PLSC 503: "Multivariate Analysis for Political Research"

Multivariate Regression, II: Inference
February 16, 2017

Multivariate Regression: Inference

As we did before, we'll spend today discussing the conduct of inference in multivariate OLS regression. We'll talk about the mathematics behind inference, then discuss some tests and practices, and finish up with some examples.

Inference, In General

First, think of a very general case. We have a rank-K vector of parameters – call it $\boldsymbol{\beta}$ – that we wish to estimate from our sample data. Following the usual convention, call our estimate $\hat{\boldsymbol{\beta}}$. Typical hypothesis testing about $\boldsymbol{\beta}$ occurs in five steps:

1. Pick some $\mathbf{H}_A : \boldsymbol{\beta} = \boldsymbol{\beta}_A$
2. Estimate $\hat{\boldsymbol{\beta}}$
3. Determine distribution of $\hat{\boldsymbol{\beta}}$ under $\mathbf{H}_A$
4. Use (2) and (3) to form a *test statistic* $\hat{\mathbf{S}} = h(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$
5. Assess $\Pr(\hat{\mathbf{S}}|\mathbf{H}_A)$

Step (1) is straightforward enough. For example, if we want to test the usual "null hypothesis," we'd choose $\mathbf{H}_A : \boldsymbol{\beta} = \mathbf{0}$ Step (2) is also easy enough; we learned how to do that last time. And we began to get to step (3) last time, when we noted that one of our usual OLS assumptions is that

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}). \tag{1}$$

Steps (3), (4) and (5) are what we'll talk about today.

The Importance of $\mathbf{V}(\hat{\boldsymbol{\beta}})$

As in the bivariate case, to conduct inference we first need to arrive at a statement of the relative variances and covariances of our parameter estimates $\hat{\boldsymbol{\beta}}$. In a multivariate context, the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ (which we'll denote $\mathbf{V}(\hat{\boldsymbol{\beta}})$) is:

$$\begin{aligned}\mathbf{V}(\hat{\boldsymbol{\beta}}) &= \mathrm{E}[\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})]^2 \\ &= \mathrm{E}\{[\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})][\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})]'\}\end{aligned}$$

Since we know from last time that $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, we can rewrite (2) as:

$$\begin{aligned}
\mathbf{V}(\hat{\boldsymbol{\beta}}) &= \mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \\
&= \mathrm{E}\{[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}]'\} \\
&= \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \qquad (2)
\end{aligned}$$

Taking expectations, we get:

$$\begin{aligned}
\mathbf{V}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\mathbf{u}\mathbf{u}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \qquad (3)
\end{aligned}$$

Note that:

- We use the assumption of homoscedastic, uncorrelated errors to get this result (that is, that $\mathrm{E}(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I}$). As we'll see later, if we muck around with this assumption, inference gets complicated.
- The actual values of the variances for each estimate $\hat{\beta}_j$ are just like their bivariate analogue: the error variance "divided by" the variability in $\mathbf{X}_j$.
- A key, important difference, however, is that (3) is now a $K \times K$ matrix that includes a function of the amount of *covariance* among the $\mathbf{X}$s as its off-diagonal elements.
- That is, $(\mathbf{X}'\mathbf{X})^{-1}$ tells us both the *variability* of the $\mathbf{X}$s and the *covariance* of those variables; among others, this will become very important when we discuss multicollinearity.

Estimating $\mathbf{V}(\hat{\boldsymbol{\beta}})$

As was the case for bivariate regression, we would need to know σ^2 to make any use of (3). In practice, we don't know σ^2, and so we have to use a consistent estimate instead. In the scalar / bivariate context, that estimate $\hat{\sigma}^2$ was just the sum of squared errors, divided by the degrees of freedom. It's the same thing here: remembering that we can calculate the sum of squares as $\mathbf{u}'\mathbf{u}$, we can write:

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{N - K} \qquad (4)$$

This provides a direct estimate of the error variance (and thus the standard error of the estimate), and we can thus use it to estimate the variance-covariance matrix of the coefficient estimates:

$$\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \qquad (5)$$

Note a few things about $\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}$:

- The elements on the main diagonal are the estimated variances of the K parameter estimates, including the constant term. That means that
- The square roots of the diagonal elements of $\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}$ are the estimated "standard errors" of our coefficient estimates $\hat{\boldsymbol{\beta}}$.
- The off-diagonal elements of (5) are the *covariances* of the parameter estimates. (Remember: the $\hat{\boldsymbol{\beta}}$s are random variables too!).

Hypothesis Testing: Single Coefficient Estimates

The most common usage of $\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}$ is to do hypothesis testing on the estimated parameters $\hat{\boldsymbol{\beta}}$. If all our assumptions about the linear regression model hold, then it can be shown that:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}] \tag{6}$$

This means that our estimates are normally distributed around their "true" population values, and that the variance of a particular estimate $\hat{\beta}_k$ is given by σ^2 times the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

In practice, the use of $\hat{\sigma}^2$ in place of σ^2 means that our hypothesis testing can no longer use a Normal distribution, but instead follows a t distribution with $N - K$ degrees of freedom.[1]

So the standard way of engaging in hypothesis testing is essentially identical to that in the bivariate case:

1. Choose a value of β_k that you want to test (say, $\beta_k = 0$),
2. Calculate the t-statistic for the coefficient associated with X_k, which is:

$$\hat{t}_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\widehat{\mathbf{V}(\hat{\beta}_k)}}}$$

3. Compare $\hat{t}_k$ to a t distribution with $N - K$ degrees of freedom.

The larger the value of $\hat{t}_k$, the greater the confidence with which we can reject the stated hypothesis about β_k (assuming the directionality, if any, is also correct). We'll do an example a bit later.

[1] Note also that this means that, as N increases relative to K, our hypothesis testing distributions come to more and more closely resemble a Normal distribution.

Multivariate Hypothesis Testing

Once we enter the realm of multivariate regression, it also becomes of interest to conduct tests on *groups* of estimates. In the most general case, we might want to see if:

$$\mathrm{H}_0 : \beta_1 = \beta_2 = ... = \beta_K = 0,$$

that is, if *all* the variables' effects on $\mathbf{Y}$ are (jointly) zero. Alternatively, we may be interested in testing only a subset of those estimates:

$$\mathrm{H}_0 : \beta_3 = \beta_6 = 0,$$

or some other set of linear restrictions – say, that one variable's influence is the same as another's:

$$\mathrm{H}_j : \beta_2 = \beta_4.$$

There is a relatively simple, general way to do these sorts of tests: the **F-test**. The intuition behind the F-test is to estimate "restricted" and "unrestricted" models, with the former imposing the constraints that the hypotheses we wish to test impose and the latter removing those constraints. If we then see that the "unrestricted" model is a better "fit" to the data than is the "restricted" model, we can reject the hypotheses that the restrictions are based upon. The F-test is nothing more than a formalization of this logic.

F-tests: Math and Intuition

Consider once again the decomposition of the variation in Y that we talked about when we learned about R^2:

$$\mathbf{TSS} = \mathbf{MSS} + \mathbf{RSS}$$

Now, think about this in the context of a regression model with four covariates:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_{Ui} \tag{7}$$

and suppose we were interested in testing the hypothesis that the joint effect of X_2 and X_4 on Y was zero:

$$\mathrm{H}_a : \beta_2 = \beta_4 = 0$$

That is, that neither X_2 nor X_4 nor the combination of the two had any significant impact on Y. If that is in fact the case, then we can think of the model as:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + 0X_{2i} + \beta_3 X_{3i} + 0X_{4i} + u_i \\ &= \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + u_{Ri} \qquad (8) \end{aligned}$$

We'll refer to the model in (7) as the *unrestricted* model, and that in (8) as the *restricted* model, and denote them with subscripts $_U$ and $_R$ respectively.

We can estimate both (7) and (8), and from those estimates recover the extent to which each model explains the dependent variable Y. More specifically, we can consider the estimated residual sums of squares for each model:

$$\text{RSS}_U \equiv \hat{\mathbf{u}}'_U \hat{\mathbf{u}}_U = \sum_{i=1}^{N} \hat{u}^2_{Ui} \tag{9}$$

and

$$\text{RSS}_R \equiv \hat{\mathbf{u}}'_R \hat{\mathbf{u}}_R = \sum_{i=1}^{N} \hat{u}^2_{Ri} \tag{10}$$

Now, if the unrestricted model fits no better than the restricted model, then the difference between the two RSSs should not be statistically distinguishable from zero – that is, we would not be able to reject the difference as being due to anything other than chance / random error.

To make this comparison formally, we use the F statistic:

$$F = \frac{(\text{RSS}_R - \text{RSS}_U)/q}{\text{RSS}_U/(N-K)} \tag{11}$$

where:

- q is the number of linear restrictions imposed,
- K is the number of variables in the unrestricted (full) model, and
- N is the number of observations.

Think for a moment about the distribution of this "thing" called F...

- It is a ratio of two other "things."
- The "thing" in the numerator is essentially an indicator of the degree of *difference* between the fits of the two models. It is a (normed-by-q) difference between yet two more things, each of which is a sum of squared Normal variates.
- This means that the numerator is itself a χ^2 variable with q degrees of freedom.
- The denominator is an indication of the relative fit of the unrestricted model; it serves as something of a "baseline" for the test statistic. It is also a (normed) sum of a bunch of squared Normal variates, and so is a χ^2 variable with $N - K$ degrees of freedom.

- And the ratio of two χ^2 variables with q and $N - K$ degrees of freedom, respectively, is distributed according to an F distribution with q and $N - K$ degrees of freedom.

In addition, think for a second about what this F thing will "look like..."

- Since RSS_U will always be less than or equal to RSS_R, the numerator will always be positive.
- Since RSS_U is a sum of squares, the denominator always will too.
- $F \rightarrow 0$ as $\text{RSS}_R - \text{RSS}_U \rightarrow 0$; that is, F gets smaller as the difference between the "fit" of the two models declines.
- Conversely, F grows when $\text{RSS}_R - \text{RSS}_U$ gets large, which in turn happens when the model in (7) fits much better than the one in (8).
- And F also gets bigger when – all else equal – the unrestricted model fits "better" (that is, when RSS_U is small).

Also – FYI (that's a pun... get it?) – the F-test for the significance of a single variable can be shown to be the same as the square of the t-statistic for that coefficient estimate.

Note finally that the value of F itself is more or less meaningless, except to the extent that larger ones are (*ceteris paribus*) "better" than smaller ones. Note as well that, because it is purely a function of the residuals of the two models, we can rewrite F in terms of the R^2s of the two models:

$$F = \frac{(R^2_U - R^2_R)/q}{(1 - R^2_U)/N - K} \tag{12}$$

As we'll see below, this formulation can be useful for testing other kinds of linear restrictions in models.

The Amazing Versatility and Usefulness of F

The F statistic is a very general way of testing for any number of linear restrictions on the coefficient estimates of an OLS model. Suppose, for example, that we had some theoretical reason to believe that sum of the effects of two of the variables in (7) on Y was equal to 1.0 – that is:

$$\text{H}_b : \beta_1 + \beta_4 = 1$$

This is a different form of linear restriction on the vector of coefficients $\hat{\boldsymbol{\beta}}$, in that it forces the estimates of $\hat{\beta}_1$ and $\hat{\beta}_4$ to lie on a particular horizon in the parameter space. We can rewrite this particular restriction equivalently as:

$$\mathrm{H}_b : \beta_1 = 1 - \beta_4$$

Call this restriction R'. With this restriction, we can rewrite the model in (7) as:

$$\begin{aligned}
Y_i &= \beta_0 + (1-\beta_4)X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_{R'i} \\
&= \beta_0 + X_{1i} - \beta_4 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_{R'i} \\
&= \beta_0 + X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 (X_{4i} - X_{1i}) + u_{R'i}
\end{aligned}$$

which leads to:

$$Y_i - X_{1i} = \beta_0 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 (X_{4i} - X_{1i}) + u_{R'i} \tag{13}$$

Now we can estimate the models in (7) and (13) and use the formula in (12) to calculate the F statistic associated with this particular restriction. The latter statistic will be distributed according to an F distribution with 1 and $N - K$ degrees of freedom.

Had we decided to impose an additional linear restriction (say, that $\beta_2 = \beta_3$, which is equivalent to $\beta_2 - \beta_3 = 0$), we could have transformed the model accordingly and calculated that F-test as well. We'll get to an example a little later.

Confidence Regions

Just as we discussed confidence intervals in the bivariate case, we can consider *confidence ellipses* in the multivariate one. Confidence ellipses are a way of displaying the estimated value and confidence intervals of two model parameters (that is, two $\hat{\beta}$s) simultaneously in a two-dimensional space:

- Each axis represents one value of $\hat{\beta}_k$,
- a central point indicates the estimates of $\hat{\beta}$ for each of the two variables, and
- an ellipse denotes the 95% confidence region around those estimates.

Creating such regions allows us to consider pairs of possible values for $\hat{\boldsymbol{\beta}}$. That is, for two possible values of $\hat{\beta}_j$ and $\hat{\beta}_k$, any value of the pair falling outside the (say) 95% confidence region would be rejected. Put differently, the region is constructed such that, in repeated random sampling, a certain percentage of those regions will contain the true parameter values.

Formally, we can think of this as an inversion of the F-test we just discussed. Call q the number of restrictions we are interested in testing (so, in the example immediately above, $q = 1$), and let $\hat{\boldsymbol{\beta}}_q$ be the q-length vector of parameter estimates for the variables of interest and $\hat{\mathbf{V}}_q$ is the $q \times q$ matrix of the extracted elements of $\hat{\mathbf{V}}$ associated with the q variables of interest. With this notation, it can be shown that another way of thinking about (and calculating) an F-test is as:

$$F = \frac{(\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q^H)'\hat{\mathbf{V}}_q^{-1}(\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q^H)}{q\hat{\sigma}^2} \tag{14}$$

where $\boldsymbol{\beta}_q^H$ is a vector of hypothesized values for $\boldsymbol{\beta}_q$. Under the null hypothesis, this test statistic is (like all the rest) distributed according to an F distribution with q and $N-K$ degrees of freedom. In the simplest case, for example, we might want to consider the (null) hypothesis that the effect of some set of q variables in $\boldsymbol{\beta}$ is jointly zero. This corresponds to $\boldsymbol{\beta}_q^H = \mathbf{0}$, so that Equation (14) becomes

$$F = \frac{\hat{\boldsymbol{\beta}}_q'\hat{\mathbf{V}}_q^{-1}\hat{\boldsymbol{\beta}}_q}{q\hat{\sigma}^2}.$$

As Fox notes in his textbook, if α is our desired confidence level, then it must be the case that:

$$\Pr\left[\frac{(\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q^H)'\hat{\mathbf{V}}_q^{-1}(\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q^H)}{q\hat{\sigma}^2} \leq F_{q,N-K}\right] = 1-\alpha. \tag{15}$$

This equality can be used to construct a "region" around our parameter estimates $\hat{\boldsymbol{\beta}}$; that region consists of all points in the parameter space that satisfy:

$$(\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q^H)'\hat{\mathbf{V}}_q^{-1}(\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q^H) \leq q\hat{\sigma}^2 F_{q,N-K}.$$

Most commonly, this is done by testing against the null alternative that $\boldsymbol{\beta} = \mathbf{0} \, \forall \, k \in K$ – that is, considering the "null model" F-test described above; the region in question is then an ellipse in K-dimensional space around $\hat{\boldsymbol{\beta}}$. This is, in its entirety, a bit hard to get one's brain around, so we normally restrict our examination of it to two variables at a time.

The main value of such plots is that they account for the fact that the more highly two variables' estimates of $\hat{\beta}$ are correlated, the more "dependent" their confidence intervals will be. If the two variables are uncorrelated, the ellipse will be perpendicular/parallel to the X- and Y-axes of the plot. The greater the correlation of the two, the more correlated (diagonal) the confidence ellipse around the parameter estimates will be, *but in the opposite direction of the correlation between the Xs.*

Multivariate Prediction

It's easy to generate a prediction from a multivariate regression model. For an observation (either in- or out-of-sample) with a specific covariate vector $\mathbf{X}_j$, the predicted value of Y is just:

$$\hat{Y}_j = \mathbf{X}_j\hat{\boldsymbol{\beta}} \tag{16}$$

Equally important as the prediction itself is the variability of that prediction. As in the scalar case, since the prediction is based on random variables (that is, the estimated $\hat{\beta}$s), it also has some variability, and that variability needs to be accounted for when talking about predicted values.

So, we need to come up with the variance of the estimated $\hat{Y}_j$s. And we know that that variance is a function of (a) the amount of variability in $\mathbf{X}_j$, (b) the variability in the $\hat{\beta}$s, and (c) the random variation in $\mathbf{Y}$ itself (that is, the variation in $\mathbf{u}$). So,

- Q: What is the variation (and covariation) in $\mathbf{X}_j$?

  A: $\mathbf{X}_j\mathbf{X}_j'$.

- Q: What is the estimated variability in the $\hat{\beta}$s?

  A: $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$.

- Q: What is the estimated variation in $\mathbf{u}$?

  A: $\hat{\sigma}^2$.

Combining these terms, the variance of the prediction $\hat{Y}_j$ is:

$$\widehat{\mathbf{V}(\hat{Y}_j)} = \hat{\sigma}^2[1 + \mathbf{X}_j(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_j'] \tag{17}$$

What does this mean?

- The term $\mathbf{X}_j(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_j'$ is sometimes called the *leverage* of observation j...
  - $\mathbf{X}'\mathbf{X}$ is the $K \times K$ variance-covariance matrix of *all* the $\mathbf{X}$s.
  - Multiplying the two together yields a *scalar* that is also sometimes called the "diagonal projection of the 'hat' matrix."
  - Inverting $\mathbf{X}'\mathbf{X}$ means that we're "discounting" the variability in $\mathbf{X}_j$ depending on the amount of overall variability in $\mathbf{X}$.
  - So, the greater the variability in $\mathbf{X}$, the smaller our prediction variance (all other things equal).

  Among other things leverage can be useful for detecting outliers – observations that have a disproportionate influence on our estimates (more on this later in the term...).

- The estimated $\hat{\sigma}^2$ is the variability of the errors – that is, a measure of the overall "fit" of the model.
  - As this increases, the variability of our prediction goes up.

- ○ Makes sense: Large $\hat{\sigma}^2$ implies that the model is not doing a very good job of "explaining" Y, and so we can't make very precise predictions of $\hat{Y}_j$.

The standard error of the prediction is just the square root of $\widehat{\mathbf{V}(\hat{Y}_j)}$:

$$\widehat{\text{s.e.}(\hat{Y}_j)} = \sqrt{\hat{\sigma}^2[1 + \mathbf{X}_j(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_j']} \tag{18}$$

This is useful when we want to draw confidence intervals around our predicted values, as we should always do. Those, in turn, can tell us whether our predictions are statistically different from some particular value of interest.

Example: Africa Data

The slides illustrate an analysis of the Africa data ($N = 42$), where we model the *Adult HIV Prevalence Rate* as a function of *POLITY*, an indicator for *subsaharan* countries, the *percent Muslim* in the population, and each country's *literacy* rate. The example illustrates model fitting, hypothesis testing, generating confidence ellipses, and various uses of predicted values. The code is available on the github repo.