Estimation of Misspecified Models

Stas Kolenikov

Department of Statistics
University of Missouri-Columbia

February 5, 2007
Joint reading group in advanced econometrics

Outline

Problem

- A very typical statistical/econometric model assumes something like

$$y_t \sim \text{i.i.d. } f(y, x, \theta) \tag{1}$$

  where $f(\cdot)$ is a parametric family known up to parameters θ.
- Parameter estimation: maximum likelihood

$$\hat{\theta}_n = \arg\max \sum_t \ln f(Y_t, X_t, \theta) \tag{2}$$

- What if the basic model assumptions of (1) are violated? The parametric family may not contain the true model $f_0(x, y)$ that generated the data; or the data may not be i.i.d.; etc.

Huber's (1967) framework

Let $X_1, X_2, \ldots$ are independent random variables with values in $\mathcal{X}$ having the common probability distribution P.
Huber (1967) considers two situations:

- (near-)minimization of an objective function

$$\frac{1}{n}\sum_t \rho(X_t, \hat\theta_n) - \inf_\Theta \frac{1}{n}\sum_t \rho(X_t, \theta) \to 0 \tag{3}$$

- estimating equations

$$\frac{1}{n}\sum_t \psi(X_t, \hat\theta_n) \to 0 \in \mathbf{R}^p \tag{4}$$

The estimating equations may be the derivatives of the objective function from (3), or may come as (exactly identified) system of equations (method of moments, instrumental variables, ...)

Estimating equations

- Linear regression: $\sum_t (y_t - X_t\beta)^2 \to \min \Rightarrow$ normal equations:

$$-2\sum_t (y_t - X_t\beta)X_t = 0 \in \mathbf{R}^p \tag{5}$$

- Instrumental equations (exactly ID):

$$\sum_t (y_t - X_t\beta)Z_t = 0 \tag{6}$$

- Logit model:

$$\ln L(y_t, X_t, \beta) = \sum_t y_t \ln \Lambda(X_t\beta) + (1 - y_t)\ln\bigl(1 - \Lambda(X_t\beta)\bigr),$$
$$\Lambda(z) = \bigl(1 + \exp(-z)\bigr)^{-1} \tag{7}$$

Likelihood scores:

$$\frac{\partial \ln L}{\partial \beta} = \sum_t X_t\bigl(y_t - \Lambda(X_t\beta)\bigr) = 0 \tag{8}$$

Huber's (1967) results

Under certain regularity conditions (next slide),

- the sequence of estimators $\hat{\theta}_n$ a.s. stays in a compact set
- the estimators $\hat{\theta}_n$ are strongly consistent
- the estimators $\hat{\theta}_n$ are asymptotically normal

Regularity conditions

Regularity conditions usually refer to the conditions on objective functions, their derivatives, parameter spaces, etc., that are necessary for all mathematical expressions to be well defined, and all derivations to be fully justified.
E.g., in order to take Taylor series expansion to apply the delta method, one needs that

1. the point θ_0 where expansion is to be taken is contained in the parameter space along with some neighborhood, so there is "enough room to step around" (i.e., θ_0 is an interior point of the parameter space)
2. the function to be expanded such as $l(Y, X, \theta)$ is defined in a neighborhood of the expansion point θ_0
3. the function is sufficiently smooth in a neighborhood of θ_0

Some of those conditions would need to hold with probability 1, or with probability tending to 1 as $n \to \infty$.

Huber's (1967) regularity

- Local compactness of the space of θ
- Measurability and separability of the objective function $\rho(\cdot)$ and estimating functions $\psi(\cdot)$ over the space of X
- Sufficient smoothness with respect to θ: lower semicontinuity of $\rho(\cdot)$; a.s. continuity of $\psi(\cdot)$.
- Boundedness of the objective function $\rho(\cdot)$, estimating functions $\psi(\cdot)$, and their expectations
- Lower boundedness of $\gamma(\theta) = \mathbb{E}\,\rho(X,\theta)\forall\theta$
- Uniqueness and sufficient separation of population minimum θ_0 of $\gamma(\cdot)$ at the interior of the parameter space
- Well defined expectation $\lambda(\theta) = \mathbb{E}\,\psi(X,\theta)$ with a unique zero at θ_0
- Lipschitz conditions on expectations of $u(x,\theta,d) = \sup_{|\tau-\theta|\le d} |\psi(x,\tau) - \psi(x,\theta)|$ and its square
- Finite $\mathbb{E}\,|\psi(x,\theta_0)|^2$

Huber (1967): idea of proof

- Consistency: ϵ-manipulations with $\inf \frac{1}{n}\sum_t \rho(X_t,\theta)$ or $\sup \frac{1}{n}\sum_t |\psi(X_t,\theta)-\lambda(\theta)|$
- Asymptotic normality:
  1. bound in probability the differences $\psi(x,\theta)-\lambda(\theta)$ (tail conditions for CLT)
  2. show asymptotic equivalence of $\frac{1}{\sqrt{n}}\sum_t \psi(X_t,\theta_0)$ and $\sqrt{n}\lambda(\hat\theta_n)$
  3. asymptotic normality of $\hat\theta_n$ then follows by the standard delta method argument:

$$\sqrt{n}(\hat\theta_n-\theta_0) \xrightarrow{d} N\bigl(0, A^{-1}BA^{-T}\bigr),$$
$$A=\mathbb{E}\,\partial\psi(X,\theta_0), \quad B=\mathbb{E}\,\psi(X,\theta_0)\psi(X,\theta_0)^T \qquad (9)$$

Other proofs based on Brower's fixed point theorem are available (Maronna 1976).

Origins of the sandwich

Estimating equations: $\psi(X, \hat{\theta}_n) = 0$.
Taylor series expansion of $\psi(\cdot)$:

$$\sqrt{n}\big(\psi(X, \hat{\theta}_n) - \psi(X, \theta_0)\big) =$$
$$= \partial\psi(X, \theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(\sqrt{n}\|\hat{\theta}_n - \theta_0\|)$$
$$\sim N\big(0, \mathbb{V}[\psi(X_1, \theta_0)]\big)$$

as the sum of i.i.d. terms $\psi(X_i, \theta_0)$. Hence,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \approx -\sqrt{n}A^{-1}\psi(X, \theta_0) + o_p(\sqrt{n}\|\hat{\theta}_n - \theta_0\|)$$
$$\xrightarrow{d} N(0, A^{-1}BA^{-T})$$

M-estimates

- Later work by Huber: foundations of robust statistics; M-estimates defined through optimization of a certain criteria (aka *extremum estimators* in econometrics)
- Huber (1974), Hampel, Ronchetti, Rousseeuw & Stahel (2005), Maronna, Martin & Yohai (2006)
- Huber (1967) is still the cornerstone paper! It gives the most general conditions for consistency and asymptotic normality of M-estimates

White (1982)

White (1982) is a culmination of his preceding work on misspecified models.

- Interpretation of θ_0: the quasi-MLEs define the density that minimizes the Kullback-Leibler distance between the distributions $\mathbb{E}\left[\ln P(x)/f(x,\theta_0)\right]$
- Weaker regularity conditions that are easier to verify
- Information matrix test for misspecification
- Hausman test for misspecification

Notation

Quasi-log-likelihood:

$$l_n(X,\theta) = \frac{1}{n}\sum_t \ln f(X_t,\theta) \tag{10}$$

$$\hat{\theta}_n = \arg\max_{\Theta} l_n(X,\theta) \tag{11}$$

$$A_n(\theta) = n^{-1}\sum_i \partial^2 \ln f(X_i,\theta),$$

$$B_n(\theta) = n^{-1}\sum_i \partial \ln f(X_i,\theta)\partial' \ln f(X_i,\theta)$$

$$A(\theta) = \mathbb{E}\,\partial^2 \ln f(X,\theta),$$

$$B(\theta) = \mathbb{E}\,\partial \ln f(X,\theta)\partial' \ln f(X,\theta)$$

$$C_n(\theta) = A_n(\theta)^{-1}B_n(\theta)A_n(\theta)^{-T}, \quad C(\theta) = A(\theta)^{-1}B(\theta)A(\theta)^{-T} \tag{12}$$

White's (1982) regularity conditions

- The independent random vectors X_t have a distribution with (Radon-Nykodym) density $g(\cdot)$, and the parametric family of distribution functions all have densities $f(u, \theta)$
- $f(u, \theta)$ and $\partial \ln f / \partial \theta$ are measurable in u and continuous in θ
- $|\partial^2 \ln f / \partial \theta_i \, \partial \theta_j|$, $|\partial \ln f / \partial \theta_i \cdot \partial \ln f / \partial \theta_j|$ and $|\partial (f \partial / \partial \theta_i) \, / \, \partial \theta_j|$ are dominated by functions integrable in u
- The parameter space is compact
- $\mathbb{E} \, |\ln g(X)| < \infty$, $|\ln f(x, \theta)|$ is bounded uniformly in θ
- Kullback-Leibler info $I(g : f, \theta) = \mathbb{E} \ln g(X) / f(X, \theta)$ has a unique minimum at θ_0
- $\theta_0 \in \operatorname{int} \Theta$; $|B(\theta)| \neq 0$; $\operatorname{rk} A(\theta)$ is constant in a neighborhood of θ_0
- $\operatorname{supp} f$ does not depend on θ

White's (1982) results

Some theorems require only a subset of regularity conditions

- Theorem 2.1: existence of QMLE $\hat{\theta}_n$
- Theorem 2.2: strong consistency: $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$
- Theorem 3.1: identifiability of the model
- Theorem 3.2: asymptotic normality:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\big(0, C(\theta_0)\big); \quad C_n(\hat{\theta}_n) \xrightarrow{a.s.} C(\theta_0) \quad (13)$$

- Theorem 3.3: if the model is correctly specified, then

$$-A(\theta_0) = B(\theta_0) = C^{-1}(\theta_0) \quad (14)$$

  (information matrix identity in max likelihood)
- Theorems 3.4 and 3.5: Wald and Lagrange multiplier (score) tests

Information matrix test

- Under correct specification, $H_0 : A(\theta_0) + B(\theta_0) = 0$. Can we test it?
- Let $d_l(x, \theta) = \frac{\partial \ln f}{\partial \theta_i} \frac{\partial \ln f}{\partial \theta_j} + \frac{\partial^2 \ln f}{\partial \theta_i \partial \theta_j}$, where l enumerates the pairs $(i, j), l = 1, \dots, q \le p(p+1)/2$ (subset of interest?)
- Indicators $D_{ln}(\hat{\theta}_n) = n^{-1} \sum_t d_l(X_t, \hat{\theta}_n)$ are asymptotically normal $\Leftarrow$ regularity
- Define $q \times p$ Jacobian $\nabla D = \mathbb{E}\, \partial d_l(X, \theta)/\partial \theta_k$, $V(\theta) = \mathbb{E}\big[$ outer product of $d(X, \theta) - \nabla D(\theta) A(\theta)^{-1} \nabla \ln f(X, \theta)\big]$ and their sample analogues
- Theorem 4.1: $H_0 : g(x) = f(x, \theta_0)$, $V(\theta_0) > 0 \Rightarrow$

$$\text{(i) } \sqrt{n} D_n(\hat{\theta}_n) \xrightarrow{d} N(0, V(\theta_0))$$

$$\text{(ii) } V_n(\hat{\theta}_n) \xrightarrow{a.s.} V(\theta_0)$$

$$\text{(iii) } \mathcal{J}_n = n D_n(\hat{\theta}_n) V_n(\hat{\theta}_n)^{-1} D_n(\theta_n) \xrightarrow{d} \chi^2_q \qquad (15)$$

- Should have good power against alternatives that render the usual ML inference invalid

Hausman test

- Need an alternative estimator $\gamma = (\beta, \alpha)$ of (a subset of) the same parameters $\theta = (\beta, \psi)$, $\dim \beta = k$
- γ needs to be consistent under misspecification
- Estimates $\hat{\theta}_n = (\hat{\beta}_n, \hat{\psi}_n)$, $\tilde{\gamma} = (\tilde{\beta}_n, \tilde{\alpha}_n)$
- Under $H_0 : f(x, \theta) = g(x)$, $\sqrt{n}(\hat{\beta}_n - \tilde{\beta}_n) \xrightarrow{d} N(0, S)$ where S involves the information matrices for both estimators, as well as outer products of scores within and between the models
- Test statistic:

$$\mathcal{H}_n = n(\hat{\beta}_n - \tilde{\beta}_n)' S_n(\hat{\theta}_n, \hat{\gamma}_n)^{-1} (\hat{\beta}_n - \tilde{\beta}_n) \xrightarrow{d} \chi^2_k \qquad (16)$$

- Should have good power against alternatives leading to parameter inconsistency
- LM form of the test is also available

White's (1982) recommendations

1. Estimate the model by the maximum likelihood
2. Apply the overall misspecification IM test (15)
3. Pass: use the standard ML inference
4. Fail: investigate inconsistency (local misspecification?) by using Hausman test (16)
    1. Pass, no evidence of bias: apply specification robust inference with the sandwich estimator (13) aka (9)
    2. Fail: the model is badly misspecified, model specification must be re-examined

GMT (1984) theory

- Gourieroux, Monfort & Trognon (1984*a*) consider estimation of the parameters of the first and the second moments in a model $y_t = h(x_t, \theta) + e_t$
- Strong consistency of pseudo-MLE $\hat{\theta}_n \Leftrightarrow$ the likelihood is of linear exponential family form

$$f(x, \theta) = \exp\{A(\theta) + B(x) + C(\theta)x\} \tag{17}$$

- Asymptotic normality of $\hat{\theta}_n$
- Nice properties of the exp families have long been known (Brown 1987)! We teach exponential families in Stat 7760 and Stat 9710.
- Lower bound on variance (in the sandwich form) is attainable provided the nuisance parameters/variance structure is modeled correctly (QGPML)
- Similar strong consistency and asymptotic normality results for quadratic exponential family:

$$f(x, \theta, \Sigma) = \exp\{A(\theta, \Sigma) + B(x) + C(\theta, \Sigma)x + x' D(\theta, \Sigma)x\} \tag{18}$$

GMT (1984) application

Gourieroux, Monfort & Trognon (1984*b*) provide an application of the above theory:

- Poisson model: $y_t|x_t \sim \text{Poi}\big[\exp(x_t b)\big]$
- Overdispersed Poisson: $y_t|x_t \sim \text{Poi}\big[\exp(x_t b + \epsilon_t)\big]$
- $\epsilon_t \sim \Gamma \Rightarrow y_t|x_t \sim$ negative binomial
- What if the distribution of ϵ_t is misspecified, but is known to have $\mathbb{E}(\exp \epsilon_t) = 1$, $\mathbb{V}(\exp \epsilon_t) = \eta^2$?
- PML with normal (nonlinear least squares), Poisson, negative binomial, gamma families: estimators $\hat{b}$ are consistent and asymptotically normal, asymptotic covariance matrices derived; relative efficiency?
- QGPML: need consistent estimators of b, η^2 for the first stage; then plug the estimated nuisance parameter $\hat{\eta}^2$ into the regular PML objective function; efficiency gains wrt PML
- Simultaneous estimation of b and η^2: quadratic exponential family

Sandwich estimator

The sandwich estimator of asymptotic variance $A^{-1}BA^{-T}$ is a common feature. It appears in many ways and in many areas of applied statistics and econometrics:

- *M*-estimates (Huber 1974)
- Non-linear regression (Gallant 1987, White 1981)
- Heteroskedastic regression models (Eicker 1967, White 1980)
- Autocorrelated error terms (West & Newey 1987)
- Survey statistics (Binder 1983, Skinner 1989)
- Longitudinal data and generalized estimating equations (Diggle, Heagerty, Liang & Zeger 2002)
- Covariance structure/SEM models (Browne 1984, Satorra 1990, Satorra & Bentler 1994, Yuan & Hayashi 2006)

Review of history of the sandwich estimator: Hardin (2003)

Linear regression

$$y_t = X_t\beta + \epsilon_t, \quad \mathbb{V}[\epsilon_t] = \sigma_t^2 \neq \text{const}$$

Sandwich variance estimator, aka heteroskedasticity-consistent estimator, aka "robust" (which I don't like) estimator:

$$\mathbb{V}_e = (X'X)^{-1}\Bigl(\sum_t x_t x_t' e_t^2\Bigr)(X'X)^{-1} \tag{19}$$

A lot is known about it: Eicker (1967), White (1980), MacKinnon & White (1985), Kauermann & Carroll (2001), Bera, Suprayitno & Premaratne (2002), ...

- Scale $n/(n-p)$ to correct some of the small sample bias (Hinkley 1977)
- Use $e_t^2/(1-h_t)$ in the "meat" of the sandwich (MacKinnon & White 1985)
- MINQUE estimator of Bera et al. (2002): unbiased under heteroskedasticity!

Finite sample performance

Kauermann & Carroll (2001) consider linear regression, quasi-likelihood and generalized estimating equations. In linear regression,

- bias of a simple sandwich depends on the kurtosis of X
- corrected versions have bias of $O(n^{-1})$ under heteroskedasticity
- sandwich is less efficient under the null than $s^2(X'X)^{-1}$; efficiency also depends on kurtosis (3 times more variable with normal X's; 6 times more variable with Laplace X's)
- CI undercoverage is proportional to $\mathbb{V}[\hat{\sigma}^2]$ and variability of sandwich estimator, and cannot be corrected by the use of t- rather than z-quantiles

Consistency of sandwich comes with a price: high variability in finite samples and CI undercoverage — a typical robustness vs. efficiency trade-off.

Implementation: Stata's _robust

Stata's _robust (aka robust option of many estimation commands):

- Implements an empirical version of (9) = (13)
- Available for all estimation commands that allow straightforward computation of likelihood scores:
- Observation-by-observation likelihood: numerical derivatives
- Complex likelihoods: analytical derivatives supplied by the programmer
- Variation: cluster — the summation is over clusters of (possibly dependent) observations
- Inference for complex surveys
- Some of the aforementioned corrections for linear regression model

Target of inference?

Freedman (2006): what is the use of (asymptotically) accurate standard errors if the point estimates are absurd?
True model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

Fitted model:

$$y = b_0 + b_1 x + \text{error}$$

The misspecification (omitted nonlinearity) won't be detected through the use of sandwich estimator. Estimates of b_0 and b_1 will be biased relative to β_0 and β_1.

My personal interest

Kolenikov & Bollen (2006) study misspecification in structural equation/covariance structure models.

- Distinguish structural and distributional misspecification
- A bold evidence of misspecification: negative estimates of variances — significance?
- Examples of gross structural misspecification with Heywood cases in population
- Behavior of the sandwich estimator vs. some other popular variance estimators
- Earlier work by Bollen (1996): an alternative estimator consistent under milder conditions, natural instruments, and Hausman structural misspecification test

Conclusions

- Model misspecification: a common rather than a rare phenomenon?
- Extensive statistical and especially econometrics literature
- Estimates are still consistent and asymptotically normal, although interpretation may suffer
- Variance estimation: information sandwich
- Corrections will be useful for small samples

References

Bera, A. K., Suprayitno, T. & Premaratne, G. (2002), 'On some heteroskedasticity-robust estimators of variance-covariance matrix of the least-squares estimators', *Journal of Statistical Planning and Inference* **108**(1–2), 121–136.

Bera, A. K., Suprayitno, T. & Premaratne, G. (2002), 'On some heteroskedasticity-robust estimators of variance-covariance matrix of the least-squares estimators', *Journal of Statistical Planning and Inference* **108**(1–2), 121–136.

Binder, D. A. (1983), 'On the variances of asymptotically normal estimators from complex surveys', *International Statistical Review* **51**, 279–292.

Bollen, K. A. (1996), 'An alternative two stage least squares (2SLS) estimator for latent variable models', *Psychometrika* **61**(1), 109–121.

Brown, L. D. (1987), *Fundamentals of statistical exponential families: with applications in statistical decision theory*, Vol. 9 of *IMS Lecture Notes*, Institute of Mathematical Statistics, California.

Browne, M. W. (1984), 'Asymptotically distribution-free methods for the analysis of the covariance structures', *British Journal of Mathematical and Statistical Psychology* **37**, 62–83.

Diggle, P., Heagerty, P., Liang, K.-Y. & Zeger, S. (2002), *Analysis of Longitudinal Data*, 2nd edn, Oxford University Press.

Eicker, F. (1967), Limit theorems for regressions with unequal and dependent errors, *in* 'Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, University of California Press, Berkeley, pp. 59–82.

Freedman, D. A. (2006), 'On the so-called "Huber sandwich estimator" and "robust standard errors"', *The American Statistician* **60**(4), 299–302.

Gallant, A. R. (1987), *Nonlinear Statistical Models*, John Wiley and Sons, New York.

Gourieroux, C., Monfort, A. & Trognon, A. (1984*b*), 'Pseudo maximum likelihood methods: Theory', *Econometrica* **52**(3), 681–700.

Gourieroux, C., Monfort, A. & Trognon, A. (1984*a*), 'Pseudo maximum likelihood methods: Applications to Poisson models', *Econometrica* **52**(3), 701–720.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (2005), *Robust Statistics: The Approach Based on Influence Functions*, Wiley Series in Probability and Statistics, revised edn, Wiley-Interscience, New York.

Hardin, J. W. (2003), The sandwich estimator of variance, *in* T. B. Fomby & R. C. Hill, eds, 'Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later', Elsevier, New York.

Hinkley, D. V. (1977), 'Jackknifing in unbalanced situations', *Technometrics* **19**(3), 285–292.

Huber, P. (1967), The behavior of the maximum likelihood estimates under nonstandard conditions, *in* 'Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, University of California Press, Berkeley, pp. 221–233.

Huber, P. (1974), *Robust Statistics*, Wiley, New York.

Kauermann, G. & Carroll, R. J. (2001), 'A note on the efficiency of sandwich covariance matrix estimation', *Journal of the American Statistical Association* **96**(456), 1387–1396.

Kolenikov, S. & Bollen, K. A. (2006), A specification test for negative error variances. Working paper.

MacKinnon, J. G. & White, H. (1985), 'Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties', *Journal of Econometrics* **29**(3), 305–325.

Maronna, R. A. (1976), 'Robust *M*-estimators of multivariate location and scatter', *Annals of Statistics* **4**(1), 51–67.

Maronna, R. A., Martin, D. R. & Yohai, V. J. (2006), *Robust Statistics: Theory and Methods*, John Wiley and Sons, New York.

Satorra, A. (1990), 'Robustness issues in structural equation modeling: A review of recent developments', *Quality and Quantity* **24**, 367–386.

Satorra, A. & Bentler, P. M. (1994), Corrections to test statistics and standard errors in covariance structure analysis, *in* A. von Eye & C. C. Clogg, eds, 'Latent variables analysis', Sage, Thousands Oaks, CA, pp. 399–419.

Skinner, C. J. (1989), Domain means, regression and multivariate analysis, *in* C. J. Skinner, D. Holt & T. M. Smith, eds, 'Analysis of Complex Surveys', Wiley, New York, chapter 3, pp. 59–88.

West, K. D. & Newey, W. K. (1987), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix', *Econometrica* **55**(3), 703–708.

White, H. (1980), 'A heteroskedasticity-consistent covariance-matrix estimator and a direct test for heteroskedasticity', *Econometrica* **48**(4), 817–838.

White, H. (1981), 'Consequences and detection of misspecified nonlinear regression models', *Journal of the American Statistical Association* **76**, 419–433.

White, H. (1982), 'Maximum likelihood estimation of misspecified models', *Econometrica* **50**(1), 1–26.

Yuan, K.-H. & Hayashi, K. (2006), 'Standard errors in covariance structure models: Asymptotics versus bootstrap', *British Journal of Mathematical and Statistical Psychology* **59**, 397–417.