Testing for Omitted Variables

Jeroen Weesie

Department of Sociology
University of Utrecht
The Netherlands

email `J.weesie@fss.uu.nl`
tel +31–30–2531922
fax +31–30–2534405

Prepared for

North American Stata users meeting
Boston, March 2001

The three classic likelihood-based approaches to test smooth hypotheses about parameters $H : g(\theta) = 0$,

- LR test:

  estimate model with and without constraint $g(\theta) = 0$. A large difference between fit statistics (e.g., deviance) is evidence against H.

- Wald test:

  estimate the model without the constraint. Test whether the parameters satisfy a linearized version of the constraint.

- (efficient) Score/Lagrange Multiplier test:

  estimate the restricted model. If the fit criterion (log-lik) sharply increases in directions away from the constraint, this is evidence against the constraint.

- Methods are often asymptotically equivalent (under the null).
- Likely, the higher order asymptotic properties of LR are better.
- Little is known in general about small sample properties.
- Computations may vary widely
  - It may be hard to estimate the *restricted* model (e.g., non-linear constraints g)
  - It may be hard to estimate the *unrestricted* model (e.g., in random effects/coef models, in which the restriction effectively eliminates the random effects/coefs)

Did I use the right set of predictor variables?

- non-linear transformations of an included x-var (e.g., a squared term)
- is it right to treat a variable as an 'interval variable', or should it be treated as a categorical variable (e.g., level of education)
- interactions between x-variables
- What about some of the variables that I did not enter in the model? (To hell with theory!)

Sometimes this may involve ancillary parameters, e.g.

- the scale parameter in regression-type models
- the between-equation correlation in selection models
- the cutpoints in an ordinal regression model

We typically assume that these parameters are constant between subjects, but there is considerable attention to heteroscedasticty issues in regression-style models, not so in other regression-type models.

- Parameters
  θ a parameter-vector partitioned as $\theta = (\theta_1, \theta_2)$,
  θ_1 are the parameters of the restricted model
  θ_2 are associated with the omitted variables.
  and $\hat{\theta}_0 = (\hat{\theta}_1, 0)$.
- Linear predictor:

$$\mathrm{lp}_i = x_i'\theta = x_{i1}'\theta_1 + x_{i2}'\theta_2$$

- l_i is log-likelihood contribution of i-th observations
- The score statistic

$$U_i(\theta) = \frac{\partial\, l_i(\theta)}{\partial\, \theta} = \frac{\partial\, l_i(\theta)}{\partial\, \mathrm{lp}_i} x_i = s_i x_i$$

  Stata calls s_i a "score variable".
- Let $U(\theta) = \Sigma\, U_i(\theta) = \Sigma\, s_i x_i$
  It depends on the estimator only via s_i

Score tests are based on the large sample distribution under H of the quadratic form

$$U(\hat{\theta}_0)'\mathrm{var}(U)^{-1}U(\hat{\theta}_0) \sim \chi^2_{\mathrm{k}}$$

The "score variable" $\frac{\partial\, l_i(\theta)}{\partial\, \mathrm{lp}_i}$ has to be evaluated under $\hat{\theta}_0$. And so it is computed if a `score()` option is specified while estimating the restricted model.

How to estimate $\mathrm{var}(U)$? The classic model-based estimator uses the fact under under regularity conditions,

$$\mathrm{var}(U) = E\left(\frac{\partial^2\ \sum l_i(\theta)}{\partial\,\theta\,\partial\,\theta'}\right) = I(\theta)$$

and so this requires additional information about the model that was estimated, namely the (expected) Fisher information. An alternative based on the hessian / observed information is feasible.

Yet another alternative is the outer-product of gradients estimator,

$$\sum_i U_i(\hat{\theta}_0)U_i(\hat{\theta}_0)' = \sum_i s_i^2 x_i x_i'$$

This requires only the score variable s. The modification of the OPG estimator to the case of clustered observations and complex survey data is straightforward.

- Language to specify potentially omitted variables
  - variables not yet in model (lp),
  - transformations of variables already in model (lp)
  - factorial versions of vars in model (lp)
  - Quadratic extension of the current model (lp)
- Different types of tests
  - Likelihood ratio test
  - Wald
  - Score test, with three estimator of the variance of the scores
- Univariate as well as simultaneous tests.
  Adjusted P-values (Bonferroni, Holm, Sidak, . . .)

Continuation

The presentation continues with the presenations of the command (boston.do)