Multivariate Statistics
Lecture Notes

Stephen Ansolabehere

Spring 2004

TOPICS

1. **The Basic Regression Model**

2. **Regression Model in Matrix Algebra**

3. **Estimation**

4. **Inference and Prediction**

5. **Logit and Probit**

6. **Panel Models**

7. **Structural Models and Simultaneous Equations**

1. The Basic Regression Model

Multivariate regression is the primary statical tool for applied statistics in the social sciences. Consider two applications.

Example 1. Capital Asset Pricing Model. In value an investment we assess the expected return and the risk associated with that investment. The risk is how much the stock may deviate from the expected return if we cash out the investment at some future date. The expected return is the average return. How risky is a stock? There are two views. First, there is the "overall" risk, which is measured as the variance of the stock. Second, there is the risk associated with the stock itself, and not with the greater market. It may be possible to reduce the overall risk by holding a diverse portfolio so we need to know how risky is a stock relative to the market.

A simple model of this is captured in the regression of the rate of return to a given firm's stock on the rate of return of the market. Let Y_t be the percentage change in the daily stock price of a firm; let X_t be the percentage change in the market (e.g., Dow Jones Industrial Average); let ϵ_t represent a random error term; and let t index date.

$$Y_t = \alpha + \beta X_t + \epsilon_t$$

The coefficient β measures the risk relative to the market. If $\beta > 1$ the firm price is more volatile than the market; if $\beta < 1$ the firm price is less volatile than the market. The figure shows the CAPM regression for two arbitrarily chosen Fortune 500 firms – WalMart and General Motors.

Example 2. Voting Technologies and Uncounted Votes. Votes are counted using at least 5 different technologies for tabulation in the United States – hand-counted paper, lever machines, punch cards (at the center of the controversial recount in Florida in 2000), optically scanned paper, and electronic voting machines. A standard measure for the percent of ballots uncounted is the percentage difference between total ballots cast and the votes counted for any given office, say president. This is termed the "residual vote." Other factors, such as demographics, also affect the rate of uncounted ballots.

The rate of residual votes ought to be independent of the technology used. For simplicity, suppose we wish to know whether punch cards are worse than alternative technologies, ignoring what exact alternative technology is used. Technologies used vary from county to county, sometimes town to town. Let j index counties. Let D_j indicate whether a county uses punch cards; it equals 1 if the county uses punchards and 0 otherwise. Let Y_j represent the residual vote rate in each county, and let X_j represent the socio-economic status of the county; let ϵ_j represent a random error.

$$Y_j = \alpha + \beta_1 D_j + \beta_2 X_j + \epsilon_j$$

Essentially, the regression is two parallel lines, both with slope β_2, but one with intercept α and the other with intercept $\alpha + \beta_1$. Make a graph Y on X.

The coefficient β_1 is the expected difference between Punch Cards and other technologies, holding constant a county's socio-economic condition.

Why hold X_j constant? Without X_j, the regression is mathematically identical to a difference of means. One worry about the data, though, is that the counties with punch cards might differ systematically from the counties without punchcards *and* the factors that account for the difference also affect the residual vote. For example, rural counties have more administrative problems and thus higher residual vote rates than urban counties, but urban counties are more likely to use punch card equipment (as an accident of history). We might miss an important effect of punch cards if we ignore county demographics. (Try to show this in a graph.)

The coefficient β_1 measures the difference between punch card and non-punch card counties, holding other factors constant.

1.A. Two Conceptions of Regression

i. The functional relationship of Y to $X_1, X_2, ..., X_k$.

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} ... + \beta_k X_{k,i} + \epsilon_i$$

ii. Conditional Mean

The random variables $Y, X_1, X_2, ..., X_k$ have joint density $f(y, x_1, x_2, ..., x_k)$. The conditional distribution $f(y|x_1, x_2, ..., x_k)$ can be characterized by its mean and conditional variance. We begin with the assumption that the conditional mean is a linear function of the values of the random variables and that the error variance is constant:

$$E[Y|X_1, X_2, ..., X_k] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 ... + \beta_k X_k$$

$$V[Y|X] = \sigma^2_\epsilon$$

Both assumptions are readily relaxed, as we will see later in the course.

As a specific case we may begin with the formula for the joint normal distribution. This is of historical interest, but I also find that it helps make these ideas more explicit. For simplicity I will deal with the case of 2 variables. In fact, I will often work with just 2 variables in this course to develop the basic concept or intuition.

$$f(y,x) = \frac{1}{2\pi\sigma_y\sigma_x\sqrt{(1-\rho^2)}} e^{\frac{-1}{2(1-\rho^2)}[(\frac{y-\mu_y}{\sigma_y})^2 + (\frac{x-\mu_x}{\sigma_x})^2 - 2\rho(\frac{y-\mu_y}{\sigma_y})(\frac{x-\mu_x}{\sigma_x})]}$$

$$f(y|x) = \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{(1-\rho^2)}} e^{\frac{-1}{2\sigma_y^2(1-\rho^2)}[(y-(\mu_y-\beta\mu_x)-\beta x)^2]},$$

where $\beta = \frac{\sigma_{xy}}{\sigma_x^2}$.

The conditional mean and variance are:

$$E[Y|X=x] = \int yf(y|x)dy = (\mu_y - \beta\mu_x) + \beta x$$

$$V[Y|X=x] = \sigma_y^2(1-\rho^2)$$

To develop the core concepts we begin with the bivariate regression model. It is good to get your footing with a simple model and then build up. Here I will develop the basic properties of regression in the bivariate setting.

1.B. Assumptions

1. Linearity
$$y_i = \alpha + \beta_1 X_{1,i} + ...\beta_k X_{k,i} + \epsilon_i$$
2. $E[\epsilon] = 0$.
3. Independence/Exogeneity: $E[\epsilon X] = 0$.
4. No Measurement Error in X.
5. Homoskedasticity: $E[\epsilon_i] = \sigma^2$ for all i.
6. No Autocorrelation: $E[\epsilon_i \epsilon_j] = 0$.

The first four assumptions affect the bias of the estimates. Of particular concern is assumption 3, which means that there are no ommitted variables and the relationship between Y and X is not simultaneous.

The last two assumptions concern efficiency.

1. C. Estimation

Estimation consists of deriving formulae for converting data into guesses about the unknown parameters of the model, i.e., α, β, and σ_ϵ^2.

Method of Moments. We wish to find a and b, parameters of the estimated regression line also known as the *predicted values*: $\hat{y} = a + bx$. Define the residual as the vertical distance of any point y from the estimated regression line: $e_i = y_i - \hat{y_i} = y_i - a - bx_i$. The residual can be thought of as a realization of the error term ϵ.

The second and third assumptions are in fact theoretical moments describing the conditional distribution. Assume that those hold. That is, let us impose those restrictions on the residuals.

$$\frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})(x_i - \bar{x}) = 0$$

Substituting the definition of e_i in each equation and solving for a and b yields:

$$a = \bar{y} - b\bar{x}$$

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

Least Squares. Another way to derive estimates is to minimize the error variance to find the "best fitting" line. Define the sum of squares as:

$$S = \sum_{i=1}^{n} \epsilon^2$$

We wish to find the values of α and β that minimize this function. Those values are determined by the first derivatives. At a minimum both first derivatives of the function with respect to the unknown parameters must equal 0.

$$\frac{\partial S}{\partial \alpha} = 0$$

$$\frac{\partial S}{\partial \beta} = 0$$

After some algebraic manipulations, we find that least squares yields the following estimates:

$$a = \bar{y} - b\bar{x}$$

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

Maximum Likelihood.

1. D. Inference

Once we have taken data and calculated a and b and $\hat{\sigma}_\epsilon^2$, we would like to know how to use the estimates to draw inferences about theoretical conjectures and to make predictions. We will build on the basic concepts of sampling theory.

Sampling Theory. Recall from basic statistics that the average value of a random variable X in any data set or sample is thought of as an estimate or guess about the population mean.

1. The sample mean, $\bar{x}$, is expected to yield an unbiased estimate of the population mean, μ. That is, $E[\bar{x}] = \mu$.
2. From sample to sample or study to study, the mean varies at a rate of σ^2/n, which is the variance of $\bar{x}$. The square root of the variabe of $\bar{x}$ is called the *standard error*.
3. Because the mean is the sum of random variables, the distribution of the mean is normal with mean μ and variance σ^2/n. This is an application of the Central Limit Theorem.
4. As $n \to \infty$, $V(\bar{x}) \to \infty$ and $\bar{x} \to \mu$ in probability. This is called consistency of $\bar{x}$ as an estimate of μ. It is an application of the Law of Large Numbers.

The same ideas apply in regression analysis, which is, afterall, the estimate of a conditional mean $E[Y|X]$.

1.D.i. Properties of Estimates

We wish to know whether estimates are consistent or unbiased. Consistency means that as the number of cases becomes large the estimate becomes extremely close to the true value. This is the application of the Law of Large Numbers from probability theory to our statistical estimation problem. Unbiasedness means that the average value is equal to the true value. Both are statistical definitions of the "correctness" or "accuracy" of the estimator.

We also wish to know whether estimates are efficient – that is, do they use the information available in a way the minimizes the variability or uncertainty about the estimate. Frequetly, in choosing the best statistical specification and model we must choose between a specification that is consistent (or unbiased) but inefficient and one that is potentially inconsistent. We face this choice, for example, when we decide whether to include a variable or not in a regression analysis.

1.D.ii. Confidence Intervals

The degree of confidence that we have in an estimate will be expressed as the standard error, which measures the rate at which the estimates will vary from sample to sample, if we had the luxury of repeated sampling. The variance of the estimator (the square of the standard error) is the weighted average of the squared distance estimator from its mean (the true parameter value).

1.D.iii. Prediction.

Consider any specific value of X, say x_0. We wish to know what value of y is predicted when $X = x_0$. This can be computed simply from the predicted values of the regression line

$$\hat{y}_0 = a + bx_0$$

Predictions are uncertain