1 Testing

There are several main uses for hypothesis testing. Firstly, to evaluate the prediction of a scientific theory or some model thought to be true or approximately so: for example, that the interest elasticity of demand for money is zero or that aggregate production is constant returns to scale. Secondly, we may wish to establish whether certain assumed properties of the data hold or not. Put another way, we want to test whether our specified model is correct. Examples of misspecifications include omitted variables, wrong functional form, distributional specification, exogeneity, parameter variation, serial correlation, and heteroskedasticity. Incorrect specification may lead to inefficient estimates or even worse inconsistent estimates. This latter outcome is very common in the sort of nonlinear models used in Microeconometrics and is one good reason for carrying out specification tests.[1] A final application of testing is as part of model building process; this is often given the name pretesting.

Suppose that we have data $y = (y_1, \ldots, y_n)$, and a hypothesis $\mathbf{H}_0$ about its density $f \in \mathcal{F}$. An hypothesis is simple if it completely specifies $f(\cdot)$ and is composite otherwise, i.e., when there are nuisance parameters not referred to by $\mathbf{H}_0$. We also distinguish between single and multiple restrictions and whether the alternative $\mathbf{H}_A$ is one-sided or two-sided. General Notations (Fisher, Neyman–Pearson). Want to test $\mathbf{H}_0$ vs. $\mathbf{H}_A$. Let T (data) be a test statistic, typically scalar valued. Choose $\alpha \in [0,1]$ size of test. Find C_α critical region of size α such that $\text{pr}[T \notin C_\alpha | \mathbf{H}_0] = \alpha$. Rule is to reject $\mathbf{H}_0$ if (exact or approximate) $T \notin C_\alpha$ (choose T so that C_α is easy to find). Type I error is to reject a true hypothesis. Type II error is of accepting a false hypothesis. Power of test: $\pi = \text{pr}[T \notin C_\alpha | \mathbf{H}_A] = 1 -$ Type II error. An unbiased test has $\pi \geq \alpha$. The p–value of a test T is

$$\alpha_{\text{obs}} = \text{pr}[T \geq T_{\text{obs}} | \mathbf{H}_0],$$

where the subscript obs refers to observed values. In two-sided case, take $|T| \geq |T_{\text{obs}}|$. Low α_{obs} is evidence against the null hypothesis. Confidence interval is just critical region centered not at $\mathbf{H}_0$, but at a function of parameter estimates.

There are some critiques of hypothesis testing procedures

- It is just a primitive decision-making apparatus; can't we do better?
- α is arbitrary. Some argue that α should be made a function of sample size, because in practice as n gets large almost any hypothesis is rejected. In reality hypotheses are neighborhoods not points.

[1] or using estimation procedures robust to such misspecifications.

2 Likelihood Tests

Suppose we have a likelihood $L_n(y;\theta)$ and a null hypothesis concerning the p-dimensional parameters:

$$\mathbf{H}_0 \, : \, r(\theta) = 0_{J\times 1}.$$

It is often the case that θ can be partitioned into θ_1 and θ_2, where the null hypothesis specifies some aspects of θ_2, for example, $\theta_2 = 0$, but leaves θ_1 unrestricted. The alternative may be two-sided, i.e.,

$$\mathbf{H}_A \, : \, r(\theta) \neq 0$$

or one-sided, e.g.

$$\mathbf{H}_A \, : \, r(\theta) \geq 0.$$

When $J > 1$, the theory of testing against one-sided alternatives is complicated and relatively incomplete relative to that for two-sided alternatives, see Andrews (1993).

Let $\widehat{\theta}$ be the Maximum likelihood estimator of θ and let θ^* be the restricted MLE subject to $r(\theta) = 0$. For the general case $J > 1$, we use the non-directional test statistics

$$\begin{aligned} \text{Log} - \text{Likelihood Ratio} &= 2\{l_n(\widehat{\theta}) - l_n(\theta^*)\} \\ \text{Wald} &= \widehat{r}' \left\{ \widehat{R}\mathcal{I}(\widehat{\theta})^{-1}\widehat{R}' \right\}^{-1} \widehat{r} \\ \text{Score Test} &= \frac{\partial l_n}{\partial \theta'}(\theta^*)\mathcal{I}^{-1}(\theta_*)\frac{\partial l_n}{\partial \theta}(\theta^*) \end{aligned}$$

where $\widehat{r} = r(\widehat{\theta})$ and $\widehat{R} = \frac{\partial r}{\partial \theta}(\widehat{\theta})$. The Score test can be rewritten in terms of the Lagrange multiplier using that $\frac{\partial l_n}{\partial \theta}(\theta^*) = R^*\lambda^*$, where $R^* = \frac{\partial r}{\partial \theta}(\theta^*)$, which is why it is often called the LM test. Note that in the last two tests, $\mathcal{I}(\widehat{\theta})$ and $\mathcal{I}(\theta_*)$ can be replaced by any consistent (under the null) estimate thereof. Thus one can evaluate at $\widehat{\theta}$ and θ_* with little or no effect on the outcome as we see below. Also, the expected information matrix $\mathcal{I}$, which is often hard to calculate, can be replaced by the observed information $-\frac{\partial^2 l_n}{\partial \theta \partial \theta'}$; with i.i.d data one can also use the outer product of the scores $\sum_i \frac{\partial l_i}{\partial \theta}\frac{\partial l_i}{\partial \theta'}$. One may wish to protect the test against partial misspecification, such as omitted heteroskedasticity by using robust estimates of the asymptotic variance matrix as in the following robust version of the Wald test

$$\text{Wald}_* = \widehat{r}' \left[\widehat{R} \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial \theta \partial \theta'}(\widehat{\theta}) \right\}^{-1} \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta}\frac{\partial l_i}{\partial \theta'}(\widehat{\theta}) \right\} \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial \theta \partial \theta'}(\widehat{\theta}) \right\}^{-1} \widehat{R}' \right]^{-1} \widehat{r}.$$

Under the null hypothesis,

$$\mathrm{pr}\left(T \leq x\right) \rightarrow \mathrm{pr}\left(\chi^2_J \leq x\right), \quad \text{as } n \rightarrow \infty,$$

where χ^2_J is a chi-squared random variable with J degrees of freedom. Here, T refers to any of the above test statistics. The above approximation is used to calculate approximate critical values.

When $J = 1$, one can also test specifically against the one-sided alternative by using

$$\begin{aligned} \text{Signed Log} - \text{Likelihood Ratio} &= sign\left\{r(\widehat{\theta}) - r(\theta^*)\right\}\left|2\left\{l_n(\widehat{\theta}) - l_n(\theta^*)\right\}\right|^{1/2} \\ \mathrm{t} &= \widehat{r}\left\{\widehat{R}\mathcal{I}(\widehat{\theta})^{-1}\widehat{R}'\right\}^{-1/2}, \end{aligned}$$

which are both asymptotically $N(0,1)$ under the null hypothesis.

The Wald statistic is advantageous when the unrestricted model is simpler to estimate, as with complicated cross equation restrictions in a simultaneous equation model, for example. The LM test is simpler when the restricted model is simpler to estimate, as in a test for homoskedasticity.[2] The likelihood ratio is invariant to reparameterizations which property is not shared by the other tests. Note that we can always re-express nonlinear hypotheses in many different ways. For example, $\theta^2 = 1$ is equivalent to $\theta^4 = 1$. The Wald statistics corresponding to these alternative representations are different, and although they lead to the same results asymptotically, the small sample difference between them can be substantial, see Phillips and Park (1988).

The usual likelihood-based approach to specification testing is to embed the chosen model in a larger parametric model which has a set of departures from the null model. For example, if serial correlation is suspected in a regression model, one might choose as a general model a regression whose errors follow an AR(1) process. Inside this general model, the null hypothesis that the autoregression parameter is zero corresponds to a correctly specified initial model. This particular test is most conveniently carried out by means of Lagrange Multiplier test. Tests for homoskedasticity can be constructed similarly. See Harvey (1981), Engle (1984) and Godfrey (1988) for many examples.

Example. Linear regression with linear hypothesis, $R\beta = r$. Let $\theta = (\beta, \sigma^2)$ and

$$\begin{aligned} l_n(\theta) &= \frac{-n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}u(\beta)'u(\beta) \\ \frac{\partial l_n}{\partial\beta}(\theta) &= \frac{1}{\sigma^2}X'u(\beta) \\ \frac{\partial^2 l_n}{\partial\beta\partial\beta'}(\theta) &= \frac{-1}{\sigma^2}X'X, \end{aligned}$$

[2]The LM test can often be written as nR^2, where R^2 is from a certain regression.

where $u(\beta) = y - X\beta$. Let $Q = u(\widehat{\beta})'u(\widehat{\beta})$ and $Q^* = u(\beta^*)'u(\beta^*)$.

The Wald test is

$$W = (R\widehat{\beta} - r)' \left\{ R(X'X)^{-1}R'\widehat{\sigma}^2 \right\}^{-1} (R\widehat{\beta} - r) = \frac{Q^* - Q}{(Q/n)}$$

The Score test is

$$\frac{u^{*\prime} X}{\sigma^{*2}} \left\{ \frac{X'X}{\sigma^{*2}} \right\}^{-1} \frac{X'u^*}{\sigma^{*2}} = \frac{Q^* - Q}{(Q^*/n)}$$

The Likelihood Ratio test is

$$n \left\{ \ln(Q^*) - \ln(Q) \right\} = n \ln \left\{ 1 + \frac{Q^* - Q}{Q} \right\}$$

3 GMM Tests

There is a corresponding trinity of tests based on the GMM criterion, see Newey and McFadden (1994). As before, let $\widehat{\theta}$ and θ^* refer to unrestricted and restricted estimates respectively and let $Q_n(\theta)$ be the optimal GMM criterion defined above. Then

$$\begin{aligned} \text{Distance Metric} &= n\{Q_n(\widehat{\theta}) - Q_n(\theta^*)\} \\ \text{Wald} &= \widehat{r}' \left\{ \widehat{R} \left(\widehat{D}'\widehat{S}^{-1}\widehat{D} \right)^{-1} \widehat{R}' \right\}^{-1} \widehat{r} \\ \text{LM} &= \frac{\partial Q_n}{\partial \theta'}(\theta^*) \widehat{D}'\widehat{S}^{-1}\widehat{D} \frac{\partial Q_n}{\partial \theta}(\theta^*), \end{aligned}$$

where $\left(\widehat{D}'\widehat{S}^{-1}\widehat{D} \right)^{-1}$ is an estimate of the asymptotic variance matrix of the unrestricted GMM estimator. Under the null hypothesis,

$$\text{pr}(T \leq x) \rightarrow \text{pr}\left(\chi^2_J \leq x \right), \quad \text{as } n \rightarrow \infty,$$

where χ^2_J is a chi-squared random variable with J degrees of freedom and T refers to any of the above test statistics.

An important example here is in testing the overidentifying restrictions. In this case the restricted estimate uses only the specified subset that exactly identifies the system, and $Q_n(\theta^*) = 0$. Therefore, the test is just based on the value of the criterion function at the estimated values and the degrees of freedom is equal to the number of overidentifying restrictions.

4 Other Tests

4.1 Hausman Test

This is a widely used specification test. It is applicable in many situations, it is simple to compute, and appears to require less structure than the likelihood-based tests. Let $\widehat{\theta}$ and $\widetilde{\theta}$ be two different $p \times 1$ random vectors (possibly estimating the parameters θ). Suppose there are two states of the world which we label, not coincidentally, $\mathbf{H}_0$ and $\mathbf{H}_A$, and suppose that there exist possibly random $p \times p$ matrices A_n, B_n, with

$$\widehat{\theta} - \widetilde{\theta} = O_p(A_n), \quad \text{under } \mathbf{H}_0,$$

and

$$\widehat{\theta} - \widetilde{\theta} = O_p(B_n), \quad \text{under } \mathbf{H}_A,$$

where the $O_p(.)$ notation means that, for example, $A_n^{-1}(\widehat{\theta} - \widetilde{\theta})$ is stochastically bounded, possibly converging in distribution to some random variable. Provided the rates of convergence A_n and B_n are different between the two states of the world, we should be able to distinguish between them. This general setup does not require the alternative states of nature to be parametrically specified.

In the usual case, $n^{1/2}(\widehat{\theta} - \widetilde{\theta}) \xrightarrow{D} N(0, V)$ under $\mathbf{H}_0$, and $\widehat{\theta} - \widetilde{\theta} = O_p(1)$ under $\mathbf{H}_A$. Let

$$T = n(\widehat{\theta} - \widetilde{\theta})' \widehat{V}^{-} (\widehat{\theta} - \widetilde{\theta}),$$

where $\widehat{V}$ is a consistent estimate of V under $\mathbf{H}_0$ and the minus superscript denotes a generalized inverse. Then, T is widely called a Hausman or Hausman-Wu test after the two papers of these authors. The distribution of T can be approximated, in large samples, by χ_q^2, where q is the rank of V. Usually, one selects subvectors or appropriate linear combinations such that V is of full rank, although the above result applies to the more general case. When one estimate, for example $\widehat{\theta}$, is efficient under the null hypothesis, then the asymptotic variance of the scaled difference is actually the difference of the asymptotic variances of the two estimates under the null, i.e., $V = V_2 - V_1$, where $n^{1/2}(\widehat{\theta} - \theta) \xrightarrow{D} N(0, V_1)$ and $n^{1/2}(\widetilde{\theta} - \theta) \xrightarrow{D} N(0, V_2)$. This is computationally advantageous, because otherwise one would have to also calculate the covariance between the two estimates.

Note that the setting here is very general and no parametric model is explicitly needed, only the separation property. This test is widely used in panel data analysis to discriminate between the situations where the uncorrelated random effects and fixed effects procedures are appropriate.

4.2 Information Matrix Test

White (1987) proposed making a specification test out of the information matrix equality. Specifically, if the likelihood is misspecified in a substantive way, then the information matrix equality is violated. Again, the precise form of departure is not specified here. A test is made from

$$T = \left\{ n^{-1/2} \sum_{i=1}^{n} \widehat{q}_i \right\}' \left\{ \frac{1}{n} \sum_{i=1}^{n} \widehat{q}_i \widehat{q}_i' \right\}^{-} \left\{ n^{-1/2} \sum_{i=1}^{n} \widehat{q}_i \right\},$$

where

$$\widehat{q}_i = \text{vech} \left\{ \frac{\partial^2 l_i}{\partial \theta \partial \theta'}(\widehat{\theta}) + \frac{\partial l_i}{\partial \theta} \frac{\partial l_i}{\partial \theta'}(\widehat{\theta}) \right\}$$

is a $J = p(p+1)/2$ by 1 vector. Critical values are taken from a χ^2_J distribution. Some practical problems can arise due when p is large due to the singularity of $\frac{1}{n} \sum_{i=1}^{n} \widehat{q}_i \widehat{q}_i'$.

4.3 Distribution Free Tests

In some cases one wants to make simple comparisons between two variables. For example, you may want to test for differences such as stochastic dominance between one income distribution from another. A similar type of problem is to test for independence of two random variables. These hypotheses can be stated in general terms and do not require some parametric specification such as normality. They can also be tested outside of the usual parametric framework considered previously. Let $y_1, \ldots, y_n$ be i.i.d continuous random variables. Define the ranks $R_1, \ldots, R_n$, where R_j is the integer rank (in decreasing order) of y_j in the sample, and the empirical distribution function

$$F_n(y) = n^{-1} \sum_{i=1}^{n} 1 \left\{ y_i \leq y \right\}.$$

The empirical distribution function consistently estimates the population distribution F.

Distribution-free tests are often based on either the ranks or the empirical distribution function, see Cox and Hinkley (1975).

1. Testing for dominance of one distribution. Suppose that $y_1, \ldots, y_{n_1}, y_{n_1+1}, \ldots, y_n$, where $n = n_1 + n_2$, are available, and that it is believed that the population distribution from which the first n_1 observations were sampled dominates the distribution of the succeeding n_2 observations. Then we expect that the ranks of $y_1, \ldots, y_{n_1}$ in the full sample to be small. A general class of test statistics based on weighted ranks

$$T = \sum_{j=1}^{n} w_j R_j$$

are used in such situations. The advantage of working with ranks is primarily robustness to distributional specification.

2. Testing for a specified distribution such as normality. Define the Kolmogorov-Smirnov and Cramér-von Mises tests

$$T_1 = n^{1/2} \sup_y |F_n(y) - F_0(y)| \quad ; \quad T_2 = n \int \{F_n(y) - F_0(y)\}^2 \, dy,$$

where F_0 is the specified distribution, e.g. the standard normal c.d.f. In this latter case, we first scale the data $y_i \to \frac{y_i - \overline{y}}{s}$, where $\overline{y}$ is the sample mean and s^2 is the sample variance. The asymptotic null distributions of the these statistics are nonstandard but are tabulated. These tests have power against *all* alternative distributions, unlike the parametric - for example, with a normal null one might use a combination of the third and fourth cumulants - tests

3. Testing for independence. Our null hypothesis is that Y is independent of X, but their distributions are otherwise unspecified. This is a fully nonparametric hypothesis. Let

$$T_1 = n^{1/2} \sup_{y,x} |G_n(y,x) - F_n(y)H_n(x)| \quad ; \quad T_2 = n \int \{G_n(y,x) - F_n(y)H_n(x)\}^2 \, dydx,$$

where G_n is the joint empirical distribution of Y and X, and H_n is the empirical distribution of X. When Y and X are independent, T_1 and T_2 have a limiting distribution, but under the alternative they tend to infinity. The asymptotic null distributions of the these statistics are nonstandard but are tabulated.

5 Optimality

One general optimality criterion is maximum power controlling for size. There is a classical theory that applies only to some very special cases: the Neyman-Pearson optimality theory for likelihood ratio test. A special criterion which is widely used criterion is local asymptotic optimality. Suppose that the null hypothesis is that $\theta_2 = 0$, and consider the local alternative $\theta_{2n} = \delta n^{-1/2}$ for some constant δ. The likelihood-based tests all asymptotically achieve optimal power against such alternatives. See Cox and Hinkley (1975) for further discussion.

References

[1] Andrews, D.W.K. (1993): "Hypothesis testing with a restricted parameter space," *Cowles Foundation Discussion Paper* No. 1060.

[2] Cox, D.R., and D.V. Hinkley (1979) *Theoretical Statistics*, Cambridge: Chapman and Hall.

[3] Davidson, R., and J.G. Mackinnon (1993) *Estimation and Inference in Econometrics*, Oxford University Press.

[4] Engle, R.F. (1984): "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics," in *Handbook of Econometrics, Vol. 2.* ed. Z. Griliches and M.D. Intriligator. North Holland.

[5] Godfrey, L.G. (1988): *Misspecification Tests in Econometrics*, Econometric Society Monograph No. 16, Cambridge University Press.

[6] Harvey, A.C. (1981): *Econometric Analysis of Time Series,* Phillip Allan.

[7] Hausman, J.A. (1978): "Specification tests in econometrics," *Econometrica* **46**, 1251-1271.

[8] Newey, W.K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *The Handbook of Econometrics*, Vol. 4, pp2111-2245.

[9] Phillips, P.C.B., and J. Park (1988): "On the formulation of Wald tests of nonlinear restrictions," *Econometrica* **56**, 1065-83.

[10] White, H. (1987): "Specification testing in dynamic models," Ch. 1 in *Advances in Econometrics - Fifth World Congress*, Vol 1, ed T. Bewley, Cambridge University Press.

[11] Wu, D. (1973): "Alternative tests of independence between stochastic regressors and disturbances," *Econometrica* **41**, 733-750.