17.874 Lecture Notes
Part 3: Regression Model

3. Regression Model

Regression is central to the social sciences because social scientists usually want to know the effect of one variable or set of variables of interest on an outcome variable, holding all else equal. Here are some classic, important problems:

What is the effect of knowledge of a subject on preferences about that subject? Do uninformed people have any preferences? Are those preferences short-sighted? Do people develop "enlightened" interests, seeing how they benefit from collective benefit?

What is the effect of increasing enforcement and punishment on deviant behavior? For example, how much does an increase in the number of police on the streets on the crime rate? Does the death penalty deter crime?

How are markets structure? How does the supply of a good change with the price people are willing to pay? How does people's willingness to pay for a good change as a good becomes scarce?

To whom are elected representatives responsible? Do parties and candidates converge to the preferred policy of the median voter, as predicted by many analytical models of electoral competition? Or do parties and interest groups strongly influence legislators' behavior?

In each of these problems we are interested in measuring how changes in an input or independent variable produce changes in an output or dependent variable.

For example, my colleagues interested in the study of energy wish to know if increasing public understanding of the issue would change attitudes about global warming. In particular, would people become more likely to support remediation. I have done a couple of surveys for this initiative. Part of the design of these surveys involves capturing the attitude toward global warming, the dependent variable. One simple variable is the willingness to pay.

> If it solved global warming, would you be willing to pay \$5 more a month on your electricity bill? [of those who answered yes] \$10 more a month? \$25 more a month? \$50 more a month? \$100 more a month?

We also wanted to distinguish the attitudes of those who know a lot about global warming and carbon emissions and those who do not. We asked a battery of questions. People could have gotten up to 10 factual questions right. The average was 5. Finally, we controlled many other factors, such as income, religiosity, attitude toward business regulation, size of electricity bill and so forth.

The results of the analysis are provided on the handout. The dependent variable is coded 0 for those unwilling to pay \$5, 1 for those willing to pay no more than \$5, 2 for those willing to pay no more than \$10, 3 for those willing to pay more than \$25, 4 for those willing to pay no more than \$50, and 5 for those who said they would be willing to pay \$100 (roughly double the typical electricity bill).

What is estimated by the coefficient on information (**gotit**) is the effect of knowing more about carbon dioxide on willingness to pay for a solution to global warming. Moving from no information to complete information (from 0 to 10) increases willingness to pay by .65 units along the willingness to pay scale, which ranges from 0 to 5 and has a mean of 1.7 and standard deviation of 1.3

More generally, we seek to measure the effect on Y of increasing X_1 from x_1^0 to x_1^1. The expected change in the random variable Y is:

$$E[y|x_1^1, x_2, ...x_k] - E[y|x_1^0, x_2, ...x_k],$$

where the values of all independent variables other than x_1 are held constant.

Unfortunately, the world is a mess. We usually deal with observational data rather than carefully constructed experiments. It is exceedingly difficult to hold all else constant in order to isolate the partial effect of a variable of interest on the outcome variable. In stead of trying to isolate and control the independent variable of interest in the design of a study; we attempt to measure its independent effect in the analysis of the data collected.

Multivariate regression is the primary data analytic tool with which we attempt to hold other things constant.

3.1. Basic Model

In general, we can treat y and X as *random* variables. At times we will switch into a simpler problem – X fixed. What is meant by this distinction? Fixed values of X would arise if the researcher chose the specific values of X and *assigned* them to the units. Random or stochastic X arises when the researcher does not control the assignment – this is true in the lab as well as the field.

Fixed X will be easier to deal with analytically. If X is fixed then it is guaranteed to be have no correlation with ϵ because the vectors of the matrix of X consist of fixed values of the variables, not random draws from the set of possible values of the variables. More generally, X is random, and we must analyze the conditional distributions of $\epsilon|X$ carefully.

The basic regression model consists of several sets of assumptions.

1. Linearity

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

2. 0 mean of the errors.

$$E[\epsilon|\mathbf{X}] = \mathbf{0}$$

3. No correlation between X and ϵ:

$$E[\epsilon\mathbf{X}|\mathbf{X}] = \mathbf{0}$$

4. Spherical Errors (constant error variance and no autocorrelation):

$$E[\epsilon\epsilon'|\mathbf{X}] = \sigma_\epsilon^2\mathbf{I}$$

5. Normally distributed errors:

$$\epsilon|\mathbf{X} \sim \mathbf{N}(\mathbf{0}, \sigma_\epsilon^2\mathbf{I})$$

Indeed, these properties are implied by (sufficient but not necessary), the assumption that $\mathbf{y}, \mathbf{X}$ are jointly normally distributed. See the properties of the conditional distributions in section 2.6.

It is easy to see that this model implements the definition of an effect above. Indeed, if all of the assumptions hold we might even say the effect captured by the regression model is causal.

Two failures of causality might emerge.

First, there may be omitted variables. Any variable X_j that is not measured and included in a model is captured in the error term ϵ. An included variable might appear to cause (or not cause) y, but we have in fact missed the true relationship because we did not hold X_j constant. Of course, there are a very large number of potential omitted variables, and the struggle in any field of inquiry is to speculate what those might be and come up with explicit measures to capture those effects or designs to remove them.

Second, there may be simultaneous causation. X_1 might cause y, but y might also cause X_1. This is a common problem in economics where prices at which goods are bought and the quantities purchased are determined through bargaining or through markets. Prices and quantities are simultaneously determined. This shows up, in complicated ways, in the error term. Specifically, the error term becomes recursive $\epsilon|X_1$ necessarily depends on $u|y$, the error term from the regression of X on y.

Social scientists have found ways to solve these problems, involving "quasi-experiments" and sometimes real experiments. Within areas of research there is a lot of back and forth about what specific designs and tools really solve the problems of omitted variables and simultaneity. The last third of the course will be devoted to these ideas.

Now we focus on the tools for holding constant other factors directly.

3.2. Estimation

The regression model has $K + 1$ parameters – the regression coefficients and the error variance. But, there are n equations. The parameters are overdetermined, assumning $n > K + 1$. We must some how reduce the data at hand to devise estimates of the unknown parameters in terms of data we know.

Overdetermination might seem like a nuisance, but if n were smaller than $k + 1$ we could not hope to estimate the parameters. Here lies a more general lesson about social inquiry. Be wary of generalizations from a single or even a small set of observations.

3.2.1. The Usual Suspects

There are, as mentioned before, three general ideas about how use data to estimate parameters.

1. The Method of Moments. (a) Express the theoretical moments of the random variables as functions of the unknown parameters. The theoretical moments are the means, variances, covariances, and higher powers of the random variables. (b) Use the empirical moments as estimators of the theoretical moments – these will be unbiased estimates of the theoretical moments. (c) Solve the system of equations for the values of the parameters in terms of the empirical moments.

2. Minimum Variance/Minimum Mean Squared Error (Minimum χ^2). (a) Express the objective function of the sum squared errors as a function of the unknown parameters. (b) Find values of the parameters that minimize that function.

3. Maximum Likelihood. (a) Assume that the random variables follw a particular density function (usually normal). (b) Find values of the parameters that maximize the density function.

We have shown that for the regression model, all three approaches lead to the same estimate of the vector of coefficients:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

An unbiased estimate of the variance of the error term is

$$s_e^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$$

The maximum likelihood estimate is

$$\hat{\sigma}_\epsilon^2 = \frac{\mathbf{e}'\mathbf{e}}{n}$$

We may also estimate the variance decomposition. We wish to account for the total sum of squared deviations in y:

$$\mathbf{y}'\mathbf{y}$$

, which has mean square $\mathbf{y}'\mathbf{y}/(\mathbf{n-1})$ (the est. variance of y).

The residual or error sum of squares is

$$\mathbf{e}'\mathbf{e}$$

, which has mean square $s_e^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$.

The model or explained sum of squares is the difference between these:

$$\mathbf{y}'\mathbf{y} - \mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - (\mathbf{y}' - \mathbf{b}'\mathbf{X}')(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}.$$

which has mean square $\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}/(\mathbf{K-1})$. $R^2 = \frac{\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}}{\mathbf{y}'\mathbf{y}}$.

3.2.2. Interpretation of the Ordinary Least Squares Regression

An important interpretation of the regression estimates is that they estimate the partial derivative, β_j. The coefficient from a simple bivariate regression of y on X_j measures the TOTAL effect of a X_j on y. This is akin to the total derivative. Let us break this down into component parts and see where we end up.

The total derivative, you may recall, is the partial derivative of y with respect to X_j plus the sum of partial derivatives of y with respect to all other variables, X_k, times the partial derivative of X_k with respect to X_j. If y depends on two X variables, then:

$$\frac{dy}{dX_1} = \frac{\partial y}{\partial X_1} + \frac{\partial y}{\partial X_2}\frac{dX_2}{dX_1}$$

$$\frac{dy}{dX_2} = \frac{\partial y}{\partial X_2} + \frac{\partial y}{\partial X_1}\frac{dX_1}{dX_2}$$

Using the idea that the bivariate regression coefficients measure the total effect of one variable on another, we can write the bivariate regression estimates in terms of the partial regression coefficient (from a multivariate regression) plus other partial regression coefficients and bivariate regression coefficients. For simplicity consider a case with 2 X variables. Let a_1 and a_2 be the slope coefficients from the regressions of y on X_1 and y on X_2. Let b_1 and b_2 be the partial regression coefficients from the multivariate regression of y on X_1 and X_2. And let c_1 be the slope coefficient from the regression of X_1 on X_2 and c_2 be the slope from the regression of X_2 on X_1.

We can express our estimates of the regression parameters in terms of the other coefficients. Let r be the correlation between X_1 and X_2.

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} \frac{(\mathrm{x}_2'\mathrm{x}_2)(\mathrm{x}_1'\mathrm{y})-(\mathrm{x}_1'\mathrm{x}_2)(\mathrm{x}_2'\mathrm{y})}{(\mathrm{x}_1'\mathrm{x}_1)(\mathrm{x}_2'\mathrm{x}_2)-(\mathrm{x}_1'\mathrm{x}_2)^2} \\ \\ \frac{(\mathrm{x}_1'\mathrm{x}_1)(\mathrm{x}_2'\mathrm{y})-(\mathrm{x}_1'\mathrm{x}_2)(\mathrm{x}_1'\mathrm{y})}{(\mathrm{x}_1'\mathrm{x}_1)(\mathrm{x}_2'\mathrm{x}_2)-(\mathrm{x}_1'\mathrm{x}_2)^2} \end{pmatrix} = \begin{pmatrix} \frac{a_1-a_2c_1}{1-r^2} \\ \\ \frac{a_2-a_1c_2}{1-r^2} \end{pmatrix}$$

We can solve these two equations to express a_1 and a_2 in terms of b_1 and b_2.

$$a_1 = b_1 + b_2c_1$$

$$a_2 = b_2 + b_1c_2.$$

This is estimated total effect (bivariate regression) equals the estimated direct effect of X_j (partial coefficient) on y plus the indirect effect of X_j through X_k.

Another way to express these results is that multivariate regression can be simplified to bivariate regression, once we partial out the effects of the other variables on y and on X_j.

Suppose we wish to measure and display the effect of X_1 on y. Regress X_1 on $X_2, ...X_k$. Regress y on $X_2, ...X_k$. Take the residuals from each of these regressions, $\mathbf{e}(\mathbf{x_1}|\mathbf{x_2}, ...\mathbf{x_k})$ and $\mathbf{e}(\mathbf{y}|\mathbf{x_2}, ...\mathbf{x_k})$. The first vector of residuals equals the part of X_1 remaining after subtracting out the other variables. Importantly, this vector of residuals is independent of $X_2, ...X_k$. The second vector equals the part of y remaining after subtratcting the effects of $X_2, ...X_k$ – that is the direct effect of X_1 and the error term ϵ.

Regress $\mathbf{e}(\mathbf{y}|\mathbf{x_2}, ...\mathbf{x_k})$ on $\mathbf{e}(\mathbf{x_1}|\mathbf{x_2}, ...\mathbf{x_k})$. This bivariate regression estimates the partial regression coefficient from the multivariate regression, β_1.

Partial regression is very useful for displaying a particular relationship. Plotting the first vector of residuals against the second vector displays the partial effect of X_1 on y, holding all other variables constant. In STATA partial regression plots are implmented through the command **avplot**. After performing a regression type the command **avplot x1**, where **x1** is the variable of interest.

A final way to interpret regression is using prediction. The predicted values of a regression are the values of the regression plane, $\hat{y}$. These are estimates of $E[\mathbf{y}|\mathbf{X}]$. We may vary values of any X_j to generate predicted values. Typically, we consider the effect on $\hat{y}$ of a one-standard deviation change in X_j, say from 1/2 standard deviation below the mean of X to 1/2 standard deviation above the mean.

Denote $\hat{y}_1$ as the predicted value at $x_j^1 = \bar{x_j} + .5s_j$ and $\hat{y}_0$ as the predicted value at $x_j^0 = \bar{x_j} - .5s_j$. The matrix of all variables except x_j is $X_{(j)}$ and $\mathbf{b_{(j)}}$ is the vector of coefficients except b_j. Choose any vector $\mathbf{x_{(j)}}$. A standard deviation change in the predicted values is:

$$\hat{y}_1 - \hat{y}_0 = (b_0 + b_j x_j^1 + \mathbf{x}'_{(j)}\mathbf{b}_{(j)}) - (\mathbf{b_0} + \mathbf{b_j x_j^0} + \mathbf{x}'_{(j)}\mathbf{b}_{(j)}) = \mathbf{b_j}(\mathbf{x_j^1} - \mathbf{x_j^0}) = \mathbf{b_j s_j}.$$

For non-linear fucntions such variations are harder to analyze, because the effect of any one variable depends on the values of the other variables. We typically set other variables equal to their means.

3.2.3. An Alternative Estimation Concept: Instrumental Variables.

There are many other ways we may estimate the regression parameters. For example, we could minimize the Mean Absolute Deviation of the errors.

One important alternative to ordinary least squares estimation is instrumental variables estimation. Assume that there is a variable or matrix of variables, $\mathbf{Z}$, such that $\mathbf{Z}$ does not directly affect $\mathbf{y}$, i.e., it is uncorrelated with ϵ, but it does correlate strongly with $\mathbf{X}$.

The instrumental variables estimator is:

$$b_{IV} = (X'Z)^{-1}Z'y.$$

This is a linear estimator. It is a weighted average of the y's, where the weights are of the form $\frac{z_i}{(z_i-\bar{z})(x_i-\bar{x})}$, in stead of $\frac{x_i}{(x_i-\bar{x})(x_i-\bar{x})}$.

In the bivariate case, this esimator is the ratio of the slope coefficient from the regression of y on z to the slope coefficient from the regression of x on z.

This estimator is quite important in quasi-experiments, because, if we can find valid instruments, the estimator will be unbiased because it will be independent of omitted factors in the least squares regression. It will, however, come at some cost. It is a noisier estimator than least squres.

3.3. Properties of Estimates

Parameter estimates themselves are random variables. They are functions of random variables that we use to make guesses about unknown constants (parameters). Therefore, from one study to the next, parameter estimates will vary. It is our hope that a given study has no bias, so if the study were repeated under identical conditions the results would vary around the true parameter value. It is also hoped that estimation uses all available data as efficiently as possible.

We wish to know the sample properties of estimates in order to understand when we might face problems that might lead us to draw the wrong conclusions from our estimates, such as spurious correlations.

We also wish to know the sample properties of estimates in order to perform inferences. At times those inferences are about testing particular theories, but often our inferences concern whether we've built the best possible model using the data at hand. In this regard, we are concerned about potential bias, and will try to error on the side of getting unbiased estimates.

Erring on the side of safety, though, can cost us efficiency. Test for the appropriateness of one model versus another, then, depend on the tradeoff between bias and efficiency.

3.3.1. Sampling Distributions of Estimates

We will characterize the random vector **b** with its mean, the variance, and the frequency function. The mean is β, which means that it is unbiased. The variance is $\sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}$, which is the matrix describing the variances and covariances of the estimated coefficients. And the density function $f(\mathbf{b})$ is approximated by the Normal distribution as n becomes large.

The results about the mean and the variance stem from the regression assumptions.

First, consider the mean of the parameter estimate. Under the assumption that $\epsilon' X = 0$, the regression estimates are unbiased. That is, $E[\mathbf{b}] = \beta$

$$E[\mathbf{b}|\mathbf{X}] = \mathbf{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = \mathbf{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon)] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + \mathbf{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon = \beta$$

This means that before we do a study we expect the data to yield a $\mathbf{b}$ of β. Of course, the density of any one point is zero. So another way to think about this is if we do repeated sampling under identical circumstances, then the coefficients will vary from sample to sample. The mean of those coefficients, though, will be β.

Second, consider the variance.

$$\begin{aligned} V[\mathbf{b}] &= \mathbf{E}[(\mathbf{b}-\beta)(\mathbf{b}-\beta)'] = \mathbf{E}[(\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon - \beta)(\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon - \beta)'] \\ &= E[(\mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\epsilon)((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon)'] = \mathbf{E}[\mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\epsilon\epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= \sigma^2(\mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) = \sigma^2\mathbf{X}'\mathbf{X}^{-1} \end{aligned}$$

Here we have assumed the sphericality of the errors and the exogeneity of the independent variables.

The sampling distribution function can be shown to follow the joint normal. This follows from the multivariate version of the central limit theorem, which I will not present because of the complexity of the mathematics. The result emerges, however, from the fact that the regression coefficients are weighted averages of the values of the random variable $\mathbf{y}$, where the weights are of the form $\frac{x_i}{(x_i - \bar{x})^2}$. More formally, let $\mathbf{b_n}$ be the vector of regression coefficients estimated from a sample of size n.

$$\mathbf{b_n} \rightarrow^{\mathrm{d}} \mathbf{N}(\beta, \sigma^2_\epsilon(\mathbf{X}'\mathbf{X}^{-1}))$$

Of course, as $n \rightarrow \infty$, $(\mathbf{X}'\mathbf{X}^{-1}) \rightarrow \mathbf{0}$. We can see this in the following way. The elements of $(\mathbf{X}'\mathbf{X}^{-1})$ are, up to a signed term, $\frac{\mathsf{x}_j'\mathsf{x}_k}{|\mathsf{X}'\mathsf{X}|}$. The elements of the determinant are multiples of the cross products, and as observations are added the determinant grows faster than any single cross product. Hence, each element approaches 0.

Consider the 2x2 case.

$$(\mathbf{X}'\mathbf{X}^{-1}) = \frac{\mathbf{1}}{\mathsf{x}_1'\mathsf{x}_1\mathsf{x}_2'\mathsf{x}_2 - (\mathsf{x}_1'\mathsf{x}_2)^2}\begin{pmatrix} \mathsf{x}_2'\mathsf{x}_2, -\mathsf{x}_1'\mathsf{x}_2 \\ -\mathsf{x}_2'\mathsf{x}_1, \mathsf{x}_1'\mathsf{x}_1 \end{pmatrix}.$$

This may be rewritten as

$$\begin{pmatrix} \frac{1}{\mathrm{x}_1'\mathrm{x}_1-(\mathrm{x}_1'\mathrm{x}_2)^2/\mathrm{x}_2'\mathrm{x}_2}, & \frac{-1}{(\mathrm{x}_1'\mathrm{x}_1\mathrm{x}_2'\mathrm{x}_2/\mathrm{x}_1'\mathrm{x}_2)-(\mathrm{x}_1'\mathrm{x}_2)} \\ \frac{-1}{(\mathrm{x}_1'\mathrm{x}_1\mathrm{x}_2'\mathrm{x}_2/\mathrm{x}_1'\mathrm{x}_2)-(\mathrm{x}_1'\mathrm{x}_2)}, & \frac{1}{\mathrm{x}_1'\mathrm{x}_1-(\mathrm{x}_1'\mathrm{x}_2)^2/\mathrm{x}_2'\mathrm{x}_2} \end{pmatrix}$$

Taking the limit of this matrix as $n \rightarrow \infty$ amounts to taking the limit of each element of the matrix. Considering each term, we see that the sum of squares in the denominators grow with the addition of each observation, while the numerators remain constant. Hence, each element approaches 0.

Combined with unbiasedness, this last result proves *consistency*. Consistency means that the limit as n grows of the probability that an estimator deviates from the parameter of interest approaches 0. That is,

$$lim_{n\rightarrow\infty} Pr(|\hat{\theta}_n - \theta|) = 0$$

A sufficient condition for this is that the estimator be unbiased and that the variance of the estimator shrink to 0. This condition is called convergence in mean squared error, and is an immediate application of Chebychev's Inequality.

Of course, this means that the limiting distribution of $f(\mathbf{b})$ shrinks to a single point, which is not so good. So the normality of the distribution of $\mathbf{b}$ is sometimes written as follows:

$$\sqrt{n}(\mathbf{b}_\mathrm{n} - \beta) \sim \mathbf{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{Q}^{-1}),$$

where $\mathbf{Q} = \frac{1}{\mathrm{n}}(\mathbf{X'X})$, the asymptotic variance covariance matrix of X's.

We may consider in this the distribution of a single element of $\mathbf{b}$.

$$b_j \sim N(\beta_j, \sigma_\epsilon^2 a_{jj}),$$

where a_{jj} is the jth diagonal element of $(\mathbf{X'X})^{-1}$.

Hence, we can construct a 95 percent confidence interval for any parameter β_j using the normal distribution and the above formula for the variance of b_j. The standard error is $\sqrt{\sigma_\epsilon^2 a_{jj}}$. So, a 95 percent confidence interval for a single parameter is:

$$b_j \pm 1.96\sqrt{\sigma_\epsilon^2 a_{jj}}$$

.

The instrumental variables estimator provides an interesting contrast to the least squares estimator. Let us consider the sampling properties of the Instrumental Variables estimator.

$$E[\mathbf{b}_{\mathrm{IV}}] = \mathbf{E}[(\mathbf{X}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{y})] = \mathbf{E}[(\mathbf{X}'\mathbf{Z})^{-1}(\mathbf{Z}'(\mathbf{X}\beta+\epsilon)]$$

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

The instrumental variables estimator is an unbiased estimator of β.

The variance of the instrumental variables estimator is:

$$V[\mathbf{b}_{\mathrm{IV}}] = \mathbf{E}[(\mathbf{b}_{\mathrm{IV}} - \beta)(\mathbf{b}_{\mathrm{IV}} - \beta)'] = \mathbf{E}[(\mathbf{X}'\mathbf{Z})^{-1}(\mathbf{Z}'\epsilon))((\mathbf{X}'\mathbf{Z})^{-1}(\mathbf{Z}'\epsilon))']$$

$$= E[(\mathbf{X}'\mathbf{Z})^{-1}(\mathbf{Z}'\epsilon\epsilon\mathbf{Z}((\mathbf{X}'\mathbf{Z})^{-1}] = (\mathbf{X}'\mathbf{Z})^{-1}(\mathbf{Z}'\sigma_\epsilon^2\mathbf{I}\mathbf{Z}((\mathbf{X}'\mathbf{Z})^{-1}$$

$$= \sigma_\epsilon^2(\mathbf{X}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Z}((\mathbf{X}'\mathbf{Z})^{-1}.$$

As with the least squares estimator the instrumental variables estimator will follow a normal distribution because the IV estimator is a (weighted) sum of the random variables, y.

3.3.2. Bias versus Efficiency

3.3.2.1. General Efficiency of Least Squares

An important property of the Ordinary Least Squares estimates is that they have the lowest variance of all linear, unbiased estimators. That is, they are the most efficient unbiased estimators. This result is the Gauss-Markov Theorem. Another version arises as a property of the maximum likelihood estimator, where the lower bound for the variances of all possible consistent estimators is $\sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}$.

This implies that the Instrumental Variables estimator is less efficient than the Ordinary Least Squares estimator. To see this, consider the bivariate regression case.

$$V[b] = \frac{\sigma_\epsilon^2}{\sum(x_i - \bar{x})^2}$$

$$V[b_{IV}] = \frac{\sigma_\epsilon^2 \sum(z_i - \bar{z})^2}{\sum(x_i - \bar{x})(z_i - \bar{z})}$$

Comparing these two formulas: $V[b_{IV}]/V[b] = (\frac{\sum(z_i-\bar{z})^2 \sum(x_i-\bar{x})^2}{\sum(x_i-\bar{x})(z_i-\bar{z})}$. This ratio is the inverse of the square of the correlation between X and Z. Since the correlation never exceeds 1, we know the numerator must be larger than the denominator. (The square of the correlation is known to be less than 1 because of the Cauchy-Schwartz inequality.)

3.3.2.2. Sources of Bias and Inefficiency in Least Squares

There are four primary sources of bias and inconsistency in least squares estimates: measurement error in the independent variables, omitted variables, non-linearities, and simultaneity. We'll discuss two of these cases here – specification of regressors (omitted variables) and measurement error.

Measurement Error.

Assume that X^* is the true variable of interest but we can only measure $X = X^* + u$, where u is a random error term. For example, suppose we regress the actual share of the vote for the incumbent president in an election on the job approval rating of the incumbency president, measured in a 500 person preelection poll the week before the election. Gallup has measured this since 1948. Each election is an observation. The polling data will have measurement error based on random sampling inherent in surveys. Specifically, the variance of the measurement error is $\frac{p(1-p)}{n}$, where p is the percent approving of the president.

Another common source of measurement error arises from typographical errors in datasets. Keypunching errors are very common, even in data sets distributed publicly through reputable sources such as the Interuniversity Consortium for Political and Social Research. For example, Gary King, Jim Snyder, and others who have worked with election data estimate that about 10 percent of the party identification codes of candidates are incorrect in some of the older ICPSR datasets on elections.

Finally, some data sources are not very reliable, or estimates must be made. This is common in social and economic data in developing economies and in data on wars.

If we regress y on X, the coefficient is a function of both the true variable X^* and the error term. That is

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i^* - \bar{x^*} + u_i - \bar{u})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i^* - \bar{x^*} + u_i - \bar{u})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i^* - \bar{x^*})(y_i - \bar{y}) + \sum_{i=1}^{n}(u_i - \bar{u})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i^* - \bar{x^*})^2 + \sum_{i=1}^{n}(u_i - \bar{u})^2 + \sum_{i=1}^{n}(x_i^* - \bar{x}^*)(u_i - \bar{u})}$$

This looks like quite a mess.

A few assumptions are made to get some traction. First, it is usually assumed that u and ϵ and that u and X^* are uncorrelated. If they are correlated, things are even worse. Second, the u_i's is assumed to be uncorrelated with one another and to have constant variance.

Taking expected values of b will be very difficult, because b is a function of the ratio of random variables. Here is a situation where Probability Limits (plim's) make life easier. Because *plims* are limits, they obey the basic rules of limits. The limit of a sum is the sum of the limit and the limit of a ratio is the ratio of the limits. Divide the top and bottom of b by n. Now consider the probability limit of each element of b:

$$plim\frac{1}{n}\sum_{i=1}^{n}(x_i^* - \bar{x^*})^2 = \sigma_x^2$$

$$plim\frac{1}{n}\sum_{i=1}^{n}(x_i^* - \bar{x^*})(y_i - \bar{y}) = plim\frac{1}{n}\sum_{i=1}^{n}(x_i^* - \bar{x^*})(\beta x_i^* - \bar{x^*}) = \beta\sigma_x^2$$

$$plim\frac{1}{n}\sum_{i=1}^{n}(x_i^* - \bar{x^*})(u_i - \bar{u}) = 0$$

$$plim\frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})^2 = \sigma_u^2$$

We can pull these limits together as follows:

$$plimb = \frac{\beta\sigma_x^2}{\sigma_x^2 + \sigma_u^2} = \beta\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} < \beta$$

Thus, in a bivariate regression, measurement error in the X variable biases the estimate toward 0. In a multivariate setting, the bias cannot generally be signed. Attentuation is typical, but it is possible also to reverse signs or inflate the coefficients. In non-linear models, such as those using the square of X, the bias terms become quite substantial and even more troublesome.

The best approach for eliminating measurement error is cleaning the dataset. However, it is possible to fix some measurement error using instrumental variables. One approach is to use the quantiles or the ranks of the observed X to predict the observed X and then use the predicted values in the regression predicting y. The idea behind this is that the ranks of X are correlated with the underlying true values of X, i.e., X^* but not with u or ϵ.

These examples of measurement error assume that the error is *purely* random. It may not be. Sometimes measurement error is systematic. For example, people underreport socially undesirable attitudes or behavior. This is an interesting subject that is extensively studied in public opinion research, but often understudied in other fields. A good survey researcher, for example, will tap into archives of questions and even do question wording experiments to test the validity and reliability of instruments.

Choice of Regressors: omitted and Included Variables.

The struggle in most statistical modeling is to specify the most appropriate regression model using the data at hand. The difficulty is deciding which variables to include and which to exclude. In rare cases we can be guided by a specific theory. Most often, though, we have gathered data or will gather data to test specific ideas and arguments. From the data at hand what is the best model?

There are three important rules to keep in mind.

1. Omitted Variables That Directly Affect Y And Are Correlated With X Produce Bias.

The most common threat to the validity of estimates from a multivariate statistical analysis is omitted variables. omitted variables affect both the consistency and efficiency of our estimates. First and foremost, they create bias. Even if they do not bias our results, we often want to control for other factors to improve efficiency.

To see the bias due to omitted variables assume that $\mathbf{X}$ is a matrix of included variables and $\mathbf{Z}$ is a matrix of variables not included in the analysis. The full model is

$$\mathbf{y} = \mathbf{X}\beta_{\mathrm{X}} + \mathbf{Z}\beta_{\mathrm{z}} + \epsilon.$$

Suppose that $\mathbf{Z}$ is omitted. Obviously we can't estimate the coefficient β_z. Will the other coefficients be biased? Is there a loss of efficiency?

Let $\mathbf{b}_{\mathrm{x}}$ be the parameter vector estimated when only the variables in the matrix $\mathbf{X}$ are included. Let β_{X} bet the subset of coefficients from the true model on the included variables, X. The model estimated is

$$\mathbf{y} = \mathbf{X}\beta_{\mathrm{X}} + \mathbf{u},$$

where $\mathbf{u} = \mathbf{Z}\beta_{\mathrm{z}} + \epsilon$.

$$E[\mathbf{b}_{\mathrm{X}}] = \mathbf{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = \mathbf{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta_{\mathrm{X}} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}]$$
$$= \beta_{\mathrm{X}} + \mathbf{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\beta_{\mathrm{z}}] + \mathbf{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon] = \beta_{\mathrm{X}} + \mathbf{\Pi}'_{\mathrm{zx}}\beta_{\mathrm{z}},$$

where $\mathbf{\Pi}_{\mathrm{zx}}$ is a matrix of coefficients from the regression of the columns of $\mathbf{Z}$ on the variables in $\mathbf{X}$.

This is an extremely useful formula. There are two important lessons to take away.

First, omitted variables will bias regression estimates of the included variables (and lead to inconsistency) if (1) those variables directly affect Y and (2) those variables are correlated with the included variables (X). It is not enough, then, to object to an analysis that there are variables that have not been included. That is always true. Rather, science advances by conjecturing (and then gathering the data on) variables that affect y directly and are correlated with X. I think the latter is usually hard to find.

Second, we can generally sign the bias of omitted variables. When we think about the potential problem of an omitted variable we usually have in mind the direct effect that it likely has on the dependent variable and we might also know or can make a reasonable guess about the correlation with the included variables of interest. The bias in an included variable will be the direct effect of the omitted variable on y times the effect of the included variable on the excluded variable. If both of those effects are positive or both are negative then the estimated effect of X on Y will be biased up – it will be too large. If one of these effects is negative and the other positve then the estimate effect of X on Y will be biased downward.

2. Efficiency Can Be Gained By Including Variables That Predict Y But Are Uncorrelated With X.

A straightforward analysis reveals that the estimated $V[\mathbf{b}_{\mathrm{X}}] = \mathbf{s}_{\mathbf{e}}^2(\mathbf{X'X})^{-1}$. So far so good. But the variance of the error term is inflated. Specifically, $s_e^2 = \frac{1}{n-K}\mathbf{u'u}$. Because $\mathbf{u} = \mathbf{Z}\beta_{\mathrm{Z}} + \epsilon$, $E[\mathbf{u'u}/(\mathbf{n}-\mathbf{K})] = \beta'_{\mathrm{Z}}\mathbf{Z'Z}\beta_{\mathrm{Z}}/(\mathbf{n}-\mathbf{K}) + \sigma_\epsilon^2 > \sigma_\epsilon^2$. In fact, the estimated residual variance is too large by the explained or model sums of squared errors for the omitted variables.

This has an interesting implication for experiments. Randomization in experiments guarantees unbiasedness. But, we still want to control for other factors to reduce noise. In fact, combining regression models in the data analysis is a powerful way to gain efficiency (and reduce the necessary sample sizes) in randomized experiments.

Even in observational studies we may want to keep in a model a variable whose inclusion does not affect the estimated value of the parameter of a variable of interest if the included variable strongly affects the dependent variable. Keeping such a variable captures some of the otherwise unexplained error variance, thereby reducing the estimated variance of the residuals. As a result, the size of confidence intervals will narrow.

3. Including Variables That Are Unrelated To Y and X Loses Efficiency (Precision), And There Is No Bias From Their Exclusion

That there is no bias can be seen readily from the argument we have just made.

The loss of efficiency occurs because we use up degrees of freedom. Hence, all variance estimates will be too large.

Simply put, parsimonious models are better.

COMMENTS:

a. Thinking through the potential consequences of omitted variables in this manner is very useful. It helps you identify what other variables will matter in your analysis and why, and it helps you identify additional information to see if this could be a problem. It turns out that the ideological fit with the district has some correlation with the vote, but it is not that strong. The reason is there is relatively little variation in ideological fit that is not explained by simple knowledge of party. So this additional information allows us to conjecture (reasonably safely) that, although ideological fit could be a problem, it likely does not explain the substantial bias in the coefficients on spending.

b. The general challenge in statistical modeling and inference is deciding how to balance possible biases against possible inefficiencies in choosing a particular specification. Naturally, we usually wish to err on the side of inefficiency. But, these are choices we make *on the margin.* As we will see statistical tests measure whether the possible improvement in bias from one model outweighs the loss of efficiency compared to another model. This should not distract from the main objective of your research, which is to find phenomena and relationships of large magnitude and of *substantive* importance. Concern about omitted Variable Bias should be a seed of doubt that drives you to make the estimates that you make as good as possible.

3.3.3. Examples

Model building in Statistics is really a progressive activity. We usually begin with interesting or important observations. Sometimes those originate in a theory of social behavior, and sometimes they come from observation of the world. Statistical analyses allow us to refine those observations. And sometimes they lead to a more refined view of how social behavior works. Obvious problems of measurement error or omitted variables exist when the implications of an analysis are absurd. Equally valid, though, are arguments that suggest a problem with a plausible result. Here we'll consider four examples.

1. Incumbency Advantages

The observation of the incumbency advantage stems from a simple difference of means. From 1978 to 2002, the average Demcoratic vote share of a typical U.S. House Democratic incumbent is 68%; the average Democratic vote share a typical U.S. House Republican incumbent is 34

The incumbency advantage model is specified as follows. The vote for the Democratic candidate in district i in election t equals the normal party vote, N_i, plus a national party tide, α_t, plus the effect of incumbency. Incumbency is coded $I_{it} = +1$ for Democratic Incumbents, $I_{it} = -1$ for Republican Incumbents, and $I_{it} = 0$ for Open Seats.

$$V_{it} = \alpha_t + N_i + \beta I_{it} + \epsilon_{it}$$

Controlling for the normal vote and year tides reduces the estimated incumbency effect to about 7 to 9 percentage oints.

2. Strategic Retirement and the Incumbency Advantage

An objection to models of the incumbency advantage is that incumbents choose to step down only when they are threatened, either by changing times, personal problems, or an unusually good challenger. The reasons that someone retires, then, might depend on factors that the researcher cannot measure but that predict the vote – omitted variables. This would cause I to be correlated to the regression error u. The factors that are thought to affect the retirement decisions and the vote are negatively correlated with V and negatively correlated with I (making it less likely to run). Hence, the incumbency advantage may be inflated.

3. Police/Prisons and Crime

The theory of crime and punishment begins with simple assumptions about rational behavior. People will comit crimes if the likelihood of being caught and the severity of the punishment are lower than the benefit to the crime. A very common observation in the sociology of crime is that areas that have larger numbers of police or more people in prison have higher crime rates.

4. Campaign Spending and Votes

Researchers measuring the factors that explain House election outcomes include various measures of electoral competition in explaining the vote. Campaign Expenditures, and the advertising they buy, are thought to be one of the main forces affecting election outcomes.

A commonly used model treats the incumbent party's share of the votes as a function of the normal party division in the congressional district, candidate characteristics (such as experience or scandals), and campaign expenditures of the incumbent and the challenger. Most of the results from such regressions make sense: the coefficient on challenger spending and on incumbent party strength make sense. But the coefficient on incumbent spending has the wrong (negative) sign. The naive interpretation is that the more incumbents spend the worse they do.

One possibile explanation is that incumbents who are unpopular and out of step with their districts have to spend more in order to remain in place. Could this explain the incorrect sign? In this account of the bias in the spending coefficients there is a positive correlation between the omitted variable, "incumbent's fit with the district," and the included variable, "incumbent spending." Also, the incumbent's fit with the district likely has a negative direct effect on the vote. The more out-of-step an incumbent is the worse he will do on election day. Hence, lacking a measure of "fit with the district" might cause a downward bias.

3.3.3. General strategies for Correcting for Omitted Variables and Measurement Error Biases

1. More Data, More Variables. Identify relevant omitted variables and then try to collect them. This is why many regression analyses will include a large number of variables that do not seem relevant to the immediate question. They are included to hold other things constant, and also to improve efficiency.

2. Multiple Measurement.

Two sorts of use of multiple measurement are common.

First, to reduce measurement error researchers often average repeated measures of a variable or construct an index to capture a "latent" variable. Although not properly a topic for this course, factor analysis and muti-dimensional scaling techniques are very handy for this sort of data reduction.

Second, to eliminate bias we may observe the "same observation" many times. For example, we could observe the crime rate in a set of cities over a long period of time. If the omitted factor is one that due to factors that are constant within Panel models. omitted Variables as Nuisance Factors. Using the idea of "control in the design."

3. Instrumental Variables.

Instrumental variables estimates allow researchers to purge the independent variable of interest with its correlation with the omitted variables, which cause bias. What is difficult is finding suitable variables with which to construct instruments. We will deal with this matter at length later in the course.

3.4. Prediction

3.4.1. Prediction and Interpretation

We have given one interpretation to the regression model as an estimate of the partial derivatives of a function, i.e., the effects of a set of independent variables holding constant the values of the other independent variables. In constructing this definition we began with the definition of an effect as the difference in the conditional mean of Y across two distinct values of X. And, a causal effect assumes that we hold all else constant.

Another important way to interpret conditional means and regressions is as predicted values. Indeed, sometimes the goal of an analysis is not to estimate effects but to generate predictions. For example, one might be asked to formulate a prediction about the coming presidential election. A common sort of election forecasting model regresses the incumbent president's vote share on the rate of growth in the economy plus a measure of presidential popularity plus a measure of party identification in the public. Based on elections since 1948, that regression has the following coefficients:

$$Vote = xxx + xxxGrowth + xxxPopularity + xxxxParty.$$

We then consider plausible values for Growth, Popularity, and Partisanship to generate predictions about the Vote.

For any set of values of X, say $\mathbf{x}_0$, the most likely value or expected value of Y is $y_0 = E[Y|\mathbf{x} = \mathbf{x}_0]$. This value is calculated in a straightforward manner from the estimated regression. Let $\mathbf{x}_0$ be a row vector of values of the independent variables for which a prediction is made. I.e., $\mathbf{x}_0 = (\mathbf{x}_1^0, \mathbf{x}_2^0, ...\mathbf{x}_k^0)$. The predicted value is calculated as

$$\hat{y}^0 = \mathbf{x}_0\mathbf{b} = \mathbf{b}_0 + \sum_{j=1}^{K} \mathbf{b}_j\mathbf{x}_j^0.$$

It is standard practice to set variables equal to their mean value if no specific value is of interest in a prediction. One should be somewhat careful in the choice of predicted values so that the value does not lie too far out of the set of values on which the regression was originally estimated.

Consider the presidential election example. Assume a growth rate of 2 percent, a Popularity rating of 50 percent, and a Republican Party Identification of 50 Percent (equal split between the parties), then Bush is predicted to receive xxx percent of the two-party presidential vote in 2004.

The predicted value or forecast is itself subject to error. To measure the forecast error we construct the deviation of the observed value from the "true" value, which we wish to predict. The true value is itself a random variable: $y^0 = \mathbf{x}_0\beta + \epsilon^0$. The prediction or forecast error is the deviation of the predicted value from the true value:

$$e^0 = y^0 - \hat{y}^0 = \mathbf{x}_0(\beta - \mathbf{b}) + \epsilon^0$$

The varince of the prediction error is

$$V[e^0] = \sigma_\epsilon^2 + V[(\mathbf{x}_0(\beta - \mathbf{b})] = \sigma_\epsilon^2 + (\mathbf{x}_0\mathbf{E}[(\beta - \mathbf{b})(\beta - \mathbf{b})']\mathbf{x}_0{}') = \sigma_\epsilon^2 + \sigma_\epsilon^2(\mathbf{x}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0{}')$$

We can use this last result to construct the 95 percent confidence interval for the predicted value:

$$\hat{y}^0 \pm 1.96\sqrt{V[e^0]}$$

As a practical matter this might become somewhat cumbersome to do. A quick way to generate prediction confidence intervals is with an "augmented" regression. Suppose we have estimated a regression using n observations, and we wish to construct several different predicted values based on different sets of values for $\mathbf{X}$, say $\mathbf{X}_0$. A handy trick is to add the matrix of values $\mathbf{X}_0$ to the bottom of the $\mathbf{X}$ matrix. That is add n^0 observations to your data set for which the values of the indepedent variables are the appropriate values of $\mathbf{X}_0$. Let the dependent variable equal 0 for all of these values. Finally, add n^0 columns to your dataset that equal -1 for each new observation. Now regress y on X and the new set of dummy variables.

The resulting estimates will reproduce the original regression and will have coefficient estimates for each of the independent variables. The coefficients on the dummy variables equal the predicted values and the standard errors of these estimates are the correct prediction standard errors.

As an example, consider the analysis of the relationship between voting weights and posts in parliamentary governments. Let's let regression calculate the predicted values and standard errors for 6 distinct cases: Voting Weight = .25 and Formateur, Voting Weight = .25 and Not Formateur, Voting Weight = .35 and Formateur, Voting Weight = .35 and Not Formateur, Voting Weight = .45 and Formateur, and Voting Weight = .45 and Not Formateur. First, I ran the regression of Share of Posts on Share of Voting Weight plus an Indicator variable of The Party that formed the government (Formateur). I then ran the regression with 6 additional observations, constructed as described above. Below are the estimated coefficients and standard errors (constants not reported).

Using Augmented Regression To Calculate Predicted Values

Variable	Coeff. (SE)	Coeff. (SE)
Voting Weight	.9812 (.0403)	.9812 (.0403)
Formateur	.2300 (.0094)	.2300 (.0094)
D_1 (.25, 1)	–	.5572 (.0952)
D_2 (.25, 0)	–	.3272 (.0952)
D_3 (.35, 1)	–	.6554 (.0953)
D_4 (.35, 0)	–	.4253 (.0955)
D_5 (.45, 1)	–	.7535 (.0955)
D_6 (.45, 0)	–	.5235 (.0960)

3.4.2. Model Checking

Predicted values allow us to detect deviations from most of the assumptions of the regression model. The one assumption we cannot validate is the assumption that X and ϵ are uncorrelated. The residual vector, $\mathbf{e}$, is defined to be orthogonal to the independent variables, $\mathbf{X}$: $\mathbf{e}'\mathbf{X} = \mathbf{0}$. This implies that $\mathbf{e}'\hat{\mathbf{y}} = \mathbf{e}'\mathbf{X}\mathbf{b} = \mathbf{0}$. This is a restriction, so we cannot test how well the data approximate or agree with this assumption.

The other assumptions – linearity, homoskedasticity, no autocorrelation, and normality – are readily verified, and fixed. A useful diagnostic tool is a residual plot. Graph the residuals from a regression against the predicted values. This plot will immediately show many problems, if they exist. We expect an elliptical cloud centered around $e = 0$.

If the underlying model is non-linear, the residual plot will reflect the deviations of the data from a straightline fit through the data. Data generated from a quadratic concave function will have negative residuals for low values of $\hat{y}$, then positive for intermediate values of $\hat{y}$, then negative for large values of $\hat{y}$.

If there is heteroskedasticity, the residual plot will show deviations non-constant deviations around $e = 0$. A common case arises when the residual plot looks like a funnel. This situation means that the effects are multiplicative. That is the model is $y = \alpha X^{\beta}\epsilon$ (where ϵ takes only positive values), not $y = \alpha + \beta X + epsilon$. This is readily fixed by taking logarithms, so that the model becomes: $y = log(\alpha) + \beta log(X) + log(\epsilon)$.

A further plot for measuring heteroskedasticity is of e^2 against $\hat{y}$, against a particular X variable, or against some other factor, such as the "size" of the unit. In this plot e^2 serves as the estimate of the variance. This is an enlightening plot when the variance is a function of a particular X or when we are dealing with aggregated data, where the aggregates consist of averages of variables across places of different populations.

To detect autocorrelation we use a slightly different plot. Suppose the units are indexed by time, say years, t. We may examine the extent of autocorrelation by taking the correlations between observations that are s units apart:

$$r_s = \frac{\sum_{t=1}^{T} e_t e_{t-s}}{\sum_{t=1}^{T} e_t^2}$$

The correlation between an observation and the previous observation is r_1. This is called first-order autocorrelation. Another form of autocorrelation, especially in monthly economic data, is seasonal variation, which is captured with $s = 12$.

It is instructive to plot the estimated autocorrelation against s, where s runs from 0 to a relatively large number, say 1/10th of T.

What should this plot look like? At $s = 0$, the autocorrelation parameter is just $\sigma_{\epsilon}^2 = \frac{\sigma_u^2}{1-\rho^2}$. The values of r_s for $s > 0$ depends on the nature of the autocorrelation structure.

Let's take a closer look.

The most basic and commonly analyzed autocorrelation structure involves an autoregression of order 1 (or AR-1). Let u_t be an error term that is independent of ϵ_t. First-order autocorrelation in ϵ is of the form:

$$\epsilon_t = \rho\epsilon_{t-1} + u_t$$

The variance of ϵ_t follows immediately from the definition of the variance:

$$\sigma^2_\epsilon = E[(\rho\epsilon_{t-1} + u_t)(\rho\epsilon_{t-1} + u_t)] = \rho^2\sigma^2_{epsison} + \sigma^2_u$$

Solving for $\sigma^2_{epsilon}$:

$$\sigma^2_\epsilon = \frac{\sigma^2_u}{(1-\rho^2)}.$$

To derive, the correlation between t and $t-s$, we must derive the covariance first. For $s = 1$, the covariance between two observations is

$$E[\epsilon_t\epsilon_{t-1}] = E[(\rho\epsilon_{t-1} + u_t)\epsilon_{t-1}] = \rho\sigma^2_\epsilon$$

Using repeated substitutions for ϵ_t we find that for an AR-1:

$$E[\epsilon_t\epsilon_{t-s}] = \rho^s\sigma^2_\epsilon$$

Now, we can state what we expect to observe in the r_s when the residuals contain first-order autocorrelation:

$$\rho_s = \frac{Cov(\epsilon_t, \epsilon_{t-s})}{\sqrt{V(\epsilon_t)V(\epsilon_{t-1})}} = \frac{\rho^s\sigma^2_\epsilon}{\sigma^2_\epsilon} = \rho^s$$

There are two patterns that may appear in the autocorrelation plot, depending on the sign of ρ. If $\rho > 0$, the plot should decline exponentially toward 0. For examle, suppose $\rho = .5$, then we expect $r_0 = 1$, $r_1 = .5$, $r_2 = .25$, $r_3 = .125$, $r_4 = .0625$, $r_5 = .03125$. If $\rho < 0$, the plot will seesaw, converging on 0. For example, suppose $\rho = -.5$, then we expect $r_0 = 1$, $r_1 = -.5$, $r_2 = .25$, $r_3 = -.125$, $r_4 = .0625$, $r_5 = -.03125$.

Higher order auto-correlation structures – such as $\epsilon_t = \rho_1\epsilon_{t-1} + \rho_2\epsilon_{t-2} + u_t$ – lead to more complicated patterns. We may test for the appropriateness of a particular structure by comparing the estimated autocorrelations, r_s, with the values implied by a structure. For example, we may test for first order autocorrelation by comparing the observed r_s's with those implied by the AR-1 model when $\hat{\rho} = r_1$.

A simple rule of thumb applies for all autocorrelation structures. If $r_s < .2$ for all s, then there is no significant degree of autocorrelation.

3.5. Inference

3.5.1. General Framework.

Hypotheses.

What is an hypothesis? An hypothesis is a statement about the data derived from an argument, model, or theory. It usually takes the form of a claim about the behavior of parameters of the distribution function or about the effect of one variable on another.

For example, a simple argument about voting holds that in the absence of other factors, such as incumbency, voters use party to determine their candidate of choice. Therefore, when no incumbent is on the ticket, an additional one-percent Democratic in an electoral district should translate into one-percent higher Democratic vote for a particular office. In a regression model, controlling for incumbency, the slope on the normal vote ought to equal one. Of course there are a number of reasons why this hypothesis might fail to hold. The argument itself might be incorrect; there may be other factors beside incumbency that must be included in the model; the normal vote is hard to measure and we must use proxies, which introduce measurement error.

In classical statistics, an hypothesis test is a probability statement. We reject an hypothesis if the probability of observing the data given that the hypothesis is true is sufficiently small, say below .05. Let $\mathbf{T}$ be the vector of estimated parameters and θ be the true parameters of the distribution. Let θ_0 be the values of the distribution posited by the hypothesis. Finally, let $\mathbf{\Sigma}$ be the variance of $\mathbf{T}$. If the hypothesis is true, then $\theta = \theta_0$. If the hypothesis is true, then the deviation of $\mathbf{T}$ from θ_0 should look like a random draw from the underlying distribution, and thus be unlikely to have occured by chance.

What we have just described is the size of a test. We also care about the power of a test. If θ_0 is not true, what is the probability of observing a sufficiently small deviation that we do not reject the hypothesis? This depends on sample size and variance of X.

Tests of a Single Parameter.

We may extend the framework for statistical tests about means and differences of means to the case of tests about a single regression coefficient. Recall that the classical hypothesis test for a single mean was:

$$Pr(|\bar{x} - \mu_0| > t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}) < \alpha$$

Because $\bar{x}$ is the sum of random variables, its distribution is approximately normal. However, because we must estimate the standard deviation of X, s, the t-distribution is used as the reference distribution for the hypothesis test. The test criterion can be rewritten as follows. We reject the hypothesis if $\frac{|\bar{x}-\mu_0|}{s/\sqrt{n}} > t_{\alpha/2,n-1}$.

There is an important duality between the test criterion above and the confidence interval. The test criterion for size .05 may be rewritten as:

$$Pr(\bar{x} - t_{.025,n-1}\frac{\sigma}{\sqrt{n}} < \mu_0 < \bar{x} + t_{.025,n-1}\frac{\sigma}{\sqrt{n}}) > .95$$

So, we can test the hypothesis by ascertaining whether the hypothesized value falls inside the 95 percent confidence interval.

Now consider the regression coefficient, β. Suppose our hypothesis is $H : \beta = \beta_0$. A common value is $\beta_0 = 0$ – i.e., no effect of X on Y. Like the sample average, a single regression parameter follows the normal distribution, because the regression parameter is the sum of random variables. The mean of this distribution is β and the variance $\frac{\sigma_\epsilon^2}{\sum(x_i-\bar{x})^2}$, in the case of a bivariate regression, or, more generally, $\sigma_\epsilon^2 a_{jj}$, where a_{jj} is the jth diagonal element of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$.

The test criterion for the hypothesis states the following. If the null hypothesis is true, then we expect that the probability of a large standardized deviation b from β_0 will be unlikely to have occurred by chance:

$$Pr(|b - \beta_0| > t_{\alpha/2,n-K}s\sqrt{a_{jj}}) < \alpha$$

As with the sample mean, this test criterion can be expressed as follows. We reject the hypothesized value if $\frac{|b-\beta_0|}{s\sqrt{a_{jj}}} > t_{\alpha/2,n-K}$.

The Wald Criterion.

The general framework for testing in the multivariate context is a Wald Test. Construct a vector of estimated parameters and a vector of hypothesized parameters. Generically, we will write these as $\mathbf{T}$ and θ_0, because they may be functions of regression coefficients, not just the coefficients themselves. Define the vector of deviations of the observed data from the hypothesized data as $\mathbf{d} = \mathbf{bfT} - \theta_0$. This is a vector of length J.

What is the distribution of $\mathbf{d}$? Under the null hypothesis, $E[\mathbf{d}] = \mathbf{0}$ and $V[\mathbf{d}] = \mathbf{V}[\mathbf{T}] = \mathbf{\Sigma}$. The Wald statistic is:

$$W = \mathbf{d}'\mathbf{\Sigma}^{-1}\mathbf{d}.$$

Assuming $\mathbf{d}$ is normally distributed, the Wald Statistic follows the χ^2 distribution with J degrees of freedom.

Usually, we do not know $\mathbf{\Sigma}$, and must estimate it. Substituting the estimated value of $\mathbf{\Sigma}$ and making an appropriate adjustment for the degrees of freedom, we arrive at a new statistic that follows the F distribution. The degrees of freedom are J and $n - K$, which are the number of restrictions in $\mathbf{d}$ and the number of free pieces of information available after estimation of the model.

In the regression framework, we may implement a test of a set of linear restrictions on the parameters as follows. Suppose the hypotheses take the form that linear combinations of various parameters must equal specific constants. For example,

$$\beta_1 = 0$$

$$\beta_2 = 3$$

and

$$\beta_4 + \beta_3 = 1$$

. We can specify the hypothesis as

$$\mathbf{R}\beta = \mathbf{q}$$

Where R is a matrix of numbers that correspond to the parameters in the set of linear equations found in the hypothesis; β is the vector of coefficients in the full model (length K), and q is an appropriate set of numbers. In the example, suppose there are 5 coefficints in the full model:

$$\mathbf{R}\beta = \begin{pmatrix} 1\ 0\ 0\ 0\ 0 \\ 0\ 1\ 0\ 0\ 0 \\ 0\ 0\ 1\ 1\ 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} \begin{pmatrix} 0 \\ 3 \\ 1 \end{pmatrix}$$

The deviation of the estimated value from the true value assuming the hypothesis is true is

$$\mathbf{d} = \mathbf{Rb} - \mathbf{R}\beta = \mathbf{Rb} - \mathbf{q}$$

If the hypothesis is correct: $E[\mathbf{d}] = \mathbf{E}[\mathbf{Rb} - \mathbf{q}] = \mathbf{0}$.

$$V[\mathbf{d}] = \mathbf{E}[\mathbf{dd}'] = \mathbf{E}[(\mathbf{Rb} - \mathbf{R}\beta)(\mathbf{Rb} - \mathbf{R}\beta)'] = \mathbf{E}[\mathbf{R}(\mathbf{b} - \beta)(\mathbf{b} - \beta)'\mathbf{R}']$$

Because $\mathbf{R}$ is a matrix of constants:

$$V[\mathbf{d}] = \mathbf{RE}[(\mathbf{b} - \beta)(\mathbf{b} - \beta)']\mathbf{R}' = \mathbf{R}\sigma_\epsilon^2\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$$

Because we usually have to estimate the variance of ϵ the Wald statistic is

$$F = \frac{(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'](\mathbf{Rb} - \mathbf{q})/\mathbf{J}}{\mathbf{e}'\mathbf{e}/(\mathbf{n} - \mathbf{K})}$$

Another way to write the F-statistic for the Wald test is as the percentage loss of fit. Suppose we estimate two models. In one model, the Restricted Model, we impose the hypothesis, and on the other model, the Unrestricted Model, we impose no restrictions. The F-test for the Wald criterion can be written as

$$F = \frac{\mathbf{e}_\mathrm{R}'\mathbf{e}_\mathrm{R} - \mathbf{e}_\mathrm{u}'\mathbf{e}_\mathrm{u}/\mathbf{J}}{\mathbf{e}_\mathrm{u}'\mathbf{e}_\mathrm{u}/(\mathbf{n} - \mathbf{K})}$$

Denote the residuals from these models as $\mathbf{e}_\mathrm{R}$ and $\mathbf{e}_\mathrm{U}$. We can write the residuals from the restricted model in terms of the residuals from the unrestricted model: $\mathbf{e}_\mathrm{R} = \mathbf{y} - \mathbf{X}\mathbf{b}_\mathrm{R} =$

$\mathbf{y} - \mathbf{X}\mathbf{b}_{\mathrm{U}} - \mathbf{X}(\mathbf{b}_{\mathrm{R}} - \mathbf{b}_{\mathrm{U}}) = \mathbf{e}_{\mathrm{U}} - \mathbf{X}(\mathbf{b}_{\mathrm{R}} - \mathbf{b}_{\mathrm{U}})$. The sum of squares residuals from the restricted model is $\mathbf{e}'_{\mathrm{R}}\mathbf{e}_{\mathrm{R}} = (\mathbf{e}_{\mathrm{U}} - \mathbf{X}(\mathbf{b}_{\mathrm{R}} - \mathbf{b}_{\mathrm{U}}))'(\mathbf{e}_{\mathrm{U}} - \mathbf{X}(\mathbf{b}_{\mathrm{R}} - \mathbf{b}_{\mathrm{U}}))$
$= \mathbf{e}'_{\mathrm{U}}\mathbf{e}_{\mathrm{U}} - (\mathbf{b}_{\mathrm{R}} - \mathbf{b}_{\mathrm{U}})'\mathbf{X}'\mathbf{X}(\mathbf{b}_{\mathrm{R}} - \mathbf{b}_{\mathrm{U}}) = \mathbf{e}'_{\mathrm{U}}\mathbf{e}_{\mathrm{U}} - (\mathbf{R}\mathbf{b}_{\mathrm{U}} - \mathbf{q})'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')(\mathbf{R}\mathbf{b}_{\mathrm{U}} - \mathbf{q})$ So, the difference in the sum of squares between the restricted and the unrestricted model equals the numerator of the Wald test.

There are two other important testing criteria – likelihood ratio statistics and lagrange multiplier tests. These three are asymptotically the same. The Wald criterion has better small sample properties, and it is easy to implement for the regression model. We will discuss likelihood ratio statistics as part of maximum likelihood estimation.

3.5.2. Applications

Three important applications of the Wald Test are (1) tests of specific hypotheses, (2) verification of choice of a specification, and (3) tests of bias reduction in causal estimates, specifically OLS versus IV. All three amount to comparing the bias due to the more parsimonious hypothesized model against the efficiency gain with that model.

i. Hypothesis Tests: Seats and Votes in England, 1922-2002.

The Cube Law provides an excellent example of the sort of calculation made to test a concrete hypothesis. The Cube Law states that the proportion of seats won varies as the cube of the proportion of votes won:

$$\frac{S}{(1-S)} = \left(\frac{V}{1-V}\right)^3 .$$

We may implement this in a regression model as follows. This is a multiplicative model which we will linearize using logarithms. There are really two free parameters – the constant term α and the exponent β.

$$log(\frac{S}{(1-S)}) = \alpha + \beta log\left(\frac{V}{1-V}\right)$$

For English Parliamentary elections from 1922 to 2002, the graph of the log of the odds of the Conservatives winning a seat (labelled LCS) versus the log of the odds of the Conservatives winning a vote (labelled LCV) is shown in the graph. The relationship looks fairly linear.

Results of the least squares regression of LCS on LCV are shown in the table. The slope is 2.75 and the intercept is -.07. Tests of each coefficient separately show that there we would safely accept the hypothesis that $\beta_0 = 0$ and $\beta_1 = 3$, separately. Specifically, the t-test of whether the slope equals 3 is $\frac{2.75-3}{.20} = 1.25$, which has a p-value of .25. The t-test of whether the constant equals 0 is $\frac{-.07-0}{.05} = -1.4$, which has a p-value of .18.

Regression Example:
Cube Law in England, 1922-2002
Y = Log of the Odds Ratio of Conservative Seats
X = Log of the Odds Ratio of Conservative Votes

Variable	Coeff. (SE)	t-test
LCV	2.75 (.20)	.25/.2 = 1.25 (p = .24)
Constant	-0.07 (.05)	-.07/.05 = -1.4 (p = .18)
N	21	
R^2	.91	
MSE	.221	
F-test for $H : \beta_0 = 0,$ $\beta_1 = 3$	2.54 (p=.10)	

However, this is the wrong approach to testing hypotheses about multiple coefficients. The hypothesis has implications for *both* coefficients at the same time. An appropriate test measures how much the vector $\mathbf{b}' = (\mathbf{2.75}, -\mathbf{.07})$ deviates from the hypothesized vector $\beta_0 = (\mathbf{3}, \mathbf{0})$.

After estimating the regression, we can extract the variance-covariance matrix for **b**, using the command **matrixliste(V)**. If the null hypothesis is true, then $E[\mathbf{b}] = \beta_0$. The Wald criterion is $\mathbf{d}'\mathbf{V}[\mathbf{d}]^{-1}\mathbf{d}$. Because the matrix **R** is the identity matrix in this example, i.e., each of the hypotheses pertains to a different coefficient, $V(\mathbf{d})$ equals $V(\mathbf{b})$. So, the F-statistic for the Wald criterion is:

$$F = \frac{1}{2}(.246, -.070)\begin{pmatrix} 27.6405 & 38.2184 \\ 38.2184 & 427.3381 \end{pmatrix}\begin{pmatrix} .246 \\ -.070 \end{pmatrix}$$

$$= .5((-.246)^2 27.6405 + 2(-.070)(-.246)38.2184 + (-.070)^2 427.3381) = 2.54$$

The probability of observing a deviation this large for a random variable distributed F with 2 and 19 degrees of freedom is .10 (i.e., p-value = .10). This suggests that the data do deviate somewhat from the Cube Law.

We can implement such tests using the **test** command in STATA. Following a regression, you can use **test variable names = q** to test a single restriction — $f(b_1, b_2, ...) = q_1$, such as **test LCV = 3**. To test multiple restrictions, you must accumulate successive tests using the option **accum**. In our example, first type **test LCV = 3**, then type **test _cons = 0, accum**. This returns $F = 2.54, p - value = .10$.

ii. Choice of Specification: Energy Survey.

By a *specification*, we mean a specific set of independent variables included in a regression model. Because there is a tradeoff between efficiency when irrelevant variables are included and bias when relevant variables are excluded, researchers usually want to find the smallest possible model. In doing so, we often jettison variables that appear insignificant according to the rule of thumb that they have low values of the t-statistic. After running numerous regressions we may arrive at what we think is a good fitting model.

The example of the Cube Law should reveal an important lesson. The t-statistic on a single variable is only a good guide about that variable. If you've made decisions about many variables, you might have made a mistake in choosing the appropriate set.

The F-test for the Wald criterion allows us to test whether an entire set of variables in fact ought to be omitted from a model. The decision to omit a set of variables amounts to a specific hypothesis – that each of the variables can be assumed to have coefficients equal to 0.

Consider the data from the Global Warming Survey discussed earlier in the course. See handout. There are 5 variables in the full regression model that have relatively low t-statistics. To improve the estimates of the other coefficients, we ought to drop these values. Also, we might want to test whether some of these factors should be dropped as a substantive matter.

The F-test is implemented by calculating the average loss of fit in the sum of squared residuals. The information in the ANOVA can be used to compute:

$$F = \frac{(1276.4912 - 1270.5196)/5}{1.260436} = .94754514$$

One may also use the commands **test** and **test, accum**.

Specification tests such as this one are very common.

iii. OLS versus Instrumental Variables: Using Term Limits To Correct For Strategic Retirement.

We may also use the Wald criterion to test across estimators. One hypothesis of particular importance is that there are no omitted variables in the regression model. We cannot generally detect this problem using regression diagnostics, such as graphs, but we may be able to construct a test.

Let $\mathbf{d} = \mathbf{b_{IV}} - \mathbf{b_{OLS}}$. The Wald criterion is $\mathbf{d}'[\mathbf{V}[\mathbf{d}]^{-1}\mathbf{d}$.

$$V[b_{IV} - b_{OLS}] = V[b_{IV}] + V[b_{OLS}] - Cov[b_{IV}, b_{OLS}] - Cov[b_{OLS}, b_{IV}]$$

Hausman shows that $Cov[b_{OLS}, b_{IV}] = V[b_{IV}]$, so

$$V[b_{IV} - b_{OLS}] = V[b_{IV}] - V[b_{OLS}].$$

Let $\hat{\mathbf{X}}$ be the set of predicted values from regressing X on Z. The F-test for OLS versus IV is

$$H = \frac{\mathbf{d}'[(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{d}}{s^2}$$

This is distributed F with J and n-K degrees of freedom, where J is the number of variables we have tried to fix with IV.

4. Non-Linear Models

4.1. Non-Linearity

We have discussed one important violation of the regression assumption – omitted variables. And, we have touched on problems of inefficiency introduced by heteroskedasticity and autocorrelation. This and the following subsections deal with violations of the regression assumptions (other than the omitted variables problem). The current section examines corrections for non-linearity; the next section concerns discrete dependent variables. Following that we will deal briefly with weighting, heteroskedasticity, and autocorrelation. Time permitting we will do a bit on Sensitivity.

We encounter two sorts of non-linearity in regression analysis. In some problems non-linearity occurs among the X variables but it can be handled using a linear form of non-linear functions of the X variables. In other problems non-linearity is inherent in the model: we cannot "linearize" the relationship between Y and X. The first sort of problem is sometimes called "intrinsically linear" and the second sort is "intrinsically non-linear."

Consider, first, situations where we can convert a non-linear model into a linear form. In the Seats-Votes example, the basic model involved multiplicative and exponential parameters. We converted this into a linear form by taking logarithms. There are a wide range of non-linearities that can arise; indeed, there are an infinite number of transformations of variables that one might use. Typically we do not know the appropriate function and begin with a linear relationship between Y and X as the approximation of the correct relationship. We then stipulate possible non-linear relationships and implement transformations.

Common examples of non-linearities include:

Multiplicative models, where the independent variables enter the equation multiplicatively rather than linearly (such models are linear in logarithms);

Polynomial regressions, where the independent variables are polynomials of a set of variables (common in production function analysis in economics); and

Interactions, where some independent variables magnify the effects of other independent variables (common in psychological research).

In each of these cases, we can use transformations of the independent variables to construct a linear specification with with we can estimate all of the parameters of interest.

Qualitative Independent variables and interaction effects are the simplest sort of non-linearity. They are simple to implement, but sometimes hard to interpret. Let's consider a simple case. Ansolabehere and Iyengar (1996) conducted a series of laboratory experiments involving approximately 3,000 subjects over 3 years. The experiments manipulated the content of political advertisements, nested the manipulated advertisements in the commercial breaks of videotapes of local newscasts, randomly assigned subjects to watch specific tapes, and then measured the political opinions and information of the subjects. These experiments are reported in the book *Going Negative*.

On page 190, they report the following table.

Effects of Party and Advertising Exponsure on Vote Preferences: General Election Experiments

Independent	(1)	(2)
Variable	Coeff. (SE)	Coeff. (SE)
Constant	.100 (.029)	.102 (.029)
Advertising Effects		
Sponsor	.077 (.012)	.023 (.034)
Sponsor*Same Party	–	.119 (.054)
Sponsor*Opposite Party	–	.028 (.055)
Control Variables		
Party ID	.182 (.022)	.152 (.031)
Past Vote	.339 (.022)	.341 (.022)
Past Turnout	.115 (.030)	.113 (.029)
Gender	.115 (.029)	.114 (.030)

The dependent variable is +1 if the person stated an intention to vote Democratic after viewing the tape, -1 if the person stated an intention to vote Republican, and 0 otherwise. The variable Sponsor is the party of the candidate whose ad was shown; it equals +1 if the ad was a Democratic ad, -1 if the ad was a Republican ad, and 0 if no political ad was shown (control). Party ID is similarly coded using a trichotomy. Same Party was coded as +1 if a person was a Democrat and saw a Democratic ad or a Republican and saw a Republican ad. Opposite Party was coded as +1 if a person was a Democrat and saw a Republican ad or a Republican and saw a Democratic ad.

In the first column, the "persuasion" effect is estimated as the coefficient on the variable Sponsor. The estimate is .077. The interpretation is that exposure to an ad from a candidate increases support for that candidate by 7.7 percentage points.

The second column estimates interactions of the Sponsor variable with the Party ID variable. What is the interpretation of the set of three variables Sponsor, Sponsor*Same Party , and Sponsor*Opposite Party. The variables Same Party and Opposite Party encompass all party identifiers. When these variables equal 0, the viewer is a non-partisan. So, the coefficient on Sponsor in the second column measures the effect of seeing an ad among independent viewers. It increases support for the sponsor by only 2.3 percentage points. When Same Party equals 1, the coefficient on Sponsor is 11.9. This is not the effect of the ad among people of the same party. It is the difference between the Independents and those of the same party. To calculate the effect of the ad on people of the same party we must add .119 to .023, yielding an effect of .142, or a 14.2 percentage point gain.

Interactions such as these allow us to estimate different slopes for different groups, changes in trends, and other discrete changes in functional forms.

Another class of non-linear specifications takes the form of Polynomials. Many theories of behavior begin with a conjecture of an inherently non-linear function. For instance, a firm's production function is thought to exhibit decreasing marginal returns on investments, capital or labor. Also, risk aversion implies a concave utility function.

Barefoot empiricism sometimes leads us to non-linearity, too. Examination of data, either a scatter plot or a residual plot, may reveal a non-linear relationship, between Y and X. While we do not know what the right functional form is, we can capture the relationship in the data by including additional variables that are powers of X, such as $X^{1/2}$, X^2, X^3, as well as X. In other words, to capture the non-linear relationship, $y = g(x)$, we approximate the unknown function , $g(x)$, using a polynomial of the values of x.

One note of caution. Including powers of independent variables often leads to collinearity among the righthand side variables. One trick for breaking such collinearity is to deviate the independent variables from their means before transforming them.

Example: Candidate Convergence in Congressional Elections.

An important debate in the study of congressional elections concerns how well candidates represent their districts. Two conjectures concern the extent to which candidates reflect their districts. First, are candidates responsive to districts? Are candidates in more liberal districts more liberal? Second, does competitiveness lead to closer representation of district preferences? In highly competitive districts, are the candidates more "converged"? Huntington posits greater divergence of candidates in competitive areas – a sort of clash of ideas.

Ansolabehere, Snyder, and Stewart (2001) analyze data from a survey of congressional candidates on their positions on a range of issues during the 1996 election. There are 292 districts in which both of the competing candidates filled out the survey. The authors constructed a measure of candidate ideology from these surveys, and then examine the midpoint between the two candidates (the cutpoint at which a voter would be indifferent between the candidates) and the ideological gap between the candidates (the degree of convergence). To measure electoral competitiveness the authors used the average share of vote won by the Democratic candidate in the prior two presidential elections (1988 and 1992).

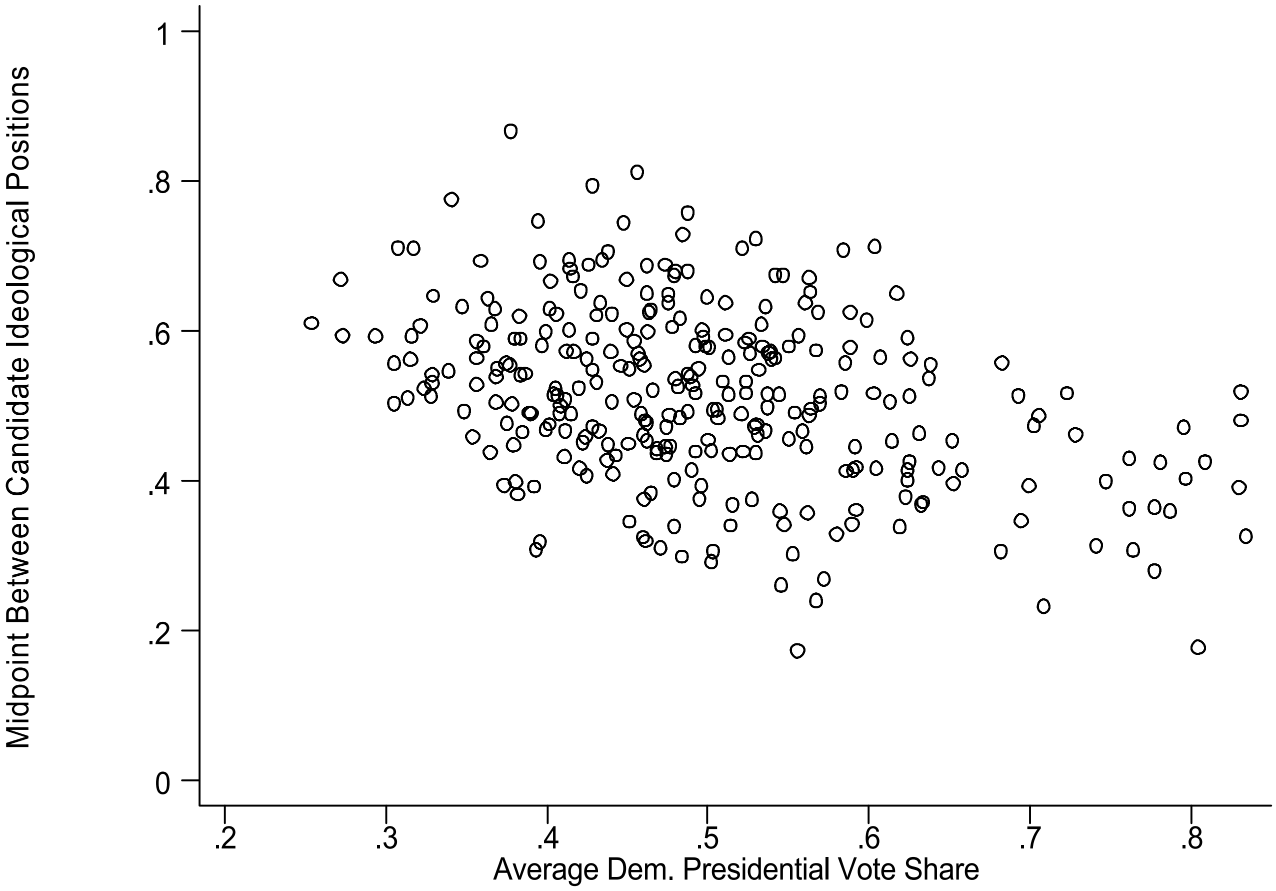
The Figures show the relationship between Democratic Presidential Vote share and, respectively, candidate midpoints and the ideological gap between competing candidates.

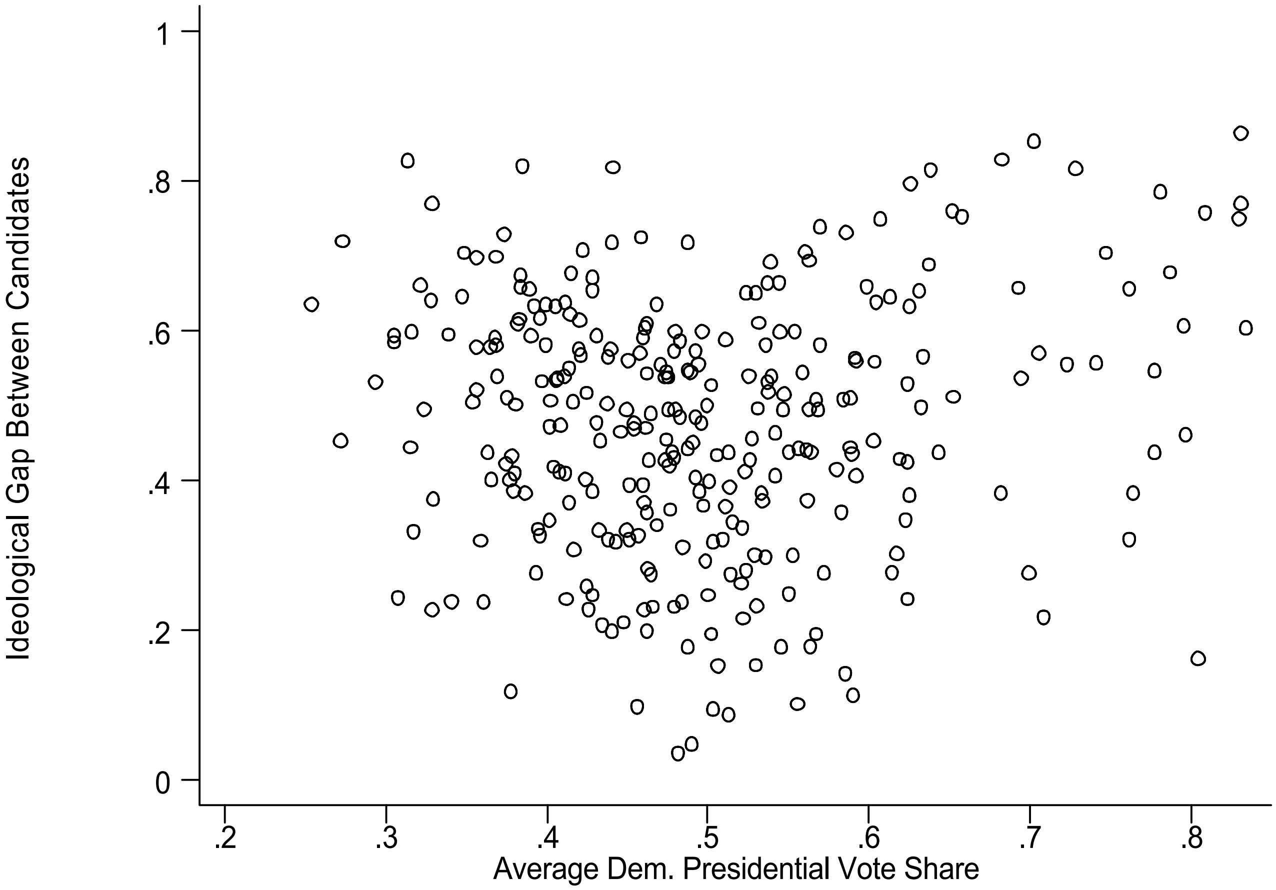
There is clear evidence of a non-linear relationship explaining the gap between the candidates

The table presents a series of regressions in which the Midpoint and the Gap are predicted using quadratic functions of the Democratic Presidential Vote plus an indicator of Open Seats, which tend to be highly competitive. We consider, separately, a specification using the value of the independent variable and its square and a specification using the value of the independent variable deviated from its mean and its square. The last column uses the absolute value of the deviation of the presidential vote from .5 as an alternative to the quadratic.

Effects of District Partisanship on Candidate Positioning
N = 292

	Dependent Variable				
	Midpoint		Gap		
Independent	(1)	(2)	(3)	(4)	(5)
Variable	Coeff.(SE)	Coeff.(SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)
Democratic Presidential Vote	-.158 (.369)	—	-2.657 (.555)	—	—
Democratic Presidential Vote Squared	-.219 (.336)	—	+2.565 (.506)	—	—
Dem. Pres. Vote Mean Deviated	—	-.377 (.061)	—	-.092 (.092)	—
Dem. Pres. Vote Mean Dev. Squared	—	-.219 (.336)	—	+2.565 (.506)	—
Absolute Value of Dem. Pres. Vote, Mean Dev.	—	—	—	—	.764 (.128)
Open Seat	.007 (.027)	.007 (.027)	-.024 (.040)	-.024 (.040)	-.017 (.039)
Constant	.649 (.098)	.515 (.008)	1.127 (.147)	.440 (.012)	.406 (.015)
R^2	.156	.156	.089	.089	.110
$\sqrt{MSE}$	.109	.109	.164	.164	.162
F (p-value)	18.0 (.00)	18.0 (.00)	9.59 (.00)	9.59 (.00)	12.46 (.00)


Midpoint Between Candidate Ideological Positions
1
.8
.6
.4
.2
0
.2
.3
.4
.5
.6
.7
.8
Average Dem. Presidential Vote Share


Ideological Gap Between Candidates
1
.8
.6
.4
.2
0
.2
.3
.4
.5
.6
.7
.8
Average Dem. Presidential Vote Share

Consider, first, the regressions explaining the midpoint (columns 1 and 2). We expect that the more liberal the districts are the more to the leftward the candidates will tend. The ideology measure is oriented such that higher values are more conservative positions. Figure 1 shows a strong relationship consistent with the argument.

Column (1) presents the estimates when we naively include Presidential vote and Presidential vote squared. This is a good example of what collinearity looks like. The F-test shows that the regression is "significant" – i.e., not all of the coefficients are 0. But, neither the coefficient on Presidential Vote or Presidential Vote Squared are signficant. Tell-tale collinearity. A trick for breaking the collinearity in this case is by deviating X from its mean. Doing so, we find a significant effect on the linear coefficient, but the quadratic coefficient doesn't change. There is only really one free coefficient here, and it looks to be linear.

The coefficients in a polynomial regression measure the partial derivatives of the unknown function evaluated at the mean of X. Taylor's approximation leads us immediately to this interpretation of the coefficients for polynomial models. Recall from Calculus that Taylor's Theorem allows us to express any function as a sum of derivatives of that function evaluated at a specific point. We may choose any degree of approximation to the function by selecting a specific degree of approximation. A first-order approximation uses the first derivatives; a second order approximation uses the second derivatives; and so on. A second order approximation of an unknown fuction, then, may be expressed as:

$$y_i \approx \alpha + \beta' \mathbf{x_i} + \frac{\mathbf{1}}{\mathbf{2}} \mathbf{x_i' H_0 x_i},$$

where

$$\mathbf{g_0} = \left[\frac{\partial \mathbf{f(x)}}{\partial \mathbf{x}} \right]_{\mathbf{x=x_0}}$$

$$\mathbf{H_0} = \left[\frac{\partial^2 \mathbf{f(x)}}{\partial \mathbf{x} \partial \mathbf{x}'} \right]_{\mathbf{x=x_0}}$$

$$\alpha = f(\mathbf{x_0}) - \mathbf{g_0' x_0} + \frac{\mathbf{1}}{\mathbf{2}} \mathbf{x_0' H_0 x_0}$$

$$\beta = \mathbf{g_0} - \mathbf{H_0 x_0}.$$

The coefficients on the squares and cross-product terms, then, capture the approximate second derivative. The coefficients on the linear values of x equal the gradient, adjusting for the quadrature around the point at which the data are evaluated (X_0). If we deviate all of the variables from their means first, then the coefficient on X in the polynomial regression can be interpreted straightforwardly as the gradient of the (unknown) function at the mean.

Example: Candidate Convergence in Congressional Elections, continued.

Consider the estimates in column (3). We may analyze these coefficients to ask several basic questions. What is the marginal rate of change? $\frac{\partial y}{\partial DP} = \beta_1 + 2\beta_2 DP$. Since DP ranges from .2 to .8, the rate of change in the Gap for a change in electoral competition ranges from -1.62 when DP = .2 to +1.42 when DP = .8. At what point does the rate of change equal 0 (what is the minimum)? Setting the partial equal to 0 reveals that $DP_0 = \frac{-\beta_1}{2\beta_2} = .517$, so the candidates are most converged near .5. What is the predicted value of the gap at this point? $1.127 - 2.657 * .517 + 2.565 * (.517)^2 = .438$. At the extremes, the predicted values are .64 (when DP = .8) and .70 (when DP = .2).

If we must choose among several transformations, such as logarithms, inverses, and polynomials, we typically cannot test which is most appropriate using the Wald test. Davidson and MacKinnon (1981) propose using the predicted values from one of the non-linear analysis to construct a test. Suppose we have two alternative models: $y = \mathbf{X}\beta$ and $y = \mathbf{Z}\gamma$. Consider a compound model

$$\mathbf{y} = (\mathbf{1} - \lambda)\mathbf{bfX}\beta + \lambda\mathbf{Z}\gamma + \epsilon$$

The J-test is implemented in two steps. First, estimate $y = \mathbf{Z}\gamma$ to get $\mathbf{Z}\hat{\gamma}$. Then, include $\mathbf{Z}\hat{\gamma}$ in the regression of $y = \mathbf{X}\beta$. The coefficient on $\mathbf{Z}\hat{\gamma}$ yields an estimate of $\hat{\lambda}$. Asymptotically, the t-ratio $\frac{\lambda\hat{-}0}{SE(\lambda}$ tests which model is appropriate.

Example: Candidate Convergence in Congressional Elections, continued.

An alternative to the quadratic specification in column (4) is an absolute value specification in column (5). Notice that the R^2 is slightly better with the absolute value. However, owing to collinearity, it is hard to test which model is more appropriate. I first regress on Gap on DP and DP^2 and $Open$ and generated predicted values, call these $GapHat$. I then regress Gap on ZG and $|DP - .5|$ and $Open$. The coefficient on $|DP - .5|$ is 1.12 with a standard error of .40, so the t-ratio is above the critical value for statistical significance. The coefficient on ZG equals -.55 with a standard error of .58. The estimated λ is not significantly different from 0; hence, we favor the absolute value specification.