EC771: Econometrics, Spring 2010

Greene, Econometric Analysis (6th ed, 2008)

Chapters 6, 7, 8: Inference, Prediction, Functional Form, Structural Change, Specification Error

Inference in the regression model

We may readily test hypotheses which involve "nested models": that is, the comparison of a general model with a special case of that model, in which one or more hypotheses on the coefficient vector are assumed to hold with certainty. If we work with purely linear restrictions on β, we may write the problem as:

$$y = X\beta + \epsilon,$$

subject to $R\beta = q$. The matrix R, with J rows, must be of full row rank, implying that $J \leq K$.

Realistically, we should also rule out the case of equality, since in that case $\beta = R^{-1}q$, and no estimation is necessary.

We can follow two approaches with this constrained system: first, we may estimate the unconstrained model, and test the degree to which the constraints would be satisfied by the data. Second, we may estimate the model subject to constraints, which must involve a loss of fit, and evaluate how serious that loss might be in relation to the unconstrained alternative.

For the first approach, consider testing the null hypothesis $R\beta = q$ versus the alternative of inequality. Several special cases:

- ANOVA "F": $R = [0 : I]$ contains a (K-1) order identity matrix with a zero vector for the constant term, with q as a (K-1) row

null vector. The null hypothesis is that all slopes are jointly zero.

- A single coefficient is zero (or a particular value): R is a zero vector with a single 1 corresponding to that coefficient's position in the β vector, $q = 0$ (or the particular value).

- Two of the coefficients are equal: R is a zero vector with 1 and -1 in positions corresponding to those coefficients' positions in the β vector, $q = 0$.

- The ratio of two coefficients is a: R is a zero vector with 1 and a in appropriate locations, and $q = 0$.

- Several coefficients sum to 1: R is a zero vector with 1s in positions corresponding to those coefficients, and $q = 1$.

- A set of coefficients are jointly zero: $R = [I : 0]$ where the first J coefficients are assumed to equal zero, and $q = 0$.

More complicated sets of linear restrictions on β may be expressed in terms of R. In any case, the restrictions must be linearly independent, so that R is of full row rank.

Given the least squares estimator of the unconstrained model, b, we are interested in the discrepancy vector $m = Rb - q$. Although it is unlikely that $m = 0$, how far from the null vector might m be (in terms of its norm) before we will conclude that the null hypothesis

should be rejected? Under that null hypothesis, m has a mean vector 0 and a covariance matrix

$$\text{Var}[Rb-q|X] = R\,\text{Var}[b|X]R' = \sigma^2 R(X'X)^{-1}R'.$$

We can thus base a test of H_0 on the Wald statistic

$$W = m'(\text{Var}[m|X])^{-1}m$$

$$W = \sigma^{-2}(Rb-q)'[R(X'X)^{-1}R']^{-1}(Rb-q)$$

which under the null hypothesis will have a $\chi^2(J)$ distribution. To make this test statistic operational, we must replace the unknown σ^2 with a consistent estimate thereof, s^2. Most commonly, we express this test statistic as an F–statistic,

$$\widehat{W} = J^{-1}(Rb-q)'[R(s^2X'X)^{-1}R']^{-1}(Rb-q)$$

which under the null hypothesis will have J and $(n-K)$ degrees of freedom. If $J = 1$,

then this F–statistic is just the square of a t–statistic with $(N-K)$ degrees of freedom. Indeed, any of the hypotheses listed above that involve a single row in R ($J=1$) are commonly expressed as t–tests, no matter how many coefficients may be involved. In contrast, tests involving $J>1$ are joint tests, and can only be expressed as an F–statistic (or χ^2 statistic). Any of these subset hypothesis tests may be performed in Stata via the `test` command. To compute a joint test, give each test to be performed as a parenthesized expression on the `test` command line. The `testparm` command is also useful in testing sets of exclusion restrictions, particularly as it (unlike `test`) supports wildcards.

The restricted least squares estimator

Alternatively, we might estimate b subject to the restrictions $Rb=q$. This may be done as the Lagrangean expression

$$L(b_0,\lambda)=(y-Xb_0)'(y-Xb_0)+2\lambda'(Rb_0-q).$$

The necessary conditions are

$$\frac{\partial L}{\partial b_0} = -2X'(y - Xb_0) + 2R'\lambda = 0$$

and

$$\frac{\partial L}{\partial \lambda} = 2(Rb_0 - q) = 0.$$

Assuming that X is of full column rank, this system may be solved for the solution in terms of the unconstrained estimator b:

$$b_0 = b - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(Rb - q)$$

in which the restricted least squares (RLS) estimator may be seen as a corrected version of the OLS estimator in which the correction factor relates to the magnitude of the discrepancy vector m. The explicit solution for λ likewise depends on the discrepancy vector, indicating the shadow cost of each constraint on the unconstrained solution. The covariance matrix for b_0 may be written as the covariance matrix of the unrestricted OLS estimator minus

a nonnegative definite matrix. To compare the unconstrained and constrained (RLS) solutions, we may form an F–statistic from the expression in the difference of sums of squared residuals:

$$e_0'e_0 - e'e = (Rb - q)'[R(X'X)^{-1}R'](Rb - q)$$

where e_0 are the RLS residuals. This gives rise to the F–statistic

$$F[J, n - K] = \frac{(e_0'e_0 - e'e)/J}{e'e/(n - K)}$$

which can be transformed into

$$F[J, n - K] = \frac{(R^2 - R_0^2)/J}{(1 - R^2)/(n - K)}$$

In this context, the effect of the restrictions may be viewed as either the loss in the least squares criterion (with the numerator of the first statistic viewed as the average cost per restriction) or the reduction in R^2 from the restrictions. The numerator of either expression is non–negative, since imposition of the

restrictions cannot increase R^2 nor decrease the sum of squared residuals. This estimator is provided by Stata as the command `cnsreg` (constrained regression).

These test statistics are valid for normally distributed disturbances; without normality, we still have an asymptotic justification for the statistics, and $J \times F(J, n-K) \sim \chi^2(J)$ in large samples. Since b is asymptotically normal, irrespective of the distribution of the errors, this form of the test statistic will be appropriate for large samples.

Testing nonlinear restrictions

The apparatus we have developed here applies for sets of linear restrictions on β. What if we have nonlinear restrictions on the parameters: e.g., a constraint involving a product of parameters? Then we might write $H_0 : c(\beta) = q$, and

we would require an estimate of the variance of the expression $c(b) - q$, a nonlinear function of the LS coefficient vector b. This test statistic may be calculated via the delta method, as we have discussed earlier. In the context of Stata, the command `testnl` performs nonlinear tests on the coefficient vector.

Choosing between non–nested models

The apparatus described above works quite well for classical hypothesis testing, where one model can be viewed as a proper subset of another. Annoyingly, economic theories often are cast in the form of competing hypotheses, where neither may be expressed as a proper subset of the other. Furthermore, no theory may propound the "supermodel" which would encompass all elements of both theories: essentially the artificial nesting of both models. Tests in this context are testing one specific

model versus a hybrid model that contains elements of both, which is not proposed by either theory. Thus, what if we have a setup like

$$H_0 : y = X\beta + \epsilon_0$$
$$H_1 : y = Z\gamma + \epsilon_1$$

where some of the elements of both X and Z are unique? The Bayesian econometrician would have no difficulty with this, since she would merely ask "which of these hypotheses is more likely to have generated the data"? But examining goodness of fit, and noting that one of these models has a higher R^2 or $\bar{R}^2$, is not likely to be satisfying.

A solution to this problem was proposed by Davidson and MacKinnon as their J test: not to be confused with Hansen's J test in GMM. It relates to a very simple approach: if model 0 has explanatory power over and above that of model 1, then model 0 is superior, and vice

versa. The J test therefore is performed by generating the predicted values of each series, and including them in an augmented regression of the other model. So, for instance, include the $\widehat{y}$ from the alternative hypothesis above in the null hypothesis' model. If that $\widehat{y}$ is significant, then we reject the model of the null hypothesis. We now reverse the definitions of the two models, and include the $\widehat{y}$ from the null hypothesis in the alternative hypothesis' model. Unfortunately, all four possibilities can arise: H_0 may stand against H_1, H_1 may stand against H_0, both hypotheses may be rejected, or neither hypothesis may be rejected. Only in the first two cases does the J test deliver a definitive verdict. The Stata user–contributed command `nnest` performs the J test.

A separate test along these lines is that of Cox (1961, 1962) extended by Pesaran (1974) and Pesaran and Deaton (1978). These tests are

based on likelihood ratio tests that may be constructed from the fitted values and sums of squared residuals of the nonnested models. The Cox–Pesaran–Deaton test is also performed by the `nnest` package.

Prediction

We often use regression methods to generate predictions from the estimated model. These predictions may be in–sample: that is, for the values of X used in the estimation, or out–of–sample, for arbitrary values of X: X_0. If we wish to predict $E[y_0|X_0]$, the Gauss–Markov theorem indicates that the evaluation of the regression surface is the BLUP (best linear unbiased predictor):

$$E[y_0|X_0] = \widehat{y}_0 = X_0 b.$$

In this context, we are predicting the value of y that would be expected for every observation with its regressors equal to X_0. We must

distinguish this from predicting the value of y that would apply for a specific X_0: for instance, the value of an individual house with a particular set of characteristics. When we do the former, we may use our assumptions on the distribution of the error term to take the expected value of ϵ over the subpopulation as zero. When we want to make a specific prediction, we must acknowledge that this predicted value contains an ϵ_0.

The prediction error for the latter case is

$$y_0 - \widehat{y}_0 = (\beta - b)X_0 + \epsilon_0,$$

an expression that includes both the sampling error of b and the specific ϵ associated with that value of y. To generate a standard error for $E[y_0|X_0]$, we may form the expectation of this error, which only involves the first term, giving rise to the standard error of prediction. For the prediction of a specific value

of y (rather than the expected value from the subpopulation), we consider the variance of the prediction error, which gives rise to the standard error of forecast (including the variance of ϵ_0).

The variance of ϵ_0 in the former case may be written, for a simple y on X regression, as

$$\sigma^2 \left[n^{-1} + \frac{(X_0 - \bar{X})^2)}{\sum x^2} \right]$$

where x is the deviation from mean of X. The square root of this quantity (with σ^2 replaced with our consistent estimate s^2) is then utilized in the prediction interval estimate:

$$\widehat{y}_0 \pm t_{\lambda/2} \; se(\epsilon_0)$$

The interval prediction is then defined as a pair of parabolas, with the narrowest interval at $\bar{X}$, widening as we diverge from the multivariate point of means.

In contrast, if we wish to predict the specific value for y_0 corresponding to a value X_0, we have the forecast variance

$$\sigma^2 \left[1 + n^{-1} + \frac{(X_0 - \bar{X})^2)}{\sum x^2} \right]$$

Since the latter expression contains a σ^2 term which does not go to zero as $n \to \infty$, the forecast interval is wider than the prediction interval, and the σ^2 represents an irreducible minimum for the accuracy of the forecast.

For a multiple regression problem, the prediction variance is

$$\text{Var}[(\beta - b)X_0|X] = X_0'[\sigma^2(X'X)^{-1}]X_0,$$

whereas the forecast variance for a specific value of y_0 is

$$\text{Var}[(\beta - b)X_0|X] = \sigma^2 + X_0'[\sigma^2(X'X)^{-1}]X_0.$$

These two constructs, the standard error of prediction and standard error of forecast, may

be computed following a regression in Stata via `predict newvar, stdp` and `predict newvar, stdf` respectively.

Binary/dummy/indicator variables

One of the most useful tools in applied econometrics is the binary (or dummy, or indicator) variable. We use dummies for many purposes, and there are numerous commands in any econometrics language that work with dummy variables. We often want to use a set of dummy variables to reflect the values of a qualitative variable: one which takes on several distinct values. We may also use dummies to reflect the values of an ordinal variable: one which takes on discrete values, but has no numerical significance. E.g., a Likert scale with values "Strongly Disagree", "Disagree", "Neutral", "Agree", "Strongly Agree" might be coded as 1,2,3,4,5, but we would never want to enter

those values into a regression equation: for that would treat them as cardinal measurements, which they are not.

Stata version 11 introduced the notion of *factor variables* (see the separate slideshow on *Factor variables and marginal effects*). If we have a variable taking on nonnegative integer values (e.g., Q4), we may refer to it with the `i.` factor-variable operator to instantiate the appropriate set of indicator variables on the fly as `i.Q4`.

We often use dummies to indicate membership in a set. If we are defining a dummy for male survey respondents, we can use `generate male = (sex == "M")` if we know that there are no missing values in the sex variable. Otherwise, we should use `generate male = (sex == "M") if sex !=""` to ensure that missing values become missing in the dummy. If sex was

coded as (1,2), we would use `generate male = (sex == 2) if sex <.` to accomplish the same task. Alternatively, in Stata 11, we could just refer to `i.sex`.

For a numerically coded qualitative variable (sex=(1,2), or region=(11,21,31,41)) we could use the `tab varname, gen(stub)` command, or use the factor variable prefix: e.g., `i.region`. Either alternative will automatically take care of missing values. These strategies become particularly handy if we have a large number of categories (e.g. the 50 states).

What if the variable is a string variable, e.g., state=AK, AL, AZ, etc.? Then we may use the `encode` command. For instance, `encode state, gen(stid)` would generate a new variable `stid` which appears identical to state. In reality, it now takes on numeric values of 1–50, with value labels linking each numeric code and

the original string. The `stid` variable may now be used in any command that requires numeric values (e.g., `tsset`, or as the factor variables `i.stid`).

We define a complete set of dummy variables as mutually exclusive and exhaustive. If D is a $n \times d$ matrix of dummies, $D\ \iota_d = \iota_n$. Note that the mean of a dummy variable is the sample proportion of 1s for that category, and those proportions must sum to unity for a complete set. Since a complete set of dummies sums to an ι vector, it cannot normally be included in a regression equation; we must drop one (any one) of the dummies and include the rest. The coefficients on the included dummies are then with reference to the excluded class. Using the factor variables syntax, Stata will automatically define one value of the factor variable as the *base value*. By default, this is the lowest

value of the integer variable, but it can be defined as a particular value, the largest value, the most frequent value, etc.

One-way ANOVA

For instance, consider region, with four values. A regression of income on any three of the region dummies will yield estimates of the differences in mean income levels between those regions and the excluded region (whose income level is given, in point and interval form, by the constant term). The ANOVA F statistic for this regression is aptly named, for this model is a one–way analysis of variance: does the qualitative factor region affect income? That null hypothesis is tested by the ANOVA F, and all "slopes" being equal to zero is equivalent to all regions having statistically indistinguishable income levels. (This is analogous to the t–test performed by the `ttest` command, but

that test can only be used on two groups (e.g., M and F)).

What if we wanted point and interval estimates of the regional income levels? Since we are performing one–way ANOVA, we can include all dummies and suppress the constant term (using the options `nocons hascons`). This form of the equation will not be useful for testing differences in mean income levels, though. In Stata 11, it would be easiest to estimate the model using factor variables (e.g., `regress income i.region`) and then use the `margins` command to produce point and interval estimates: `margins i.region`.

Two-way ANOVA

Any form of ANOVA may be expressed as a regression equation with nothing but dummy

variables on the RHS. For instance, a "two-way ANOVA" takes account of two qualitative factors: e.g., gender and region. We may include all but one of the dummies for each factor in the equation; the constant term is now the mean income level of the excluded class (e.g., males in region 11). In this form, we consider that the two–way table defined by this equation may be filled out from the marginal entries, which is equivalent to assuming that the effects of gender and region on income are independent factors. In Stata 11, this could be implemented with `regress income i.gender i.region`. Even if we exclude the constant term from the equation, we cannot include more than one complete set of dummies.

We can imagine cases where the assumption that the two factors' effects are independent is not tenable. For instance, unemployment rates may be high for teenagers, high for minorities, but the unemployment rate among

minority teens may be even higher than would be predicted by the sum of the two marginal effects. Alternatively, the effects of obesity and smoking on high blood pressure (hypertension) may not be additive; an overweight person who smokes may have a much higher risk of hypertension than would be implied by independent effects.

This gives rise to a two-way ANOVA *with interactions*, in which we consider an additional set of coefficients that interact the factors. This can be done in a Boolean manner, treating the individual factors' dummies with an AND condition, or algebraically, since the product of two dummies is the AND of those factors. In Stata 11, it can be done with factor variables, as shown below.

Such a model allows the two factors to have nonlinear effects on the dependent variable. If

one variable has `m` levels and a second variable has `r` levels, then there will be `(m−1)` and `(r−1)` first-level coefficients in the regression, and `((m−1)` $\times$ `(r−1))` interaction terms. The nonlinearity is readily testable as the condition that all interaction terms have zero coefficients. Using the factor variable notation, we can estimate this model with **`regress income i.gender i.region i.gender#i.region`**,
where the factor-variable operator # indicates that all possible interactions of the two categorical variables are to be included. The joint test can be performed with
`testparm i.gender#i.region`.

Alternatively, you can use the *factorial operator*: **`regress income i.gender##i.region`**, which will include both the factor variables themselves and all appropriate interactions. Stata will automatically exclude interactions that are not needed to fully characterize the joint distribution of the variables. The same **`testparm`**

command can be used to evaluate the importance of interactions.

We may consider higher–level ANOVA models, but even in a large dataset we often run out of degrees of freedom in populating all of the cells of a three- or four-way design. Since each cell of such a design is a conditional mean, we must have a large enough N in each cell to estimate such a mean reliably. The factor-variable interaction and factorial operators can be applied to three-way, four-way,. . . interactions. Note also that a dummy variable that is unity for a single observation has the effect of removing that observation from the analysis. This should be kept in mind in a time–series data context when defining dummies which allow for special events (strikes, wars, etc.): if those events last a single period, a dummy has the effect of dropping them from the sample.

Analysis of covariance

In much applied econometric research, we are concerned with what a statistician would call "ANOCOVA": analysis of covariance, in which the regression equation contains both qualitative and quantitative factors (a combination of measurable Xs and dummies). In such a model, dummies alone serve to shift the constant term for those observations which belong to a particular category. In such a model, dummies may also be interacted with measurable variables, which allows for different slopes as well as intercepts for different categories. We can readily define a set of functions that would apply for, e.g., data containing males/females and blacks/whites. Up to four distinct functions may be defined for these data, and we may consider all of the proper subsets of those functions in which slopes, intercepts, or both are constrained across the categories. Note

that estimating the four functions (and allowing both slope and intercept to vary) is essentially the same as estimating separate regression equations for each category, with the single constraint that all of those equations are forced to have the same σ^2.

Analysis of covariance models can be easily implemented in Stata version 11 using an additional factor-variable operator: the `c.` operator. To interact a categorical variable (represented by a set of dummies) with a continuous variable, you must specify that the latter variable should be considered as continuous. For example, `regress income i.region c.education i.region#c.education` will fit the model with separate intercepts and slopes for each `region`. Just as in ANOVA models, we can use the factorial operator to estimate this model: `regress income i.region##c.education`. Although this model could be implemented using dummy variables, it is advisable to estimate it with the

factor-variables syntax, as then the `margins` command will be able to compute marginal effects at different levels of education.

The `c.` operator may also be used to define polynomial interactions. Rather than creating an x^2 variable and entering it into the regression, you may merely use `c.x c.x#c.x` to enter the level and square of `x`. Again, this is advisable to enable the use of `margins` to compute marginal effects at varying levels of `x`. This logic extends to interactions of continuous variables, such as `c.x1#c.x2`, with the same rationale.

Seasonal adjustment

A set of seasonal dummies may be used, in the context of time–series data, to deseasonalize the data. If quarterly (monthly) retail sales are regressed on three (eleven) seasonal dummies,

the residuals from this model are deseasonalized retail sales. This can be easily done using factor variables notation applied to the season (quarter, month) indicator. It is customary to add the mean of the original series to restore the scale of the data. This would be additive seasonal adjustment, in which we assume that the effect of being in Q1 is a certain dollar amount of retail sales.

If we wished to apply multiplicative seasonal adjustment, we would regress $\log y$ on the three (eleven) dummies, which would assume that the effect of being in Q1 is a certain percentage deviation in retail sales from the excluded season. This seasonal adjustment may be done as a separate step, or may be applied in the context of a regression model: if we wish to regress retail sales (NSA) on other data which are SA or SAAR, we may just include the seasonal dummies in the regression equation (as

we would a time trend; the notion of partialling off the effect of trend or seasonality is the same if we assume that the regressors are free from those effects. If the regressors might contain a trend or seasonal factors, they should be detrended (or deseasonalized) as well). In any case, we would test for the existence of seasonal factors by an F–test on the set of seasonal dummies.

Testing for structural change

Dummy variables are often used to test for structural change in a regression function in which we specify *a priori* the timing of the possible structural breakpoints. We may allow for different slopes and/or intercepts for different periods in a time–series regression (e.g., allowing for a consumption function to shift downward during wartime). If we fully interact the

regime dummies, we are in essence estimating separate regressions for each regime, apart from the constraint of a common σ^2. The test that all coefficients associated with regime 2 (3,...) are equal to zero is often termed a *Chow test*. One can perform such a test by estimating the two (three,...) regressions separately, and comparing the sum of their SSRs to that of the restricted (single-regime) regression. However, this is generally easier to implement using regime dummies interacted with all regressors. This strategy is also desirable since it allows for a variety of special cases where we assume that some coefficients are stable across regimes while others vary: those interacted with regime dummies will be allowed to vary.

In some cases a regime may be too short to perform this test: that is, to estimate the regime as a separate regression. This applies

particularly at the end of a time series, where we may want to ask whether the most recent n_2 observations are generated by the same model as the prior n_1 observations. In this case, one can construct an F-test by running the regression over all n observations, then running it over the first n_1 observations, and comparing their SSRs. The SSR for the full sample will exceed that from the first n_1 observations unless the regression fits perfectly over the additional n_2 data points. Therefore, this test has n_2 d.f. in the numerator:

$$F[n_2, n_1 - K] = \frac{(e_n' e_n - e_1' e_1)/n_2}{(e_1' e_1)/(n_1 - K)}$$

where e_n is the residual vector from the full sample. This is often termed the *Chow predictive test*.

We may also be concerned about the realistic possibility that the σ^2 has changed over

regimes. We could deal with this by computing robust standard errors for the regression with regime dummies, but we might want to estimate the differing variances for each regime. This could be handled by explicitly providing for "groupwise heteroskedasticity", as we will discuss at a later date.

In other cases, we may want to allow a profile (such as an age–earnings profile) to reflect different slopes over time, but force the resulting function to be piecewise continuous. This can be achieved by a *linear spline*: for three segments, for instance, we wish to constrain the estimated segments to join at the two points, known as *knots*, of the spline function. Say that we wish to estimate three separate segments of the earnings function, for those less than 25, between 25 and 40, and over 40. Rather than estimating six coefficients (three slopes, three intercepts), we must place two

constraints on the system: that the function evaluated at the knots is piecewise continuous. Doing the algebra for this problem will show that three regressors (and a constant term) are needed: `age, (age−25) and (age−40)`. The latter two are set to zero if they are negative. Regression of earnings on these three regressors and a constant term will result in a piecewise continuous age-earnings profile. This can be automated by Stata's `mkspline` command. See the example in Baum (2006).

Splines of higher orders are often useful in applied work; e.g. a quadratic spline will be continuous and differentiable once, and a cubic spline will be continuous and twice differentiable. Both have been widely used in financial analysis: e.g., term structure research.

Tests of model stability

The dummy variable methods described above are useful when the timing (or location, in cross-sectional terms) of a structural break is known *a priori*. However, we often are unsure as to whether a change has taken place, or in a time-series context, when a relationship may have undergone a structural shift. This is particularly problematic when a change may be a gradual process rather than an abrupt and discernable break. A number of tests have been devised to evaluate the likelihood that a change has taken place, and if so, when that break may have occurred. The "cusums" tests of Brown, Durbin and Evans (1975) are based on *recursive residuals* (or their squares), and the notion that a change will show up in the one-step ahead prediction errors if we consider all possible breakpoints in a series. A formal test may be devised by evaluating the

cumulative sum of residuals, and noting that the cumulative sum will stray outside a confidence interval in the vicinity of a structural break. Tests of this nature are available in Stata via the user-contributed routine `cusum6`. The power of cusums tests may be quite low in practice, but they will be somewhat useful in detecting a structural break in the relationship.

Specification error

We have worked with regression models under the maintained hypothesis that the model is correctly specified. What if this assumption does not hold? Although there is no necessary relation between the specification employed and the true data generating process, let us consider two alternatives: given the dependent variable y, we may omit relevant variables, or include irrelevant variables. First consider omission of relevant variables from the correctly specified model

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

with K_1 and K_2 regressors in the two subsets. Imagine that we ignore X_2, so that

$$b_1 = (X_1'X_1)^{-1}X_1'y$$
$$= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\epsilon$$

Unless $X_1'X_2 = 0$ or $\beta_2 = 0$, the estimate of b_1 is biased, so that

$$E[b_1|X] = \beta_1 + P_{1\cdot 2}\beta_2$$

where $P_{1\cdot 2} = (X_1'X_1)^{-1}X_1'X_2$ is the $K_1 \times K_2$ matrix reflecting the regression of each column of X_2 on the columns of X_1. If $K_1 = K_2 = 1$ and the X_2 variable is correlated with X_1, we may derive the direction of bias, but in general we will not be able to make any statements about the coefficients in b_1 with multiple variables in each set. It appears, then, that the cost of omitting relevant variables is high: the resulting coefficients are biased and inconsistent.

What about the inclusion of irrelevant explanatory variables: for instance, if the true DGP involves only X_1, but we mistakenly include the X_2 variables as well? In that case, we are failing to impose the restrictions that $b_2 = 0$, but since in the population, $\beta_2 = 0$, those restrictions are costless. The inclusion of X_2 leaves our estimates of b_1 unbiased and consistent, as is the estimate of σ^2. What is the cost, then, of overfitting the model and including the additional variables? We lose degrees of freedom, of course, and by ignoring the information that the X_2 variables do not belong in the model, we generate less efficient estimates of b_1 than we would with the correctly specified model. This is especially apparent if we have $K_1 = K_2 = 1$ and the correlation between X_1 and X_2 is high. Mistakenly including X_2 will lead to quite imprecise estimates of b_1.

Although there is some cost to overfitting a model, it would appear that the costs of these

two types of specification error are quite asymmetric, and that we would much rather err on the side of caution (including additional variables) to avoid the severe penalties of underfitting the model. Given this, a model selection strategy that starts with a simple specification and seeks to refine it by adding variables is flawed, and the opposite approach: David Hendry's general-to-specific methodology has much to recommend it. Although a very general specification may be plagued by collinearity, and a model with sufficient variables will run afoul of the 5% type I error probability, it is much more likely than a recursive simplification strategy will yield a usable model at the end of the specification search.

In dynamic models utilizing time-series data, this advice translates into "do not underfit the dynamics". If the time form of a dynamic relationship is not known with certainty, it would

be prudent to include a number of lagged values, and "test down" to determine whether the longer lags are necessary. This will follow a general-to-specific methodology, allowing for the more complex dynamic specification if it is warranted. Omitting higher-order dynamic terms is a common cause of apparent non–independence of the regression errors, as signalled by residual independence tests.

The RESET test

One particular approach to possible functional form misspecification is offered by Ramsey's RESET (regression specification error) test. This test is simple: after a regression of y on the regressors X, one adds polynomials in the fitted values $\widehat{y}$ to the regression: for instance, squares and cubes. Under the hypothesis that the relationship between y and X is adequately modeled as linear, the coefficients

on these additional regressors should not be significantly different from zero. Since polynomials in the regressors can approximate many functions, a rejection in the RESET test may also suggest the appropriateness of a nonlinear specification: for instance, a log-linear or double-log model. The RESET test may easily be performed in Stata via the `estat ovtest` command.

A generalized RESET test

Mark Schaffer's `ivreset` routine (available from `ssc`) extends Ramsey's RESET test to the case of instrumental variables estimation. To quote from his extensive help file:

The Ramsey (1969) RESET test is a standard test of neglected nonlinearities in the choice of functional form (sometimes, perhaps misleadingly, also described as a test for omitted variables; see `estat ovtest` and Wooldridge

(2002), pp. 124–5). The principle is to estimate $y = X\beta + W\gamma + u$ and then test the significance of γ. The Ws in a RESET test can either be powers of X or, as implemented here, powers of the forecast values of y.

As Pagan and Hall (1983) and Pesaran and Taylor (1999) point out, a RESET test for an IV regression cannot use the standard IV predicted values $X\widehat{\beta}$ because X includes endogenous regressors that are correlated with u. Instead, the RESET test needs to be implemented using "forecast values" of y that are functions of the instruments (exogenous variables) only. Denote the full set of instruments by Z (possibly including exogenous regressors also in X).

In the Pagan–Hall version of the test, the forecast values $\widehat{y}$ are the reduced form predicted

values of y, i.e., the predicted values $Z\widehat{\pi}$ from a regression of y on the instruments Z.

In the Pesaran–Taylor version of the test, the forecast values $\widehat{y}$ are the "optimal forecast" values. The optimal forecast (predictor) $\widehat{y}$ is defined as $\widehat{X}\widehat{\beta}$ where $\widehat{\beta}$ is the IV estimate of the coefficents and $\widehat{X}$ consists of the exogenous regressors in X and the reduced form predicted values of the endogenous regressors in X. The latter are the predicted values $Z\widehat{\pi}$ from regressions of the endogenous Xs on the instruments in Z. Note that if the equation is exactly identified, the optimal forecasts and reduced form forecasts coincide, and the Pesaran–Taylor and Pagan–Hall test statistics are identical.

In both the Pesaran–Taylor and Pagan–Hall versions of the RESET test, the augmented equation is $y = X\beta + W\gamma + u$, where the Ws are the powers of $\widehat{y}$. The default is to include

$\widehat{y}^2$, but 3rd and 4th powers of $\widehat{y}$ can be requested. This equation is estimated by IV, and the default test statistic is a Wald test of the significance of γ. Under the null that there are no neglected nonlinearities and the equation is otherwise well-specified, the test statistic is distributed as χ^2 with degrees of freedom equal to the number of powers of $\widehat{y}$.

Alternatively, Godfrey has suggested that a C test statistic (also known as a "GMM-distance" or "difference-in-Sargan" test) be used to test whether the powers of $\widehat{y}$ can be added to the orthogonality or moment conditions that define the IV or OLS estimator (see Pesaran and Smith, pp. 262–63). This test can be requested with the `cstat` option. Under the null that the equation is well-specified and has no neglected nonlinearities, $(J - J1)$ is distributed as χ^2 with degrees of freedom equal to the number of powers of $\widehat{y}$, where $J1$ is the

Sargan–Hansen statistic for the original IV estimation and J is the Sargan–Hansen statistic for the IV estimation using the additional orthogonality conditions provided by the powers of $\widehat{y}$.

If the equation was estimated using OLS or HOLS (heteroskedastic OLS) and there there are no endogenous regressors, `ivreset` reports a standard Ramsey RESET test using the fitted values of y, i.e., $X\widehat{\beta}$.

If the original equation was estimated using the `robust, cluster` or `bw` options, so is the augmented equation, and the RESET test statistic will be heteroskedastic-, cluster-, and/or autocorrelation-robust, respectively.