1. What is theoretical econometrics?
Theoretical econometricians investigate the properties of existing statistical tests and procedures for estimating unknowns in the model. They also seek to develop new statistical procedures that are valid (or robust) despite the peculiarities of economic data-such as their tendency to change simultaneously.

2. What are the various types of econometrics?
There are two branches of econometrics: theoretical econometrics and applied econometrics. The former is concerned with methods, both their properties and developing new ones.

3. How do econometricians proceed in their analysis of an economic problem?
Statement of theory or hypothesis.
Specification of the mathematical model of the theory
Specification of the statistical, or econometric, model
Obtaining the data
Estimation of the parameters of the econometric model
Hypothesis testing
Forecasting or prediction
Using the model for control or policy purposes.

4. What is the difference between the population and sample regression functions? Is this a distinction without a difference?
Population regression function (PRF) is the locus of the conditional mean of variable Y (dependent variable) for the fixed variable X (independent variable). The sample regression function (SRF) shows the estimated relation between explanatory or independent variable X and dependent variable Y.

5. What do you mean by a two-variable regression model?
In the simplest type of linear regression analysis, we model the relationship between 2. variables y and x, and this is assumed to be a linear relationship. In particular, we are. interested in the expected value of the random variable, y, given a specific value for x.

6. What do you understand by R-squared?
R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

7. Define a two-variable model.
In general, a solution of a system in two variables is an ordered pair that makes BOTH equations true. In other words, it is where the two graphs intersect, and what they have in common. So, if an ordered pair is a solution to one equation, but not the other, then it is NOT a solution to the system.

8. What is the Least Squares Method of Estimation?
The least squares method is a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve. Least squares regression is used to predict the behavior of dependent variables.

9. GLS Explain the application of estimating parameters of GLS. The GLS model is useful in the regionalization of hydrologic data.
GLS is also useful in reducing autocorrelation by choosing an appropriate weighting matrix.
It is one of the best methods to estimate regression models with auto-correlated disturbances and test for serial correlation.

10. State and prove the statistical properties of the estimators.

Properties of Good Estimator

A distinction is made between an estimate and an estimator. The numerical value of the sample means is said to be an estimate of the population mean figure. On the other hand, the statistical measure used, that is, the method of estimation is referred to as an estimator. A good estimator, as common-sense dictates, is close to the parameter being estimated. Its quality is to be evaluated in terms of the following properties:

1. Unbiasedness

An estimator is said to be unbiased if its expected value is identical to the population parameter being estimated. That is if θ is an unbiased estimate of θ, then we must have $E (\theta) = \theta$. Many estimators are “Asymptotically unbiased” in the sense that the biases reduce to a practically insignificant value (Zero) when n becomes sufficiently large. The estimator S_2 is an example.

It should be noted that bias in estimation is not necessarily undesirable. It may turn out to be an asset in some situations.

2. Consistency

If an estimator, say θ, approaches the parameter θ closer and closer as the sample size n increases, θ is said to be a consistent estimator of θ. Stating somewhat more rigorously, the estimator θ is said is be a consistent estimator of θ if, as n approaches infinity, the probability approaches 1 that θ will differ from the parameter θ by no more than an arbitrary constant.

The sample mean is an unbiased estimator of μ no matter what form the population distribution assumes, while the sample median is an unbiased estimate of μ only if the population distribution is symmetrical. The sample mean is better than the sample median as an estimate of μ in terms of both unbiasedness and consistency.

3. Efficiency

The concept of efficiency refers to the sampling variability of an estimator. If two competing estimators are both unbiased, the one with the smaller variance (for given sample size) is said to be relatively more efficient. Stated in a somewhat different language, an estimator θ is said to be more efficient than another estimator θ_2 for θ if the variance of the first is less than the variance of the second. The smaller the variance of the estimator, the more concentrated the distribution of the estimator around the parameter being estimated and, therefore, the better this estimator is.

4. Sufficiency

An estimator is said to be sufficient if it conveys much information as possible about the parameter which is contained in the sample. The significance of sufficiency lies in the fact that if a sufficient estimator exists, it is unnecessary to consider any other estimator; a sufficient estimator ensures that all information a sample can furnish to the estimation of a parameter is being utilized.

Many methods have been devised for estimating parameters that may provide estimators satisfying these properties. The two important methods are the least square method and the method of maximum likelihood.

11. We take the relationship between consumption and income as follows:
$Y_i = \alpha+\beta X_i + \varepsilon_i$, where Y is consumption and X is income. We take the hypothetical data for the above variables as follows:

Consumption(Y)(Rs.)	70	65	90	95	110	115	120	140	145	150
Income(X)(Rs.)	80	100	120	140	160	180	200	220	240	260

Find the estimated regression line.

12. When do you use a dummy variable?

Dummy variables can be used in regression analysis just as readily as quantitative variables. A regression model may contain only dummy explanatory variables. Regression models that contain only dummy explanatory variables are called analysis-of-variance (ANOVA) models.

Consider the following example of the ANOVA model:

Yi = B1 + B2Di + ui

where Y = annual expenditure on food ($) Di = 1 if female = 0 if male Note that the model given above is like the two-variable regression models encountered previously except that instead of a quantitative explanatory variable X, we have a qualitative or dummy variable D. As noted earlier, from now on we will use D to denote a dummy variable.

13. How to interpret dummy variables?

Once a categorical variable has been recoded as a dummy variable, the dummy variable can be used in regression analysis just like any other quantitative variable.

For example, suppose we wanted to assess the relationship between household income and political affiliation (i.e., Republican, Democrat, or Independent). The regression equation might be:

Income = $b_0 + b_1X_1 + b_2X_2$

where b_0, b_1, and b_2 are regression coefficients. X_1 and X_2 are regression coefficients defined as:

X_1 = 1, if Republican; X_1 = 0, otherwise.

X_2 = 1, if Democrat; X_2 = 0, otherwise.

The value of the categorical variable that is not represented explicitly by a dummy variable is called the reference group. In this example, the reference group consists of Independent voters.

In the analysis, each dummy variable is compared with the reference group. In this example, a positive regression coefficient means that income is higher for the dummy variable political affiliation than for the reference group; a negative regression coefficient means that income is lower. If the regression coefficient is statistically significant, the income discrepancy with the reference group is also statistically significant.

14. Explain the methods of detecting multicollinearity.

How to Measure Multicollinearity

There are two popular ways to measure multicollinearity:

(1) compute a coefficient of multiple determination for each independent variable, or

(2) compute a variance inflation factor for each independent variable.

Coefficient of Multiple Determination

We described how the coefficient of multiple determination (R^2) measures the proportion of variance in the dependent variable that is explained by all of the independent variables.

If we ignore the dependent variable, we can compute a coefficient of multiple determination (R^2_k) for each of the k independent variables. We do this by regressing the k^th independent variable on all of the other independent variables. That is, we treat X_k as the dependent variable and use the other independent variables to predict X_k.

How do we interpret R^2_k? If R^2_k equals zero, variable k is not correlated with any other independent variable; and multicollinearity is not a problem for variable k. As a rule of thumb, most analysts feel that multicollinearity is a potential problem when R^2_k is greater than 0.75; and, a serious problem when R^2_k is greater than 0.9.

15. What are the remedial measures which we can take in case of heteroskedasticity?

There are a set of heteroscedasticity tests and remedies that require an assumption about the structure of the heteroscedasticity if it exists. That is, to use these tests you must choose a specific functional form for the relationship between the error variance and the variables that you believe determine the error variance. The major difference between these tests is the functional form that each test assumes.

Breusch-Pagan Test

The Breusch-Pagan test assumes the error variance is a linear function of one or more variables.

Harvey-Godfrey Test

The Harvey-Godfrey test assumes the error variance is an exponential function of one or more variables. The variables are usually assumed to be one or more of the explanatory variables in the regression equation.

The White Test

The white test of heteroscedasticity is a general test for the detection of heteroscedasticity existence in the data set. It has the following advantages:

1. It does not require you to specify a model of the structure of the heteroscedasticity if it exists.
2. It does not depend on the assumption that the errors are normally distributed.
3. It specifically tests if the presence of heteroscedasticity causes the OLS formula for the variances and the covariances of the estimates to be incorrect.

Remedies for Heteroscedasticity

Suppose that you find evidence of the existence of heteroscedasticity. If you use the OLS estimator, you will get unbiased but inefficient estimates of the parameters of the model. Also, the estimates of the variances and covariances of the parameter estimates will be biased and inconsistent, and as a result hypothesis tests will not be valid. When there is evidence of heteroscedasticity, econometricians do one of the two things:

- Use the OLS estimator to estimate the parameters of the model. Correct the estimates of the variances and covariances of the OLS estimates so that they are consistent.
- Use an estimator other than the OLS estimator to estimate the parameters of the model.

Many econometricians choose the first alternative. This is because the most serious consequence of using the OLS estimator when there is heteroscedasticity is that the estimates of the variances and covariances of the parameter estimates are biased and inconsistent. If this problem is corrected, then the only shortcoming of using OLS is that you lose some precision relative to some other estimator that you could have used. However, to get more precise estimates with an alternative estimator, you must know the approximate structure of the heteroscedasticity. If you specify the wrong model of heteroscedasticity, then this alternative estimator can yield estimates that are worse than the OLS.

16. What are the types of heteroscedasticity?

The Types Heteroskedasticity

Unconditional

Unconditional heteroskedasticity is predictable and can relate to variables that are cyclical by nature. This can include higher retail sales reported during the traditional holiday shopping period or the increase in air conditioner repair calls during warmer months.

Changes within the variance can be tied directly to the occurrence of particular events or predictive markers if the shifts are not traditionally seasonal. This can be related to an increase in smartphone sales with the release of a new model as the activity is cyclical based on the event but not necessarily determined by the season.

Heteroskedasticity can also relate to cases where the data approach a boundary-where the variance must necessarily be smaller because the boundary restricts the range of the data.

Conditional

Conditional heteroskedasticity is not predictable by nature. There is no telltale sign that leads analysts to believe data will become more or less scattered at any point in time. Often, financial products are considered subject to conditional heteroskedasticity as not all changes can be attributed to specific events or seasonal changes.

A common application of conditional heteroskedasticity is to stock markets, where the volatility today is strongly related to the volatility yesterday. This model explains periods of persistent high volatility and low volatility.

17. Describe the Goldfeld-Quandt Test of heteroscedasticity. Give its limitations also.

In statistics, the Goldfeld–Quandt test checks for homoscedasticity in regression analyses. It does this by dividing a dataset into two parts or groups, and hence the test is sometimes called a two-group test. The Goldfeld–Quandt test is one of two tests proposed in a 1965 paper by Stephen Goldfeld and Richard Quandt. Both parametric and nonparametric tests are described in the paper, but the term "Goldfeld–Quandt test" is usually associated only with the former.

*18.*What are the various tests for autocorrelation?

Durbin-Watson Test: We usually assume that the error terms are independent unless there is a specific reason to think that this is not the case. Usually, violation of this assumption occurs because there is a known temporal component for how the observations were drawn. The easiest way to assess if there is dependency is by producing a scatterplot of the residuals versus the time measurement for that observation (assuming you have the data arranged according to a time sequence order). If the data are independent, then the residuals should look randomly scattered about 0. However, if a noticeable pattern emerges (particularly one that is cyclical) then dependency is likely an issue.

Ljung-Box Q Test: The Ljung-Box Q test (sometimes called the Portmanteau test) is used to test whether or not observations over time are random and independent.

19. What are the problems that arise from autocorrelation?

Autocorrelation refers to the degree of correlation between the values of the same variables across different observations in the data. The concept of autocorrelation is most often discussed in the context of time series data in which observations occur at different points in time (e.g., air temperature measured on different days of the month). For example, one might expect the air temperature on the 1st day of the month to be more similar to the temperature on the 2nd day compared to the 31st day. If the temperature values that occurred closer together in time are, in fact, more similar than the temperature values that occurred farther apart in time, the data would be correlated.

However, autocorrelation can also occur in cross-sectional data when the observations are related in some other way. In a survey, for instance, one might expect people from nearby geographic locations to provide more similar answers to each other than people who are more geographically distant. Similarly, students from the same class might perform more similarly to each other than students from different classes. Thus, autocorrelation can occur if observations are dependent on aspects other than time. Autocorrelation can cause problems in conventional analyses (such as ordinary least squares regression) that assume independence of observations.

In a regression analysis, the autocorrelation of the regression residuals can also occur if the model is incorrectly specified. For example, if you are attempting to model a simple linear relationship but the observed relationship is non-linear (i.e., it follows a curved or U-shaped function), then the residuals will be autocorrelated.

20. Explain the Runs Test. What are its limitations?

Run Test: This method is similar to the run test for randomness.

In this method first, the regression model is fitted using OLS method and the residuals are obtained.

The residuals are arranged according to time.

The no. of runs (R) formed by + and – signs are counted. If it exceeds the tabulated value then autocorrelation is said to be absent.

If N1 & N2 are no. of + & – signs respectively then for a large sample the test can be approximated by Wald's test using.

21. Explain the instrumental variables Method.

An instrumental variable (sometimes called an "instrument" variable) is a third variable, Z, used in regression analysis when you have endogenous variables-variables that are influenced by other variables in the model. In other words, you use it to account for unexpected behavior between variables.

22. Find the value of b_1 and b_2 using deviation method for the given data below:

X	2	3	4	5
Y	1	4	8	12

23. Discuss D-W d test decision rule.

The Durbin Watson statistic is a test statistic to detect autocorrelation in the residuals from a regression analysis. It is named after professor James Durbin, a British statistician and econometrician, and Geoffrey Stuart Watson, an Australian statistician.

The Durbin Watson statistic is a test statistic used in statistics to detect autocorrelation in the residuals from a regression analysis.

The Durbin Watson statistic will always assume a value between 0 and 4. A value of DW = 2 indicates that there is no autocorrelation.

One important way of using the test is to predict the price movement of a particular stock based on historical data.

Special consideration

A rule of thumb is that DW test statistic values in the range of 1.5 to 2.5 are relatively normal. Values outside this range could, however, be a cause for concern. The Durbin–Watson statistic, while displayed by many regression analysis programs, is not applicable in certain situations.

For instance, when lagged dependent variables are included in the explanatory variables, then it is inappropriate to use this test.

24. What is Autocorrelation?

Serial correlation, also called autocorrelation, refers to the degree of correlation between the values of variables across different data sets. It is usually used when working with time series data in which observations occur at different points in time (e.g., wind speed measured on different days of the week). If the wind speed values measured that occurred closer in time are more similar to the values that occurred farther apart in time, the data is said to be correlated.

25. What are Residuals in Statistics?

In statistics, residuals are nothing but the difference between the observed value and the mean value that a particular model predicts for that observation. Residual values are extremely useful in regression analysis as they indicate the extent to which a model accounts for the variation in the given data.

26. What is Regression Analysis?

Regression analysis is a method used in statistics that helps to identify which variables exert an impact on a particular experiment topic. The process helps determine which factors matter the most, which are to be ignored, and how the factors influence each other. Variables play an important role in regression, and it is important to understand the types of variables:

Dependent Variable: The main factor that is being understood or predicted in the experiment, dependent on other variables

Independent Variable: Variables that impact the dependent variable

27. How to Calculate the Durbin Watson Statistic

The hypotheses followed for the Durbin Watson statistic:

H(0) = First-order autocorrelation does not exist.

H(1) = First-order autocorrelation exists.

The assumptions of the test are:

- Errors are normally distributed with a mean value of 0
- All errors are stationary.

The formula for the test is:

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

Where:

- Et is the residual figure
- T is the number of observations of the experiment.

28. Interpreting the Durban Watson Statistic

The Durban Watson statistic will always assume a value between 0 and 4. A value of DW = 2 indicates that there is no autocorrelation. When the value is below 2, it indicates a positive autocorrelation, and a value higher than 2 indicates a negative serial correlation.

To test for positive autocorrelation at significance level α (alpha), the test statistic DW is compared to lower and upper critical values:

If DW < Lower critical value: There is statistical evidence that the data is positively autocorrelated

If DW > Upper critical value: There is no statistical evidence that the data is positively correlated.

If DW is in between the lower and upper critical values: The test is inconclusive.

To test for negative autocorrelation at significance level α (alpha), the test statistic 4-DW is compared to lower and upper critical values:

If 4-DW < Lower critical value: There is statistical evidence that the data is negatively autocorrelated.

If 4-DW > Upper critical value: There is no statistical evidence that the data is negatively correlated.

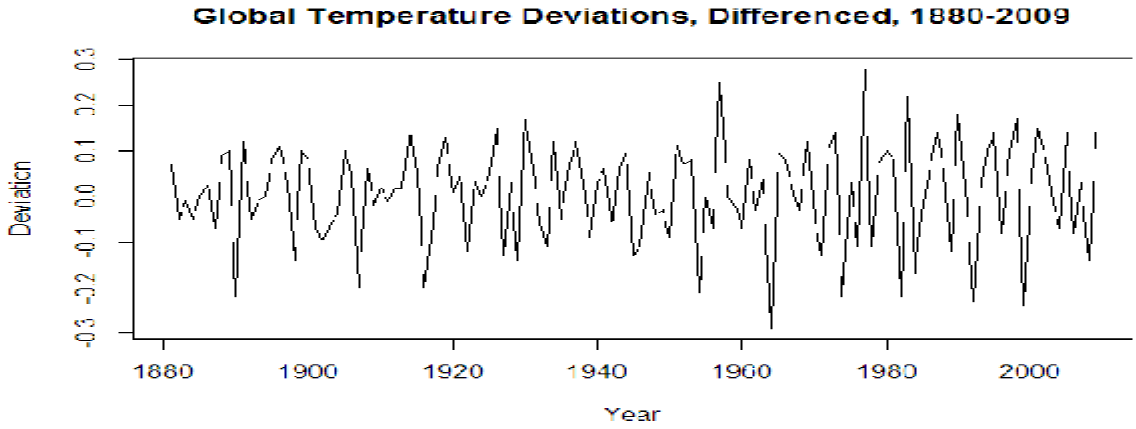
If 4-DW is in between the lower and upper critical values: The test is inconclusive.

29. Differentiate Stationary and Non-Stationary time series.

Stationary time series:

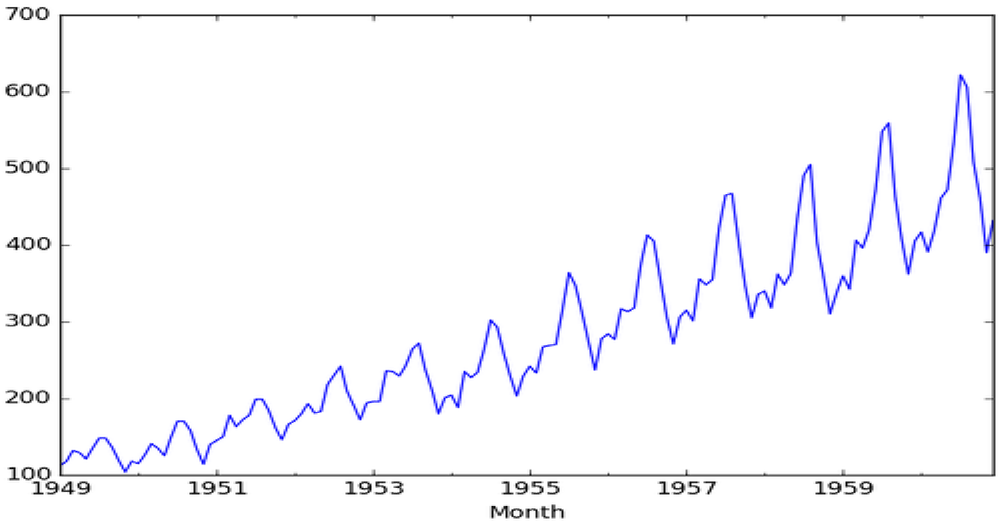
- In such a time series the statistical measures such as the mean, standard deviation, and autocorrelation are somewhat similar over time.

- It has no trend.



Non-Stationary time series:

- In such a time series the statistical measures such as the mean, standard deviation, and autocorrelation show a decreasing or increasing trend over time.
- It has a trend.
- The below plot shows an increasing trend.



30. Explain Multiple linear regression with an example.

Regression models are used to describe relationships between variables by fitting a line to the observed data. Regression allows you to estimate how dependent variable changes as the independent variable(s) change. Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. You can use multiple linear regression when you want to know:

1. How strong the relationship is between two or more independent variables and one dependent variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth).
2. The value of the dependent variable at a certain value of the independent variables (e.g. the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).

Example. You are a public health researcher interested in social factors that influence heart disease. You survey 500 towns and gather data on the percentage of people in each town who smoke, the percentage of people in each town who bike to work, and the percentage of people in each town who have heart disease.

Because you have two independent variables and one dependent variable, and all your variables are quantitative, you can use multiple linear regression to analyze the relationship.

31. Assumptions of multiple linear regression

Multiple linear regression makes all of the same assumptions as simple linear regression:

Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.

Independence of observations: the observations in the dataset were collected using statistically valid methods, and there are no hidden relationships among variables.

In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated (r2 > ~0.6), then only one of them should be used in the regression model.

Normality: The data follows a normal distribution.

Linearity: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

32. How to perform a multiple linear regression

Multiple linear regression formula

The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

- y = the predicted value of the dependent variable
- B_0 = the y-intercept (value of y when all other parameters are set to 0)
- B_1X_1 = the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
- … = do the same for however many independent variables you are testing
- B_nX_n = the regression coefficient of the last independent variable
- ϵ = model error (a.k.a. how much variation there is in our estimate of y)

To find the best-fit line for each independent variable, multiple linear regression calculates three things:

- The regression coefficients that lead to the smallest overall model error.
- The t-statistic of the overall model.
- The associated p-value (how likely it is that the t-statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true).

It then calculates the t-statistic and p-value for each regression coefficient in the model.

33. What are the assumptions of linear regression?

There are primarily five assumptions of linear regression. They are:

1. There is a linear relationship between the predictors (x) and the outcome (y)
2. Predictors (x) are independent and observed with negligible error
3. Residual Errors have a mean value of zero
4. Residual Errors have constant variance
5. Residual Errors are independent from each other and predictors (x)

34. Discuss advantages of GLMs over traditional (OLS) regression

We do not need to transform the response to have a normal distribution.

The choice of link is separate from the choice of random component, giving us more flexibility in modeling.

The models are fitted via maximum likelihood estimation, so likelihood functions and parameter estimates benefit from asymptotic normal and chi-square distributions.

All the inference tools and model checking that we will discuss for logistic and Poisson regression models apply for other GLMs too; e.g., Wald and Likelihood ratio tests, deviance, residuals, confidence intervals, and overdispersion.

There is often one procedure in a software package to capture all the models listed above, e.g. PROC GENMOD in SAS or glm() in R, etc., with options to vary the three components.

35. Discuss testing the assumptions of linear regression

There are four principal assumptions that justify the use of linear regression models for purposes of inference or prediction:

(i) linearity and additivity of the relationship between dependent and independent variables:
(a) The expected value of the dependent variable is a straight-line function of each independent variable, holding the others fixed.
(b) The slope of that line does not depend on the values of the other variables.
(c) The effects of different independent variables on the expected value of the dependent variable are additive.
(ii) statistical independence of the errors (in particular, no correlation between consecutive errors in the case of time series data)
(iii) homoscedasticity (constant variance) of the errors
(a) versus time (in the case of time series data)
(b) versus the predictions
(c) versus any independent variable
(iv) normality of the error distribution.

36. Explain inverse regression

Inverse regression refers to (inversely) predicting the corresponding value of an independent variable when one only observes the value(s) of the corresponding dependent variable, using a model that has already been established for the dependence between the two variables.