

Ec241a

Econometrics

Spring 2004

UC Berkeley Department of Economics

MAXIMUM LIKELIHOOD ESTIMATION III:
CLASSICAL TESTS AND HAUSMAN-WU TESTS

A. CLASSICAL TESTS

After estimating the exponential model for the unemployment durations, Lancaster considers an extension. Consider the hazard function or escape rate

$$h(y|x, \theta) = \lim_{\Delta \rightarrow 0} Pr(y \leq Y < y + \Delta | y \leq Y, X) / \Delta = \frac{f(y|x, \theta)}{1 - F(y|x, \theta)}.$$

The hazard function is just another way of characterizing a distribution, like the density function, the distribution function, the survivor function, the moment generating function or the characteristic function. It is just a particularly convenient and interpretable way of describing a distribution or durations. Given the hazard you can calculate the distribution function as

$$F(y|x, \theta) = 1 - \exp\left(-\int_0^y h(s|x, \theta) ds\right),$$

and hence the density function. The exponential model implies that the hazard function stays constant over the duration of the spell, equal to $\exp(x'\beta)$ in our previous specification. To see what this means, take a person and look at their chances of finding a job on the first day of being unemployed. These chances are the same as the chances that this same person would find a job on the fiftieth day given that he has been unsuccessful in finding work in the first forty-nine days. This may be reasonable, but it might also be something you do not wish to impose from the outset. Lancaster therefore considers an extension allowing the hazard function to either increase, stay constant, or decrease over time. This extension is known as the Weibull distribution:

$$h(y|x, \beta, \alpha) = (\alpha + 1) \cdot y^\alpha \exp(x'\beta).$$

Note that this reduces to the exponential distribution if $\alpha = 0$. The implied density function for the Weibull distribution is

$$f(y|x, \beta, \alpha) = (\alpha + 1) \cdot y^\alpha \exp(x'\beta) \exp\left(-y^{\alpha+1} \exp(x'\beta)\right).$$

The moments of this distribution are

$$E[Y^k|X] = \exp\left(-k \cdot x'\beta / (\alpha + 1)\right) \cdot \Gamma\left(\frac{k+1}{\alpha+1}\right),$$

where $\Gamma(a) = \int_0^\infty y^{a-1} \exp(-y) dy$. (Note that for the case with $\alpha = 0$ this reduces to the exponential case with

$$E[Y^k|X] = \exp(-k \cdot x'\beta) \cdot \Gamma(1+k),$$

and thus with $k = 1$ the mean of the exponential distribution is $E[Y|X] = \exp(-x'\beta)$.)

One can estimate this model using any of the numerical methods described before (Newton-Raphson, Davidon-Fletcher-Powell). The one (minor) complication is that numerical algorithms have to take account of the restriction that $\alpha > -1$; with $\alpha = -1$ the density is degenerate and all probability mass piles up at $y = 0$.

To test the hypothesis $\alpha = 0$ against the alternative hypothesis $\alpha \neq 0$ there are three classical tests, the likelihood ratio, the Wald and the score test. We shall consider the three tests in a general context. Suppose we have a model for a random variable Z , specifying the density function

$$f(z|\theta_0, \theta_1),$$

where we split the parameter vector θ into two parts, $\theta = (\theta'_0, \theta'_1)'$. The dimension of θ_0 is K_0 , the dimension of θ_1 is K_1 , and the dimension of θ is $K = K_0 + K_1$. We are interested in testing the null hypothesis

$$H_0 : \theta_0 = 0,$$

against the alternative hypothesis

$$H_1 : \theta_0 \neq 0.$$

Let $\hat{\theta}_u = (\hat{\theta}_{0u}, \hat{\theta}_{1u})$ denote the unrestricted mle's, and $\hat{\theta}_r = (\hat{\theta}_{r0}, \hat{\theta}_{r1}) = (0, \hat{\theta}_{r1})$ denote the estimates based on the restricted model, that is, based on the restriction $\theta_0 = 0$. Let $L(\theta_0, \theta_1)$ denote the log likelihood function

$$L(\theta_0, \theta_1) = \sum_{i=1}^N \ln f(z_i | \theta_0, \theta_1),$$

so

$$\hat{\theta}_{r1} = \operatorname{argmax}_{\theta_1} L(0, \theta_1),$$

and

$$(\hat{\theta}_{u0}, \hat{\theta}_{u1}) = \operatorname{argmax}_{\theta_0, \theta_1} L(\theta_0, \theta_1).$$

In addition let $\mathcal{S}(z, \theta_0, \theta_1)$ denote the score function:

$$\mathcal{S}(z, \theta_0, \theta_1) = \begin{pmatrix} \frac{\partial \ln f}{\partial \theta_0}(z | \theta_0, \theta_1) \\ \frac{\partial \ln f}{\partial \theta_1}(z | \theta_0, \theta_1) \end{pmatrix},$$

let $\mathcal{H}(z, \theta_0, \theta_1)$ be the Hessian:

$$\mathcal{H}(z, \theta_0, \theta_1) = \begin{pmatrix} \frac{\partial^2 \ln f}{\partial \theta_0 \partial \theta_0}(z | \theta_0, \theta_1) & \frac{\partial^2 \ln f}{\partial \theta_0 \partial \theta_1}(z | \theta_0, \theta_1) \\ \frac{\partial^2 \ln f}{\partial \theta_1 \partial \theta_0}(z | \theta_0, \theta_1) & \frac{\partial^2 \ln f}{\partial \theta_1 \partial \theta_1}(z | \theta_0, \theta_1) \end{pmatrix},$$

and, finally, let $\mathcal{I}(\theta_0, \theta_1)$ be the information matrix evaluated at (θ_0, θ_1) :

$$\mathcal{I}(\theta_0, \theta_1) = -E[\mathcal{H}(Z, \theta_0, \theta_1)].$$

We will use various estimates of $\mathcal{I}(\theta_0^*, \theta_1^*)$, depending on where we evaluate the matrix and how we calculate or approximate the expectation. For the second, there are three choices:

1. Use the average of the second derivatives:

$$\mathcal{I}(\theta) = -\frac{1}{N} \sum_{i=1}^N \mathcal{H}(z_i, \theta).$$

2. Use the average of the outer product of the first derivatives:

$$\mathcal{I}(\theta) = \frac{1}{N} \sum_{i=1}^N \mathcal{S}(z_i, \theta) \mathcal{S}(z_i, \theta)'$$

3. Use the expectation of the first or second derivatives:

$$\mathcal{I}(\theta) = -E[\mathcal{H}(Z, \theta)] = E[\mathcal{S}(Z, \theta)\mathcal{S}(Z, \theta)'].$$

For the first choice the possibilities are to evaluate the estimate at

1. The restricted estimates $\hat{\theta}_r$.
2. The unrestricted estimates $\hat{\theta}_u$.

The leading choices include the average of the second derivative evaluated at the unrestricted estimates:

$$\hat{\mathcal{I}}_1 = -\frac{1}{N} \sum_{i=1}^N \mathcal{H}(z_i, \hat{\theta}_u),$$

or the expectation itself evaluated at the restricted estimates:

$$\hat{\mathcal{I}}_2 = -E\mathcal{H}(Z, \hat{\theta}_r),$$

or estimates based on the first derivatives:

$$\hat{\mathcal{I}}_3 = \frac{1}{N} \sum_{i=1}^N \mathcal{S}(z_i, \theta_r)\mathcal{S}(z_i, \theta_r)',$$

or

$$\hat{\mathcal{I}}_4 = \frac{1}{N} \sum_{i=1}^N \mathcal{S}(z_i, \theta_u)\mathcal{S}(z_i, \theta_u)'$$

All three classical tests are based on the quadratic approximation to the log likelihood function around the true (and therefore maximizing in the limit) values (θ_0^*, θ_1^*) . The three are first-order equivalent, meaning that if the null hypothesis is correct, their difference, multiplied by N , converges to zero in probability ($N \cdot (LR - LM) \xrightarrow{p} 0$, $N \cdot (LR - WALD) \xrightarrow{p} 0$, and $N \cdot (LM - WALD) \xrightarrow{p} 0$).

First, if the null hypothesis is true and $\theta_0^* = 0$, the value of the log likelihood function at $(\theta_{0r} = 0, \theta_{1r})$ should not be much smaller than the value of the log likelihood function at $(\theta_{0u}, \theta_{1u})$. This is the basis of the Likelihood Ratio Test. Formally, the test statistic is

$$LR = 2 \cdot \left(L(\hat{\theta}_{0u}, \hat{\theta}_{1u}) - L(0, \hat{\theta}_{1r}) \right).$$

If the null hypothesis is true this statistic has for large N a Chi-squared distribution with degrees of freedom equal to the dimension of θ_0 .

Second, if the limiting log likelihood function is maximized at $\theta_0 = 0$, the derivative of the log likelihood function with respect to θ_0 at that point should be close to zero. This is the basis of Rao's Score Test or the Lagrange Multiplier Test. Formally,

$$LM = \frac{1}{N} \sum_{i=1}^N \mathcal{S}(z_i, 0, \hat{\theta}_{1r})' \cdot \hat{\mathcal{I}}^{-1} \cdot \sum_{i=1}^N \mathcal{S}(z_i, 0, \hat{\theta}_{1r}).$$

Because

$$\sum_{i=1}^N \mathcal{S}(z_i, 0, \hat{\theta}_{1r}) = \begin{pmatrix} \sum_{i=1}^N \mathcal{S}_0(z_i, 0, \hat{\theta}_{1r}) \\ 0 \end{pmatrix}$$

the LM statistic can also be written as

$$LM = \frac{1}{N} \sum_{i=1}^N \mathcal{S}_0(z_i, 0, \hat{\theta}_{1r})' \cdot \hat{\mathcal{I}}^{00} \cdot \sum_{i=1}^N \mathcal{S}_0(z_i, 0, \hat{\theta}_{1r}),$$

but that this is in general not equal to

$$\frac{1}{N} \sum_{i=1}^N \mathcal{S}_0(z_i, 0, \hat{\theta}_{1r})' \cdot \hat{\mathcal{I}}_{00}^{-1} \cdot \sum_{i=1}^N \mathcal{S}_0(z_i, 0, \hat{\theta}_{1r}).$$

Note that if θ_1 were known, then we would use $\hat{\mathcal{I}}_{00}^{-1}$ and we would use the test statistic

$$\frac{1}{N} \sum_{i=1}^N \mathcal{S}_0(z_i, 0, \theta_{1r})' \cdot \hat{\mathcal{I}}_{00}^{-1} \cdot \sum_{i=1}^N \mathcal{S}_0(z_i, 0, \theta_{1r}).$$

Any estimate of the information matrix can be used. Typically researchers do not use $\hat{\mathcal{I}}_2$ because it would require calculation of the unrestricted estimates, and the key advantage of the Lagrange Multiplier test is that it avoids calculation of the unrestricted estimates. There is some evidence that using the average of the second derivatives (and therefore $\hat{\mathcal{I}}_1$) is to be preferred over calculation of the expectation (i.e., $\hat{\mathcal{I}}_2$). If only a conditional density is specified calculation of the latter is difficult in any case because calculation of the expectation requires specification of the full density.

One particular form that is popular for the LM test is

$$LM = N \cdot \frac{\iota'_N \mathcal{S}(\mathcal{S}'\mathcal{S})^{-1} \mathcal{S}' \iota_N}{\iota'_N \iota_N} = \iota'_N \mathcal{S}(\mathcal{S}'\mathcal{S})^{-1} \mathcal{S}' \iota_N,$$

where ι_N is the N -vector of ones (so that $\iota'_N \iota_N = N$), and \mathcal{S} is the $N \times K$ matrix with the i th row equal to the score vector $\mathcal{S}(z_i, \hat{\theta}_r)'$ so that $\iota'_N \mathcal{S} = \sum_i \mathcal{S}(z_i, 0, \hat{\theta}_{1r})$. This form has the interpretation of N times the (uncentered) R^2 in a regression of a vector of ones (that is, ι_N) on the scores \mathcal{S} . The least squares coefficients are

$$\hat{\beta} = (\mathcal{S}'\mathcal{S})^{-1} \mathcal{S}' \iota_N,$$

and the R^2 is

$$R^2 = \frac{\hat{Y}'\hat{Y}}{Y'Y} = \frac{\iota'_N \hat{\iota}_N}{\iota'_N \iota_N} = \frac{\hat{\beta}' \mathcal{S}' \mathcal{S} \hat{\beta}}{\iota'_N \iota_N} = \frac{\iota'_N \mathcal{S}(\mathcal{S}'\mathcal{S})^{-1} \mathcal{S}' \iota_N}{\iota'_N \iota_N}.$$

Finally, the restricted and unrestricted estimates of θ_0 should be close together if the null hypothesis is correct, or in other words, the unrestricted estimate of θ_0 should be close to zero. This is the basis of the Wald Test. With θ_1 known the estimator for θ_0 would be approximately $\mathcal{N}(0, (\mathcal{I}_{00})^{-1})$. With θ_1 estimated this changes to $\mathcal{N}(0, \mathcal{I}^{00})$. Formally, the Wald test is defined as

$$W = N \cdot (\hat{\theta}_{0r} - \hat{\theta}_{0u})' \cdot (\hat{\mathcal{I}}^{00})^{-1} \cdot (\hat{\theta}_{0r} - \hat{\theta}_{0u}) = N \cdot \hat{\theta}'_{0u} \cdot (\hat{\mathcal{I}}^{00})^{-1} \cdot \hat{\theta}_{0u},$$

where $\hat{\mathcal{I}}^{00}$ is the top left part of the inverse of $\hat{\mathcal{I}}$. Again any of the estimates of the information matrix can be used. Here often the average evaluated at the unrestricted estimates, $\hat{\mathcal{I}}_2$ is used because of its superior properties if the null hypothesis is false.

Result 1 *As N goes to infinity, under the null hypothesis,*

$$LR \xrightarrow{d} \chi^2(\dim(\theta_0)).$$

Result 2 *As N goes to infinity, under the null hypothesis,*

$$LR - LM \xrightarrow{p} 0,$$

$$LR - W \xrightarrow{p} 0,$$

$$LM - W \xrightarrow{p} 0.$$

For formal proofs of these two results see, for example, Engle (1984) or Holly (1985).

Here is an informal argument for the case where θ_0 is a scalar and there are no nuisance parameters (no θ_1). Expand the log likelihood around the maximum likelihood estimate:

$$L(\theta) \approx L(\hat{\theta}) + \frac{\partial L}{\partial \theta}(\hat{\theta}) \cdot (\theta - \hat{\theta}) + \frac{1}{2} \frac{\partial^2 L}{\partial \theta^2}(\hat{\theta}) \cdot (\theta - \hat{\theta})^2.$$

The derivative of the log likelihood function at the maximum likelihood function is equal to zero, so

$$\begin{aligned} 2 \cdot \left(L(\hat{\theta}) - L(\theta^*) \right) &\approx (\hat{\theta} - \theta^*)^2 \cdot \frac{\partial^2 L}{\partial \theta^2}(\hat{\theta}) \\ &\approx N \cdot (\hat{\theta} - \theta^*)^2 \cdot \mathcal{I}(\theta^*). \end{aligned}$$

This is clearly very close to the test statistic from the Wald test. Given the limiting distribution

$$\sqrt{N} \cdot (\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}(\theta^*)),$$

the limiting chi-squared distribution follows immediately.

To see the link with the score or Lagrange multiplier test, expand the derivative of the log likelihood function around the true value:

$$\frac{\partial L}{\partial \theta}(\theta) \approx \frac{\partial L}{\partial \theta}(\theta^*) + (\theta - \theta^*) \cdot \frac{\partial^2 L}{\partial \theta^2}(\theta^*).$$

Evaluating this at $\theta = \hat{\theta}$, so that the first derivative is equal to zero,

$$0 \approx \frac{\partial L}{\partial \theta}(\theta^*) + (\hat{\theta} - \theta^*) \cdot \frac{\partial^2 L}{\partial \theta^2}(\theta^*),$$

and thus

$$(\hat{\theta} - \theta^*) \approx -\frac{\partial^2 L}{\partial \theta^2}(\theta^*)^{-1} \cdot \frac{\partial L}{\partial \theta}(\theta^*).$$

Renormalizing this gives

$$\sqrt{N}(\hat{\theta} - \theta^*) \approx \mathcal{I}(\theta^*)^{-1} \cdot \frac{1}{\sqrt{N}} \sum_{i=1}^N \mathcal{S}(z_i, \theta^*),$$

which by a central limit theorem has a limiting normal distribution with mean zero and variance equal to $\mathcal{I}(\theta^*)^{-1}$. Then, squaring both sides and multiplying by the information matrix gives

$$N \cdot (\hat{\theta} - \theta^*) \cdot \mathcal{I}(\theta^*) \approx \mathcal{I}(\theta^*)^{-1} \cdot \frac{1}{N} \left(\sum_{i=1}^N \mathcal{S}(z_i, \theta^*) \right)^2,$$

demonstrating the approximate equality of the Wald and Lagrange multiplier or score tests.

B. HAUSMAN-WU TESTS

Next, we consider a fourth way of testing hypothesis in the same context as before. We have a model for a random variable Z , specifying the density function

$$f(z|\theta_0, \theta_1),$$

where we split the parameter vector θ into two parts, $\theta = (\theta'_0, \theta'_1)'$. We are interested in testing the null hypothesis

$$H_0 : \theta_0 = 0,$$

against the alternative hypothesis

$$H_1 : \theta_0 \neq 0.$$

Let $\hat{\theta}_{u0}$ and $\hat{\theta}_{u1}$ denote the unrestricted mle's, and $\hat{\theta}_{r0} = 0$ and $\hat{\theta}_{r1}$ denote the restricted mle's that is, conditional on the restriction $\theta_0 = 0$.

The test we consider compares the estimates of the parameter not affected by the restriction, $\hat{\theta}_{1r}$ and $\hat{\theta}_{1u}$. If the null hypothesis is true, the two are estimating the same thing, θ_1^* and should be close to each other. The restricted estimate should be more precise because it exploits a true restriction. If the null hypothesis is false, it is likely the two estimators are estimating different things, and there is likely to be a larger difference between them. This is the basis of the Hausman–Wu Test.

Consider the unrestricted maximum likelihood estimates $\hat{\theta}_u = (\hat{\theta}'_{u0}, \hat{\theta}'_{u1})'$. In large samples:

$$\sqrt{N} \begin{pmatrix} \hat{\theta}_{u0} - \theta_0^* \\ \hat{\theta}_{u1} - \theta_1^* \end{pmatrix} \xrightarrow{d} \mathcal{N} \left(0, \begin{pmatrix} \mathcal{I}^{00} & \mathcal{I}^{01} \\ \mathcal{I}^{10} & \mathcal{I}^{11} \end{pmatrix} \right),$$

where we partition the the information matrix and its inverse as

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{00} & \mathcal{I}_{01} \\ \mathcal{I}_{10} & \mathcal{I}_{11} \end{pmatrix} \quad \mathcal{I}^{-1} = \begin{pmatrix} \mathcal{I}^{00} & \mathcal{I}^{01} \\ \mathcal{I}^{10} & \mathcal{I}^{11} \end{pmatrix}.$$

The asymptotic variance for $\sqrt{N}(\hat{\theta}_{1u} - \theta_1^*)$ is therefore

$$V(\sqrt{N}(\hat{\theta}_{1u} - \theta_1^*)) = \mathcal{I}^{11} = \mathcal{I}_{11}^{-1} + \mathcal{I}_{11}^{-1} \mathcal{I}_{10} \left(\mathcal{I}_{00} - \mathcal{I}_{01} \mathcal{I}_{11}^{-1} \mathcal{I}_{10} \right)^{-1} \mathcal{I}_{01} \mathcal{I}_{11}^{-1}.$$

(See the appendix to these lecture notes for the partioned inverse of a matrix.) Note that $\mathcal{I}^{11} - \mathcal{I}_{11}^{-1}$ is positive semidefinite. Now consider the restricted ml estimate $\hat{\theta}_{r1}$. In large samples:

$$\sqrt{N}(\hat{\theta}_{r1} - \theta_1^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{11}^{-1}).$$

The variance for this estimator is obviously smaller than the variance for the unrestricted estimator.

We can try to calculate directly the variance of the difference $\sqrt{N}(\hat{\theta}_{r1} - \hat{\theta}_{u1})$ by looking at the joint distribution of $\hat{\theta}_{r1}$ and $\hat{\theta}_{u1}$, but there is a simpler argument that goes to the heart of the testing procedure. Consider a larger set of estimators that includes both $\hat{\theta}_{r1}$ and $\hat{\theta}_{u1}$ as special cases:

$$\hat{\theta}_\lambda = (1 - \lambda) \cdot \hat{\theta}_{r1} + \lambda \cdot \hat{\theta}_{u1} = \hat{\theta}_{r1} + \lambda \cdot (\hat{\theta}_{u1} - \hat{\theta}_{r1}).$$

The variance of this estimator is

$$V(\hat{\theta}_\lambda) = V(\hat{\theta}_{r1}) + \lambda^2 \cdot V(\hat{\theta}_{u1} - \hat{\theta}_{r1}) + 2 \cdot \lambda \cdot C(\hat{\theta}_{r1}, \hat{\theta}_{u1} - \hat{\theta}_{r1}).$$

Taking the derivative of this variance with respect to λ , evaluated at $\lambda = 0$ gives

$$\left. \frac{\partial V}{\partial \lambda}(\hat{\theta}_\lambda) \right|_{\lambda=0} = 2 \cdot C(\hat{\theta}_{r1}, \hat{\theta}_{u1} - \hat{\theta}_{r1}).$$

Now, and this is crucial, this derivative must be equal to zero because the estimator with $\lambda = 0$ is efficient (as the mle its asymptotic variance is equal to the Cramér-Rao bound). Hence the covariance of $\hat{\theta}_{r1}$ and $\hat{\theta}_{r1} - \hat{\theta}_{u1}$ is zero and thus

$$V(\hat{\theta}_{u1}) = V(\hat{\theta}_{r1} + (\hat{\theta}_{u1} - \hat{\theta}_{r1})) = V(\hat{\theta}_{r1}) + V(\hat{\theta}_{u1} - \hat{\theta}_{r1}),$$

and thus

$$V(\hat{\theta}_u - \hat{\theta}_r) = V(\hat{\theta}_u) - V(\hat{\theta}_r),$$

or

$$\sqrt{N}(\hat{\theta}_{r1} - \hat{\theta}_{u1}) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{11} - \mathcal{I}_{11}).$$

Thus the general form of the Hausman–Wu test in a maximum likelihood context is

$$\begin{aligned} HW &= (\hat{\theta}_{1u} - \hat{\theta}_{1r})' \cdot (V(\hat{\theta}_{1u}) - V(\hat{\theta}_{1r}))^{-g} \cdot (\hat{\theta}_{1u} - \hat{\theta}_{1r}) \\ &= N \cdot (\hat{\theta}_{1u} - \hat{\theta}_{1r})' \cdot (\mathcal{I}^{11} - \mathcal{I}_{11}^{-1})^{-g} \cdot (\hat{\theta}_{1u} - \hat{\theta}_{1r}). \\ &= N \cdot (\hat{\theta}_{1u} - \hat{\theta}_{1r})' \cdot \left(\mathcal{I}_{11}^{-1} \mathcal{I}_{10} (\mathcal{I}_{00} - \mathcal{I}_{01} \mathcal{I}_{11}^{-1} \mathcal{I}_{10}) \mathcal{I}_{01} \mathcal{I}_{11}^{-1} \right)^{-g} \cdot (\hat{\theta}_{1u} - \hat{\theta}_{1r}), \end{aligned}$$

where the superscript “ $-g$ ” denotes the generalized inverse. Under the null hypothesis that $\theta_0 = 0$, this test statistic would have a Chi-squared distribution with degrees of freedom equal to the rank of \mathcal{I}_{01} (note that this may differ from the degrees of freedom for the other tests). The test is therefore obviously not in general going to be first-order equivalent to the other tests discussed here, although it can be if the rank of \mathcal{I}_{01} is high enough.

The test requires the estimator under the null hypothesis to be efficient, something like a maximum likelihood estimator, but the unrestricted estimator need not be efficient. In general the test can be used when the alternative hypothesis is relatively vague.

A note of caution with this test. Taking the difference in two variance matrices, even if nominally this difference should be positive semi-definite, can often lead to extremely large test statistics when the difference is close to zero. In case the variance is not positive definite,

one can take generalized inverses. As a result, however, the small sample properties of this test are not always attractive.

Example

Consider a linear regression model.

$$Y_i = X'_{0i}\beta_0 + X'_{1i}\beta_1 + \varepsilon_i,$$

or, in matrix notation:

$$\mathbf{Y} = \mathbf{X}_0\beta_0 + \mathbf{X}_1\beta_1 + \varepsilon.$$

We assume ε_i is conditionally normal with mean zero and variance σ^2 . We are interested in testing the null hypothesis

$$H_0 : \beta_0 = 0,$$

against the alternative

$$H_a : \beta_0 \neq 0.$$

We use the following notation:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_{00} & \mathbf{X}_{01} \\ \mathbf{X}_{10} & \mathbf{X}_{11} \end{pmatrix}, \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \mathbf{X}^{00} & \mathbf{X}^{01} \\ \mathbf{X}^{10} & \mathbf{X}^{11} \end{pmatrix}.$$

The restricted estimator is

$$\hat{\beta}_{r1} = \mathbf{X}_{11}^{-1}\mathbf{X}'_1\mathbf{Y},$$

with distribution

$$\hat{\beta}_{r1} \sim \mathcal{N}(\beta_1, \sigma^2 \cdot \mathbf{X}_{11}^{-1}).$$

The unrestricted estimator is

$$\hat{\beta}_{u1} = \mathbf{X}^{10}\mathbf{X}'_0\mathbf{Y} + \mathbf{X}^{11}\mathbf{X}'_1\mathbf{Y},$$

with distribution:

$$\hat{\beta}_{u1} \sim \mathcal{N}(\beta_1, \sigma^2 \cdot \mathbf{X}^{11}) = \mathcal{N}\left(\beta_1, \sigma^2 \cdot \left(\mathbf{X}_{11}^{-1} + \mathbf{X}_{11}^{-1} \mathbf{X}_{10} (\mathbf{X}_{00} - \mathbf{X}_{01} \mathbf{X}_{11}^{-1} \mathbf{X}_{10})^{-1} \mathbf{X}_{01} \mathbf{X}_{11}^{-1}\right)\right).$$

So, using the formula given above, the variance for the difference is the difference in the variances, and

$$\hat{\beta}_{r1} - \hat{\beta}_{u1} \sim \mathcal{N}\left(0, \sigma^2 \cdot \left(\mathbf{X}_{11}^{-1} \mathbf{X}_{10} (\mathbf{X}_{00} - \mathbf{X}_{01} \mathbf{X}_{11}^{-1} \mathbf{X}_{10})^{-1} \mathbf{X}_{01} \mathbf{X}_{11}^{-1}\right)\right),$$

with the rank of the variance equal to the rank of \mathbf{X}_{01} , assuming both \mathbf{X}_{00} and \mathbf{X}_{11} have full rank). In this case one can also calculate the variance directly, with the same result.

Consider the special case where X_0 and X_1 are scalars. In that case $\hat{\beta}_{r1}$ is equal to $\hat{\beta}_{u1} + \hat{\beta}_{u0} \cdot \hat{\delta}_0$, where $\hat{\delta}_0$ is the coefficient on X_0 in a regression of X_0 on X_1 . The Hausman test is testing whether $\beta_0 \cdot \delta_0 = 0$, that is whether the product of the coefficient of the excluded regressor (β_0) and the coefficient on the included regressor in a regression of the excluded regressor on the included one (δ_0) is equal to zero. Clearly if the two regressors are uncorrelated, the test has no power against the alternative that β_0 differs from zero. \square

Example

The second example is an instrumental variables model:

$$Y_i = \beta' X_i + \varepsilon_i.$$

We are concerned that X_i is not orthogonal with the disturbance ε . For that contingency there is an instrumental variable available Z_i which we are confident is independent of ε_i . We are considering the null hypothesis that X is exogenous (that is, in this case, independent of the disturbance ε). Formally:

$$H_0 : E[X_i \varepsilon_i] = 0, \text{ against the alternative } H_1 : E[X_i \varepsilon_i] \neq 0.$$

One approach is to test $\gamma = 0$ in the regression

$$Y_i = \beta' X_i + \gamma' Z_i + \varepsilon_i.$$

An alternative is to use a Hausman test. Estimate the model efficiently under the null hypothesis:

$$\hat{\beta}_r = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}),$$

with approximate variance

$$V(\hat{\beta}_r) = \sigma_\varepsilon^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}.$$

Now consider the instrumental variables estimator

$$\hat{\beta}_u = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}(\hat{\mathbf{X}}'\mathbf{Y}),$$

where

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X}),$$

with approximate variance

$$V(\hat{\beta}_u) = \sigma_\varepsilon^2 \cdot (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}.$$

Obviously for this estimator to be well-defined we need the matrix $\hat{\mathbf{X}}'\hat{\mathbf{X}}$ to be invertible, which in turn requires that the matrix $\mathbf{Z}'\mathbf{X}$ has at least full row rank. (The dimension of \mathbf{X} is $N \times K$, the dimension of \mathbf{Z} is $N \times M$, so the dimension of $\mathbf{Z}'\mathbf{X}$ is $M \times K$, and so we need at least $M \geq K$ instruments for this to work.)

Using the fact that under the null hypothesis $\hat{\beta}_r$ is efficient as the maximum likelihood estimator, combined with the fact that the estimator $\hat{\beta}_u$ is consistent under weaker conditions, the test statistic would be

$$HW = \frac{1}{\hat{\sigma}_\varepsilon^2} \cdot \left((\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) - (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}(\hat{\mathbf{X}}'\mathbf{Y}) \right) \cdot \left((\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1} \right)^{-g} \\ \cdot \left((\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) - (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}(\hat{\mathbf{X}}'\mathbf{Y}) \right).$$

Maintaining, as before, the assumption that the matrix $\mathbf{Z}'\mathbf{X}$ has full row rank, and writing

$$\mathbf{X} = \hat{\mathbf{X}} + \eta = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} + \eta,$$

we can write the factor in the middle, the inverse of the variance, as

$$\left((\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} + \eta'\eta)^{-1} \right)^{-g}$$

If $\eta'\eta$ is invertible, this reduces to (using the formula in the appendix):

$$(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1} \left((\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1} + (\eta'\eta)^{-1} \right)^{-1} (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1},$$

and the degrees of freedom of the test are equal to the number of regressors (K , the number of columns in \mathbf{X}). In general, the degrees of freedom are equal to the rank of the η (which is the same as the rank of the matrix $\eta'\eta$). That is, only regressors not included in the set of instruments, are counted in the degrees of freedom.

So for example, if $\mathbf{X} = (\mathbf{X}_0 \ \mathbf{X}_1)$, and $\mathbf{Z} = (\mathbf{X}_0 \ \mathbf{Z}_1)$, $\hat{\mathbf{X}}$ can be partitioned in $\hat{\mathbf{X}} = (\hat{\mathbf{X}}_0 \ \hat{\mathbf{X}}_1)$, with $\hat{\mathbf{X}}_0 = \mathbf{X}_0$, and thus $\eta = (\eta_0 \ \eta_1)$ with $\eta_0 = 0$. In that case the degrees of freedom is equal to the rank of η_1 \square

APPENDIX

1. Consider a matrix

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

Assuming the matrix is invertible, its inverse is

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} E & F \\ G & H \end{pmatrix},$$

with

$$E = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1},$$

$$H = D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1},$$

$$F = -(A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1})BD^{-1},$$

$$G = D^{-1}C(A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}).$$

2. Suppose both A and B are invertible matrices, and $A + B$ and $A^{-1} + B^{-1}$ are invertible.

Then

$$A^{-1} - (A + B)^{-1} = A^{-1}(A^{-1} + B^{-1})^{-1}A^{-1}.$$

REFERENCES

ENGLE, R. (1984), “Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics”, in Griliches and Intrilligator (eds), *Handbook of Econometrics*, Vol III, Elsevier, North Holland.

HAUSMAN, (1978), “Specification Tests in Econometrics”, *Econometrica*, Vol. 46, No 6, 1251–1271.

HOLLY, A., (1987), “Specification Tests: An Overview”, in Bewley, (ed), *Advances in Econometrics*, Cambridge University Press, Cambridge.