Create the sample space (Ω) — a list of all possible outcomes,
* Assign probabilities to sample space elements,
* Identify the events of interest,
* Calculate the probabilities for the events of interest.

**2, Stationary process,** In mathematics and statistics, a stationary process (or a strict/strictly stationary process or strong/strongly stationary process) is a stochastic process whose unconditional joint probability distribution does not change when shifted in time.

Intuitively, a random process {X(t),t∈J} is stationary if its statistical properties do not change by time. For example, for a stationary process

**3, Integration variable**, the Integration variable specifies which part of the integrand expression is to vary during the process of forming the continuous sum which constitutes integration. During the accumulation of a definite integral, the integration variable moves from the lower to the upper limit of integration.

The multiple integral is a definite integral of a function of more than one real variable, for instance, f(x, y) or f(x, y, z). Integrals of a function of two variables over a region in R2 are called double integrals, and integrals of a function of three variables over a region of R3 are called triple integrals.

**4, Unit root test,** In probability theory and statistics, a unit root is a feature of some stochastic processes (such as random walks) that can cause problems in statistical inference involving time series models. ... Due to this characteristic, unit root processes are also called difference stationary.

Formally, stationar- ity tests are based on testing for a unit moving average root in ∆zt. Unit root and stationarity test statistics have nonstandard.

In statistics, a unit root test tests whether a time series variable is non-stationary and possesses a unit root. The null hypothesis is generally defined as the presence of a unit root and the alternative hypothesis is either stationarity, trend stationarity or explosive root depending on the test used.

**5, An error correction model (ECM)** belongs to a category of multiple time series **models** most commonly used for data where the underlying variables have a long-run stochastic trend, also known as cointegration.

More than any other journal, Political Analysis publishes the most powerful new ideas for the empirical analysis of politics. No serious researcher in the field, and no serious research library, will want to be without a subscription.” Prof. Christopher Achen, Department of Political Science, University of Michigan “Political Analysis is the leading journal for innovative new quantitative work in political methodology, and the journal shows great promise for becoming an important voice in the ongoing effort to bridge quantitative and qualitative methodology.' Prof. David Collier, Department of Political Science, University of California, Berkeley

**Q.No.4 A**

**Regression,** Statistically, the presence of an interaction between categorical variables is generally tested using a form of analysis of variance (ANOVA). If one or more of the variables is continuous in nature, however, it would typically be tested using moderated multiple regression.

Welcome to Linear Regression in R for Public Health! Public Health has been defined as “the art and science of preventing disease, prolonging life and promoting health through the organized efforts of society”. Knowing what causes disease and what makes it worse are clearly vital parts of this. This requires the development of statistical models that describe how patient and environmental factors affect our chances of getting ill. This course will show you how to create such models from scratch, beginning with introducing you to the concept of correlation and linear regression before walking you through importing and examining your data, and then showing you how to fit models. Using the example of respiratory disease, these models will describe how patient and other factors affect outcomes such as lung function. Linear regression is one of a family of regression models, and the other courses in this series will cover two further members. Regression models have many things in common with each other, though the mathematical details differ. This course will show you how to prepare the data, assess how well the model fits the data, and test its underlying assumptions – vital tasks with any type of regression. You will use the free and versatile software package R, used by statisticians and data scientists in academia, governments and industry worldwide.

The outcome variable is also called the response or dependent variable, and the risk factors and confounders are called the predictors, or explanatory or independent variables. In regression analysis, the dependent variable is denoted "Y" and the independent variables are denoted by "X".

In statistics, main effect is the effect of one of just one of the independent variables on the dependent variable. There will always be the same number of main effects as independent variables. An interaction effect occurs if there is an interaction between the independent variables that affect the dependent variable.

Interactions in Multiple Linear Regression. Basic Ideas. Interaction: An interaction occurs when an independent variable has a different effect on the outcome depending on the values of another independent variable.

**Q.No.4 B**

**Linear Regression Models**

The regression model that we examined up to now is called a simple linear regression model. Looking at the univariate regression: Y=Xβ+ε where β=(β0, β1)⊤. When we say linear regression, we mean *linear in parameters* β. There are no restrictions on transformations of X and Y, as long as the parameters enter the equation linearly. For example, we can use log(X) and log(Y), or √X and √X etc. in the univariate regression. While transforming X and Y does not effect the linear regression specification itself, the interpretation of the coefficients depends on the transformation of X and Y. On the other hand, there are regression models, which are not regarded as *linear*, since they are not linear in their parameters: Yi=1β0+β1Xi+ϵi, i=1,...,N

Furthermore, estimation of such models is a separate issue, which covers nonlinear regression models.

**Inclusion of the constant term in the regression**

In some cases we want to impose a *restriction* that if X=0 then Y=0 as well. An example could be the relationship between income (X) and income tax revenue (Y) - if there is no income, X=0, then the expected revenue from the taxes would also be zero - E(Y|X=0)=0

Formally, we now choose a slope estimator, β1, from the following regression model: Yi=β1Xi+ϵi, i=1,...,N which is called a regression through the origin, because the conditional expected value: E(Yi|Xi)=β1Xi of the regression passes through the origin point X=0, Y=0. We can obtain the estimate of the slope parameter via OLS by minimizing the sum of squared residuals: RSS=N∑i=1ˆϵ2i=N∑i=1(Yi−ˆβ1Xi)2∂RSS∂ˆβ1=−2N∑i=1Xi(Yi−ˆβ1Xi)=0 which leads to the following expression: ˆβ1=∑Ni=1XiYi∑Ni=1X2i So, it is possible to specify a regression without a constant term, but should we opt for it? A constant β0 can be described as the mean value of Y when all predictor variables are set to zero. However, if the predictors can’t be zero, then it is impossible to interpret the constant. Furthermore, even if all other predictors can have zero-values, the constant may still be uninterpretable, as the intercept parameter β0 may be regarded as a sort of *garbage collector* for the regression model. The reason for this is the underlying assumption that the expected value of the residuals is *zero*. So the regression is estimated to a point where the mean of the residuals is (very close to) zero, which means that any bias, that is not accounted by the model, is collected in the intercept β0.

In general, inclusion of a constant in the regression model ensures that the models residuals have a mean of zero, otherwise the estimated coefficients may be biased. Consequently, and as is often the case, without knowing the true underlying model it is generally not worth interpreting the regression constant.

**Q.No.5 A**

A “**spurious regression**” is one in which the time-series variables are non stationary and independent. ... We derive corresponding results for some common tests for the normality and homoskedasticity of the errors in a spurious regression.

The spurious relationship is said to have occurred if the statistical summaries are indicating that two variables are related to each other when in fact there is no theoretical relationship between two variables. This is typical example of spurious regression

Spurious regression happens when there are similar local trends. The solid line is y and dotted line is x. Sometimes their local trends are similar, giving rise to the spurious regression. In short, two series are cointegrated if they are nonstationary and related.

ABSTRACT. A “spurious regression” is one in which the time-series variables are non-stationary and. independent. It is well-known that in this context the OLS parameter estimates and the R.

A spurious correlation wrongly implies a cause and effect between two variables. For example, the number of astronauts dying in spacecraft is directly correlated to seatbelt use in cars: Use your seatbelt and save an astronaut life!

**Q.No.5 B**

**Heteroscedasticity,** to check for heteroscedasticity, you need to assess the residuals by fitted value plots specifically. Typically, the telltale pattern for heteroscedasticity is that as the fitted values increases, the variance of the residuals also increases.

Here's how to perform a BP test:

* Estimate your model using OLS:
* Obtain the predicted Y values after estimating the model.
* Estimate the auxiliary regression using OLS:
* From this auxiliary regression, retain the R-squared value:
* Calculate the F-statistic or the chi-squared statistic:

In statistics, heteroskedasticity (or heteroscedasticity) happens when the standard deviations of a predicted variable, monitored over different values of an independent variable or as related to prior time periods, are non-constant. ... Heteroskedasticity often arises in two forms: conditional and unconditional.

Heteroskedasticity has serious consequences for the OLS estimator. Although the OLS estimator remains unbiased, the estimated SE is wrong. Because of this, confidence intervals and hypotheses tests cannot be relied on. In addition, the OLS estimator is no longer BLUE.

In statistics, the White test is a statistical test that establishes whether the variance of the errors in a regression model is constant: that is for homoskedasticity. This test, and an estimator for heteroscedasticity-consistent standard errors, was proposed by Halbert White in 1980.

**Used for,** the Chow test tells you if the regression coefficients are different for split data sets. Basically, it tests whether one regression line or two separate regression lines best fit a split set of data.

The Wald test (also called the Wald Chi-Squared Test) is a way to find out if explanatory variables in a model are significant. “Significant” means that they add something to the model; variables that add nothing can be deleted without affecting the model in any meaningful way.