ECN 215 – Some key terms in regression analysis

Estimator A formula or algorithm for generating estimates of parameters, given relevant data. For example, the Ordinary Least Squares estimator of the parameter vector, β, in the multiple regression model $y = X\beta + u$, is

$$\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y$$

Bias An estimate is unbiased if its expectation equals the value of the parameter being estimated, otherwise it is biased.

Efficiency An estimator A is more efficient than an estimator B if A has a smaller sampling variance—that is, if the particular values generated by A are more tightly clustered around their expectation.

Consistency An estimator is consistent if the estimates it produces "converge on" the true parameter value as the sample size increases without limit. Consider an estimator that produces estimates, $\hat{\theta}$, of some parameter θ, and let ϵ denote a small number. If the estimator is consistent, we can make the probability

$$P(|\theta - \hat{\theta}| < \epsilon)$$

as close to 1.0 as we like, for ϵ as small as we like, by drawing a sufficiently large sample. Note that a biased estimator may nonetheless be consistent—this occurs if the bias tends to zero in the limit. Conversely, an unbiased estimator may be inconsistent, if its sampling variance fails to shrink appropriately as the sample size increases.

Sum of Squared Residuals (SSR) The sum of the squared differences between the actual values of the dependent variable and the fitted or predicted values from a regression equation, $\sum(y_i - \hat{y}_i)^2$. This quantity is minimized by the OLS estimator.

Standard Error of the Regression (SER) An estimate of the standard deviation of the error term in a regression model. It is computed as

$$\sqrt{\frac{\text{SSR}}{n-k-1}}$$

where n is the number of observations and k is the number of independent variables, excluding the constant or intercept. (The extra -1 in the denominator reflects the assumption that an intercept is also included.)

R-squared A standardized measure of the goodness of fit for a regression model. Computed as

$$R^2 = 1 - \frac{\text{SSR}}{\text{TSS}}$$

where TSS denotes the "total sum of squares" for the dependent variable, $\sum(y_i - \bar{y})^2$. This statistic measures the proportion of the variation in the dependent variable that is "accounted for" by the regression. When comparing models with differing numbers of regressors, the adjusted R-squared is often used:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{\text{SSR}}{\text{TSS}}$$

Standard error of regression coefficient An estimate of the standard deviation of the sampling distribution for the coefficient in question. This may be used to generate a measure of our uncertainty over the true value of the corresponding parameter. For example an approximate 95 percent confidence interval for a coefficient is given (for a large enough sample) as the measured coefficient plus or minus 2 standard errors. (Note that the well-known expression $s/\sqrt{n}$ gives the standard error of the sample mean, $\bar{x}$. This is *not to be confused* with the standard error for a regression coefficient—same general concept, but different formula altogether.)

P-value The probability, supposing the null hypothesis to be true, of drawing sample data that are as adverse to the null as the data actually drawn, or more so. When a small p-value is found the two possibilities are (a) that we happened to draw a low-probability, unrepresentative sample or (b) that the null hypothesis is in fact false.

Significance level For a hypothesis test, the smallest p-value for which we will *not* reject the null hypothesis. If we choose a significance level of 1 percent, we're saying that we'll reject the null if and only if the p-value for the test is less than 0.01. The significance level is also the probability of making a Type 1 error (that is, rejecting a true null hypothesis).

t-test The t test (or z test, which is the same thing asymptotically) is a common test for the null hypothesis that a particular regression parameter, β_i, has some specific value (commonly zero, but generically β_{H_0}). The test statistic is computed as

$$\frac{\hat{\beta}_i - \beta_{H_0}}{\hat{\sigma}_{\hat{\beta}_i}}$$

where $\hat{\sigma}_{\hat{\beta}_i}$ denotes the estimated standard error of $\hat{\beta}_i$.

F-test A common procedure for testing jointly a set of linear restrictions on a regression model. In the context of OLS, the F statistic can be computed as

$$F(v_1, v_2) = \frac{(\text{SSR}_R - \text{SSR}_U)/v_1}{\text{SSR}_U/v_2}$$

where SSR_U and SSR_R are the sums of squared residuals for the unrestricted and restricted models, respectively; v_1 is the number of restrictions; and v_2 is the degrees of freedom for the unrestricted model. The last term is calculated as the number of observations minus the number of parameters in the unrestricted model.

One particular variant of the F test is commonly presented as part of the standard output for a regression equation, namely, the test for the null hypothesis that all the "slope" parameters in the model equal zero.

Multicollinearity A situation where there is a high degree of correlation among the independent variables in a regression model—or more generally, where some of the xs are close to being linear combinations of other xs. Symptom: large standard errors, inability to produce precise parameter estimates. This is not a serious problem if one is primarily interested in forecasting; it is a problem is one is trying to estimate causal influences.

Omitted variable bias Bias in the estimation of regression parameters that arises when a relevant independent variable is omitted from a model, and the omitted variable is correlated with one or more of the included variables.

Log variables A common transformation, which permits the estimation of a nonlinear model using OLS, is to substitute the natural log of a variable for the level of that variable. This can be done for the dependent variable and/or one or more of the independent variables. A key point to remember about logs is that, for small changes, the change in the log of a variable is a good approximation to the *proportional* change in the variable itself. For example, if $\log(y)$ changes by 0.04 this means that y changes by about 4 percent.

Quadratic terms Another common transformation. When both x_i and x_i^2 are included as regressors, it is important to remember that the estimated effect of x_i on y is given by the derivative of the regression equation with respect to x_i. If the coefficient on x_i is β and the coefficient on x_i^2 is γ, the derivative is $\beta + 2\gamma x_i$.

Interaction terms Pairwise products of the "original" independent variables. The inclusion of interaction terms in a regression allows for the possibility that the degree to which x_i affects y depends on the value of some other variable x_j—or in other words, x_j modulates the effect of x_i on y. For example, the effect of experience (x_i) on wages (y) might depend on the gender (x_j) of the worker.