ECON2228 Notes 4

Christopher F Baum

Boston College Economics

2014–2015

Chapter 4: Multiple regression analysis: Inference

We have discussed the conditions under which OLS estimators are unbiased, and derived the variances of these estimators under the Gauss-Markov assumptions. The Gauss-Markov theorem establishes that OLS estimators have the smallest variance of any linear unbiased estimators of the population parameters.

We must now more fully characterize the sampling distribution of the OLS estimators–beyond its mean and variance–so that we may test hypotheses on the population parameters.

To make the sampling distribution tractable, we add an assumption on the distribution of the errors:

Proposition

MLR6 (Normality) The population error u is independent of the explanatory variables $x_1, .., x_k$ and is normally distributed with zero mean and constant variance: $u \sim N\left(0, \sigma^2\right)$.

This is a much stronger assumption than we have previously made on the distribution of the errors. The assumption of normality, as we have stated it, subsumes both the assumption of the error process being independent of the explanatory variables, and that of homoskedasticity. For cross-sectional regression analysis, these six assumptions define the *classical linear model*.

The rationale for normally distributed errors is often phrased in terms of the many factors influencing y being additive, appealing to the Central Limit Theorem to suggest that the sum of a large number of random factors will be normally distributed. Although we might have reason in a particular context to doubt this rationale, we usually use it as a working hypothesis. Various transformations, such as taking the logarithm of the dependent variable, are often motivated in terms of their inducing normality in the resulting errors.

What is the importance of assuming normality for the error process? Under the assumptions of the classical linear model, normally distributed errors give rise to normally distributed OLS estimators:

$$b_j \sim N\left(\beta_j, Var\left(b_j\right)\right) \tag{1}$$

which will then imply that:

$$\frac{\left(b_j - \beta_j\right)}{\sigma_{b_j}} \sim N(0,1) \tag{2}$$

This follows since each of the b_j can be written as a linear combination of the errors in the sample. Since we assume that the errors are independent, identically distributed normal random variates, any linear combination of those errors is also normally distributed.

We may also show that any linear combination of the b_j is also normally distributed, and a subset of these estimators has a joint normal distribution. These properties will come in handy in formulating tests on the coefficient vector. We may also show that the OLS estimators will be approximately normally distributed (at least in large samples), even if the underlying errors are not normally distributed.

Testing an hypothesis on a single β_j

To test hypotheses about a single population parameter, we start with the model containing k regressors:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u \tag{3}$$

Under the classical linear model assumptions, a test statistic formed from the OLS estimates may be expressed as:

$$\frac{(b_j - \beta_j)}{s_{b_j}} \sim t_{n-k-1} \tag{4}$$

Why does this test statistic differ from (2) above? In that expression, we considered the variance of b_j as an expression including σ, the unknown standard deviation of the error term (that is, $\sqrt{\sigma^2}$). In this operational test statistic (4), we have replaced σ with a consistent estimate, s.

That additional source of sampling variation requires the switch from the standard normal distribution to the t distribution, with $(n-k-1)$ degrees of freedom. Where n is not all that large relative to k, the resulting t distribution will have considerably fatter tails than the standard normal.

Where $(n-k-1)$ is a large number–greater than 100, for instance–the t distribution will essentially be the standard normal. The net effect is to make the critical values larger for a finite sample, and raise the threshold at which we will conclude that there is adequate evidence to reject a particular hypothesis.

The test statistic (4) allows us to test hypotheses regarding the population parameter β_j : in particular, to test the null hypothesis

$$H_0 : \beta_j = 0 \tag{5}$$

for any of the regression parameters. The "t-statistic" used for this test is merely that printed on the output when you run a regression in Stata or any other program: the ratio of the estimated coefficient to its estimated standard error.

If the null hypothesis is to be rejected, the "t-stat" must be larger (in absolute value) than the critical point on the t-distribution. The "t-stat" will have the same sign as the estimated coefficient, since the standard error is always positive. Even if β_j is actually zero in the population, a sample estimate of this parameter–b_j– will never equal exactly zero.

But when should we conclude that it could be zero? When its value cannot be distinguished from zero. There will be cause to reject this null hypothesis if the value, scaled by its standard error, exceeds the threshold.

For a "two-tailed test," there will be reason to reject the null if the "t-stat" takes on a large negative value or a large positive value; thus we reject in favor of the alternative hypothesis (of $\beta_j \neq 0$) in either case. This is a two-sided alternative, giving rise to a two-tailed test. If the hypothesis is to be tested at, e.g., the 95% level of confidence, we use critical values from the t-distribution which isolate 2.5% in each tail, for a total of 5% of the mass of the distribution.

When using a computer program to calculate regression estimates, we usually are given the "*p-value*" of the estimate–that is, the tail probability corresponding to the coefficient's t-value. The p-value may usefully be considered as the probability of observing a t-statistic as extreme as that shown *if the null hypothesis is true*.

If the t-value was equal to, e.g., the 95% critical value, the p-value would be exactly 0.05. If the t-value was higher, the p-value would be closer to zero, and vice versa. Thus, we are looking for small p-values as indicative of rejection. A p-value of 0.92, for instance, corresponds to an hypothesis that can be rejected at the 8% level of confidence–thus quite irrelevant, since we would expect to find a value that large 92% of the time under the null hypothesis. On the other hand, a p-value of 0.08 will reject at the 90% level, but not at the 95% level; only 8% of the time would we expect to find a t-statistic of that magnitude if H_0 was true.

What if we have a one-sided alternative? For instance, we may phrase the hypothesis of interest as:

$$\begin{aligned} H_0 &: \beta_j > 0 \\ H_A &: \beta_j \leq 0 \end{aligned} \tag{6}$$

Here, we must use the appropriate critical point on the t-distribution to perform this test at the same level of confidence. If the point estimate b_j is positive, then we do not have cause to reject the null. If it is negative, we may have cause to reject the null if it is a sufficiently large negative value.

The critical point should be that which isolates 5% of the mass of the distribution in that tail (for a 95% level of confidence). This critical value will be smaller (in absolute value) than that corresponding to a two-tailed test, which isolates only 2.5% of the mass in that tail. The computer program always provides you with a p-value for a two-tailed test; if the p-value is 0.08, for instance, it corresponds to a one-tailed p-value of 0.04 (that being the mass in that tail).

Testing other hypotheses about β_j

Every regression output includes the information needed to test the two-tailed or one-tailed hypotheses that a population parameter equals zero. What if we want to test a different hypothesis about the value of that parameter? For instance, we would not consider it sensible for the *mpc* for a consumer to be zero, but we might have an hypothesized value (of, say, 0.8) implied by a particular theory of consumption. How might we test this hypothesis?

If the null is stated as:

$$H_0 : \beta_j = a_j \tag{7}$$

where a_j is the hypothesized value, then the appropriate test statistic becomes:

$$\frac{(b_j - a_j)}{s_{b_j}} \sim t_{n-k-1} \tag{8}$$

and we may simply calculate that quantity and compare it to the appropriate point on the t-distribution.

Most computer programs provide you with assistance in this effort; for instance, if we believed that b_j, the coefficient on *bdrms*, should be equal to \$20,000 in a regression of house prices on square footage and *bdrms* (e.g. using HPRICE1), we would use Stata's `test` command:

```
regress price bdrms sqrft
test bdrms=20000
```

We use the name of the variable as a shorthand for the name of the coefficient on that variable. Stata, in that instance, presents us with:

```
( 1) bdrms = 20000.0

F( 1, 85) = 0.26

Prob > F = 0.6139
```

making use of an F-statistic, rather than a t-statistic, to perform this test. In this particular case, of an hypothesis involving a single regression coefficient, we may show that this F-statistic is merely the square of the associated t-statistic. The p-value would be the same in either case. The estimated coefficient is 15198.19, with an estimated standard error of 9483.517.

Plugging in these values to (8) yields a t-statistic:

```
. di (_b[bdrms]-20000)/_se[bdrms]
-.50633208
```

which, squared, is the F-statistic shown by the `test` command. Just as with tests against a null hypothesis of zero, the results of the `test` command may be used for one-tailed tests as well as two-tailed tests; then, the magnitude of the coefficient matters (i.e. the fact that the estimated coefficient is about $15,000 means we would never reject a null that it is less than $20,000), and the p-value must be adjusted for one tail.

Any number of `test` commands may be given after a `regress` command in Stata, testing different hypotheses about the coefficients.

Confidence intervals

We may use the point estimate and its estimated standard error to calculate an hypothesis test on the underlying population parameter, or we may form a confidence interval for that parameter. Stata makes that easy in a regression context by providing the 95% confidence interval for every estimated coefficient.

If you want to use some other level of significance, you may either use the `level()` option on `regress` (e.g., `regress price bdrms sqrft, level(90)`) or you may change the default level for this run with `set level`. All further regressions will report confidence intervals with that level of confidence.

To connect this concept to that of the hypothesis test, consider that in the above example the 95% confidence interval for β_{bdrms} extended from -3657.581 to 34053.96; thus, an hypothesis test with the null that β_{bdrms} takes on any value in this interval (including zero) will not lead to a rejection.

Testing hypotheses about a single linear combination of the parameters

Economic theory will often suggest that a particular linear combination of parameters should take on a certain value: for instance, in a Cobb-Douglas production function, that the slope coefficients should sum to one in the case of constant returns to scale (*CRTS*):

$$\begin{aligned} Q &= AL^{\beta_1}K^{\beta_2}E^{\beta_3} \\ \log Q &= \log A + \beta_1 \log L + \beta_2 \log K + \beta_3 \log E + v \end{aligned} \tag{9}$$

where K, L, E are the factors capital, labor, and energy, respectively. We have added an error term to the double-log-transformed version of this model to represent it as an empirical relationship.

The hypothesis of *CRTS* may be stated as:

$$H_0 : \beta_1 + \beta_2 + \beta_3 = 1 \tag{10}$$

The test statistic for this hypothesis is quite straightforward:

$$\frac{(b_1 + b_2 + b_3 - 1)}{s_{b_1+b_2+b_3}} \sim t_{n-k-1} \tag{11}$$

and its numerator may be easily calculated.

The denominator, however, is not so simple; it represents the standard error of the linear combination of estimated coefficients. You may recall that the variance of a sum of random variables is not merely the sum of their variances, but an expression also including their covariances, unless they are independent. The random variables $\{b_1, b_2, b_3\}$ are not independent of one another since the underlying regressors are not independent of one another.

Each of the underlying regressors is assumed to be independent of the error term u, but not of the other regressors. We would expect, for instance, that firms with a larger capital stock also have a larger labor force, and use more energy in the production process.

The variance (and standard error) that we need may be readily calculated by Stata, however, from the variance-covariance matrix of the estimated parameters via the `test` command:

```
test cap + labor + energy = 1
```

will provide the appropriate test statistic, again as an F-statistic with a p-value. You may interpret this value directly.

If you would like the point and interval estimate of the hypothesized combination, you can compute that (after a regression) with the `lincom` (linear combination) command:

```
lincom cap + labor + energy
```

will show the sum of those values and a confidence interval for that sum.

We may also use this technique to test other hypotheses than adding-up conditions on the parameters. For instance, consider a two-factor Cobb-Douglas function in which you have only labor and capital, and you want to test the hypothesis that labor's share is 2/3. This implies that the labor coefficient should be twice the capital coefficient, or:

$$\begin{aligned} H_0 &: \beta_L = 2\beta_K, \text{or} \\ H_0 &: \frac{\beta_L}{\beta_K} = 2, \text{ or} \\ H_0 &: \beta_L - 2\beta_K = 0 \end{aligned} \tag{12}$$

Note that this does not allow us to test a nonlinear hypothesis on the parameters: but considering that a ratio of two parameters is a constant is not a nonlinear restriction. In the latter form, we may specify it to Stata's `test` command as:

```
test labor - 2*cap = 0
```

In fact, Stata will figure out that form if you specify the hypothesis as:

```
test labor = 2*cap
```

but it is not quite smart enough to handle the ratio form. It is easy to rewrite the ratio form into one of the other forms. Either form will produce an F-statistic and associated p-value related to this single linear hypothesis on the parameters which may be used to make a judgment about the hypothesis of interest.

Testing multiple linear restrictions

When we use the `test` command, an F-statistic is reported, even when the test involves only one coefficient, because in general, hypothesis tests may involve more than one restriction on the population parameters. The hypotheses discussed above, even that of CRTS, involving several coefficients, still only represent one restriction on the parameters.

For instance, if CRTS is imposed, the elasticities of the factors of production must sum to one, but they may individually take on any value. But in most applications of multiple linear regression, we concern ourselves with *joint tests* of restrictions on the parameters.

The simplest joint test is that which every regression reports: the so-called “ANOVA F” test, which has the null hypothesis that *each* of the slopes is equal to zero. Note that in a multiple regression, specifying that each slope individually equals zero is not the same thing as specifying that their sum equals zero. This “ANOVA” (ANalysis Of VAriance) F-test is of interest since it essentially tests whether the entire regression has any explanatory power.

The null hypothesis, in this case, is that the "model" is $y = \beta_0 + u$: that is, none of the explanatory variables assist in explaining the variation in y. We cannot test any hypothesis on the R^2 of a regression, but we will see that there is an intimate relationship between the R^2 and the ANOVA F:

$$\begin{aligned} R^2 &= \frac{SSE}{SST} \\ F &= \frac{SSE/k}{SSR/\left(n-(k+1)\right)} \\ \therefore F &= \frac{R^2/k}{\left(1-R^2\right)/\left(n-(k+1)\right)} \end{aligned} \tag{13}$$

where the ANOVA F, the ratio of mean square explained variation to mean square unexplained variation, is distributed as $F^k_{n-(k+1)}$ under the null hypothesis.

For a simple regression, this statistic is F^1_{n-2}, which is identical to $\left(t_{b_1,n-2}\right)^2$: that is, the square of the $t-$ statistic for the slope coefficient, with precisely the same $p-$ value as that $t-$ statistic. In a multiple regression context, we do not often find an insignificant $F-$ statistic, since the null hypothesis is a very strong statement: that *none* of the explanatory variables, taken singly or together, explain any significant fraction of the variation of y about its mean. That can happen, but it is often somewhat unlikely.

The ANOVA F tests *k exclusion restrictions*: that all *k* slope coefficients are jointly zero. We may use an F-statistic to test that a number of slope coefficients are jointly equal to zero. For instance, consider a regression of 353 major league baseball players' salaries (from MLB1).

If we regress `lsalary` (log of player's salary) on `years` (number of years in majors), several indicator variables indicating the position played (`frstbase, scndbase, shrtstop, thrdbase, catcher)` and `gamesyr` (number of games played per year), we get an R^2 of 0.6105, and an ANOVA F (with 7 and 345 d.f.) of 77.24 with a $p-$ value of zero.

The overall regression is clearly significant, and the coefficients on `years` and `gamesyr` both have the expected positive and significant coefficients. Only one of the five coefficients on the positions played, however, are significantly different from zero at the 5% level: `scndbase`, with a negative value (-0.034) and a $p-$value of 0.015. The `frstbase` and `shrtstop` coefficients are also negative (but insignificant), while the `thrdbase` and `catcher` coefficients are positive and insignificant.

Should we just remove all of these variables (except for `scndbase`)? The F-test for these five exclusion restrictions will provide an answer to that question:

```
. test frstbase scndbase shrtstop thrdbase catcher
( 1) frstbase = 0.0
( 2) scndbase = 0.0
( 3) shrtstop = 0.0
( 4) thrdbase = 0.0
( 5) catcher = 0.0
F( 5, 345) = 2.37
Prob > F = 0.0390
```

At the 95% level of significance, these coefficients are not each zero. That result, of course, could be largely driven by the `scndbase` coefficient:

```
. test frstbase shrtstop thrdbase catcher
( 1) frstbase = 0.0
( 2) shrtstop = 0.0
( 3) thrdbase = 0.0
( 4) catcher = 0.0
F( 4, 345) = 1.56
Prob > F = 0.1858
```

So perhaps it would be sensible to remove these four, which even when taken together do not explain a meaningful fraction of the variation in `lsalary`. But this illustrates the point of the joint hypothesis test: the result of simultaneously testing several hypotheses (that, for instance, individual coefficients are equal to zero) cannot be inferred from the results of the individual tests. If each coefficient is significant, then a joint test will surely reject the joint exclusion restriction; but the converse is assuredly false.

Notice that a joint test of exclusion restrictions may be easily conduced by Stata's `test` command, by merely listing the variables whose coefficients are presumed to be zero under the null hypothesis. The resulting test statistic is an F with as many numerator degrees of freedom as there are coefficients (or variables) in the list. It can be written in terms of the residual sums of squares (*SSRs*) of the "unrestricted" and "restricted" models:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \tag{14}$$

Since adding variables to a model will never increase *SSR* (nor decrease R^2), the "restricted" model—in which certain coefficients are not freely estimated from the data, but constrained—must have *SSR* at least as large as the "unrestricted" model, in which all coefficients are data-determined at their optimal values. Thus the difference in the numerator is non-negative.

If it is a large value, then the restrictions severely diminish the explanatory power of the model. The amount by which it is diminished is scaled by the number of restrictions, q, and then divided by the unrestricted model's s^2. If this ratio is a large number, then the "average cost per restriction" is large relative to the explanatory power of the unrestricted model, and we have evidence against the null hypothesis (that is, the $F-$ statistic will be larger than the critical point on an $F-$ table with q and $(n-k-1)$ degrees of freedom.

If the ratio is smaller than the critical value, we do not reject the null hypothesis, and conclude that the restrictions are consistent with the data. In this circumstance, we might then reformulate the model with the restrictions in place, since they do not conflict with the data. In the baseball player salary example, we might drop the four insignificant variables and reestimate the more parsimonious model.

Testing general linear restrictions

The apparatus described above is far more powerful than it might appear. We have considered individual tests involving a linear combination of the parameters (e.g. *CRTS*) and joint tests involving exclusion restrictions (as in the baseball players' salary example).

The "subset F" test defined in (14) is capable of being applied to any set of linear restrictions on the parameter vector: for instance, that $\beta_1 = 0$, $\beta_2 + \beta_3 + \beta_4 = 1$, and $\beta_5 = -1$. What would this set of restrictions imply about a regression of y on $\{X_1, X_2, X_3, X_4, X_5\}$? That regression, in its unrestricted form, would have $k = 5$, with 5 estimated slope coefficients and an intercept.

The joint hypotheses expressed above would state that a restricted form of this equation would have three fewer parameters, since β_1 would be constrained to zero, β_5 to -1, and one of the coefficients $\{\beta_2, \beta_3, \beta_4\}$ expressed in terms of the other two. In the terminology of (14), $q = 3$. How would we test the hypothesis? We can readily calculate SSR_{ur}, but what about SSR_r?

One approach would be to algebraically substitute the restrictions in the model, estimate that restricted model, and record its SSR_r value. This can be done with any computer program that estimates a multiple regression, but it requires that you do the algebra and transform the variables accordingly. (For instance, constraining β_5 to -1 implies that you should form a new dependent variable, $(y + X_5)$). Alternatively, if you are using a computer program that can test linear restrictions, you may use its features. Stata will test general linear restrictions of this sort with the `test` command:

```
regress y x1 x2 x3 x4 x5
test (x1) (x2+x3+x4=1) (x5=-1)
```

This `test` command will print an F-statistic for the set of three linear restrictions on the regression. For the original baseball salary model:

```
test (years) (frstbase+scndbase+shrtstop=1) (thrdbase=-1)
( 1) years = 0
( 2) frstbase + scndbase + shrtstop = 1
( 3) thrdbase = -1
F( 3, 345) = 51.39
Prob > F = 0.0000
```

The F-test will have three numerator degrees of freedom, because you have specified three linear hypotheses to be jointly applied to the coefficient vector. This syntax of `test` may be used to construct any set of linear restrictions on the coefficient vector, and perform the joint test for the validity of those restrictions. The test statistic will reject the null hypothesis (that the restrictions are consistent with the data) if its value is large relative to the underlying F-distribution.