10 Common Mistakes in Applied Econometrics

From Econometrics For Dummies

By Roberto Pedace

Avoiding mistakes when you do econometric analysis depends on your ability to apply knowledge you acquired before and during your econometrics class. Following is a rundown of common pitfalls to help you improve your application of econometric analysis.

Failing to use your common sense and knowledge of economic theory

One of the characteristics that differentiate applied research in econometrics from other applications of statistical analysis is the use of economic theory and common sense to motivate the connection between the independent and dependent variables.

In econometrics, you should be able to make a strong case for the independent variables (Xs) causing changes in the dependent variable (Y). You need sound theory and good common sense to justify your approach. Doing so allows you to provide a sensible interpretation of your results in addition to the typical measures of statistical significance and fit.

Asking the wrong questions first

Getting obsessed with the technical details of estimating econometric models can be easy. However, you should always take a step back and ask yourself why you're doing what you're doing. Why will others find my topic interesting and important?

Ignoring the work and contributions of others

Failing to connect your work with that of others who have examined your research question or something closely related to it is a serious mistake. Understanding how others have dealt with similar issues can help you figure out which model to use, may yield refinements in your work, and allows readers to better understand the relevance of your topic.

In your literature review, focus on papers or segments of papers that are directly related to your work. Summarize the approach, data, and findings of other researchers. Finally, be clear about how your work fits in with what's already been done by others, what's been improved, and/or how new dimensions of the topic have been explored.

Failing to familiarize yourself with the data

Students often assume that the data they're working with is complete for all variables and that the reported information is accurate. You can reduce your chances of getting unwelcome surprises in your results by doing some exploratory work that includes descriptive statistics, line charts (for time-series data), frequency distributions, and even listings of some individual data values.

A number of undesirable outcomes can result from failing to get familiar with your analysis data. These three examples are perhaps the most common:

- *Variables you thought were measured continuously are actually in categories or groups.*
- *Measurements that you believed were real values are actually missing values.*
- *Data values that appear perfectly legitimate are actually censored values.*

Making it too complicated

The art of econometrics lies in finding the appropriate specification or functional form to model your particular outcome of interest. In many cases, however, theory can be vague about the specific elements of a model's specification.

Given the uncertainty of choosing the "perfect" specification, many applied econometricians make the mistake of overspecifying their models (meaning they include numerous irrelevant variables) or favor complicated estimation methods over more straightforward techniques. It can result in undesirable estimator properties and difficulty interpreting the meaning of the results.

Being inflexible to real-world complications

The solutions or predictions derived by using economic theories use logical deduction and/or mathematical proof that usually rely on the *ceteris paribus* (all else constant) assumption.

The data you use to test economic hypotheses, however, are derived from a world where agents (individuals, firms, or what have you) are engaged with their surrounding environment in ways that aren't likely to satisfy the *ceteris paribus* assumption because many of the variables defining their specific circumstances vary considerably from one observation to another.

Looking the other way when you see bizarre results

Most econometric research projects contain estimation results for numerous variations of related models. You want to focus on your primary variables of interest (core variables), but make sure you examine all of your results.

That means don't ignore unreasonable results (mostly insignificant estimates, coefficients with the wrong sign, and magnitudes that are too large) and proceed to reporting and interpretation. If some results don't pass a common-sense test, then the statistical tests are likely to be meaningless and may even indicate that you've made a mistake with your variables, the estimation technique, or both.

Obsessing over measures of fit and statistical significance

After you estimate an econometric model, focus your attention and guide the reader (if you're writing a research paper) to the results that are most relevant in addressing your research question.

The importance of your results shouldn't be determined on the basis of fit (R-squared values) or statistical significance alone. Sure, statistically insignificant coefficients suggest that your independent variable isn't likely to affect your dependent variable. However, if the lack of a relationship is new or unexpected, this finding may be significant!

Forgetting about economic significance

You can use measures of statistical significance to determine which variables aren't likely to have an effect on the dependent variable, but you can't use them to determine which variables have a relevant effect.

After you've established that a variable is statistically significant, don't forget to focus your attention on the coefficient. Sometimes variables can have coefficients that are highly statistically significant even though no economic significance is associated with the result.

The most important element in the discussion of your results is the evaluation of statistical significance *and* magnitude for the primary variables of interest. If a variable has a statistically significant coefficient but the magnitude is too small to be of any importance, then you should be clear about its lack of economic significance.

Assuming your results are robust

In most cases, economic theory allows for a considerable amount of flexibility in determining the exact specification of the econometric model. You'll want to see if minor adjustments change your results.

Don't assume that only one econometric model can apply to your research question and that the results won't change with reasonable modifications to your specification. You want to perform robustness (or sensitivity) analysis to show that your model estimates aren't sensitive (are robust) to slight variations in specification.

https://www.dummies.com/education/economics/econometrics/10-common-mistakes-in-applied-econometrics/

Econometrics For Dummies Cheat Sheet

By Roberto Pedace

You can use the statistical tools of econometrics along with economic theory to test hypotheses of economic theories, explain economic phenomena, and derive precise quantitative estimates of the relationship between economic variables. To accurately perform these tasks, you need econometric model-building skills, quality data, and appropriate estimation strategies. And both economic and statistical assumptions are important when using econometrics to estimate models.

Econometric Estimation and the CLRM Assumptions

Econometric techniques are used to estimate economic models, which ultimately allow you to explain how various factors affect some outcome of interest or to forecast future events. The ordinary least squares (OLS) technique is the most popular method of performing regression analysis and estimating econometric models, because in standard situations (meaning the model satisfies a series of statistical assumptions) it produces optimal (the best possible) results.

The proof that OLS generates the best results is known as the *Gauss-Markov theorem,* but the proof requires several assumptions. These assumptions, known as the *classical linear regression model* (CLRM) *assumptions,* are the following:

- The model parameters are linear, meaning the regression coefficients don't enter the function being estimated as exponents (although the variables can have exponents).
- The values for the independent variables are derived from a random sample of the population, and they contain variability.
- The explanatory variables don't have perfect collinearity (that is, no independent variable can be expressed as a linear function of any other independent variables).
- The error term has zero conditional mean, meaning that the average error is zero at any specific value of the independent variable(s).
- The model has no heteroskedasticity (meaning the variance of the error is the same regardless of the independent variable's value).
- The model has no autocorrelation (the error term doesn't exhibit a systematic relationship over time).

If one (or more) of the CLRM assumptions isn't met (which econometricians call *failing*), then OLS may not be the best estimation technique. Fortunately, econometric tools allow you to modify the OLS technique or use a completely different estimation method if the CLRM assumptions don't hold.

Useful Formulas in Econometrics

After you acquire data and choose the best econometric model for the question you want to answer, use formulas to produce the estimated output. In some cases, you have to perform these calculations by hand (sorry). However, even if your problem allows you to use econometric software such as STATA to generate results, it’s nice to know what the computer is doing.

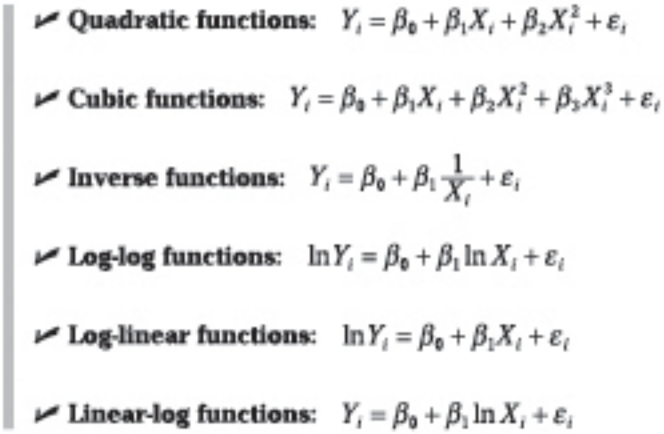
Here’s a look at the most common estimators from an econometric model along with the formulas used to produce them.

Estimate	***Formula***
Regression coefficients in a model with one independent variable	$\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$ $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$
Standard error of the estimate or mean squared error	$\hat{\sigma}_\varepsilon = \sqrt{\dfrac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-p-1}}$
Standard error of regression coefficients in a model with one independent variable	$\hat{\sigma}_{\hat{\beta}_1} = \dfrac{\hat{\sigma}_\varepsilon}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}}$ $\hat{\sigma}_{\hat{\beta}_0} = \left(\sqrt{\dfrac{\sum_{i=1}^{n}X_i^2}{n\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}}\right)\hat{\sigma}_\varepsilon$
Explained sum of squares (ESS), residual sum of squares (RSS), and total sum of squares (TSS)	$ESS = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$ $RSS = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$ $TSS = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = ESS + RSS$
Coefficient of determination; R-squared	$R^2 = \dfrac{ESS}{TSS} = 1 - \dfrac{RSS}{TSS}$
t-statistic for regression coefficients	$t = \dfrac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}}$
Confidence interval for regression coefficients	$\hat{\beta}_k \pm \left(t_{\alpha/2,n-p-1} \cdot \hat{\sigma}_{\hat{\beta}_k}\right)$

Econometric Analysis: Looking at Flexibility in Models

You may want to allow your econometric model to have some flexibility, because economic relationships are rarely linear. Many situations are subject to the “law” of diminishing marginal benefits and/or increasing marginal costs, which implies that the impact of the independent variables won’t be constant (linear).

The precise functional form depends on your specific application, but the most common are as follows:



- ✓ **Quadratic functions:** $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$
- ✓ **Cubic functions:** $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i$
- ✓ **Inverse functions:** $Y_i = \beta_0 + \beta_1 \frac{1}{X_i} + \varepsilon_i$
- ✓ **Log-log functions:** $\ln Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i$
- ✓ **Log-linear functions:** $\ln Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- ✓ **Linear-log functions:** $Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i$

Typical Problems Estimating Econometric Models

If the classical linear regression model (CLRM) doesn’t work for your data because one of its assumptions doesn’t hold, then you have to address the problem before you can finalize your analysis. Fortunately, one of the primary contributions of econometrics is the development of techniques to address such problems or other complications with the data that make standard model estimation difficult or unreliable.

The following table lists the names of the most common estimation issues, a brief definition of each one, their consequences, typical tools used to detect them, and commonly accepted methods for resolving each problem.

Problem	Definition	Consequences	Detection	Solution
High multicollinearity	Two or more independent variables in a regression model exhibit a close linear relationship.	Large standard errors and insignificant *t*-statistics Coefficient estimates sensitive to minor changes in model specification	Pairwise correlation coefficients Variance inflation factor (VIF)	1. Collect additional data. 2. Re-specify the model. 3. Drop redundant variables.

		Nonsensical coefficient signs and magnitudes		
Heteroskedasticity	The variance of the error term changes in response to a change in the value of the independent variables.	Inefficient coefficient estimates Biased standard errors Unreliable hypothesis tests	Park test Goldfeld-Quandt test Breusch-Pagan test White test	1. Weighted least squares (WLS) 2. Robust standard errors
Autocorrelation	An identifiable relationship (positive or negative) exists between the values of the error in one period and the values of the error in another period.	Inefficient coefficient estimates Biased standard errors Unreliable hypothesis tests	Geary or runs test Durbin-Watson test Breusch-Godfrey test	1. Cochrane-Orcutt transformation 2. Prais-Winsten transformation 3. Newey-West robust standard errors

Recognizing Usual Variables: Normal Distribution

By Roberto Pedace

In econometrics, a random variable with a normal distribution has a probability density function that is *continuous, symmetrical,* and *bell-shaped*. Although many random variables can have a bell-shaped distribution, the density function of a normal distribution is precisely

$$f(X)=\left(\frac{1}{\sigma_X\sqrt{2\pi}}\right)\exp\left(\frac{-(X-\mu_X)^2}{2\sigma_X^2}\right)=\left(\frac{1}{\sigma_X\sqrt{2\pi}}\right)e^{\left(\frac{-(X-\mu_x)^2}{2\sigma_x^2}\right)}$$

where

$$\mu_X$$

represents the mean of the normally distributed random variable *X*,

$$\sigma_X$$

is the standard deviation, and

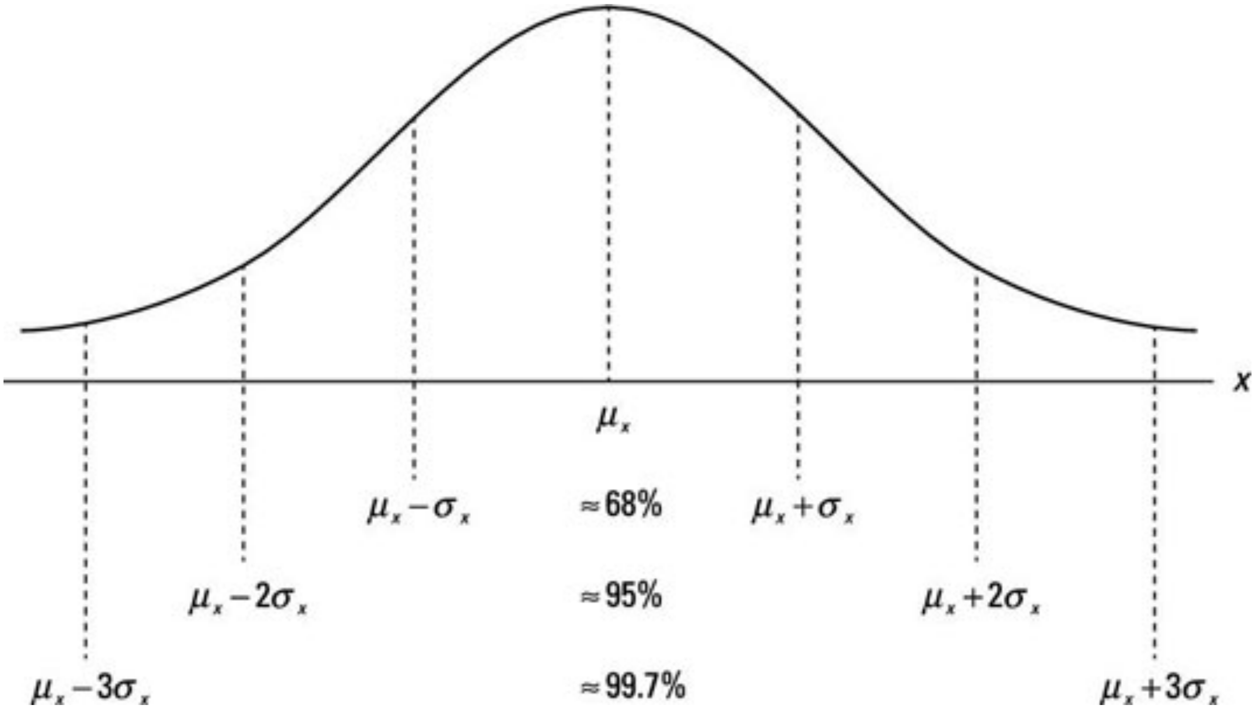
$$\sigma_X^2$$

represents the variance of the normally distributed random variable.

A shorthand way of indicating that a random variable, *X*, has a normal distribution is to write

$$X \sim N\left(\mu_X, \sigma_X^2\right)$$

A distinctive feature of a normal distribution is the probability (or density) associated with specific segments of the distribution. The normal distribution in the figure is divided into the most common intervals (or segments): one, two, and three standard deviations from the mean.



With a normally distributed random variable, approximately 68 percent of the measurements are within one standard deviation of the mean, 95 percent are within two standard deviations, and 99.7 percent are within three standard deviations.

Suppose you have data for the entire population of individuals living in retirement homes. You discover that the average age of these individuals is 70, the variance is 9

$$\left(\text{standard deviation, } \sigma_X = \sqrt{9} = 3\right)$$

and the distribution of their age is normal. Using shorthand, you could simply write this information as

$$X \sim N(70,9)$$

If you randomly select one person from this population, what are the chances that he or she is more than 76 years of age?

Using the density from a normal distribution, you know that approximately 95 percent of the measurements are between 64 and 76

$$(70 - 2\sigma_X < X < 70 + 2\sigma_X)$$

(notice that 6 is equal to two standard deviations). The remaining 5 percent are individuals who are less than 64 years of age or more than 76. Because a normal distribution is symmetrical, you can conclude that you have about a 2.5 percent (5% / 2 = 2.5%) chance that you randomly select somebody who is more than 76 years of age.

If a random variable is a linear combination of another normally distributed random variable(s), it also has a normal distribution.

Suppose you have two random variables described by these terms:

$$X \sim N\left(\mu_X, \sigma_X^2\right)$$

$$Y \sim N\left(\mu_Y, \sigma_Y^2\right)$$

In other words, random variable *X* has a normal distribution with a mean of

$$\mu_X$$

and variance of

$$\sigma_X^2$$

and random variable *Y* has a normal distribution with a mean of

$$\mu_Y$$

and a variance of

$$\sigma_X^2$$

If you create a new random variable, *W*, as the following linear combination of *X* and *Y, W* = *aX* + *bY*, then *W* also has a normal distribution. Additionally, using expected value and variance properties, you can describe the new random variable with this shorthand notation:

$$W \sim N\left(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}\right)$$

https://www.dummies.com/education/economics/econometrics/econometrics-for-dummies-cheat-sheet/