**Q1 :(a):** **Following are the main steps in methodology of econometrics**

1. Theory

A theory should have a prediction. In statistics and econometrics, we also speak of **hypothesis**. One example is **the marginal propensity to consume** (MPC) proposed by Keynes. Other examples could be that lower taxes would increase growth, or maybe that it would increase economic inequality, and that introducing a common currency has a positive effect on trade.

2. Specification of the Mathematical Model

This is where the algebra enters. We need to use mathematical skills to produce an equation. Assume a theory predicting that more schooling increases the wage. In economic terms, we say that the return to schooling is positive.

3. Specification of the Econometric Model

Here, we assume that the mathematical model is correct but we need to account for the fact that it may not be so. We add an **error term**, *u* to the equation above. It is also called a **random (stochastic) variable**. It represents other non-quantifiable or unknown factors that affect Y. It also represents mismeasurements that may have entered the data.

4. Obtain Data

We need data for the variables above. This can be obtained from government statistics agencies and other sources. A lot of data can also be collected on the Internet in these days. But we need to learn the art of finding appropriate data from the ever-increasing loads of data.

5. Estimation of the model

Here, we quantify B1 and B2  i.e. we obtain numerical estimates. This is done by statistical technique called **regression analysis**.

6. Hypothesis Testing

Now we go back to the part where we had economic theory. The prediction was that schooling is good for the wage. Does the econometric model support this hypothesis. What we do here is called **statistical inference (hypothesis testing)**. Technically speaking, the B2 coefficient should be greater than 0.

7. Forecasting

If the hypothesis testing was positive, i.e. the theory was concluded to be correct, we **forecast** the values of the wage by **predicting** the values of education. For example, how much would someone earn for an additional year of schooling? If the X variable is the years of schooling, the B2 coefficient gives the answer to the question.

8. Use for Policy Recommendation

Lastly, if the theory seems to make sense and the econometric model was not refuted on the basis of the hypothesis test, we can go on to use the theory for policy recommendation. If your theory was really good, then maybe you will earn the Nobel Prize of Economics.

Q1 (b): Regression analysis is a statistical procedure to obtain estimates. Causal analysis isn't a specific statistical procedure, it can be regression analysis, path analysis, or variance analysis. For example, if the research designs allows causal conclusions, then a regression analyses on that data will be a causal analysis.

Regression deals with dependence amongst variables within a model. But it cannot always imply causation. For example, we stated above that rainfall affects crop yield and there is data that support this. However, this is a one-way relationship: crop yield cannot affect rainfall. It means there is no cause and effect reaction on regression if there is no causation.

Q2 (a): R-squared or R2 explains the degree to which your input variables explain the variation of your output / predicted variable. So, if R-square is 0.8, it means 80% of the variation in the output variable is explained by the input variables. So, in simple terms, higher the R squared, the more variation is explained by your input variables and hence better is your model.

However, the problem with R-squared is that it will either stay the same or increase with addition of more variables, even if they do not have any relationship with the output variables. This is where “Adjusted R square” comes to help. Adjusted R-square penalizes you for adding variables which do not improve your existing model.

Hence, if you are building Linear regression on multiple variable, it is always suggested that you use Adjusted R-squared to judge goodness of model. In case you only have one input variable, R-square and Adjusted R squared would be exactly same.

Typically, the more non-significant variables you add into the model, the gap in R-squared and Adjusted R-squared increases.

For Example:

In the world of investing, R-squared is expressed as a percentage between 0 and 100, with 100 signaling perfect correlation and zero no correlation at all. The figure does not indicate how well a particular group of securities is performing. It only measures how closely the returns align with those of the measured benchmark. It is also backward-looking—it is not a predictor of future results.

Adjusted R-squared can provide a more precise view of that correlation by also taking into account how many independent variables are added to a particular model against which the stock index is measured. This is done because such additions of independent variables usually increase the reliability of that model—meaning, for investors, the correlation with the index.

Q2 (b): The disturbance term, or error term, is central in econometrics and modeling more generally.

Different models have different assumptions about the error term, which themselves reflect features of the data generating mechanism you think you have.

So what makes the error term important is that the assumptions placed on it reflect assumptions about the process you think is generating the data. Error terms with different assumptions require different types of modeling.

Q3 (a): The Least Squares Estimators as Random Variables

when the formulas for b1 and b2 are taken to be rules that are used whatever the sample data turn out to be, then b1 and b2 are random variables since their values depend on the random variable y whose values are not known until the sample is collected. In this context we call b1 and b2 the least squares estimators. When actual sample values, numbers, are substituted into the formulas, we obtain numbers that are values of random variables. In this context we call b1 and b2 the least squares estimates.

The Sampling Properties of the Least Squares Estimators

The means (expected values) and variances of random variables provide information about the location and spread of their probability distributions. As such, the means and variances of b1 and b2 provide information about the range of values that b1 and b2 are likely to take. Knowing this range is important, because our objective is to obtain estimates that are close to the true parameter values. Since b1 and b2 re random variables, they may have covariance, and this we will determine as well. These “pre- data” characteristics of b1 and b2 are called sampling properties, because the randomness of the estimators is brought on by sampling from a population.

Q3 (b): The Seven Classical OLS Assumption

1. The regression model is linear in the coefficients and the error term
2. The error term has a population mean of zero
3. All independent variables are uncorrelated with the error term
4. Observations of the error term are uncorrelated with each other
5. The error term has a constant variance (no heteroscedasticity)
6. No independent variable is a perfect linear function of other explanatory variables
7. The error term is normally distributed (optional)