17.874 Lecture Notes
Part 5: Sensitivity and Weights

5. Sensitivity and Weights

In practical data analysis, not all observations count the same. For example, unusual observations may have stronger effects on the estimated regression line than typical observations. Also, observations may be correlated with each other, leading us to believe that we have more independent observations than we do. And, some observations may be more variable than other observations, and thus we might wish to weight them less.

These examples violate the assumptions of the regression model. The latter two situations violate the assumptions that the variances of the errors are the same and the covariances are zero. The first situation introduces the possibility that with insufficient data we may encounter problems produced by outliers and other influential cases.

5.1. Generalized Least Squares and the Problems of Heteroskedasticity and Autocorrelation

Let us begin with a simple case. Suppose that the data consist of averages or sums of individual level data to the level of municipalities. The individual level data are:

$$y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + ... + \beta_k X_{kij} + \epsilon_{ij}$$

We observe aggregate measures, say the average value, summing over individuals i:

$$y_{.j} = \beta_0 + \beta_1 X_{1.j} + \beta_2 X_{2.j} + ... + \beta_k X_{k.j} + \epsilon_{.j},$$

where the subscript . reminds us that we have averaged over individuals.

If the municipalities vary greatly in size, the error variance will no longer be constant. $V[\epsilon_{.j}] = V[\frac{1}{n_j}\sum_i^{n_j} \epsilon_{ij}] = \frac{\sigma_\epsilon^2}{n_j}$. This violates the basic regression assumptions, as error variances will vary with n_j.

5.1.1. Generalized Linear Model

It is possible to relax the assumptions of homoskedasticity and no autocorrelation. The generalized linear regression model may be expressed as follows:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{1}$$

$$E[\epsilon|\mathbf{X}] = \mathbf{0} \tag{2}$$

$$E[\epsilon\epsilon'|\mathbf{X}] = \sigma^2 \quad = \mathbf{\Sigma} \tag{3}$$

where is a matrix, possibly with varying values on the diagonal and non-zero terms in the off-diagonals.

In the example above,

$$= \begin{pmatrix} \frac{1}{n_1}, 0, 0, 0, ..., 0 \\ 0, \frac{1}{n_2}, 0, 0, ..., 0 \\ 0, 0, \frac{1}{n_3} 0, ..., 0 \\ ... \\ 0, 0, 0, 0, ..., \frac{1}{n_1} \end{pmatrix}$$

Heteroskedasticity, generally is of the form:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2, 0, 0, 0, ..., 0 \\ 0, \sigma_2^2, 0, 0, ..., 0 \\ 0, 0, \sigma_3^2, 0, ..., 0 \\ ... \\ 0, 0, 0, 0, ..., \sigma_n^2 \end{pmatrix}$$

Autocorrelation takes the general form:

$$\mathbf{\Sigma} = \sigma^2 \begin{pmatrix} 1, \rho_1, \rho_2, \rho_3, ..., \rho_{n-1} \\ \rho_1, 1, \rho_1, \rho_2, ..., \rho_{n-2} \\ \rho_2, \rho_1, 1, \rho_1, ..., \rho_{n-3} \\ ... \\ \rho_1, \rho_2, \rho_3, \rho_4, ..., 1 \end{pmatrix}$$

5.1.2. Exact Nature of the Problem with Ordinary Least Squares

Ordinary Least Squares will not be biased under the generalized linear model, but it will be inefficient, possibly leading to incorrect inferences.

If the generalized linear model is correct,

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

Two caveats to this derivation arises with dynamic models. (1) The least squares estimates may be biased if lagged values of y are right-hand side variables. (2) For some autocorrelation structures, the least squares estimates may be inconsistent. For instance, inconsistency arises when the autocorrelation is not decreasing as one gets closer to the current value.

Inefficiencies will arise with OLS.

$$V[b|X] = E[(b-\beta)(b-\beta)'] = E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}] = \sigma^2(X'X)^{-1}X'\text{-}\ X(X'X)^{-1}]$$

Hence, $V[b|X] \neq \sigma^2(X'X)^{-1}$.

The estimated error variance will also be off.

A simple test for autocorrelation is the Durbin-Watson test

$$d = \frac{\sum(e_t - e_{t-1})^2}{\sum e_t^2} = 2(1-r) - \frac{e_1^2 + e_T^2}{\sum e_t^2} \approx 2(1-r).$$

Is this statistic close to 2?

5.1.3. Estimation

<u>Generalized Least Squares</u>

While the generalized model expresses the variance-covariance of the error structure quite generically, it is not possible to estimate this model because there are 2n + K parameters and only n observations. However, when we encounter a specific problem, such as the averaging problem above, or if we do no care about the estimate of each specific variance-covariance parameter, solutions are readily available.

One approach to estimation is Generalized Least Squares. We wish to characterize the matrix - , which can be thought of as deviations from spherically distributed errors. We may write any square matrix as the product of two matrices: - $= P'P$. We use this decomposition of - to develop a scheme for weighting observations.

Let us suppose that there exists such a matrix P, and that we transform all of our variables by weighting them by P as follows:

$$Py = PX\beta + P\epsilon.$$

Then perform least squares on this rescaled data.

$$b_* = (X'P'PX)^{-1}(X'P'Py)]$$

This estimator yields unbiased estimates that are efficient.

$$E[b_*] = E[(X'P'PX)^{-1}(X'P'(PX\beta + P\epsilon))] = \beta + E[(X'\text{-}\ X)^{-1}(X'\text{-}\ \epsilon))] = \beta$$

$$V[b_*] = E[(X'\text{-}\ X)^{-1}(X'\text{-}\ \epsilon\epsilon'\text{-}\ X)(X'\text{-}\ X)^{-1})]$$

$$= \sigma^2 E[(X'\text{-}\ X)^{-1}(X'\text{-}\ \text{-}\ \text{-}\ X)(X'\text{-}\ X)^{-1})] = \sigma^2(X'\text{-}\ X)^{-1} = \sigma^2(X_*'X_*)^{-1}$$

To correct for autocorrelation, we stipulate a specific structure and ascertain what that structure implies about the matrix - . Assume that we have an AR-1 error structure. This reduces the problem to a single additional parameter, ρ, in stead of n additional parameters.

$$\mathbf{\Sigma} = \sigma^2 \begin{pmatrix} 1, \rho, \rho^2, \rho^3, ..., \rho^{n-1} \\ \rho^1, 1, \rho^1, \rho^2, ..., \rho^{n-2} \\ \rho^2, \rho^1, 1, \rho^1, ..., \rho^{n-3} \\ ... \\ \rho^{n-1}, \rho^{n-2}, \rho^{n-3}, ..., \rho, 1 \end{pmatrix}$$

An appropriate matrix P is

$$\mathbf{\Sigma} = \sigma^2 \begin{pmatrix} \sqrt{1-\rho}, 0, 0, 0, ..., 0 \\ -\rho, 1, 0, 0, ..., 0 \\ 0, -\rho, 1, 0, ..., 0 \\ ... \\ 0, 0, 0, ..., -\rho, 1 \end{pmatrix}$$

Example. WalMart Stocks

This procedure suggests an immediate solution to the weighting problem above in which municipalities have very different populations. Let $P = [\frac{1}{\sqrt{n_j}}]$. Thus, bigger cases receive more weight and the amount of weight grows in squareroot of the population. This transformation also converts the unit of observation from the "typical municipality" to the "typical person."

Example. State Legislative Representation and the Distribution of Public Expenditures.

Analytical Weights

Analytical weights calculate the percent of the total weight attributable to a specific observation and reweight the data. This is a somewhat different approach that often has similar effects on the data.

Consistent Standard Errors

5.2. Messy Data

5.2.1. Outliers and Influential Observations

Reestimate the regression with and without each observation. How much change occurs? The observations whose deletion produces the largest change in the coefficients is referred to as the most influential observation.

Sometimes this reflects a problem of weighting. See handout.

There are many approaches to dealing with these cases. Omit them. Include a dummy variable. "Weight" them. Use heteroskedastic consistent standard errors.

5.2.2. Quantile Regression

Heteroskedasticity sometimes reflects a subtantively meaningful problem: we may care about the effects of X on the distribution of Y, not only the mean of Y.

Median regression consists of minimizing

$$LAD = \sum_i |y_i - \mathbf{x_i'beta}|$$

More generally we can minimize

$$LQD = \sum_i \rho_q(y_i - \mathbf{x_i'beta})$$

, where $\rho_q()$ is the absolute value "tilted" to yield the qth sample quantile. Huber (1977) shows that this can be further simplified to be expressed as a weighted LAD estimator. Define the weight $h_i = 2q$ if $e_i > 0$ and $h_i = w(1-q)$ if $e_i < 0$.

$$LQD = \sum_i h_i |y_i - \mathbf{x_i'beta}|$$

Thus, if we wish to estimate the 75th percentile, we weight the negiatve residuals by .5 and the postive residuals by 1.50.

5.2.2. Estimators

We may get a good first approximation by doing weighted least squares, as discussed last lecture, in which the weights are h_i. We can improve this with iterative fitting along the lines of MLE in which we minimize the sum of absolute deviations.