Quantitative Methods, Basic Assumptions

The use of quantitative methods in political science generally means the application of a statistical model to political science data, and a statistical model is simply a set of compatible probabilistic assumptions. Fundamentally, assumptions are modelling choices made by a researcher concerning the distribution of the data to be modelled, how the parameters of that distribution change over observations or time, and the dependence of one observation on another. The assumptions serve the dual purpose of reducing the number of parameters in the model that must be estimated and imbuing potential estimators with certain properties. The goals of the modelling process are description and inference, and how well a model accomplishes these goals is a direct function of how appropriate its assumptions are for a particular data set.

The structure of the entry proceeds as follows. First, the relationship between assumptions, models, and estimators is discussed. Second, the assumptions of the linear regression model are discussed in detail as more complex models are often defined as departures from this set of assumptions. Particular attention is paid to the assumptions necessary for attaining correct coefficient estimates and standard errors. Threats to these core assumptions are assessed in terms of their effects, both in theory and in practice. A brief discussion of possible remedies follows. The final sections discuss the degree to which the assumptions of the linear regression model are modified for use in the generalized linear model, and how the assumption of a random sample is dealt with in a discipline where random samples are rare.

Assumptions, Models, and Estimators

A statistical model is a mathematical representation of the actual process in the world that generated the data (known as a data generating process or DGP). The point of creating a statistical model is both to describe the data produced by the DGP and to make inferences about features of the DGP that are unknown. As noted above, the model, itself, consists of a set of assumptions. This account makes clear that assumptions are characteristics of models and not characteristics of data, and the question to ask of a model is not whether it is true or false, but how descriptively useful is it. Models are more or less useful in describing a given data set, and when an assumption fails to be useful in the process of description, the error lies with the model, not the data.

Once a model has been chosen, then an estimator (a function of the sample data that provides an estimated value for an unknown parameter) is chosen. Many common models can be estimated by any number of estimators, and the choice between estimators is driven by the model's assumptions, which imbue the estimators with various properties. An estimator with good (or better) properties under the assumptions of the model is chosen over an estimator with bad (or worse) properties under the assumptions of the model. The reference to “good estimators” and not to “good estimates” is by design. In frequentist statistics, a good estimate, by definition, is one produced by a good estimator. Properties of estimators that are considered good include unbiasedness (the estimator is on average neither high nor low), efficiency (the estimator has a small variance around the true value), and consistency (the estimator is near the true value almost all of the time when the sample size is large). Often these properties are assessed in the aggregate as a slightly biased estimator with a small variance may be preferred over an unbiased estimator with a larger variance.

The practice and art of data analysis lies in understanding which assumptions, and therefore which models and estimators, are appropriate for a given data set. This understanding comes from three sources: the data themselves, theory, and substantive knowledge. Some assumptions can be tested, but such tests are often inconclusive, work only under specific conditions, and rarely provide more than vague recommendations. Information drawn from theory and substantive knowledge of the process being modelled are far more reliable guides. The discussion to follow necessarily abstracts from the particulars of any one data set, but the importance of using information from outside the data should be kept in mind.

Assumptions and the Linear Regression Model

Most discussions of model assumptions begin with the linear regression model. Although linear regression is no longer the workhorse of political science—pride of place goes to the generalized linear model—the linear regression model serves as a good starting point as the various roles that the assumptions play in the model are clear, and some simple results can be established.

The discussion of the linear model centers on the equation,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is a $n \times 1$ vector of observations on the dependent variable, $\mathbf{X}$ is an $n \times K$ matrix of regressors (including a constant), $\boldsymbol{\beta}$ is a $K \times 1$ vector of coefficients, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of disturbances. The data represent a random sample from a population of interest.

The discussion also centers on the most common estimator of the linear regression model, the least squares estimator. Replacing the vector of unknown coefficients $\boldsymbol{\beta}$ with an estimate $\hat{\boldsymbol{\beta}}$ defines a vector of residuals,

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}.$$

The least squares estimator is found by choosing $\hat{\boldsymbol{\beta}}$ to minimize the sum of squared residuals or $\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}$. The first step in this process is to take the derivative of the sum of squared residuals with respect to $\hat{\boldsymbol{\beta}}$,

$$\begin{aligned}
\frac{\partial \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{\partial \hat{\boldsymbol{\beta}}} &= \frac{\partial (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} \\
&= \frac{\partial \left(\mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}' + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}\right)}{\partial \hat{\boldsymbol{\beta}}} \\
&= 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - 2\mathbf{X}'\mathbf{y}.
\end{aligned}$$

The second and final step is to set the derivative equal to 0 and solve for $\hat{\boldsymbol{\beta}}$,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

On its own, the result of the least squares procedure has no properties; it is simply a method for fitting a line to data. The least squares method produces an estimator when paired with a statistical model that makes certain assumptions. Five assumptions commonly comprise the linear regression model. These are:

1. no exact linear relationships exist among the regressors ($\mathbf{X}$ has full column rank);

2. $\mathbf{X}$ is nonstochastic (regressors are fixed in repeated samples);

3. the expectation of the disturbance term is zero ($E[\boldsymbol{\epsilon}] = \mathbf{0}$);

4. homoscedasticity and no autocorrelation (spherical disturbances, $E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'] = \sigma^2\mathbf{I}$);
5. the disturbances are normally distributed ($\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2)$).

The linear regression model makes these assumptions, not because they are substantively likely to describe a data set (more on this later), but because they imbue the least squares estimator with certain properties that are considered good, namely unbiasedness and efficiency. The technical reasons for these assumptions are discussed first to reinforce the point that this particular set of assumptions is driven by the choice of the estimator and not by the demands of the data.

Technical Reasons for the Assumptions

The five assumptions of the linear regression model are chosen for technical reasons, but they also have substantive implications. This section describes what part each assumption plays in deriving the good properties of the estimator, and the next details the substantive commitments their use demands.

The first assumption, that no exact linear relationships exist among the regressors, allows the computation of the least squares estimator,

$$\begin{aligned} \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{y} \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \end{aligned}$$

If an exact linear relationship did exist among the regressors, it would be impossible to solve for $\hat{\boldsymbol{\beta}}$ by premultipling both sides of the above equation by the inverse of $(\mathbf{X}'\mathbf{X})$. The inverse of a square matrix exists only if the columns and rows of that matrix are linearly independent.

Assumptions 2 and 3—$\mathbf{X}$ is nonstochastic and the expectation of the disturbance term is zero—allow the claim that the least squares estimator is unbiased (technically, the expected value of the estimator equals the parameter or $E[\hat{\theta}] = \theta$). To see this, substitute $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ for $\mathbf{y}$ in the estimator above and write it in the following form,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}] \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}. \end{aligned}$$

The estimator is unbiased if the second term above goes to zero. Taking the expectation of the estimator gives

$$E[\hat{\boldsymbol{\beta}}] = E[\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}].$$

The expectations operator passes through $\boldsymbol{\beta}$ because it is a constant. Assumption 2 allows the operator to pass through the $\mathbf{X}$ as well, giving

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\boldsymbol{\epsilon}].$$

Finally, assumption 3 sends the second term above to zero, and the claim that the least squares estimator is unbiased follows, $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

Assumption 4—homoscedasticity and no autocorrelation—allows estimation of the estimator's variance. As variance is defined as $E[(X - E[X])^2]$, it is necessary to get the estimator in this form. Begin with the estimator in the form above and move the parameter to the left-hand side of the equation,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}. \end{aligned}$$

Square both sides (remembering that the inner product is the sum of squares and the outer product is variance/covariance matrix), and take the expectation (remembering that $\mathbf{X}$ is nonstochastic),

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}) &= E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

The expectation on the right-hand side includes n variances and $n(n-1)/2$ covariances,

$$E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix},$$

as $E[\epsilon_i^2] = \sigma_i^2$ and $E[\epsilon_i \epsilon_j] = \sigma_{ij}$. It is not possible to estimate this many variances and covariances with only n observations. Assumption 4, however, reduces the number of items that must be estimated from $n + n(n-1)/2$ to one, making estimation possible. Assumption 4 states that the variances are uniform, $E[\sigma_i^2] = \sigma^2$, and that the covariances are zero, $E[\epsilon_i \epsilon_j] = \sigma_{ij} = 0$. The expectation above then simplifies to

$$E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'] = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I},$$

and the expression for the variance is then

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

These four assumptions are enough to prove the Gauss-Markov theorem, which states that the least squares estimator is the most efficient estimator among the class of linear and unbiased estimators. The theorem does not say that the least squares estimator is the most efficient estimator; there exist estimators that are more efficient, but they happen to be biased or nonlinear or both.

Finally, assumption 5—normally distributed disturbances—allows derivation of the sampling distribution of the least squares estimates. Because the estimated coefficients are linear combinations of normally distributed disturbances, the estimated coefficients are normally distributed,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

Correct t-tests, F-tests, and χ^2 tests follow.

Substantive Implications of the Assumptions

While the five assumptions discussed above are primarily used to imbue the least squares estimator with good properties, these same assumptions have very specific substantive implications for the data being modelled. Rarely, however, do these assumptions accurately describe the data used by political scientists.

The assumption that no exact linear relationships exist among the regressors means that no regressor or set of regressors can be a linear combination of any other regressor or set of regressors. Not violating this assumption is mostly a matter of asking meaningful questions of the data. Problems with this assumption most often result from the careless use of dummy variables. As an example, consider a regression of the vote difference between two Presidential candidates in a district on an intercept and a set of independent variables. The regressors include a dummy variable for districts in southern states and a dummy variable for districts in non-southern states. The regression, in part, asks the question, "how do districts in southern states and districts in non-southern states differ from other districts?" The substantive problem, of course, is that there are no other districts (all districts have to be in the South or the non-South). The question is not a meaningful one. (The technical problem is that these dummy variables would be collinear with the intercept.)

The assumption that the regressors are fixed in repeated samples is equivalent to assuming that the analyst has performed an experiment where she has control over the inputs to the experiment. Thus, the experiment could be rerun many times with exactly the same values for the regressors, but with new values for the disturbances and therefore new values for the dependent variable. Each time the regression was rerun, the different values for the dependent variable would produce different estimated coefficients. Discussion of the distribution of the estimated coefficients is therefore possible.

The "fixed $\mathbf{X}$" assumption fails to describe all but a very few political science data sets. The assumption can easily be relaxed, however, either by assuming the regressors to be exogenous and conditioning on them or by assuming stochastic regressors. What is important is how the "fixed $\mathbf{X}$" assumption interacts with the assumption that the expected value of the disturbances is zero, $E[\boldsymbol{\epsilon}] = \mathbf{0}$. Combined, the two assumptions imply that the regressors are uncorrelated with the disturbances. The lack of correlation between the regressors and the disturbances is the key assumption in regression analysis. When the assumption holds, the least squares estimator is unbiased. When the assumption does not hold, the least squares estimator is biased. Unfortunately, there are a variety of substantive reasons why this key assumption is unlikely to hold. The three biggest concerns—measurement error, omitted variables, and simultaneity— are discussed in depth in the next section.

The assumption of homoscedasticity and no autocorrelation is perhaps the assumption least likely to describe a political science data set. The assumption implies that all the unmeasured factors in the disturbance affect different observations in exactly the same way. It implies that observations drawn from physically adjacent areas are no more alike each other than observations drawn from non-adjacent areas. It implies that political processes that are observed over time are not sticky or sluggish. Finally, the assumption implies that the model has not been misspecified in important ways. Few of these implications accurately describe the majority of political science data sets. A detailed discussion is below.

Getting the Coefficients Right

When estimating a statistical model, political scientists are concerned with getting the estimated coefficients right. That is, a researcher would like his or her estimated coefficients to be close to the population parameter values. Unfortunately, there is no way to know whether an estimate is close to the true parameter value (short of adopting a Bayesian perspective, which has yet to be done by the majority of political scientists). All that can be said is whether the estimator is good, and then assume that a good estimator produces a good estimate.

As noted above, there are two ways to think about a good estimator in this sense. One way is to say that an estimator is good if it is on average neither high nor low, a property that is known as unbiasedness. Formally, an estimator is unbiased if its expectation is equal to the population parameter, $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. The other way is to say that an estimator is good is if it is near the true value almost all of the time when the sample size is large, a property that is known as consistency. Formally, an estimator is consistent if its probability limit or plim is equal to the population parameter, $\operatorname{plim}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

Of the two definitions of the term good, the latter is closer to what political scientists actually mean. An unbiasedness estimator only informs about the average estimate, not the estimate actually calculated. Consistency, on the other hand, is a minimum requirement for any estimator; with a large sample size, the calculated estimate should be close to the truth.

A definitive statement choosing between unbiasedness and consistency is not necessary, however, as the major condition for achieving both is the same. The key assumption is that the disturbances have mean zero and are uncorrelated with each regressor, or $E[\mathbf{X}'\boldsymbol{\epsilon}] = \mathbf{0}$. (This condition is weaker than assuming the independence of $\mathbf{X}$ and $\boldsymbol{\epsilon}$. Independence implies a correlation of zero, whereas a correlation of zero does not imply independence.) Given the importance of this assumption, it would seem that testing this assumption would be the first step in marshalling evidence for the statistical adequacy of a proposed model. The unknown disturbances, $\boldsymbol{\epsilon}$, could be estimated with the least squares residuals, $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, and then correlated with each regressor. Unfortunately, this test is uninformative. The least squares residuals are, *by construction*, uncorrelated with the regressors. A simple manipulation of the least squares estimator demonstrates the point. Multiply both sides of the least squares estimator by $\mathbf{X}'\mathbf{X}$ and then substitute $\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\epsilon}}$ for

$\mathbf{y}$ on the right-hand side,

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}'(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\epsilon}}) \\
\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\hat{\boldsymbol{\epsilon}} \\
\mathbf{X}'\hat{\boldsymbol{\epsilon}} &= \mathbf{0}.
\end{aligned}$$

As this critical assumption is not directly testable, we need to consider the ways in which the assumption could be violated. The three most common violations include measurement error, omitted variables, and simultaneity. We examine each in turn.

Measurement Error

Measurement error in political science arises from the difficulty of measuring the theoretical constructs to which political scientists often refer in their work. Concepts such as power, democracy, and ideology can be measured in a variety of different ways, but each measure is an approximation of an abstract idea. Good measures include significant parts of the construct of interest, but invariably, they also include parts of other constructs (systematic bias) as well as random error. Measurement error is likely the rule, and not the exception, in most political science data.

Theory

Let the data generating process include a single regressor

$$\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\epsilon}.$$

Assume that the single regressor has been measured with some error. That is, the variable that is actually available to us is not $\mathbf{x}$, but $\mathbf{w}$,

$$\mathbf{w} = \mathbf{x} + \mathbf{u},$$

where $\mathbf{u}$ is random error. $\mathbf{w}$ thus comprises the true value, $\mathbf{x}$, plus an error component. In addition, assume that $\mathbf{u}$ has mean 0, is uncorrelated with $\boldsymbol{\epsilon}$, and is uncorrelated with the single regressor, $\mathbf{x}$. If we write $\mathbf{x}$ as a function of $\mathbf{w}$ and $\mathbf{u}$, the model we actually estimate is

$$\begin{aligned} \mathbf{y} &= (\mathbf{w} - \mathbf{u})\beta + \boldsymbol{\epsilon} \\ &= \mathbf{w}\beta + \boldsymbol{\nu}, \end{aligned}$$

where $\boldsymbol{\nu} = \boldsymbol{\epsilon} - \mathbf{u}\beta$. The least squares estimator for this model is

$$\begin{aligned}\hat{\beta} &= (\mathbf{w}'\mathbf{w})^{-1}\mathbf{w}'\mathbf{y} \\ &= \beta + (\mathbf{w}'\mathbf{w})^{-1}\mathbf{w}'\boldsymbol{\nu}.\end{aligned}$$

The problem here is that $\mathbf{w}$ and $\boldsymbol{\nu}$ are both functions of the random error, $\mathbf{u}$. Thus, $\mathbf{w}$ and $\boldsymbol{\nu}$ are correlated, which violates the assumption that guarantees consistency and unbiasedness.

In the special case just discussed (a regression with a single mismeasured regressor), the effect of measurement error is straightforward. $\mathbf{w}$ consists of a systematic part, $\mathbf{x}$, and a random part, $\mathbf{u}$. The random part of $\mathbf{w}$ is uncorrelated with the dependent variable by construction. The larger the variance of the random part grows relative to the systematic part, the closer the estimated coefficient is to zero. This effect is seen in the expression for $\hat{\beta}$'s inconsistency,

$$\text{plim}(\hat{\beta} - \beta) = -\frac{\beta\sigma_u^2}{\sigma_x^2 + \sigma_u^2}.$$

The inconsistency is negative, which accounts for the bias toward zero. The inconsistency is small only if σ_x^2 is large relative to σ_u^2. As the latter term increases in size, the greater the attenuation.

Practice

The case just discussed is highly stylized and is unlikely to be encountered in practice. The comforting conclusions of these kinds of theoretical discussions, such as effects are attenuated, generally do not hold in more realistic situations. First, most regressions are likely to include multiple regressors with some subset being measured with significant error. In this situation, it is difficult to know the effects of measurement error. Even if the measurement error is confined to one variable, it affects the estimated coefficients of the variables measured without error. Second, the error in the above discussion is random error; it is uncorrelated with the disturbance, $\boldsymbol{\epsilon}$, and uncorrelated with the lone regressor $\mathbf{x}$. Actual measurement error is far more likely to be systematic. Any operational definition of a difficult-to-measure theoretical construct such as power or democracy is likely to pick up elements of other, unwanted constructs. The problem can be severe. It has been demonstrated, for example, that correlated measurement error between two variables can lead to an incorrect sign on an estimated coefficient. Including long lists of unevenly measured control variables in a regression therefore makes little sense. Third, measurement error can also appear in the dependent variable. This case is often ignored, as random error in the dependent variable is simply added to the error component of the model, and consistency is unaffected. Again, systematic error that is correlated with included regressors is more likely, and the result is inconsistency and bias.

Omitted Variables

Hidden within the assumption that the regressors must be uncorrelated with the error term is the claim that we have all the important variables that affect the dependent variable accounted for in the model. More specifically, the claim is that the model includes all of the important regressors that are correlated with the other included regressors. Omitted variables, like measurement error, are probably the rule rather than the exception in political science research.

Theory

Let the true data generating process be

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{x}_2\beta_2 + \boldsymbol{\epsilon},$$

where $\mathbf{X}_1$ is an $n \times k - 1$ matrix of regressors and $\mathbf{x}_2$ is a single regressor. Assume that $\mathbf{x}_2$ is omitted from the model either because it cannot be measured or because it is unknown. The model actually estimated, then, is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}^*,$$

where $\boldsymbol{\epsilon}^* = \mathbf{x}_2\beta_2 + \boldsymbol{\epsilon}$. This expression makes the problem clear. The new error term, $\boldsymbol{\epsilon}^*$, includes $\mathbf{x}_2$. If that variable is correlated with the included variables, $\mathbf{X}_1$, the assumption that the regressors are uncorrelated with the disturbances no longer holds.

To see this, begin with the least squares estimator of the misspecified model, and substitute the right-hand side of the true model for $\mathbf{y}$,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_1 &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} \\
&= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{x}_2\beta_2 + \boldsymbol{\epsilon}) \\
&= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{x}_2\beta_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{\epsilon}.
\end{aligned}$$

Taking the expectation under the assumption that the included variables are uncorrelated with the disturbance term, the result is

$$E[\hat{\boldsymbol{\beta}}_1] = \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{x}_2\beta_2.$$

Thus, the difference between the expectation of $\hat{\boldsymbol{\beta}}_1$ and the truth depends on two kinds of values. The first are the coefficients from the regression of the excluded variable on the included variables, $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{x}_2$. The second is the true effect of $\mathbf{x}_2$, β_2. Using this information, it is possible to identify the direction of the inconsistency and bias.

Including irrelevant variables in a regression is considered a lesser problem. Let the true data generating process be

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon},$$

and the estimated equation be

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}.$$

The variables in $\mathbf{X}_2$ are irrelevant and thus have no effect on the dependent variable. When this equation is estimated, least squares correctly estimates $\hat{\boldsymbol{\beta}}_2$ as $\mathbf{0}$, and the estimator of $\boldsymbol{\beta}_1$ is consistent. The only downside usually noted is that the variance of $\boldsymbol{\beta}_1$ in the estimated equation,

$$\sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1},$$

where $\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$, is always as large as, or larger than, the true variance,

$$\sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}.$$

This increase in the variance is considered to be a fair price to pay for the promise of avoiding inconsistency and bias caused by omitted variables.

Practice

As is true of measurement error, the theory and practice of omitted variables diverge, and comforting conclusions regarding the direction of the bias or inconsistency apply only in very special cases. For example, it is unlikely that a single variable is omitted from a regression. A set of omitted variables is more likely making the expectation above,

$$E[\hat{\boldsymbol{\beta}}_1] = \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_{\mathbf{2}}.$$

Even more likely than a set of omitted variables is the situation where the researcher is deciding whether to include a known group of previously omitted variables while still omitting a group of unknown or unmeasured variables. This situation can be expressed using the following data generating process and two misspecified models. Let the data generating process be in scalar notation,

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 X_{t3} + \epsilon_t,$$

and the two misspecified models be,

$$\begin{aligned} &\text{Model 1:} \qquad Y_t = \beta_{01} + \beta_{11} X_{t1} + \epsilon_{t1}, \\ &\text{Model 2:} \qquad Y_t = \beta_{02} + \beta_{12} X_{t1} + \beta_{22} X_{t2} + \epsilon_{t2}. \end{aligned}$$

Model 1 omits both X_2 and X_3, and model 2 omits just X_3. The prevalent view in political science is that the bias on $\hat{\beta}_{11}$, the estimated coefficient on X_1 in model 1, is always greater than the bias on $\hat{\beta}_{12}$, the estimated coefficient on X_1 in model 2. Letting the bias on $\hat{\beta}_{11}$, $E[\hat{\beta}_{11}] - \beta_1$, be denoted as $b(\hat{\beta}_{11}, \beta_1)$, and the bias on $\hat{\beta}_{12}$, $E[\hat{\beta}_{12}] - \beta_1$, be denoted as $b(\hat{\beta}_{12}, \beta_1)$, the mathematical argument is that

$$|b(\hat{\beta}_{11}, \beta_1)| \geq |b(\hat{\beta}_{12}, \beta_1)|.$$

This conclusion, however, does not hold in general. It can be demonstrated that the inclusion of additional relevant variables can increase or decrease the bias on the X_1 coefficient. Short of knowing the effects of the still omitted variables on the newly included variables, it is impossible to know whether the newly included variables increase or decrease the bias and inconsistency.

Simultaneity

Simultaneity is the third major way in which the key assumption necessary for unbiasedness and consistency—the regressors and the disturbances are uncorrelated—can be violated. The problem of simultaneity arises when one of the right-hand side regressors is determined simultaneously with the dependent variable.

Theory

Let the true data generating process consist of two equations with two dependent variables, $\mathbf{y}_1$ and $\mathbf{y}_2$, each of which is a function of the other. The equations also include two exogenous regressors, $\mathbf{x}_1$ and $\mathbf{x}_2$, and two sets of disturbances, $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$,

$$\begin{aligned}\mathbf{y}_1 + \beta_{12}\mathbf{y}_2 + \gamma_{11}\mathbf{x}_1 &= \boldsymbol{\epsilon}_1, \\ \mathbf{y}_2 + \beta_{21}\mathbf{y}_1 + \gamma_{12}\mathbf{x}_2 &= \boldsymbol{\epsilon}_2.\end{aligned}$$

To see the problem in estimating these equations consistently, rewrite the first equation as

$$\mathbf{y}_1 = \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon}_1,$$

where $\mathbf{Z} = [\mathbf{y}_2\ \mathbf{x}_1]$ and $\boldsymbol{\delta}' = [-\beta_{12}\ \ -\gamma_{11}]$. The usual OLS estimator is

$$\hat{\boldsymbol{\delta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}_1.$$

Substituting $\mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon}_1$ for $\mathbf{y}_1$ into the above equation and taking expectations, it is easy to see that the expected value of $\hat{\boldsymbol{\delta}}$ does not equal $\boldsymbol{\delta}$,

$$E[\hat{\boldsymbol{\delta}}] = \boldsymbol{\delta} + E[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\epsilon}_1].$$

The expected value on the right-hand side of the above equation does not go to zero because $\mathbf{Z}$ contains an endogenous variable, $\mathbf{y}_2$, that is jointly determined with $\mathbf{y}_1$ and thus is correlated of $\boldsymbol{\epsilon}_1$. The result is bias and inconsistency because the main condition for the achieving unbiased and consistent estimates, that the columns of $\mathbf{Z}$ be uncorrelated with the disturbance, $\boldsymbol{\epsilon}_1$, is violated.

Faced with the above situation, it is possible to estimate an equation known as the reduced form by rewriting the original two equations as

$$\mathbf{B}\mathbf{y} + \boldsymbol{\Gamma}\mathbf{x} = \boldsymbol{\epsilon},$$

where $\mathbf{B} = \begin{bmatrix} 1 & \beta_{12} \\ \beta_{21} & 1 \end{bmatrix}$, $\mathbf{y}' = \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 \end{bmatrix}$, $\boldsymbol{\Gamma}' = \begin{bmatrix} \gamma_{11} & \gamma_{21} \end{bmatrix}$, $\mathbf{x}' = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix}$, and $\boldsymbol{\epsilon}' = \begin{bmatrix} \boldsymbol{\epsilon}_1 & \boldsymbol{\epsilon}_2 \end{bmatrix}$. Provided $\mathbf{B}$ is nonsingular, the equation can be solved in the following way,

$$\begin{aligned} \mathbf{y}_t &= -\mathbf{B}^{-1}\boldsymbol{\Gamma}\mathbf{x}_t + \mathbf{B}^{-1}\boldsymbol{\epsilon} \\ &= \boldsymbol{\Pi}\mathbf{x} + \boldsymbol{\nu}, \end{aligned}$$

where $\boldsymbol{\Pi} = -\mathbf{B}^{-1}\boldsymbol{\Gamma}$ and $\boldsymbol{\nu} = \mathbf{B}^{-1}\boldsymbol{\epsilon}$.

Most of the time, however, political scientists are interested in obtaining estimates of the original coefficients, not of the reduced form coefficients. The big question is whether estimates of the original coefficients can be recovered from the reduced form estimates. Consider the model given above, but this time premultiplied by a nonsingular matrix $\mathbf{G}$,

$$\mathbf{GBy} + \mathbf{G\Gamma x} = \mathbf{G}\boldsymbol{\epsilon}.$$

Solving this equation for the reduced form coefficients produces exactly the same coefficient estimates as the model that is not premultiplied by the $\mathbf{G}$ matrix,

$$\begin{aligned} \mathbf{y}_t &= -\mathbf{B}^{-1}\mathbf{G}^{-1}\mathbf{G\Gamma x}_t + \mathbf{B}^{-1}\mathbf{G}^{-1}\mathbf{G}\boldsymbol{\epsilon} \\ &= \mathbf{\Pi x} + \boldsymbol{\nu}, \end{aligned}$$

where $\mathbf{\Pi} = -\mathbf{B}^{-1}\mathbf{\Gamma}$ and $\boldsymbol{\nu} = \mathbf{B}^{-1}\boldsymbol{\epsilon}$.

It is impossible, then, to recover estimates of the original coefficients because both models produce exactly the same reduced form coefficients. The equivalence of these two models is known as the identification problem, and much of the practice of simultaneous equations is devoted to solving it. Systems of equations where estimates of the original coefficients can be recovered from the reduced form are identified. Systems where recovery is not possible are unidentified.

Practice

In the context of simultaneous equations, the identification problem concerns the ability to recover estimates of the coefficients of interest from the reduced form coefficients. Ensuring that a system of equations is identified generally requires the use of *a priori* nonsample information. Such information most frequently comes in the form of exclusion restrictions—a specification that certain endogenous variables and certain exogenous variables do not appear in certain equations. The point is to make it harder, and eventually impossible, to find a **G** matrix that produces the same set of reduced form coefficients as the original model. If no such matrix is found, then the system of equations is identified.

Where does the nonsample information come from? As the necessary information cannot be deduced from the data, it has to come from theory. Unfortunately, little theory in political science is detailed enough to provide justifiable exclusion restrictions. Decisions about restrictions, then, are often made on *ad hoc* grounds. The consequences of making false exclusion are the familiar ones of bias and inconsistency.

Finally, it should be noted that the three major sources of endogeneity—measurement error, omitted variables, and simultaneity—do not occur in isolation from one another. Systems of equations use more variables than single equation models and are therefore more likely, *ceteris paribus*, to suffer from measurement error. The use of false exclusion restrictions leads to omitted variable bias. Any real-life data analysis situation is likely to suffer from all three problems, making any attempts to determine the direction of bias futile.

What Does Not Need to be Assumed

Getting the coefficients "right" requires one major assumption, the disturbance has mean zero and is uncorrelated with each regressor, $E[\mathbf{X}'\boldsymbol{\epsilon}] = 0$, and one minor assumption, no regressor or set of regressors is a linear combination of any other regressor or set of regressors or $\mathbf{X}$ has full column rank. The list of assumptions that do not need to be made is much longer. For example, it is not necessary to assume normality of any variables, dependent or independent. It is not necessary to assume that the independent variables are uncorrelated with one another or correlated at low levels. It is also unnecessary to assume anything about the variance/covariance matrix of the disturbances. Neither heteroscedasticity nor autocorrelation affect the bias or consistency of the estimator.

Getting the Standard Errors Right

Getting the standard errors right is not necessary for getting the coefficients right. The reverse, however, is not true. The problems that can plague estimates of the coefficients—measurement error, omitted variables, and simultaneity—can affect the standard errors. When trying to get the coefficients right, it is often difficult in practice to know the direction of the bias or inconsistency. The same is not true for the standard errors, however. In most cases involving endogeneity, the estimated standard errors are too narrow, leading to over confidence in the results of the analysis.

Consider the omitted variable case. When the true data generating process is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon},$$

the true variance of $\boldsymbol{\beta}_1$ is $\sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}$. If the variables in $\mathbf{X}_2$ are omitted from the estimated model, the variance of $\boldsymbol{\beta}_1$ is $\sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$, which is always as small as, or smaller than, the true variance. A similar demonstration can be made in the case of measurement error. As the amount of random error in $\mathbf{x}$ increases, the estimated standard error on the coefficient of $\mathbf{x}$ decreases. The problem is that the analyst thinks that she has more information than she actually does.

The discussion to follow assumes that endogeneity problems do not exist, in order to examine the other issues that lead to incorrect standard errors. It is important to remember, however, that estimated standard errors are likely to be too small even before these additional problems are discovered.

Getting estimates of the regression coefficients right is a matter of making a single, albeit very important, assumption, $E[\mathbf{X}'\boldsymbol{\epsilon}] = \mathbf{0}$. Getting the standard errors right requires making another very important assumption involving the disturbances. This time the assumption is that the squared disturbances are uncorrelated with each regressor, its square, and all cross-products. Formally, the assumption is $E[\epsilon^2 \mathbf{X}'\mathbf{X}] = \sigma^2 E[\mathbf{X}'\mathbf{X}]$, where σ^2 is the expectation of ϵ. This assumption is implied by the more common assumption that the expectation of the squared disturbances is constant, $E[\epsilon^2] = \sigma^2$. Under this assumption, as well as the two previous assumptions—$\mathbf{X}$ has full column rank and $E[\mathbf{X}'\boldsymbol{\epsilon}] = \mathbf{0}$—the OLS estimator of $\hat{\boldsymbol{\beta}}$ is normally distributed, and the standard errors, t-statistics, and F-statistics are asymptotically valid.

As with the previous big assumption, however, violations of the new assumption are both common and likely.

Heteroscedasticity

The assumption of homoscedasticity means that each of the disturbances is drawn from a distribution with the same variance as the other disturbances. In conjunction with previous assumptions, the homoscedasticity assumption is $E[\epsilon_i^2] = \sigma^2$. Violations of this assumption can occur in one of two ways: as a result of misspecification or as a result of the data themselves.

Misspecification

Heteroscedasticity due to misspecification can arise in a number of different ways. If the true data generating process is

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon},$$

and the analyst omits the variables in $\mathbf{X}_2$ from her model,

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}^*,$$

then the error term is a function of the omitted variables, $\boldsymbol{\epsilon}^* = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$. As the omitted variables vary, so does the error variance.

An incorrect functional form can also cause heteroscedasticity. If the true data generating process is nonlinear,

$$y_i = \beta_0 + \beta_1 x_i^2 + \epsilon_i,$$

and the analyst estimates a linear model,

$$y_i = \beta_0 + \beta_1 x_i + \nu_i,$$

the disturbance is again a function of x, $\nu_i = f(x_i^2, \epsilon_i)$. As x varies, so does the error variance.

A coefficient that varies across observations can also be a source of heteroscedasticity. Consider a DGP where

$$y_i = \beta_0 + \beta_i x_i + \epsilon_i,$$

and β_i varies randomly around some fixed β, $\beta_i = \beta + \nu_i$. Application of

OLS estimates the model

$$\begin{aligned} y_i &= \beta_0 + (\beta + \nu_i)x_i + \epsilon_i \\ &= \beta_0 + \beta x_i + (\nu_i x_i + \epsilon_i), \end{aligned}$$

and once again, the error term is a function of x.

A multiplicative error term or an incorrect data transformation are other specification issues that could cause heteroscedasticity.

Data

The data can also be a source of heteroscedasticity. In the simplest case, an influential outlier can cause heteroscedasticity. By the same token, a heavily skewed regressor can also cause a nonconstant error variance.

Specific kinds of data can be a source of the problem. Aggregate data is a common source of heteroscedasticity. If the dependent variable is an average or a proportion, the variance is a function of the number of observations being aggregated in the different units. If the estimated model is based on an average,

$$\bar{y}_j = \beta_0 + \beta \bar{x}_j + \bar{\epsilon}_j,$$

where $\bar{\epsilon}_j$ is an average for the jth unit over i individuals, the variance of the error term is $\mathrm{Var}(\bar{\epsilon}_j) = \sigma^2/N_j$. The error variance, then, clearly varies with the size of the group. A similar demonstration can be performed for a dependent variable based on a proportion.

Finally, heteroscedasticity most commonly occurs with cross-sectional data where units of different sizes, whether they be individuals, firms, industries, or states, are taken together. It is easy to imagine larger units having larger absolute error terms. A $900 billion economy is more likely to see an absolute error of $10 billion than a $100 billion economy.

Autocorrelation

The assumption of no autocorrelation means that the disturbance associated with one observation is unrelated to the disturbance associated with any other observation. Formally, the assumption is $E[\epsilon_i \epsilon_j] = 0$ for $i \neq j$. As is the case with heteroscedasticity, violations are the result of misspecification or the data themselves.

Misspecification

Like heteroscedasticity, autocorrelation can be a result of omitted variables, incorrect functional form, or data manipulation. An omitted variable that happens to be autocorrelated causes the disturbances to be autocorrelated. A special case of this result occurs when the omitted variable is the lagged value of the dependent variable. Autocorrelation arises in this case due to the influence of the lagged value on the current value. If a linear functional form is used to model a nonlinear relationship, there will be sections of the relationship where the estimated regression line consistently underestimates or overestimates the relationship. Finally, smoothing techniques such as moving averages can induce a periodicity in the disturbances that did not exist prior to the smoothing.

Data

Autocorrelation is, somewhat obviously, most likely to occur in data that are observed over time. Such variables often display a "stickiness" and only change in small increments over time. State budgets, for example, only change marginally from year to year. Successive observations, then, are likely to be correlated. Similarly, random shocks to a system, such as war or a market crash, have prolonged effects across time periods leading to correlated disturbances.

A more subtle form of autocorrelation arises from cross-sectional data. Spatial autocorrelation can occur when observations are taken from units that are physically adjacent to one another, such as states or countries. The behavior of individuals from adjacent West European countries is likely to be affected by similar unmeasured factors, which leads to correlated disturbances.

What To Do

The conclusion to draw from the discussion above is not that regression can never be trusted. All of the problems that have been detailed have solutions that are examined elsewhere in this encyclopedia. The discussion to follow outlines a few broad approaches to addressing the common assumption violations encountered by political scientists using the linear regression model.

Endogeneity problems can often be dealt with through the use of instrumental variables regression. Mismeasured variables, for instance, can be replaced with "instruments," which are alternative variables that are uncorrelated with the disturbance term, and yet are still correlated with the mismeasured regressors. Any regressor that is uncorrelated with the disturbance term can serve as its own instrument.

Let the matrix of instruments be **Z**, which has at least as many columns (regressors) as the original matrix of regressors, **X**. Next, premultiply the usual linear regression equation by $\mathbf{Z}'$,

$$\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}'\boldsymbol{\epsilon},$$

where the variance of $\mathbf{Z}'\boldsymbol{\epsilon}$ is $\sigma^2(\mathbf{Z}'\mathbf{Z})$. The instrumental variables estimator is then,

$$\hat{\boldsymbol{\beta}}_{IV} = (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

with variance, $\text{Var}(\hat{\boldsymbol{\beta}}_{IV}) = \sigma^2(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')\mathbf{X})^{-1}$. Estimates are easily calculated by noticing that $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ are the fitted values from the regression of $\mathbf{X}$ on $\mathbf{Z}$. The instrumental variables estimator, then, is just

$$\hat{\boldsymbol{\beta}}_{IV} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}.$$

It can be easily shown that the IV estimator is consistent and therefore, it is possible to get correct coefficients. The tradeoff for this gain in consistency is a loss of precision.

Problems relating to the error term such as heteroscedasticity and autocorrelation can be dealt with either through feasible generalized least squares or through calculation of robust standard errors. In the case of heteroscedasticity, the challenge is to estimate the correct variance/covariance matrix, which is

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \frac{1}{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'[\sigma^2\boldsymbol{\Omega}]\mathbf{X}\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}. \end{aligned}$$

Estimating $\sigma^2\boldsymbol{\Omega}$, the dimensions of which are $n \times n$, is impossible with n observations. Estimating $\left(\frac{1}{n}\mathbf{X}'[\sigma^2\boldsymbol{\Omega}]\mathbf{X}\right)$, however, is not impossible as its dimensions are only $K \times K$, where K is the number of regressors in $\mathbf{X}$. It can be shown that this expression is equal to

$$\frac{1}{n}\Sigma\sigma_i^2\mathbf{x}_i\mathbf{x}_i',$$

and the above can be consistently estimated by

$$\mathbf{S} = \frac{1}{n}\Sigma\hat{\epsilon}_i^2\mathbf{x}_i\mathbf{x}_i',$$

where $\hat{\epsilon}_i$ is the residual for the ith observation. The correct variance/covariance is consistently estimated by

$$\text{Var}(\hat{\boldsymbol{\beta}}) = n(\mathbf{X}'\mathbf{X})^{-1}\mathbf{S}(\mathbf{X}'\mathbf{X})^{-1}.$$

It is important to remember that these techniques bring with them their own sets of assumptions, which may or may not be more plausible than the assumptions of ordinary least squares. The key assumption of the instrumental variables procedure is that the instruments are uncorrelated with the disturbance. This assumption cannot be tested, and the use of "quasi-instruments," instruments that are only approximately uncorrelated with the disturbance, can produce incorrect inferences. Strong theory is required to make such assumptions plausible.

Assumptions and the Generalized Linear Model

As noted in the introduction to the linear regression model discussion, the generalized linear model (GLM) has replaced the linear regression model as the workhorse of political science. A generalized linear model is one that assumes a distribution included in the linear exponential family. These distributions include, among others, the Normal, the Bernoulli, the Exponential, and the Poisson. These models are estimated by maximum likelihood, and their good properties only hold asymptotically.

GLMs require very strong distributional assumptions. In addition, they require many of the same assumptions as the linear regression model. Some of these assumptions, however, work very differently in this setting. Consider the probit model. The model is derived by assuming that a latent or unobserved variable, $\mathbf{y}^*$, is a function of some regressors and an error term,

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

All that is observed of the latent variable is whether it is greater than zero,

$$y_i = \begin{cases} 1 \text{ if } y_i^* > 0 \\ 0 \text{ if } y_i^* \leq 0. \end{cases}$$

Given the above, the probability of observing a one is

$$\begin{aligned}
\Pr(y_i = 1) &= \Pr(y_i^* > 0) \\
&= \Pr(\mathbf{x}_i\boldsymbol{\beta} + \epsilon_i > 0) \\
&= \Pr(\epsilon_i > -\mathbf{x}_i\boldsymbol{\beta}) \\
&= \Pr(\epsilon_i < \mathbf{x}_i\boldsymbol{\beta}) \\
&= F(\mathbf{x}_i\boldsymbol{\beta}),
\end{aligned}$$

where $F()$ is the cumulative distribution function of ϵ. The probit model comes from assuming that ϵ is distributed normally.

The probit model makes the same key assumption that the linear regression model does: $\boldsymbol{\epsilon}$ and the regressors must be uncorrelated with one another. Unfortunately, not all of the lessons learned in the linear regression case are applicable to probit models. For example, the linear regression model does not require normality for consistency. The probit model, however, does. The ϵ_i must be drawn from independent and identically distributed Normal distributions. For the linear regression model, the OLS estimator is unbiased and consistent when a relevant variable is omitted from the regression, provided that the omitted variable is uncorrelated with the included variables. This result does not hold for probit models. When a relevant variable is omitted from a probit specification, the estimated coefficients on the included variables are biased and inconsistent whether or not the omitted variable is correlated with the included variables. By the same token, robust standard errors are often a useful way of dealing with heteroscedasticity in the linear regression model. Robust standard errors, however, make little sense for the probit model. If the model is correct, so are the standard errors. If the model is seriously misspecified, use of robust standard errors provides asymptotically correct variance estimates on wrong coefficient estimates. Robust standard errors, then, are of little help in this context.

The lessons learned from the linear regression model have, at best, heuristic value when it comes to more complex models such as GLMs. New assumptions have to be made, and old assumptions often have to be revisited and reevaluated. Solutions to problems that work for the linear regression case may not work for GLMs. It is important to remember that complex models are more than just fancy linear regressions.

Assumptions and Inference

Of the two goals of quantitative modelling in political science mentioned in the Introduction, description and inference, the latter is the more difficult one to achieve. The reason for this inequity is that inference depends on assumptions that are far stronger than the assumptions upon which consistency and unbiasedness rest, and, as argued above, these assumptions rarely hold in political science data sets. Although these assumptions can be relaxed with the use of more complicated statistical models, the fact remains that the model has to be nearly right for correct inference, and that is a tall order in political science.

Of all the assumptions needed for correct inferences, the most fundamental has yet to be discussed. When introducing the linear regression model, a random sample of size n from a population of interest was assumed. This assumption is standard in statistical analysis and allows traditional statistical inference to work. In very few situations, however, do political science data sets resemble anything even close to a random sample drawn from a population. The question then becomes, "to what are political scientists making inferences?" Whether the data comprise a set of international wars, American states, or Western European democracies, there is no actual population to make inferences to, and these samples are rarely treated as populations in and of themselves. Political scientists pay little attention to this issue, and most articles that make use of quantitative modelling in the discipline make no mention of the population of interest, how the sample was generated, or to what the statistical inferences refer.

The real question that needs to be addressed concerns the source of the randomness in the data. In what is known as design-based inference, the population is seen as fixed, and the sample is the result of a stochastic process such as simple random sampling or stratified random sampling. Inferences are made in traditional fashion from the sample to the population from which it was drawn. In what is known as model-based inference, the observed values are seen as realizations of random variables and therefore constitute realizations of a random process. Most political scientists practice model-based inference as true random samples from known populations are relatively rare in the discipline.

Model-based inference comes in two versions. Political scientists, when asked what population they are making inferences to, often respond by talking about a superpopulation. A superpopulation is an imaginary population from which the data could have been randomly drawn had the imaginary population existed. A researcher first assumes the existence of an imaginary population (the superpopulation) and then assumes that the sample at hand is drawn randomly from the imaginary population. There are situations in the sciences where assuming a superpopulation makes sense. To borrow an example, it might be sensible to consider the hurricanes generated in the Atlantic Ocean in a given year under certain meteorological conditions as a random draw from the population of hurricanes that could have been produced in the Atlantic Ocean in that year. There is, in such a circumstance, a real stochastic process that generates the observations. The same cannot be said of, for instance, election data. Political scientists are fond of saying that the world could have turned out differently, and thus the data are like a random draw from the superpopulation of elections across different worlds. Unlike the process that generates Atlantic hurricanes, there is no actual stochastic process that could generate these imaginary elections.

The second version of model-based inference is closer to what political scientists actually do. In this version, there is no population, and no pretense is made about making inferences to a population. Instead, a model is proposed that accounts for the way that nature produces the data. Inferences are then made back to features of the model. The sample design is irrelevant under model-based inference, which is why it fits well with political science. The downside of model-based inference is its dependence on a model, which may be misspecified in any of the ways discussed above. If the model is bad, so are the inferences based on it. An awareness on the part of political scientists that they are engaged in model-based inference is important because it highlights just how fundamental models and their constituent assumptions are to the practice of data analysis.

See also Categorical Response Data; Data Analysis, Aggregate; Data Analysis, Exploratory; Endogeneity; Epistemological Foundations; Garbage Can Model; Generalized Least Squares; Inference, Classical and Bayesian; Logit/Probit Analysis; Maximum Likelihood; Measurement; Misspecification; Models, Statistical; Sampling ,Random and Nonrandom; Simultaneous Equation Models; Statistics; Time Series Analysis; Variables, Instrumental; Variables, Latent;

Further Readings

Achen, C. H. (1982). *Interpreting and using regression.* Thousand Oaks, CA: Sage.

Berk, R. A. (2004). *Regression analysis: a constructive critique.* Thousand Oaks, CA: Sage.

Greene, W. H. (2003). *Econometric analysis.* 5th ed. Upper Saddle River, NJ: Prentice Hall.

Johnston, J. & DiNardo J. (1997). *Econometric methods.* 4th ed. New York: McGraw-Hill.

Judge, G. G., Griffith W. E., Hill R. C., Lutkepohl H., & Lee, T. C. (1985). *The theory and practice of econometrics.* 2nd ed. New York: Wiley.

Kennedy, P. (2008). *A guide to econometrics*, 6th ed. Malden, MA: Blackwell Publishing.

Spanos, A. (1999). *Probability theory and statistical inference: econometric modeling with observational data.* Cambridge: Cambridge University Press.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data.* Cambridge, MA: The MIT Press.