# Questions about terminology, notation (syntax) and process (logic)

1. If the explanatory variable and the error term are correlated then the response will be \_\_\_\_\_\_ . biased
2. An \_\_\_\_\_\_ variable is one that is correlated with the error term . endogenous
3. An \_\_\_\_\_\_ variable is one that is uncorrelated with the error term . exogenous
4. If Question 1 is true then the least squares estimator is said to be \_\_\_\_\_ . inconsistent
5. In these cases, the explanatory variable is \_\_\_\_\_\_ and the least squares estimator is NOT an \_\_\_\_\_\_ estimator of the population parameters. endogenous unbiased
6. One cause of endogeneity is when explanatory variables are measured with or correlated with \_\_\_\_\_\_ (see pictogram below). errors

y

x

e

Figure 1

1. In Figure 1 the error term is only associated with the response (dependent) variable. True/False
2. Another cause of endogeneity is reverse \_\_\_\_\_\_, i.e. when the explanatory both causes and is caused by the dependent variable. causality
3. A \_\_\_\_\_\_ estimator goes by several names including a method of moments estimator or a two-stage lease squares estimator. Instrumental variables

Quote from Principles of Econometrics by Hill, Griffiths and Lim, “We offer fair warning, however, that this area of econometrics is filled with practical and theoretical difficulties. Our search turns from finding an estimator that is “best” to one that is “adequate,” and unfortunately producing convincing research applications requires knowledge, skill, and patience. In order for you to begin properly you should reread (right now!) Section 2.10 on the exogeneity concept and Section 5.7 on the large sample, or asymptotic, properties of the least squares estimator” (Hill, Griffiths, Lim, 2019).

However, you should not think that instrumental variables or instruments are the only controversial method in data analytics. Other controversial topics include feature selection and latent factor analysis which are both methods to reduce dimensionality in machine learning. Latent factor analysis is particularly fraught with challenges in much the same way that the use of instrumental variables is. That is, it is very subjective and much is left to the researcher’s expertise.

1. In a perfect world, controlled experiments would be used to determine the effect of on . Everything else, e.g. any random factors are included in the \_\_\_\_\_\_ and are statistically independent of . error term

y

e

x

Figure 2

1. In Figure 2 above the error term is only associated with the response (dependent) variable. True/False
2. Associated with the error term is an assumption of strict \_\_\_\_\_\_\_ and the expected value is zero. exogeneity
3. A “weaker” assumption is that the explanatory variable and the random error are contemporaneously \_\_\_\_\_\_. uncorrelated
4. \_\_\_\_\_\_ variables (IV) regression breaks x into two parts: one part that might be \_\_\_\_\_\_ with the error term e and one part that is not. By isolating the part that is NOT correlated with e, it is possible to make an \_\_\_\_\_\_ estimate . Instrumental correlated unbiased
5. An instrumental variable, z, is associated with \_\_\_\_\_ but not with the \_\_\_\_\_ term. (See pictogram below) x error

y

e

x

z

Figure 3

1. In Figure 3 the instrumental variable is associated with both the explanatory (predictor or independent) variable and the error term. True/False
2. In Figure 3 the instrumental variable is intended to separate the exogenous part of the explanatory (predictor or independent) variable from the rest (endogenous) of the explanatory (predictor or independent) variable. True/False
3. The hardest part of instrumental variables regression is finding an appropriate instrument or appropriate instruments. True/False
4. IV regression focuses on the case that the explanatory variable is \_\_\_\_\_\_ and there is an instrument which is \_\_\_\_\_\_. endogenous exogenous
5. An instrumental variable, , is \_\_\_\_\_\_ with the explanatory variables , i.e. or in other words z predicts or causes . correlated
6. For an instrumental variable (an “instrument”) Z to be valid, it must satisfy two conditions. First, it must be \_\_\_\_\_\_, i.e. relevant
7. For an instrumental variable (an “instrument”) Z to be valid, it must satisfy two conditions. Second, it must be \_\_\_\_\_\_, i.e. exogenous
8. The condition of \_\_\_\_\_\_, , ensures that you don’t divide by zero. relevance
9. In Two-Stage Least Squares (TSLS) regression the original OLS linear model, is rewritten as the structural equation, , where is the dependent variable, splits out and represents the endogenous variables and are the exogenous variables. True/False
10. In Two-Stage Least Squares (TSLS) regression, the first stage (in reduced form) is represented with an equation that only contains exogenous regressors (exogenous variables). That is, . True/False
11. To solve the Two-Stage Least Squares (TSLS) regression, the structural equation model must involve a combined set of both endogenous and exogenous variables, i.e. . True/False
12. To solve the Two-Stage Least Squares (TSLS) regression, we need to find an instrument or set of instruments, of only exogenous variables where is the instrument for itself and is the instrument for . True/False
13. In Two-Stage Least Squares (TSLS) regression when the estimated coefficients in the OLS estimation are biased, that is , and ., Therefore, the predicted values are calculated from the rewritten structural equation model . True/False

**Overall Two-Stage Least Squares estimation procedure**

1. The Two-Stage Least Squares (TSLS) procedure replaces the endogenous variable(s) with predicted values of this endogenous variable when regressed on instruments. True/False
2. In the first stage of the TSLS regression a reduced form model (equation) including predicted values is estimated using only exogenous regressors. True/False
3. In the second stage of the TSLS regression, the predicted values are substituted in the structural equation to remove any endogeneity and solve for the dependent variable. True/False
4. The best case for TSLS regression is to have strong instruments and a large sample size. True/False

**Identification Issues**

1. The “order condition” means that the number of instrumental variables must be at least as large as the number of endogenous regressors. True/False
2. There should be at least one instrument for each endogenous variable. True/False
3. The “rank condition” means that the matrices represented by must have full rank in order to be inverted. Notice that the “ ‘ “ or prime designation in these equations represents an inverted matrix. (This is a condition from linear algebra.) True/False
4. An Instrumental Variable (IV) model is minimally or just identified if there is one instrument for each endogenous variable. True/False
5. The minimally identified or “just” identified model for the unbiased estimator can be written as:

because and therefore . That is, the instrument is not correlated with the error term. True/False

1. The under-identified model exists when there are fewer instruments than endogenous variables . In this case, no consistent estimator exists. True/False
2. If there are not at least as many instruments as there are endogenous variables then no specific model can be estimated. That is, an infinite number of solutions exists. True/False
3. The over-identified model exists if there are more instruments than endogenous variables. True/False
4. In all cases, the error term should be independent and identically distributed (iid) and homoscedastic (or homoscedastic). True/False

# Instrumental variables tests

**Hausman test for endogeneity**

1. The Hausman test for endogeneity works by comparing the OLS and IV estimates for significant differences. That is, if significant differences exist then the regressor (independent variable) is endogenous. If no significant differences exist then the regressor (independent variable) is exogenous. True/False
2. In the Hausman test, if no significant difference between the OLS and IV estimates exists that means that the instruments used in the IV regression are not making much difference so the regressor (independent variable) is assumed to be exogenous. True/False

**Durbin-Wu-Hausman test for exogenous regressors**

1. The Durbin-Wu-Hausman test for exogenous regressors works by determining whether or not the covariance is not zero, i.e. . True/False
2. The Durbin-Wu-Hausman test considers whether the residuals from the first-stage of TSLS regression are significantly different from zero, or not. True/False

**Test for over-identifying restrictions**

1. You can also test for over-identifying restrictions by estimating a model using the generalized method of moments (GMM). True/False
2. In the procedure for testing for over-identifying restrictions using GMM, the model is distributed as chi-square with degrees of freedom equaling the number of over-identifying restrictions. True/False
3. If the procedure for testing for over-identifying restrictions results in the rejection of the null hypothesis then at least one instrument is not valid. True/False
4. You can run several tests for over-identifying restrictions to determine which (instrumental) variables are best for the problem you are working on. True/False

**Test for Weak Instruments**

1. A weak instrument does not help much in overcoming endogeneity. True/False
2. A weak instrument has a low or very low correlation with the endogenous variable. True/False
3. When several instruments exist for one endogenous variable, the best way to test for the weakness of the instruments is to consider the partial or partial F-statistic from the first stage of the TSLS regression. True/False
4. When several instruments used for one endogenous variable are weak, the partial F-statistic (of the joint significance of the coefficients of the instruments ) is less than 10. True/False