Inference for Health Econometrics: Inference, Model Tests, Diagnostics, Multiple Tests, and Bootstrap

A. Colin Cameron
Department of Economics, University of California - Davis.

July 2013

Abstract

This paper presents a brief summary of classical statistical inference for many commonly-used regression model estimators that are asymptotically normally distributed. The paper covers Wald confidence intervals and hypothesis tests based on robust standard errors; tests of model adequacy and model diagnostics; family-wise error rates and false discovery rates that control for multiple testing; and bootstrap and other resampling methods.

Keywords: inference, m-estimators, robust standard errors, cluster-robust standard errors, diagnostics, multiple tests, multiple comparisons, family-wise error rate, false discovery rate, bootstrap, asymptotic refinement, jackknife, permutation tests.

JEL Classification: C12, C21, C23.

Prepared for J. Mullahy and A. Basu (eds.), *Health Econometrics*, in A.J. Culyer ed., *Encyclopedia of Health Economics*. Email address: accameron@ucdavis.edu

This chapter presents inference for many commonly-used estimators – least squares, generalized linear models, generalized method of moments, and generalized estimating equations – that are asymptotically normally distributed. Section 1 focuses on Wald confidence intervals and hypothesis tests based on estimator variance matrix estimates that are heteroskedastic-robust and, if relevant, cluster-robust. Section 2 summarizes tests of model adequacy and model diagnostics. Section 3 presents family-wise error rates and false discovery rates that control for multiple testing such as subgroup analysis. Section 4 presents bootstrap and other resampling methods that are most often used to estimate the variance of an estimator. Bootstraps with asymptotic refinement are also presented.

1 Inference

Most estimators in health applications are m-estimators that solve estimating equations of the form

$$\sum\nolimits_{i=1}^{N} \mathbf{g}_i(\widehat{\boldsymbol{\theta}}) = \mathbf{0}, \tag{1}$$

where $\boldsymbol{\theta}$ is a $q \times 1$ parameter vector, i denotes the i^{th} of N observations, $\mathbf{g}_i(\cdot)$ is a $q \times 1$ vector, and often $\mathbf{g}_i(\boldsymbol{\theta}) = \mathbf{g}_i(y_i, \mathbf{x}_i, \boldsymbol{\theta})$ where y denotes a scalar dependent variable and $\mathbf{x}$ denotes the regressors or covariates. For ordinary least squares, for example, $\mathbf{g}_i(\boldsymbol{\beta}) = (y_i - \mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i$. Nonlinear least squares, maximum likelihood (ML), quantile regression and just-identified instrumental variables estimators are m-estimators. So too are generalized linear model estimators, extensively used in biostatistics, that are quasi-ML estimators based on exponential family distributions, notably Bernoulli (logit and probit), binomial, gamma, normal, and Poisson.

The estimator $\widehat{\boldsymbol{\theta}}$ is generally consistent if $\mathrm{E}[\mathbf{g}_i(\boldsymbol{\theta})] = \mathbf{0}$. Statistical inference is based on the result that $\widehat{\boldsymbol{\theta}}$ is asymptotically normal with mean $\boldsymbol{\theta}$ and variance matrix $\mathrm{V}[\widehat{\boldsymbol{\theta}}]$ that is estimated by

$$\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}] = \widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^{-1\prime}, \tag{2}$$

where $N^{-1}\widehat{\mathbf{A}}$ and $N^{-1}\widehat{\mathbf{B}}$ are consistent estimates of $\mathbf{A} = \mathrm{E}[N^{-1}\sum_i \mathbf{H}_i(\boldsymbol{\theta})]$, where $\mathbf{H}_i(\boldsymbol{\theta}) = \partial \mathbf{g}_i(\boldsymbol{\theta})/\partial \boldsymbol{\theta}'$, and $\mathbf{B} = \mathrm{E}\Big[N^{-1}\sum_i \sum_j \mathbf{g}_i(\boldsymbol{\theta})\mathbf{g}_j(\boldsymbol{\theta})'\Big]$. The variance is said to be of "sandwich form", since $\widehat{\mathbf{B}}$ is sandwiched between $\widehat{\mathbf{A}}^{-1}$ and $\widehat{\mathbf{A}}^{-1\prime}$. The estimate $\widehat{\mathbf{A}}$ is the observed Hessian $\sum_i \mathbf{H}_i(\widehat{\boldsymbol{\theta}})$, or in some cases the expected Hessian $\mathrm{E}\left[\sum_i \mathbf{H}_i(\boldsymbol{\theta})\right]|_{\widehat{\boldsymbol{\theta}}}$. By contrast, the estimate $\widehat{\mathbf{B}}$, and hence $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}]$ in (2), can vary greatly with the type of data being analyzed and associated appropriate distributional assumptions.

Default estimates of $\mathrm{V}[\widehat{\boldsymbol{\theta}}]$ are based on strong distributional assumptions, and are typically not used in practice. For ML estimation with density assumed to be correctly specified $\mathbf{B} = -\mathbf{A}$, so the sandwich estimate simplifies to $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}] = -\widehat{\mathbf{A}}^{-1}$. Qualitatively similar simplification occurs for least squares and instrumental variables estimators when model errors are independent and homoskedastic.

More generally, for data independent over i, $\widehat{\mathbf{B}} = \frac{N}{N-q}\sum_i \mathbf{g}_i(\widehat{\boldsymbol{\theta}})\mathbf{g}_i(\widehat{\boldsymbol{\theta}})'$, where the multiple $N/(N-q)$ is a commonly-used finite sample adjustment. Then the variance matrix estimate in (2) is called the Huber, White, or robust estimate – a limited form of robustness as independence of observations is assumed. For OLS, for example, this estimate is

valid even if independent errors are heteroskedastic whereas the default requires errors to be homoskedastic.

Often data are clustered, with observations correlated within cluster but independent across clusters. For example, individuals may be clustered within villages or hospitals, or students clustered within class or within school. Let c denote the typical cluster, and sum $\mathbf{g}_i(\boldsymbol{\theta})$ for observations i in cluster c to form $\mathbf{g}_c(\boldsymbol{\theta})$. Then $\widehat{\mathbf{B}} = \frac{C}{C-1}\sum_{c=1}^{C}\mathbf{g}_c(\widehat{\boldsymbol{\theta}})\mathbf{g}_c(\widehat{\boldsymbol{\theta}})'$, where C is the number of clusters, and the variance matrix estimate in (2) is called a cluster-robust estimate. The number of clusters should be large as the asymptotic theory requires $C \to \infty$, rather than $N \to \infty$. The clustered case also covers short panels with few time periods and data correlated over time for a given individual but independent across individuals. Then the clustering sums over time periods for a given individual. Wooldridge (2003) and Cameron and Miller (2011) survey inference with clustered data.

Survey design can lead to clustering. Applied biostatisticians often use survey estimation methods that explicitly control for the three complex survey complications of weighting, stratification and clustering. Econometricians instead usually assume correct model specification conditional on regressors (or instruments), so that there is no need to weight; ignore the potential reduction in standard error estimates that can occur with stratification; and conservatively control for clustering by computing standard errors that cluster at a level such as state (region) that is usually higher than the primary sampling unit.

For time series data, observations may be correlated over time. Then the heteroskedastic and autocorrelation consistent (HAC) variance matrix estimate is used; see Newey and West (1987). A similar estimate can be used when data are spatially correlated, with correlation depending on distance and independence once observations are more than a given distance apart. This leads to the spatial HAC estimate; see Conley (1999).

Note that in settings where robust variance matrix estimates are used, additional assumptions may enable more efficient estimation of $\boldsymbol{\theta}$ such as feasible generalized least squares and generalized estimating equations, especially if data are clustered.

Given $\widehat{\boldsymbol{\theta}}$ asymptotic normal with variance matrix estimated using (2), the Wald method can be used to form confidence intervals and perform hypothesis tests.

Let θ be a scalar component of the parameter vector $\boldsymbol{\theta}$. Since $\widehat{\boldsymbol{\theta}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\theta},\, \widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}]]$, we have $\widehat{\theta} \overset{a}{\sim} \mathcal{N}\left[\theta,\, s^2_{\widehat{\theta}}\right]$, where the standard error $s_{\widehat{\theta}}$ is the square root of the relevant diagonal entry in $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}]$. It follows that $(\widehat{\theta}-\theta)/s_{\widehat{\theta}} \overset{a}{\sim} \mathcal{N}[0,\, 1]$. This justifies use of the standard normal distribution in constructing confidence intervals and hypothesis tests for sample size $N \to \infty$. A commonly-used finite-sample adjustment uses $(\widehat{\theta}-\theta)/s_{\widehat{\theta}} \overset{a}{\sim} T(N-q)$, where $T(N-q)$ is the students T distribution with $(N-q)$ degrees of freedom, N is the sample size, and K parameters are estimated.

A 95% confidence interval for θ gives a range of values that 95% of the time will include the unknown true value of θ. The Wald 95% confidence interval is $\widehat{\theta} \pm c_{.025} \times s_{\widehat{\theta}}$, where the critical value $c_{.025}$ is either $z_{[.025]} = 1.96$, the .025 quantile of the standard normal distribution, or $t_{[.025]}$ the .025 quantile of the $T(N-q)$ distribution. For example, $c_{.025} = 2.042$ if $N-q = 30$.

For two-sided tests of $H_0 : \theta = \theta^*$ against $H_a : \theta \neq \theta^*$, the Wald test is based on how far $|\widehat{\theta}-\theta^*|$ is from zero. Upon normalizing by the standard error, the Wald statistic $w = (\widehat{\theta}-\theta^*)/s_{\widehat{\theta}}$ is asymptotically standard normal under H_0, though again a common finite sample correction is to use the $T(N-q)$ distribution. We reject H_0 at the 5% significance

level if $|w| > c_{.025}$. Often $\theta^* = 0$, in which case w is called the t-statistic and the test is called a test of statistical significance. Greater information is conveyed by reporting the p-value, the probability of observing a value of w as large or larger in absolute value under the null hypothesis. Then $p = \Pr[|W| > |w|]$, where W is standard normal or $T(N-q)$ distributed. We reject $H_0 : \theta = \theta^*$ against $H_0 : \theta \neq \theta^*$ at level 0.05 if $p < 0.05$.

More generally we may be interested in performing joint inference on more than one parameter, such as a joint test of statistical significance of several parameters, or on functions(s) of the parameters. Let $\mathbf{h}(\boldsymbol{\theta})$ be an $h \times 1$ vector function of $\boldsymbol{\theta}$, possibly nonlinear, where $h \leq q$. A Taylor series approximation yields $\mathbf{h}(\widehat{\boldsymbol{\theta}}) \simeq \mathbf{h}(\boldsymbol{\theta}) + \widehat{\mathbf{R}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, where $\widehat{\mathbf{R}} = \partial \mathbf{h}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}'|_{\widehat{\theta}}$ is assumed to be of full rank h (the nonlinear analog of linear dependence of restrictions). Given $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \overset{a}{\sim} \mathcal{N}[\mathbf{0}, \widehat{\mathbb{V}}[\widehat{\boldsymbol{\theta}}]]$ this yields $\mathbf{h}(\widehat{\boldsymbol{\theta}}) \overset{a}{\sim} \mathcal{N}[\mathbf{h}(\boldsymbol{\theta}), \widehat{\mathbf{R}}\widehat{\mathbb{V}}[\widehat{\boldsymbol{\theta}}]\widehat{\mathbf{R}}']$. The term delta method is used as a first derivative is taken in approximating $\mathbf{h}(\widehat{\boldsymbol{\theta}})$.

Confidence intervals can be formed in the case that $h(\cdot)$ is a scalar. Then we use $h(\widehat{\boldsymbol{\theta}}) \pm c_{.025} \times [\widehat{\mathbf{R}}\widehat{\mathbb{V}}[\widehat{\boldsymbol{\theta}}]\widehat{\mathbf{R}}']^{1/2}$. A leading example is a confidence interval for a marginal effect in a nonlinear model. For example, for $\mathrm{E}[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\theta})$ the marginal effect for the j^{th} regressor is $\partial \mathrm{E}[y|\mathbf{x}]/\partial x_j = \exp(\mathbf{x}'\boldsymbol{\theta})\theta_j$. When evaluated at $\mathbf{x} = \mathbf{x}^*$ this equals $\exp(\mathbf{x}^{*\prime}\widehat{\boldsymbol{\theta}})\widehat{\theta}_j$ which is a scalar function $h(\widehat{\boldsymbol{\theta}})$ of $\widehat{\boldsymbol{\theta}}$; the corresponding average marginal effect is $\sum_i \exp(\mathbf{x}_i'\widehat{\boldsymbol{\theta}})\widehat{\theta}_j$.

A Wald test of $H_0 : \mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$ against $H_a : \mathbf{h}(\boldsymbol{\theta}) \neq \mathbf{0}$ is based on the closeness of $\mathbf{h}(\widehat{\boldsymbol{\theta}})$ to zero, using

$$w = \mathbf{h}(\widehat{\boldsymbol{\theta}})'[\widehat{\mathbf{R}}\widehat{\mathbb{V}}[\widehat{\boldsymbol{\theta}}]\widehat{\mathbf{R}}']^{-1}\mathbf{h}(\widehat{\boldsymbol{\theta}}) \overset{a}{\sim} \chi^2(h) \tag{3}$$

under H_0. We reject H_0 at level 0.05 if $w > \chi^2_{.95}(h)$. An F version of this test is $F = w/h$, and we reject at level 0.05 if $w > F_{.95}(h, N-q)$. This is a small sample variation, analogous to using the $T(N-q)$ rather than the standard normal.

For ML estimation the Wald method is one of three testing methods that may be used. Consider testing the hypothesis that $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$. Let $\widetilde{\boldsymbol{\theta}}$ denote the ML estimator obtained by imposing this restriction, while $\widehat{\boldsymbol{\theta}}$ does not impose the restriction. The Wald test uses only $\widehat{\boldsymbol{\theta}}$ and tests the closeness of $\mathbf{h}(\widehat{\boldsymbol{\theta}})$ to zero. The log likelihood ratio test is based on the closeness of $L(\widehat{\boldsymbol{\theta}})$ to $L(\widetilde{\boldsymbol{\theta}})$ where $L(\boldsymbol{\theta})$ denotes the log-likelihood function. The score test uses only $\widetilde{\boldsymbol{\theta}}$ and is based on the closeness to zero of $\partial L(\boldsymbol{\theta})/\partial \boldsymbol{\theta}|_{\widetilde{\boldsymbol{\theta}}}$, where $L(\boldsymbol{\theta})$ here is the log-likelihood function for the unrestricted model.

If the likelihood function is correctly specified, a necessary assumption, these three tests are asymptotically equivalent. So the choice between them is one of convenience. The Wald test is most often used, as in most cases $\widehat{\boldsymbol{\theta}}$ is easily obtained. The score test is used in situations where estimation is much easier when the restriction is imposed. For example, in a test of no spatial dependence versus spatial dependence it may be much easier to estimate $\boldsymbol{\theta}$ under the null hypothesis of no spatial dependence. The Wald and score tests can be robustified. If one is willing to make the strong assumption that the likelihood function is correctly specified, then the likelihood ratio test is preferred due to the Neyman-Pearson lemma and because, unlike the Wald test, it is invariant to reparameterization.

Generalized method of moments (GMM) estimators are based on a moment condition of the form $\mathrm{E}[\mathbf{g}_i(\boldsymbol{\theta})] = \mathbf{0}$. If there are as many components of $\mathbf{g}(\cdot)$ as of $\boldsymbol{\theta}$ the model is said to be just-identified and the estimate $\widehat{\boldsymbol{\theta}}$ solves $\sum_i \mathbf{g}_i(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$, which is (1). Leading examples in the biostatistics literature are generalized linear model estimators and generalized estimating

equations estimators. If instead there are more moment conditions than parameters there is no solution to (1). Instead we make $\sum_i \mathbf{g}_i(\widehat{\boldsymbol{\theta}})$ as close to zero as possible using a quadratic norm. The method of moments estimator minimizes

$$Q(\boldsymbol{\theta}) = \left(\sum\nolimits_{i=1}^{N} \mathbf{g}_i(\boldsymbol{\theta})\right)' \mathbf{W} \left(\sum\nolimits_{i=1}^{N} \mathbf{g}_i(\boldsymbol{\theta})\right),$$

where $\mathbf{W}$ is a symmetric positive definite weighting matrix and the best choice of $\mathbf{W}$ is the inverse of a consistent estimate of the variance of $\sum_i \mathbf{g}_i(\boldsymbol{\theta})$.

The leading example of this is two-stage least squares (2SLS) estimation for instrumental variables estimation in overidentified models. Then $\mathbf{g}_i(\boldsymbol{\beta}) = \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})$, and it can be shown that the 2SLS estimator is obtained if $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$. The estimated variance matrix is again of sandwich form (2), though the expressions for $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ are more complicated. For instrumental variables estimators with instruments weakly correlated with regressors an alternative asymptotic theory may be warranted. Bound, Jaeger and Baker (1995) outline the issues and Andrews, Moreira, and Stock (2007) compare several different test procedures.

2 Model Tests and Diagnostics

The most common specification tests imbed the model under consideration into a larger model and use hypothesis tests (Wald, likelihood ratio or score) to test the restrictions that the larger model collapses to the model under consideration. A leading example is test of statistical significance of a potential regressor.

A broad class of tests of model adequacy can be constructed by testing the validity of moment conditions that are imposed by a model but have not already been used in constructing the estimator. Suppose a model implies the population moment condition

$$H_0 : \mathrm{E}[\mathbf{m}_i(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}, \tag{4}$$

where $\mathbf{w}$ is a vector of observables, usually the dependent variable y, regressors $\mathbf{x}$ and, possibly, additional variables $\mathbf{z}$. An m-test, in the spirit of a Wald test, is a test of whether the corresponding sample moment

$$\widehat{\mathbf{m}}(\widehat{\boldsymbol{\theta}}) = \frac{1}{N}\sum\nolimits_{i=1}^{N} \mathbf{m}_i(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}), \tag{5}$$

is close to zero. Under suitable assumptions, $\widehat{\mathbf{m}}(\widehat{\boldsymbol{\theta}})$ is asymptotically normal. This leads to the chisquared test statistic

$$\mathrm{M} = \widehat{\mathbf{m}}(\widehat{\boldsymbol{\theta}})'\widehat{\mathbf{V}}_{\mathbf{m}}^{-1}\widehat{\mathbf{m}}(\widehat{\boldsymbol{\theta}}) \overset{a}{\sim} \chi^2(rank(\mathbf{V}_{\mathbf{m}}), \tag{6}$$

if the moment conditions (4) are correct, where $\widehat{\mathbf{V}}_{\mathbf{m}}$ is a consistent estimate of the asymptotic variance of $\widehat{\mathbf{m}}(\widehat{\boldsymbol{\theta}})$. The challenge is obtaining $\widehat{\mathbf{V}}_{\mathbf{m}}$. In some leading examples an auxiliary regression can be used, or a bootstrap can be applied.

Especially for fully parametric models there are many candidates for $\mathbf{m}_i(\cdot)$. Examples of this approach are White's information matrix test to test correct specification of the likelihood function; a regression version of the chisquared goodness of fit test; Hausman tests

such as that for regressor endogeneity; and tests of overidentifying restrictions in a model with endogenous regressors and an excess of instruments. Such tests are not as widely used as they might be for two reasons. First, there is usually no explicit alternative hypothesis so rejection of H_0 may not provide much guidance as to how to improve the model. Second, in very large samples with actual data any test at a fixed significance level such as 0.05 is likely to reject the null hypothesis, so inevitably any model will be rejected.

Regression model diagnostics need not involve formal hypothesis tests. A range of residual diagnostic plots can provide information on model nonlinearity and observations that are outliers and have high leverage. In the linear model a small sample correction divides the residual $y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ by $\sqrt{1-h_{ii}}$, where h_{ii} is the i^{th} diagonal entry in the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$. Since $\mathbf{H}$ has rank K, the number of regressors, the average value of h_{ii} is K/n and values of h_{ii} in excess of $2K/N$ are viewed as having high leverage. This result extends to generalized linear models where a range of residuals have been proposed; McCullagh and Nelder (1989) provides a summary. Econometricians place less emphasis on residual analysis, compared to biostatisticians. If data sets are small then there is concern that residual analysis may lead to overfitting of the model. And if the data set is large then there is a belief that residual analysis may be unnecessary as a single observation will have little impact on the analysis. But even then diagnostics may help detect data miscoding and unaccounted model nonlinearities.

For linear models, R^2 is a well understood measure of goodness of fit. For nonlinear models a range of pseudo-R^2 measures have been proposed. One that is easily interpreted is the squared correlation between y and $\widehat{y}$, though in nonlinear models this is not guaranteed to increase as regressors are added.

Model testing and diagnostics may lead to more than one candidate model. Standard hypothesis tests can be implemented for models that are nested. For nonnested models that are likelihood based, one can use a generalization of the likelihood ratio test due to Vuong (1989), or use information criteria such as Akaike's information criteria based on fitted log-likelihood with a penalty for the number of model parameters. For nonnested models that are not likelihood based one possibility is artificial nesting that nests two candidate models in a larger model, though this approach can lead to neither model being favored.

3 Multiple Tests

Standard theory assumes that hypothesis tests are done once only and in isolation, whereas in practice final reported results may follow much pretesting. Ideally reported p values should control for this pretesting.

In biostatistics it is common to include as control variables in a regression only those regressors that have $p < 0.05$. By contrast, in economics it is common to have a preselected candidate set of control regressors, such as key socioeconomic variables, and include them even if they are statistically insignificant. This avoids pretesting, at the expense of estimating larger models.

A more major related issue is that of multiple testing or multiple comparisons. Examples include testing the statistical significance of a key regressor in several subgroups of the sample (subgroup analysis); testing the statistical significance of a key regressor in regressions

on a range of outcomes (such as use of a range of health services); testing the statistical significance of a key regressor interacted with various controls (interaction effects); and testing the significance of a wide range of variables on a single outcome (such as various genes on a particular form of cancer). With many such tests at standard significance levels one is clearly likely to find spurious statistical significance.

In such cases one should view the entire battery of tests as a unit. If m such tests are performed, each at statistical significance level α^*, and the tests are statistically independent, then the probability of finding no statistical significance in all m tests is $(1-\alpha^*)^m$. It follows that the probability of finding statistical significance in at least one test, called the family-wise error rate (FWER), equals $\alpha = 1-(1-\alpha^*)^m$. In order to test at FWER α, each individual test should be at level $\alpha^* = 1-(1-\alpha)^m$, called the Sidak correction. For example, if $m = 5$ tests are conducted with FWER of $\alpha = 0.05$, each test should be conducted at level $\alpha^* = 0.01021$. The simpler Bonferroni correction sets $\alpha^* = \alpha/m$. The Holm correction uses a stepdown version of Bonferroni, with tests ordered by p-value from smallest to largest, so $p_{(1)} < p_{(2)} < \cdots < p_{(m)}$, and the j^{th} test rejects if $p_{(j)} < \alpha_j^* = \alpha/(m-j+1)$. A stepdown version of the Sidak correction uses $\alpha_j^* = 1-(1-\alpha)^{m-j+1}$. These corrections are quite conservative in practice, as the multiple tests are likely to be correlated rather than independent.

Benjamini and Hochberg (1995) proposed an alternative approach to multiple testing. Recall that test size is the probability of a type I error, i.e. the probability of incorrectly rejecting the null hypothesis. For multiple tests it is natural to consider the proportion of incorrectly rejected hypotheses, the false discovery proportion (FDP), and its expectation E[FDP], called the false discovery rate (FDR). Benjamini and Hochberg (1995) argue that it is more natural to control FDR than FEWR. They propose doing so by ordering tests by p-value from smallest to largest, so $p_{(1)} < p_{(2)} < \cdots < p_{(m)}$, and rejecting the corresponding hypotheses $H_{(1)}, ..., H_{(k)}$ where k is the largest j for which $p(j) \leq \alpha j/m$, where α is the prespecified FDR for the multiple tests. If the multiple tests are independent then the FDR equals α.

In practice tests are not independent. Farcomeni (2008) provides an extensive guide to the multiple testing literature. A recent article on estimating the FDR when tests are correlated is Schwartzman and Lin (2011). Duflo, Glennerster, and Kremer (2008) provide a good discussion of practical issues that arise with multiple testing and consider the FEWR but not the FDR. White (2001) presents simulation-based methods for the related problem of testing whether the best model encountered in a specification search has better predictive power than a benchmark model.

4 Bootstrap and Other Resampling Methods

Statistical inference controls for the uncertainty that the observed sample of size N is just one possible realization of a set of N possible draws from the population. This typically relies on asymptotic theory that leads to limit normal and chi-squared distributions. Alternative methods based on Monte Carlo simulation are detailed in this section.

4.1 Bootstrap

Bootstraps can be applied to a wide range of statistics. We focus on the most common use of the bootstrap, to estimate the standard error of an estimator when this is difficult to do using conventional methods.

Suppose 400 random samples from the population were available. Then we could obtain 400 different estimates of $\widehat{\theta}$, and let the standard error of $\widehat{\theta}$ be the standard deviation of these 400 estimates. In practice, however, only one sample from the population is available. The bootstrap provides a way to generate 400 samples by resampling from the current sample. Essentially the observed sample is viewed as the population and the bootstrap provides multiple samples from this population.

Let $\widehat{\boldsymbol{\theta}}^*_{(1)}, ..., \widehat{\boldsymbol{\theta}}^*_{(B)}$ denote B estimates where, for example, $B = 400$. Then in the scalar case the bootstrap estimate of the variance of $\widehat{\theta}$ is

$$\widehat{\mathrm{V}}_{\mathrm{Boot}}[\widehat{\theta}] = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\theta}^*_{(b)} - \overline{\widehat{\theta}^*})^2, \tag{7}$$

where $\overline{\widehat{\theta}^*} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^*_{(b)}$ is the average of the B bootstrap estimates. The square root of $\widehat{\mathrm{V}}_{\mathrm{Boot}}[\widehat{\theta}]$, denoted $\mathrm{se}_{\mathrm{Boot}}[\widehat{\theta}]$, is called the bootstrap estimate of the standard error of $\widehat{\theta}$. In the case of several parameters

$$\widehat{\mathrm{V}}_{\mathrm{Boot}}[\widehat{\boldsymbol{\theta}}] = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\boldsymbol{\theta}}^*_{(b)} - \overline{\widehat{\boldsymbol{\theta}}^*})(\widehat{\boldsymbol{\theta}}^*_{(b)} - \overline{\widehat{\boldsymbol{\theta}}^*})',$$

and even more generally the bootstrap may be used to estimate the variance of functions $\mathbf{h}(\widehat{\boldsymbol{\theta}})$, such as marginal effects, not just $\widehat{\boldsymbol{\theta}}$ itself.

There are several different ways that the resamples can be obtained. A key consideration is that the quantity being resampled should be i.i.d.

The most common bootstrap for data $(y_i, \mathbf{x}_i)$ that are i.i.d. is a paired bootstrap or nonparametric bootstrap. This draws with replacement from $(y_1, \mathbf{x}_1), ..., (y_N, \mathbf{x}_N)$ to obtain a resample $(y^*_1, \mathbf{x}^*_1), ..., (y^*_N, \mathbf{x}^*_N)$ for which some observations will appear more than once, while others will not appear at all. Estimation using the resample yields estimate $\widehat{\boldsymbol{\theta}}^*$. Using B similarly generated resamples yields $\widehat{\boldsymbol{\theta}}^*_{(1)}, ..., \widehat{\boldsymbol{\theta}}^*_{(B)}$. This bootstrap variance estimate is asymptotically equivalent to the White or Huber robust sandwich estimate.

If data are instead clustered with C clusters, a clustered bootstrap draws with replacement from the entire clusters, yielding a resample $(\mathbf{y}^*_1, \mathbf{X}^*_1), ..., (\mathbf{y}^*_C, \mathbf{X}^*_C)$. This bootstrap variance estimate is asymptotically equivalent to the cluster-robust sandwich estimate.

Other bootstraps place more structure on the model. A residual or design bootstrap in the linear regression model fixes the regressors and only resamples the residuals. For models with i.i.d. errors the residual bootstrap samples with replacement from $\widehat{u}_1, ..., \widehat{u}_N$ to yield residual resample $\widehat{u}^*_1, ..., \widehat{u}^*_N$. Then the typical data resample is $(y^*_1, \mathbf{x}_1), ..., (y^*_N, \mathbf{x}_N)$ where $y^*_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{u}^*_i$. If errors are heteroskedastic one should instead use a wild bootstrap; the simplest example lets $\widehat{u}^*_i = \widehat{u}_i$ with probability 0.5 and $\widehat{u}^*_i = -\widehat{u}_i$ with probability 0.5.

For a fully parameterized model one can generate new values of the dependent variable from the fitted conditional distribution. The typical data resample is $(y_1^*, \mathbf{x}_1), ..., (y_N^*, \mathbf{x}_N)$ where y_i^* is a draw from $F(y|\mathbf{x}_i, \widehat{\boldsymbol{\theta}})$.

Whenever a bootstrap is used in applied work the seed, the initial value of the random number generator used in determining random draws, should be set to ensure replicability of results. For standard error estimation $B = 400$ should be more than adequate.

The bootstrap can also be used for statistical inference. A Wald 95% confidence interval for scalar θ is $\widehat{\theta} \pm 1.96 \times \text{se}_{\text{Boot}}[\widehat{\theta}]$. An asymptotically equivalent alternative interval is the percentile interval $(\widehat{\theta}^*_{[.025]}, \widehat{\theta}^*_{[.975]})$ where $\widehat{\theta}^*_{[\alpha]}$ is the α^{th} quantile of $\widehat{\theta}^*_{(1)}, ..., \widehat{\theta}^*_{(B)}$. Similarly, in testing $H_0 : \theta = 0$ against $H_a : \theta \neq 0$ the null hypothesis may be rejected if $|w| = |\widehat{\theta}/\text{se}_{\text{Boot}}[\widehat{\theta}]| > 1.96$, or if $\widehat{\theta} < \widehat{\theta}^*_{[.025]}$ or $\widehat{\theta} > \widehat{\theta}^*_{[.975]}$.

Care is needed in using the bootstrap in nonstandard situations as, for example, $\text{V}[\widehat{\boldsymbol{\theta}}]$ may not exist, even asymptotically, yet it is always possible to (erroneously) compute a bootstrap estimate of $\text{V}[\widehat{\boldsymbol{\theta}}]$. The bootstrap can be applied if $\widehat{\boldsymbol{\theta}}$ is root-N consistent and asymptotically normal, and there is sufficient smoothness in the cumulative distribution functions of the data generating process and of the statistic being bootstrapped.

4.2 Bootstrap with Asymptotic Refinement

The preceding bootstraps are asymptotically equivalent to the conventional methods of section 1. Bootstraps with asymptotic refinement, by contrast, provide a more refined asymptotic approximation that may lead to better performance (truer test size and confidence interval coverage) in finite samples. Such bootstraps are emphasized in theory papers, but are less often implemented in applied studies.

These gains are possible if the statistic bootstrapped is asymptotically pivotal, meaning its asymptotic distribution does not depend on unknown parameters. An estimator $\widehat{\theta}$ that is asymptotically normal is not usually asymptotically pivotal as its distribution depends on an unknown variance parameter. But the studentized statistic $t = (\widehat{\theta} - \theta_0)/s_{\widehat{\theta}}$ is asymptotically $\mathcal{N}[0, 1]$ under $H_0 : \theta = \theta_0$, so is asymptotically pivotal. We therefore compute in each bootstrap resample $t^* = (\widehat{\theta}^* - \widehat{\theta})/s_{\widehat{\theta}^*}$, and use quantiles of $t^*_{(1)}, ..., t^*_{(B)}$ to compute critical values and p-values. Note that t^* is centered around $\widehat{\theta}$ because the bootstrap views the sample as the population, so $\widehat{\theta}$ is the population value.

A 95% percentile-t confidence interval for scalar θ is $(\widehat{\theta} + t^*_{[.025]} s_{\widehat{\theta}}, \widehat{\theta} + t^*_{[.975]} s_{\widehat{\theta}})$ where $t^*_{[\alpha]}$ is the α^{th} quantile of $t^*_{(1)}, ..., t^*_{(B)}$. And a percentile-t Wald test rejects $H_0 : \theta = \theta_0$ against $H_a : \theta \neq \theta_0$ at level 0.05 if $t = (\widehat{\theta} - \theta_0)/s_{\widehat{\theta}}$ falls outside the interval $(t^*_{[.025]}, t^*_{[.975]})$.

Two commonly-used alternative methods to obtain confidence intervals with asymptotic refinement are the following. The bias-corrected method is a modification of the percentile method that incorporates a bootstrap estimate of the finite-sample bias in $\widehat{\theta}$. For example, if the estimator is upward biased, as measured by estimated median bias, then the confidence interval is moved to the left. The bias-corrected accelerated confidence interval is an adjustment to the bias-corrected method that adds an acceleration component that permits the asymptotic variance of $\widehat{\theta}$ to vary with θ.

Theory shows that bootstrap methods with asymptotic refinement outperform conven-

tional asymptotic methods as $N \to \infty$. For example, a nominal 95% confidence interval with asymptotic refinement has coverage rate of $0.95+O(N^{-1})$ rather than $0.95+O(N^{-1/2})$. This does not guarantee better performance in typical sized finite samples, but Monte Carlo studies generally confirm this to be the case. Bootstraps with refinement require a larger number of bootstraps than recommended in the previous subsection, since the critical values lie in the tails of the distribution. A common choice is $B = 999$, with B chosen so that $B+1$ is divisible by the significance level 100α.

4.3 Jackknife

The jackknife is an alternative resampling scheme used for bias-correction and variance estimation that predates the bootstrap.

Let $\widehat{\boldsymbol{\theta}}$ be the original sample estimate of $\boldsymbol{\theta}$, let $\widehat{\boldsymbol{\theta}}_{(-i)}$ denote the parameter estimate from the sample with the i^{th} observation deleted, $i = 1, ..., N$, and let $\overline{\widehat{\boldsymbol{\theta}}} = N^{-1}\sum_{i=1}^{N}\widehat{\boldsymbol{\theta}}_{(-i)}$ denote the average of the N jackknife estimates. The bias-corrected jackknife estimate of $\boldsymbol{\theta}$ equals $N\widehat{\boldsymbol{\theta}} - (N-1)\overline{\widehat{\boldsymbol{\theta}}}$, the sum of the N pseudo-values $\widehat{\boldsymbol{\theta}}^{*}_{(-i)} = N\widehat{\boldsymbol{\theta}} - (N-1)\widehat{\boldsymbol{\theta}}_{(-i)}$ that provide measures of the importance or influence of the i^{th} observation estimating $\widehat{\boldsymbol{\theta}}$.

The variance of these N pseudo-values can be used to estimate $\mathrm{V}[\widehat{\boldsymbol{\theta}}]$, yielding the leave-one-out jackknife estimate of variance

$$\widehat{\mathrm{V}}_{\mathrm{Jack}}[\widehat{\boldsymbol{\theta}}] = \left[\frac{1}{N(N-1)}\sum\nolimits_{i=1}^{N}(\widehat{\boldsymbol{\theta}}^{*}_{(-i)} - \overline{\widehat{\boldsymbol{\theta}}})(\widehat{\boldsymbol{\theta}}^{*}_{(-i)} - \overline{\widehat{\boldsymbol{\theta}}})'\right].$$

A variation replaces $\overline{\widehat{\boldsymbol{\theta}}}$ with $\widehat{\boldsymbol{\theta}}$.

The jackknife requires N resamples, requiring more computation than the bootstrap if N is large. The jackknife does not depend on random draws, unlike the bootstrap, so is often used to compute standard errors for published official statistics.

4.4 Permutation Tests

Permutation tests derive the distribution of a test statistic by obtaining all possible values of the test statistic under appropriate rearrangement of the data under the null hypothesis.

Consider scalar regression, so $y_i = \beta_1 + \beta_2 x_i + u_i$, $i = 1, .., N$, and Wald test of $H_0 : \beta_2 = 0$ based on $t = \widehat{\beta}_2/s_{\widehat{\beta}_2}$. Regress each of the $N!$ unique permutations of $(y_1, ..., y_N)$ on the regressors $(x_1, ..., x_N)$ and in each case calculate the t-statistic for $H_0 : \beta_2 = 0$. Then the p-value for the original test statistic is obtained directly from the ordered distribution of the $N!$ t-statistics.

Permutation tests are most often used to test whether two samples come from the same distribution, using the difference in means test. This is a special case of the preceding where x_i is an indicator variable equal to one for observations coming from the second sample.

Permutation methods are seldom used in multiple regression, though several different ways to extend this method have been proposed. Anderson and Robinson (2001) review these methods and argue that it is best to permute residuals obtained from estimating the model under H_0, a method proposed by Freedman and Lane (1983).

5 Conclusion

This survey is restricted to classical inference methods for parametric models. It does not consider Bayesian inference, inference following nonparametric and semiparametric estimation, or time series complications such as models with unit roots and cointegration.

The graduate-level econometrics texts by Cameron and Trivedi (2005), Greene (2012) and Wooldridge (2010) cover especially sections 1 and 2; see also Jones (2000) for a survey of health econometrics models and relevant chapters in this volume. The biostatistics literature for nonlinear models emphasizes estimators for generalized linear models; the classic reference is McCullagh and Nelder (1989). For the resampling methods in section 5, Efron and Tibsharani (1993) is a standard accessible reference; see also Davison and Hinkley (1997) and, for implementation, Cameron and Trivedi (2010).

References

Anderson, M.J., and J. Robinson (2001), "Permutation Tests for Linear Models," *Australian and New Zealand Journal of Statistics*, 43, 75-88.

Andrews, D.W.K., M.J. Moreira, and J.H. Stock (2007), "Performance of Conditional Wald tests in IV Regression with Weak Instruments," *Journal of Econometrics,* 139, 116-132.

Benjamini, Y., and Y. Hochberg (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society B*, 57, 289-300.

Bound, J., D.A. Jaeger, and R.M. Baker (1995), "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90, 443-450.

Cameron, A.C., and D.A. Miller (2011), "Robust Inference with Clustered Data," in A. Ullah and D.E. Giles (Eds.), *Handbook of Empirical Economics and Finance*, 1-28, Boca Raton, CRC Press.

Cameron, A.C., and P.K. Trivedi (2005), *Microeconometrics: Methods and Applications,* Cambridge, Cambridge University Press.

Cameron, A.C., and P.K. Trivedi (2010), *Microeconometrics using Stata*, First revised edition, College Station, TX, Stata Press.

Conley, T.G. (1999), "GMM Estimation with Cross Sectional Dependence," *Journal of Econometrics*, 92, 1-45.

Davison, A.C., and D. V. Hinkley (1997), *Bootstrap Methods and Their Application*, Cambridge, Cambridge University Press.

Duflo, E., R. Glennerster, and M. Kremer (2008), "Using Randomization in Development Economics Research: A Toolkit," in *Handbook of Development Economics,* T.P. Shultz and

J.A. Strauss (Eds.), Volume 4, 3896-3962, Amsterdam, North-Holland.

Efron, B., and J. Tibsharani (1993), *An Introduction to the Bootstrap*, London, Chapman and Hall.

Farcomeni, A. (2008), "A Review of Modern Multiple Hypothesis Testing, with Particular Attention to the False Discovery Proportion," *Statistical Methods in Medical Research*, 17, 347–88.

Freedman, D., and D. Lane (1983), "A Nonstochastic Interpretation of Reported Significance Levels," *Journal of Business and Economic Statistics*, 1, 292-298.

Greene, W.H. (2012), *Econometric Analysis*, Seventh edition, Upper Saddle River, Prentice Hall.

Jones, A.M. (2000), "Health Econometrics," in *Handbook of Health Economics,* A.J. Culyer and J.P. Newhouse (Eds.), Volume 1, 265-344, Amsterdam, North-Holland.

McCullagh, P., and J. A. Nelder (1989), *Generalized Linear Models*, Second edition, London, Chapman and Hall.

Newey, W.K., and K.D. West (1987), "A Simple, Positive Semi-Definite, Heteroscedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703-708.

Schwartzman, A. and X. Lin (2011), "The Effect of Correlation in False Discovery Rate Estimation," *Biometrika*, 98, 199-214.

Vuong, Q.H. (1989), "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses," *Econometrica*, 57, 307-333.

White, H. (2001), "A Reality Check for Data Snooping," *Econometrica*, 68, 1097-1126.

Wooldridge, J. M. (2003), "Cluster-Sample Methods in Applied Econometrics," *American Economic Review,* 93, 133-138.

Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, Second edition, Cambridge, MA, MIT Press.