Section 5 Inference in the Multiple-Regression Model

Kinds of hypothesis tests in a multiple regression

There are several distinct kinds of hypothesis tests we can run in a multiple regression. Suppose that among the regressors in a Reed Econ 201 grade regression are variables for SAT-math and SAT-verbal: $g_i = \beta_1 + \beta_2 SATM_i + \beta_3 SATV_i + \ldots + e_i$

- We might want to know if math SAT matters for Econ 201: $H_0 : \beta_2 = 0$
    - Would it make sense for this to be a one-tailed or a two-tailed test?
    - Is it plausible that a higher math SAT would *lower* Econ 201 performance?
    - Probably a one-tailed test makes more sense.
- We might want to know if either SAT matters. This is a joint test of two simultaneous hypotheses: $H_0 : \beta_2 = 0, \quad \beta_3 = 0.$
    - The alternative hypothesis is that *one or both* parts of the null hypothesis fails to hold. If $\beta_2 = 0$ but $\beta_3 \neq 0$, then the null is false and we want to reject it.
    - The joint test is *not* the same as separate individual tests on the two coefficients.
        - In general, the two variables are correlated, which means that their coefficient estimators are correlated. That means that eliminating one of the variables from the equation affects the significance of the other.
        - The joint test tests whether we can delete *both* variables at once, rather than testing whether we can delete one variable given that the other is in (or out of) the equation.
        - A common example is the situation where the two variables are highly and positively correlated (imperfect but high multicollinearity).
            - In this case, OLS may be able to discern that the two variables are collectively very important, but not which variable it is that is important.
            - Thus, individual tests of $\beta_j = 0$ may not be rejected. (OLS cannot tell for sure that either coefficient is non-zero.) However, the joint test would be strongly rejected.
            - Here, the strong positive correlation between the variables leads to a strong negative correlation between the coefficient estimators. Assuming that the joint effect is positive, then leaving one coefficient out (setting it to zero and therefore decreasing it) increases the value of the other.
    - In the case of joint hypotheses, we always use two-tailed tests.

- We might also want to know if the effect of the two scores is the same. The null hypothesis in this case is $H_0 : \beta_2 = \beta_3$ against a two-tailed alternative. Note that if this null hypothesis is true, then the model can be written as $g_i = \beta_1 + \beta_2 (SATM_i + SATV)_i + \ldots + e_i$ and we can use the SAT composite rather than the two separate scores, saving one degree of freedom.

Hypothesis tests on a single coefficient

- Hypothesis testing for a single coefficient is identical to the bivariate regression case:
  - $t^{act} = \dfrac{b_j - c}{s.e.(b_j)}$ is the test statistic
  - It is asymptotically $N(0, 1)$ under assumptions MR1–MR5.
  - It is distributed as t with $N - K$ degrees of freedom if e is normal.
  - Two-tailed test: reject the null of $\beta_j = c$ if p-value = $2\Phi(-|t^{act}|) < \alpha$, the chosen level of significance (using asymptotic normal distribution) or reject if $|t^{act}| > |t_{\alpha/2}|$ (using small-sample distribution under normality assumption).
    - Note that Stata uses t distribution to calculate p values, not normal.
    - Which is better?
      - Both are flawed in small samples
      - Normal is off because sample is not large enough for convergence to have occurred.
      - t is off because if true distribution of e is not normal, then don't know the small-sample distribution
      - ($t \rightarrow$ normal as sample gets large)
- Single-coefficient confidence intervals are also identical to the bivariate case:
  - Using the normal asymptotic (normal) distribution,
    $$\Pr\left[\beta_j \in \left(b_j - \Phi^{-1}\left(-\frac{\alpha}{2}\right) \cdot s.e.(b_j),\quad b_j + \Phi^{-1}\left(-\frac{\alpha}{2}\right) \cdot s.e.(b_j)\right)\right] = \alpha.$$
  - If we use the t distribution, all we change is drawing the critical value from the t distribution rather than the normal.
  - Again, Stata uses classical standard errors and the t distribution by default.

Simple hypotheses involving multiple coefficients

- Suppose that we want to test the hypothesis $\beta_2 = \beta_3$, or $\beta_2 - \beta_3 = 0$.
  - We can use a t test for this.
  - The estimator of $\beta_2 - \beta_3$ is $b_2 - b_3$, which has variance of
    $$\operatorname{var}(b_2 - b_3) = \operatorname{var}(b_2) + \operatorname{var}(b_3) - 2\operatorname{cov}(b_2, b_3).$$

- The standard error of $b_2 - b_3$ is the square root of the estimated variance, which can be calculated from the estimated covariance matrix of the coefficient vector.
- The test statistic is $t = \frac{(b_2 - b_3) - 0}{s.e.(b_2 - b_3)}$.
- If has the usual distributions, either t_{N-K} or (asymptotically) standard normal.

Testing joint hypotheses

It is often useful to test joint hypotheses together. This differs from independent tests of the coefficients. An example of this is the joint test that math and verbal SAT scores have no effect on Econ 201 grades against the alternative that one or both scores has an effect.

- **Some new probability distributions.** Tests of joint hypotheses have test statistics that are distributed according to either the F or χ^2 distributions. These tests are often called Wald tests and may be quoted either as F or as χ^2 statistics. (The F converges to a χ^2 asymptotically, so the χ^2 is more often used for asymptotic cases and the F—under the right assumptions—for small samples.)
  - Just as the t distribution varies with the number of degrees of freedom: t_{N-K}, the F distribution has *two* degree of freedom parameters, one the number of restrictions being tested (J) and one the number of degrees of freedom in the unrestricted model ($N - K$). The former is often called the "numerator degrees of freedom" and the latter the "denominator degrees of freedom" for reasons we shall see soon.
  - When there is only one numerator degree of freedom, we are testing only a single hypothesis and it seems like this should be equivalent to the usual t test. Indeed, if a random variable t follows the t_{N-K} distribution, then its square t^2 follows the $F_{(1,N-K)}$ distribution.
  - Since squaring the t statistic obliterates its sign, we lose the option of the one-tailed test when using the F distribution.
  - Similarly, if z follows a standard normal distribution, then z^2 follows a χ^2 distribution with one degree of freedom.
  - Finally, as the number of denominator degrees of freedom goes to infinity, if a random variable F follows the $F_{(J, N-K)}$ distribution, then JF converges in distribution to a χ^2 with J degrees of freedom.
  - Both the F and χ^2 distributions assign positive probability only to positive values. (Both involve squared values.)
    - Both are humped with long tails on the right, which is where our rejection region lies.
    - The mean of the F distribution is always 1.
    - The mean of the χ^2 distribution is J, the number of degrees of freedom.
- **General case in matrix notation**

- Suppose that there are J linear restrictions in the joint null hypothesis. These can be written as a system of linear equations $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, where $\mathbf{R}$ is a $J \times K$ matrix and $\mathbf{r}$ is a $J \times 1$ vector. Each restriction is expressed in one row of this system of equations. For example, the two restrictions $\beta_2 = 0$ and $\beta_3 = 0$ would be expressed in this general matrix notation as

$$\begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \vdots \\ \beta_K \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

- The test statistic is $F = \frac{1}{J}(\mathbf{Rb} - \mathbf{r})'\left(\mathbf{R}\hat{\boldsymbol{\Sigma}}_{\mathbf{b}}\mathbf{R}'\right)^{-1}(\mathbf{Rb} - \mathbf{r})$, with $\hat{\boldsymbol{\Sigma}}_{\mathbf{b}}$ equal to the estimated covariance matrix of the coefficient vector. Under the OLS assumptions MR1-MR6, this is distributed as an $F_{(J,\infty)}$. Multiplying the test statistic by J (eliminating the fraction in front) gives a variable that is asymptotically distributed as χ^2_J, so the Wald test can be done either way.
- If the restrictions implied by the null hypothesis are perfectly consistent with the data, then the model fits equally well with and without the restrictions, $\mathbf{Rb} - \mathbf{r} = \mathbf{0}$ holds exactly, and the F statistic is zero. This, obviously, implies acceptance of the null.
  - We reject the null when the (always positive) F statistic is larger than the critical value.
  - The Stata test command gives you a p value, which is the smallest significance level at which you can reject the null.
  - The same rejection conditions apply if the χ^2 distribution is used: reject if the test statistic exceeds the critical value (or if the p value is less than the level of significance).

- **Alternative calculation of *F* under classical assumptions**
  - If the classical homoskedastic-error assumption holds, then we can calculate the F statistic by another equivalent formula that has intuitive appeal. To do this, we run the regression with and without the restrictions (for example, leaving out variables whose coefficients are zero under the restrictions in the restricted regression). Then we calculate F as

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_{RU}/(N-K)}$$

    - This shows why we think of J as "numerator" degrees of freedom and ($N - K$) as the "denominator" degrees of freedom.

        - The numerator in the numerator is the difference between the SSE when the restrictions are imposed and the SSE when the equation is unrestricted.
            - The numerator is always non-negative because the unrestricted model always fits at least as well as the restricted one.
            - This difference is large if the restrictions make a big difference and small if they don't. Thus, other things equal, we will have a larger *F* statistic if the equation fits much less well when the restrictions are imposed.
    - This *F* statistic (which is the same as the one from the matrix formula as long as $\hat{\boldsymbol{\Sigma}}_{\mathbf{b}} = s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$) follows the $F_{(J,\, N-K)}$ distribution under classical assumptions.
    - By default, the test command in Stata uses the classical covariance matrix and in either case uses the $F_{(J,\, N-K)}$ distribution rather than the $F_{(J,\, \infty)}$ or the χ^2_J to compute the *p* value.
- **"Regression *F* statistic"**
    - A common joint significance test is the test that all coefficients except the intercept are zero: $H_0 : \beta_2 = \beta_3 = \ldots = \beta_K = 0$
    - This is the "regression *F* statistic" and it printed out by many regression packages (including Stata).
    - In bivariate regression, this is the square of the *t* statistic on the slope coefficient.
    - If you can't reject the null hypothesis that all of your regressors have zero effect, then you probably have a pretty weak regression!

Simple hypotheses involving multiple coefficients by alternative methods

- The matrix formula $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ clearly includes the possibility of:
    - Single rather than multiple restrictions, and
    - Restrictions involving more than one coefficient.
    - For example, to test $H_0 : \beta_2 = \beta_3$, we could use $\begin{pmatrix} 0 & 1 & -1 & \ldots & 0 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{pmatrix} = 0.$
    - This is how Stata does such tests and is a perfectly valid way of doing them.
- An alternative way to test such a simple linear hypothesis is to transform the model into one in which the test of interest is a zero-test of a single coefficient, which will then be printed out by Stata directly.

- For the SAT example, the restricted case is one in which only the sum (composite) of the SAT scores matters. Let $SATC \equiv SATM + SATV$. Then the model is

$$\begin{aligned} g_i &= \beta_1 + \beta_2 SATM_i + \beta_3 SATV_i + \ldots + e_i \\ &= \beta_1 + \beta_2 (SATM_i + SATV_i) + (\beta_3 - \beta_2) SATV_i + \ldots + e_i \\ &= \beta_1 + \beta_2 (SATC_i) + (\beta_3 - \beta_2) SATV_i + \ldots + e_i. \end{aligned}$$

  - Thus, we can regress g_i on $SATC$, and $SATV$ and test the hypothesis that the coefficient on $SATV$ equals zero. This null hypothesis is H_0: $\beta_3 - \beta_2 = 0$, which is equivalent to $\beta_3 = \beta_2$.
  - This alternative method gives us a t statistic that is exactly the square root of the F statistic that we get by the matrix method, and should have exactly the same test result.
  - We can *always* reformulate the model in a way that allows us to do simple tests of linear combinations of coefficients this way. (This allows us to use the standard t test printed out by Stata rather than using the test command.)
  - Again, we can use either the classical covariance matrix or the robust one. Stata will use the classical one unless the robust option is specified.
- This method can be used to calculate **restricted least-squares estimates** that impose the chosen restrictions.

Some χ^2 alternative tests

There are several tests that are often used as alternatives to the F test, especially for extended applications that are not LS. Sometimes these are more convenient to calculate; sometimes they are more appropriate given the assumptions of the model.

- **Lagrange multiplier test**
  - The Lagrange multiplier test is one that can be easier to compute than the F test. It does not require the estimation of the complete unrestricted model, so it's useful in cases where the unrestricted model is very large or difficult to estimate.
  - Recall that the effects of any omitted variables will be absorbed into the residual (or into the effects of correlated included regressors).
    - Thus it makes sense to test whether an omitted variable should be added by asking whether it is correlated with the residual of the regression from which it has been omitted.
  - Suppose that we have K regressors, of which we want to test whether the last J coefficients are jointly zero:

$$\begin{aligned} y_i &= \beta_1 + \beta_2 X_{i,2} + \ldots + \beta_{K-J} X_{i,K-J} + \beta_{K-J+1} X_{i,K-J+1} + \ldots + \beta_K X_{i,K} + e_i \\ H_0 &: \quad \beta_{K-J+1} = \beta_{K-J+2} = \ldots = \beta_K = 0. \end{aligned}$$

- For the LM test, we regress y on the first $K - J$ regressors, then regress the residuals from that regression on the last J regressors.
- NR^2 from the latter regression is asymptotically distributed as a χ^2 statistic with J degrees of freedom.

- **Likelihood-ratio test**
  - In maximum-likelihood estimation, the likelihood-ratio test is the predominant test used.
  - If L_u is the maximized value of the likelihood function when there are no restrictions and L_r is the maximized value when the restrictions are imposed, then $2(\ln L_u - \ln L_r)$ is asymptotically distributed as a χ^2 statistic with J degrees of freedom.
  - Most maximum-likelihood based procedures (such as logit, probit, etc.) report the likelihood function in the output, so computing the LR test is very easy: just read the numbers off the restricted and unrestricted estimation outputs and multiply the difference by two.

Multivariate confidence sets

- Multivariate confidence sets are the multivariate equivalent of confidence intervals for coefficients. For two variables, they are generally ellipse-shaped.
- As with confidence intervals, if the confidence set of two variables excludes the origin, we reject the joint null hypothesis that the two coefficients are jointly zero. Moreover, we reject the joint null hypothesis that the two coefficients equal any point in the space the is outside the confidence set.
- There doesn't seem to be a way of doing these in Stata.

Goodness of fit

- Standard error of the regression is similar to bivariate case, but with $N - K$ degrees of freedom.
  - There are N pieces of information in the dataset. We use K of them to minimally define the regression function (estimate the K coefficients). There are $N - K$ degrees of freedom left.
  - $SER = s_{\hat{e}} = \sqrt{s_{\hat{e}}^2} = \sqrt{\frac{1}{N-K}\sum_{i=1}^{N}\hat{e}_i^2} = \sqrt{\frac{SSE}{N-K}}.$
- R^2 is defined the same way: the share of variance in y that is explained by the set of explanatory variables:

- $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$.
- However, adding a new regressor to the equation *always* improves R^2 (unless it is totally uncorrelated with the previous residuals), so we would expect an equation with 10 regressors to have a higher R^2 than one with only 2. To correct for this, we often use an **adjusted R^2** that corrects for the number of degrees of freedom:

$$\bar{R}^2 = 1 - \frac{N-1}{N-K}\frac{SSE}{SST} = 1 - \frac{\frac{1}{N-K}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2} = 1 - \frac{s_{\hat{e}}^2}{s_y^2}.$$

- Three properties of $\bar{R}^2$:
  - $\bar{R}^2 < R^2$ whenever $K > 1$.
  - Adding a regressor generally decreases *SSE*, but also increases *K*, so the effect on $\bar{R}^2$ is ambiguous. Choosing a regression to maximize $\bar{R}^2$ is not recommended, but it's better than maximizing R^2.
  - $\bar{R}^2$ can be negative if *SSE* is close to *SST*, because $\frac{N-1}{N-K} > 1$.

Some specification issues

In practice, we never know the exact specification of the model: what variables should be included and what functional form should be used. Thus, we almost always end up trying multiple alternative models and choosing among them based on the results.

- **Specification search is very dangerous!**
  - If you try 20 independent variables that are totally uncorrelated with one another and with the dependent variable, on average one (5%) will have a statistically significant *t* statistic.
  - The maximum of several candidate *t* statistics *does not* follow the *t* or normal distribution. If you searched five variables and found one that had an apparently significant *t*, you cannot conclude that it truly has an effect.
  - This process is called data mining or specification searching. Though we all do it, it is *very* dangerous and inconsistent with a basic assumption of econometrics, which is that we know the model specification before we approach the data.
  - We shall have more to say about this later in the course.
- Interpreting R^2 and $\bar{R}^2$
  - Adding any variable to the regression that has a non-zero estimated coefficient increases R^2.

- Adding any variable to the regression that has a *t* statistic greater than one in absolute value increases $\bar{R}^2$.
  - Given that the conventional levels of significance suggest critical values much bigger than one, adopting a max $\bar{R}^2$ criterion would lead us to keep many regressions for which we can't reject the null hypothesis that their effect is zero.
- R^2 tells us nothing about causality; it is strictly a correlation-based statistic.
- One cannot infer from a high R^2 that there are no omitted variables or that the regression is a good one.
- One cannot infer from a low R^2 that one has a poor regression or that one has omitted relevant variables.

- Including irrelevant variables vs. omitting relevant ones
  - If we include an irrelevant variable that doesn't need to be in the regression, the expected value of its coefficient is zero.
    - In this case, our regression estimator is inefficient because we are "spending a degree of freedom" on estimating an unnecessary parameter.
    - However, the estimators of the other coefficients are still unbiased and consistent.
  - If we omit a variable that belongs in the regression, the estimators for the coefficients of any variables correlated with the omitted variable are biased and inconsistent.
  - This asymmetry suggests erring on the side of including irrelevant variables rather than omitting important ones, especially if the sample is large enough that degrees of freedom are not scarce.
- **Information criteria**
  - These are statistics measuring the amount of information captured in a set of regressors.
  - Two are commonly used:
    - **Akaike information criterion**
      - $AIC = \ln\left(\frac{SSE}{N}\right) + \frac{2K}{N}$
    - **Schwartz criterion (Bayesian information criterion)**
      - $SC = \ln\left(\frac{SSE}{N}\right) + \frac{K\ln(N)}{N}$
  - In both cases, we choose a regression (among nested sets) that minimizes the criterion.
  - Both give a penalty to higher *K* given *N* and SSE. (Schwartz more so.)
- **RESET test**
  - One handy test that can indicate misspecification (especially nonlinearities among the variables in the regression) is the RESET test.

- To use the RESET test, first run the linear regression, then re-run the regression with squares (and perhaps cubes) of the predicted values from the first regression and test the added term(s).
- Powers of the predicted value will contain powers and cross-products of the *x* variables, so it may be an easy way of testing whether higher powers of some of the *x* variables belong in the equation.

Multicollinearity

- If one of the *x* variables is highly correlated with a linear combination of others, then the **X′X** matrix will be nearly singular and its inverse will tend to "explode."
- It is important to realize that near-multicollinearity is *not* a violation of the OLS assumptions.
- If **X′X** is nearly singular, then the diagonal elements are "small" relative to the off-diagonal elements.
  - Remember that the diagonal elements are proportional to sample variances of the *x* variables and the off-diagonal elements are covariances. If the correlations among the *x* variables are high, then the covariances are large relative to variances.
  - If **X′X** is "near zero," then its inverse will be "very large." The variances of the regression coefficients are proportional to the diagonal elements of this matrix, so near-perfect multicollinearity leads to ***very imprecise estimators***.
  - This makes sense: if two regressors are highly correlated with each other, then the OLS algorithm won't be able to figure out which one is affecting *y*.
- Symptoms
  - Low *t* statistics but a high regression *F* statistic implies that coefficients are collectively, but not individually, significantly different from zero
  - Could have high *F* statistic on a few variables jointly but not individually: something affects *y* but can't tell which one.
- **Variance-inflation factor**
  - Measure of how unreliable a coefficient estimate is
  - $\text{var}(b_j) = \dfrac{\sigma^2}{(N-1)s_j^2}\dfrac{1}{1-R_j^2}$, where

    $s_j^2 = \widehat{\text{var}}(X_j)$, R_j^2 is from reg of X_j on other X
  - $VIF_j = \dfrac{1}{1-R_j^2}$.
  - Can do manually, or download vif and install command from Stata Web site
  - VIF > 10 (5) means that 90% (80%) of variance of X_j is explained by remaining X variables.

        - These are commonly cited thresholds for worrying about multicollinearity.
- What to do about multicollinearity?
    - Get better data in which the two regressors vary independently.
    - If no additional data are available, one variable might have to be dropped, or can report the (accurate) results of the regression.