Review of *Maximum Likelihood Estimation of Misspecified Models* by Halbert White: Results

Jim Harmon

University of Washington

May 29, 2012

Remind me...

Answers questions about the following in a unified framework:

- does MLE converge? (interpretation?)
- if yes, is MLE asymptotically normal?
- can properties of MLE determine model truth?

What White (Re)Proved

Convergence

- Kullback-Leibler Information Criterion Minimizing Parameter
- Asymptotically Normal (with Sandwich Covariance)

Inference Results

- Wald Test
- Lagrange Multiplier Test (Score Test)

Misspecification Results

- Information Matrix Test
- Hausman Test
- Gradient Test

Model Used for Testing

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\
E[Y_i \mid \beta_0, \beta_1, X_i] &= \beta_0 + \beta_1 X_i \\
X_i &\sim Unif(-2, 2) \\
\beta_0 &= 2 \\
\beta_1 &= 3
\end{aligned}
$$

To Err is Human

$$
\begin{aligned}
\epsilon_i &\sim N(0,1) \\
&\sim N(0, \sigma_i^2 = 1 + |X_i|) \\
&\sim (1/3) * N(1,1) + (2/3) * N(-1/2, 1) \\
&\sim \sqrt{28/30}\,(t_{30}) \\
&\sim Cauchy(0,1) \\
&\sim Unif(-\sqrt{3}, \sqrt{3}) \\
&\sim skew - N(0, 1, 1.5)
\end{aligned}
$$

Set-up

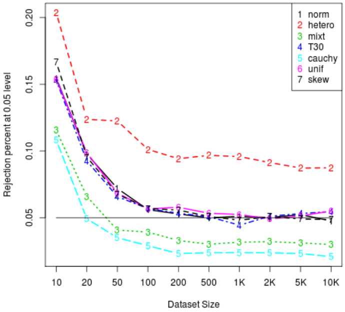
Sample sizes from 10 to 10,000

10,000 simulated datasets per size

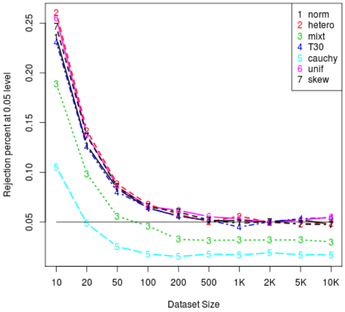
Same X-covariates within a sample size

Wald Test

$H_0 : (\beta_0, \beta_1) = (2, 3),\ H_1 : (\beta_0, \beta_1) \neq (2, 3)$

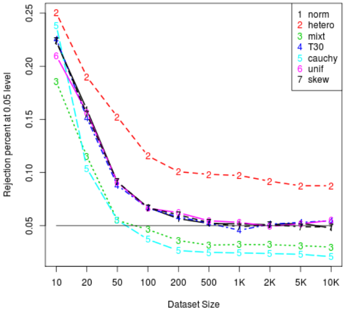


(a) MLE results

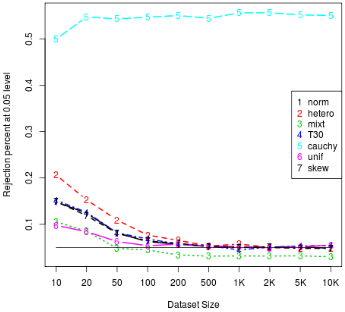


(b) Sandwich results

$H_0 : (\beta_0, \beta_1) = (2, 3),\ H_1 : (\beta_0, \beta_1) \neq (2, 3)$



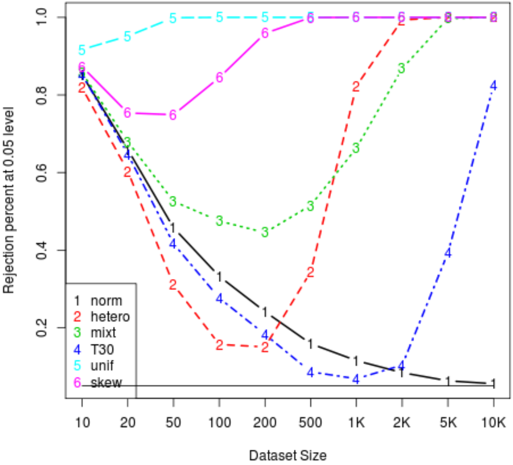
(c) MLE results



(d) Sandwich results

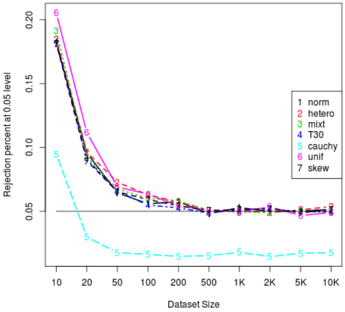
Information Matrix Test

$H_0 : \epsilon_i \sim N(\mu, \sigma^2)$, $H_1 : \epsilon_i \not\sim N(\mu, \sigma^2)$

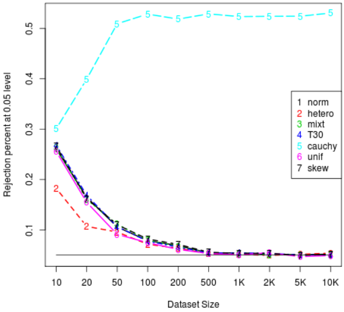


Hausman and Gradient Tests

$H_0 : \epsilon_i \sim N(\mu, \sigma^2),\ H_1 : \epsilon_i \not\sim N(\mu, \sigma^2)$



(e) Hausman Test



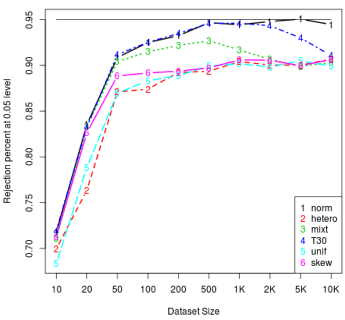
(f) Gradient Test

What Do We Do With All This???

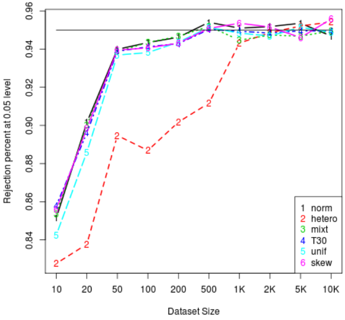
THE METHOD (according to White)

- Step 1: Perform Information Matrix test.
- Step 2a: If you "do not reject", MLE away!
- Step 2b: If you "reject", perform one of Hausman or Gradient tests.
- Step 3a: If you "do not reject", use sandwich inference.
- Step 3b: If you "reject", reconsider your model choice.

The Method

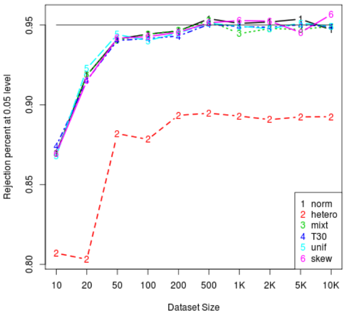


(g) β_1 Confidence Interval coverage

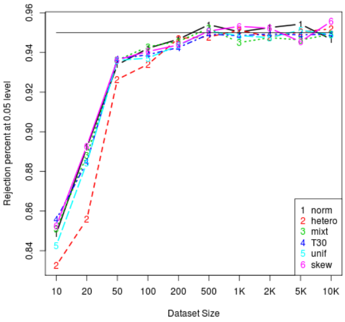


(h) β_1 Confidence Interval coverage (adjusted)

Naive MLE and Savvy Sandwich



(i) β_1 Confidence Interval coverage (MLE)



(j) β_1 Confidence Interval coverage (Sandwich)

Takeaway

The method works pretty well...

But the Sandwich works just as well

(War, Hunh!) What is it good for?

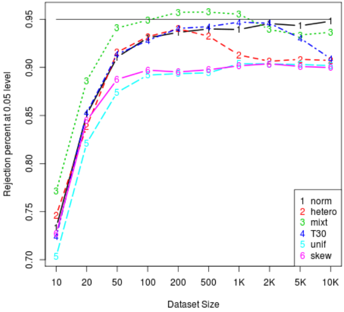
A model check based on decision theory

A reminder that the Sandwich works

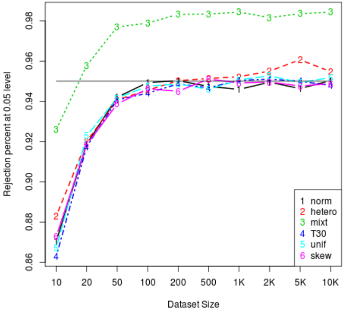
(A warning of five words)

This slide intentionally left blank.

Coverage of β_0 - METHOD

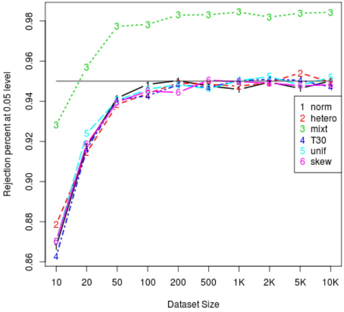


(k) β_0 Confidence Interval coverage

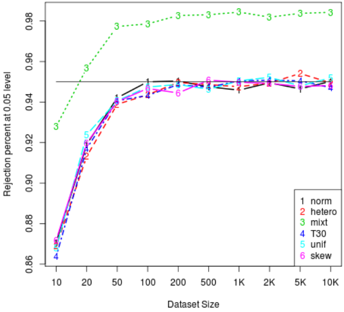


(l) β_0 Confidence Interval coverage (adjusted)

Coverage of β_0 - SIMPLE



(m) β_0 Confidence Interval coverage (MLE)



(n) β_0 Confidence Interval coverage (Sandwich)

Underlying Assumptions

A1: true density function $g(u)$ for data U_t, with distribution function G

A2: family of distributions $F(u, \theta)$, with density $f(u, \theta)$, measurable in u

for all $\theta \in \Theta$, and continuous in θ for all $u \in \Omega$

Define $L_n(U, \theta) = n^{-1} \sum_{t=1}^{n} \log f(U_t, \theta)$.

Define QMLE $= \arg\max_\theta L_n(U, \theta)$ (quasi-MLE)

Theorem

Given A1 and A2, for all n there exists a measurable QMLE, $\hat{\theta}_n$.

Note: there is an underlying dominating measure ν

Further Assumptions

A3a: $E(\log g(U_t))$ exists and $|\log f(u, \theta)|$ is bounded by an integrable function of u

A3b: KLIC $I(g : f, \theta)$ has a unique minimum at $\theta_* \in \Theta$.

Theorem

Given A1-A3, $\hat{\theta}_n \to_{a.s.} \theta_*$.

Note: All expectations are taken w.r.t. the truth, g.

Sandwich Time!!!

Need a consistent estimate of the covariance matrix:

$$\mathbf{A}(\theta) = E\left[\frac{\partial^2 log(f(U_t, \theta))}{\partial\theta_i \partial\theta_j}\right]$$

$$\mathbf{B}(\theta) = E\left[\frac{\partial log(f(U_t, \theta))}{\partial\theta_i}\frac{\partial log(f(U_t, \theta))}{\partial\theta_j}\right]$$

$$\mathbf{C}(\theta) = \mathbf{A}(\theta)^{-1}\mathbf{B}(\theta)\mathbf{A}(\theta)^{-1}$$

Guess what? – More Assummptions

A4: $\partial \log f(u, \theta)/\partial\theta_i$ with $i = 1, \ldots, p$ are measurable functions of u for each θ and continuously differentiable functions of θ for each u.

A5: $\mid \partial^2 \log f(u, \theta)/\partial\theta_i\partial\theta_j \mid$ and $\mid \partial \log f(u, \theta)/\partial\theta_i \cdot \partial \log f(u, \theta)/\partial\theta_j \mid$ with $i, j = 1, \ldots, p$ are dominated by functions integrable w.r.t. G for u and θ.

A6a: θ_* is interior to Θ

A6b: $\mathbf{B}(\theta_*)$ is nonsingular

A6c: $\mathbf{A}(\theta)$ has constant rank in some open neighborhood of θ_* (regular point)

Two Theorems

Theorem (Identification)

i: Given A1-A3a, A4-A6a, if θ_* is a unique minimum for $I(g:f,\theta)$ in an open neighborhood of Θ, and if θ_* is a regular point of $\mathbf{A}(\theta)$, then $\mathbf{A}(\theta_*)$ is negative definite.
ii: Given A1-A3a, A4-A6a, if $\mathbf{A}(\theta_*)$ is negative definite and if θ_* minimizes $I(g:f,\theta)$ in an open neighborhood of Θ, then there is an open neighborhood of Θ where θ_* is a unique minimum of $I(g:f,\theta)$.

Theorem (Asymptotic Normality)

Given A1-A6, $\sqrt{n}(\hat{\theta}_n - \theta_*) \rightarrow_d N(0, \mathbf{C}(\theta_*))$. Moreover, $\mathbf{C}_n(\hat{\theta}_n) \rightarrow_{a.s.} \mathbf{C}(\theta_*)$ element by element.

More Assumptions and Theorems?!?!

A7: $|\ \partial[\partial f(u,\theta)/\partial\theta_i \cdot f(u,\theta)]/\partial\theta_j\ |$ with $i, j = 1, \ldots, p$ are dominated by functions integrable with respect to ν for all θ in Θ and the minimal support of $f(u,\theta)$ does not depend on θ.

Theorem (Information Matrix Equivalence)

Given A1-A7, if $g(u) = f(u,\theta_0)$ for $\theta_0 \in \Theta$, then $\theta_* = \theta_0$ and $\mathbf{A}(\theta_0) = -\mathbf{B}(\theta_0)$, so that $\mathbf{C}(\theta_0) = -\mathbf{A}(\theta_0)^{-1} = \mathbf{B}(\theta_0)^{-1}$ where $-\mathbf{A}(\theta_0)^{-1}$ is Fisher's Information Matrix.

Note: A1-A7 and $g(u) = f(u,\theta_0)$ are "usual MLE regularity conditions"

Wald Test under misspecification

Suppose we wish to test $H_0 : s(\theta_0) = 0$ vs. $H_1 : s(\theta_0) \neq 0$ where

$s : \mathbf{R}^p \to \mathbb{R}^r$ is a continuous vector function of θ s.t. its Jacobian at θ_*,

$J_s(\theta_*)$ is finite with full row rank r.

Theorem (Wald Test)

$$\mathfrak{W}_n = n \cdot s(\hat{\theta}_n)'[J_s(\hat{\theta}_n)\mathbf{C}_n(\hat{\theta}_n)J_s(\hat{\theta}_n)']^{-1}s(\hat{\theta}_n) \to_d \chi^2_r$$

Lagrange Multiplier Test under misspecification

Let $\tilde{\theta}_n$ solve the constrained maximization problem $\max_{\theta \in \Theta} L_n(U, \theta)$

subject to $s(\theta) = 0$

Theorem (Lagrange Multiplier Test)

Given A1-A6 and H_0,

$$
\begin{aligned}
\mathfrak{LM}_n = {} & \nabla L_n(U, \tilde{\theta}_n)' \mathbf{A}_n(\tilde{\theta}_n)^{-1} J_s(\tilde{\theta}_n)' \\
& \times [J_s(\tilde{\theta}_n) \mathbf{C}_n(\tilde{\theta}_n) J_s(\tilde{\theta}_n)']^{-1} \\
& \times J_s(\tilde{\theta}_n) \mathbf{A}_n(\tilde{\theta}_n)^{-1} \nabla L_n(U, \tilde{\theta}_n) \\
& \to_d \chi^2_r
\end{aligned}
$$

Moreover $\mathfrak{W}_n - \mathfrak{LM}_n \to_p 0$

More Notation

θ is a p-dimensional vector.

$$d_l(U_t,\theta) = \partial log(f(U_t,\theta))/\partial\theta_i \cdot \partial log(f(U_t,\theta))/\partial\theta_j$$
$$+ \partial^2 log(f(U_t,\theta))/\partial\theta_i\partial\theta_j$$
$$dim(d) = q \times 1 \text{ with } q \leq p(p+1)/2$$
$$D_{ln}(\hat{\theta}_n) = n^{-1}\sum_{t=1}^{n} d_l(u_t,\hat{\theta}_n)$$
$$J_D(\theta) = n^{-1}\sum_{t=1}^{n} \partial d(U_t,\theta)/\partial\theta_k$$
$$W_n(\hat{\theta}_n) = d(U_t,\hat{\theta}_n) - J_D(\hat{\theta}_n)\mathbf{A}(\hat{\theta}_n)^{-1}\nabla log(f(U_t,\hat{\theta}_n))$$
$$\mathbf{V}(\theta) = n^{-1}\sum_{t=1}^{n} W_n(\hat{\theta}_n)\cdot W_n(\hat{\theta}_n)'$$

The First Specification Test!!!

A8: $\partial d_l(u,\theta)/\partial\theta_k$ for $l = 1, \dots, q$, $k = 1, \dots, p$ exist and are continuous functions of θ for each u.

A9: $\mid d_l(u,\theta) d_m(u,\theta) \mid$, $\mid \partial d_l(u,\theta)/\partial\theta_k \mid$, and $\mid d_l(u,\theta)\partial \log f(u,\theta)/\partial\theta_k \mid$, for $l, m = 1, \dots, q$, $k = 1, \dots, p$ are dominated by functions integrable w.r.t. G for all u and θ in Θ.

A10: $\mathbf{V}(\theta_*)$ is nonsingular

Theorem (Information Matrix Test)

Given A1-A10, if $g(u) = f(u,\theta_0)$ for some $\theta_0 \in \Theta$, i)

$\sqrt{n} D_n(\hat{\theta}_n) \to_d N(0, \mathbf{V}(\theta_0))$

ii) $\mathbf{V}_n(\hat{\theta}_n) \to_{a.s.} \mathbf{V}(\theta_0)$

iii) $\mathfrak{I}_n = n D_n(\hat{\theta}_n)' \mathbf{V}_n(\hat{\theta}_n)^{-1} D_n(\hat{\theta}_n) \to_d \chi^2_q$

Alternative Consistent QMLEs

Let Θ and Γ be p- and q- dimensional compact subsets of Euclidean spaces with

$\Theta = \mathbb{B} \times \Psi$ and $\Gamma = \mathbb{B} \times \mathbb{A}$, $\mathbb{B} \subset \mathbb{R}^k$ (compact)

$\hat{\theta}'_n = (\hat{\beta}'_n, \hat{\psi}'_n)$ maximizes $n^{-1} \sum \log f(U_t, \theta)$ over Θ

$\tilde{\gamma}'_n = (\tilde{\beta}'_n, \tilde{\alpha}'_n)$ maximizes $n^{-1} \sum \log h(U_t, \gamma)$ over Γ

h is a density function satisfying

A11: h satisfies A2-A6, and if $g(u) = f(u, \theta_0)$ for any $\theta'_0 = (\beta'_0, \psi'_0) \in \Theta$, then $\gamma'_* = (\beta'_0, \alpha'_*) \in \Gamma$

Note: $\tilde{\beta}_n$ is a consistent estimator of β_0 and $\sqrt{n}(\tilde{\beta}_n - \beta_0)$ is asymptotically normal, consider $\sqrt{n}(\tilde{\beta}_n - \hat{\beta}_n)$

More Definitions

$\mathbf{A}^f(\theta) = \big(E(\partial^2 \log f(U_t, \theta)/\partial\theta_i\partial\theta_j)\big)$, dimension $p \times p$

$\mathbf{B}^f(\theta) = (E(\partial \log f(U_t, \theta)/\partial\theta_i \cdot \partial \log f(U_t, \theta)/\partial\theta_j))$, dimension $p \times p$

$\mathbf{A}^h(\gamma) = \big(E(\partial^2 \log h(U_t, \gamma)/\partial\gamma_i\partial\gamma_j)\big)$, dimension $q \times q$

$\mathbf{B}^h(\gamma) = (E(\partial \log h(U_t, \gamma)/\partial\gamma_i \cdot \partial \log h(U_t, \gamma)/\partial\gamma_j))$, dimension $q \times q$

$\mathbf{A}^{f,\beta\theta}(\theta)^{-1}$ is the matrix obtained by deleting the last $p - k$ rows from the inverse of $\mathbf{A}^f(\theta)$ above.

$\mathbf{A}^{h,\beta\gamma}(\gamma)^{-1}$ is the matrix obtained by deleting the last $q - k$ rows from the inverse of $\mathbf{A}^h(\gamma)$ above.

Even more notation

$$\mathbf{R}(\theta,\gamma) = (E(\partial \log f(U_t,\theta)/\partial\theta_i \cdot \partial \log h(U_t,\gamma)/\partial\gamma_j))$$

$$\begin{aligned}\mathbf{S}(\theta,\gamma) = {} & \mathbf{A}^{h,\beta\gamma}(\gamma)^{-1}\mathbf{B}^{h}(\gamma)\mathbf{A}^{h,\beta\gamma}(\gamma)^{-1\prime} \\ & - \mathbf{A}^{h,\beta\gamma}(\gamma)^{-1}\mathbf{R}(\theta,\gamma)'\mathbf{A}^{f,\beta\theta}(\theta)^{-1\prime} \\ & - \mathbf{A}^{f,\beta\theta}(\theta)^{-1}\mathbf{R}(\theta,\gamma)\mathbf{A}^{h,\beta\gamma}(\gamma)^{-1\prime} \\ & + \mathbf{A}^{f,\beta\theta}(\theta)^{-1}\mathbf{B}^{f}(\theta)\mathbf{A}^{f,\beta\theta}(\theta)^{-1\prime}\end{aligned}$$

A12: $\mathbf{S}(\theta_*,\gamma_*)$ is nonsingular.

First of the second round of tests

Theorem (Hausman Test)

Given A1-A6, A11, and A12, if $g(u) = f(u, \theta_0)$ for $\theta_0 \in \Theta$, then

$$\mathfrak{H}_n = n(\tilde{\beta}_n - \hat{\beta}_n)' \mathbf{S}_n(\hat{\theta}_n, \tilde{\gamma}_n)^{-1}(\tilde{\beta}_n - \hat{\beta}_n) \to_d \chi^2_k$$

Gradient Test setup

$\tilde{\gamma}'_n = (\tilde{\beta}'_n, \tilde{\alpha}'_n)$ maximizes $n^{-1} \sum \log h(U_t, \gamma)$ over Γ

$\tilde{\psi}_n$ maximizes $\nabla L_n(U, \tilde{\beta}_n, \psi)$ over Ψ.

$\tilde{\theta}'_n = (\tilde{\beta}'_n, \tilde{\psi}'_n)$

$\nabla_\beta L_n(U, \tilde{\theta}_n)$ is an indicator of model misspecification

investigate asymptotic behavior of $\sqrt{n} \nabla_\beta L_n(U, \tilde{\theta}_n)$

The Last One (I Promise!!)

$\mathbf{A}_n^{f,\beta\beta}(\theta)^{-1}$ is the $k \times k$ submatrix of $\mathbf{A}_n^f(\theta)^{-1}$ obtained by deleting the last $p-k$ columns from $\mathbf{A}_n^{f,\beta\theta}(\theta)^{-1}$ (i.e., keep the upper left block)

Theorem (Gradient Test)

Given A1-A6, A11, and A12, if $g(u) = f(u, \theta_0)$ for some $\theta_0 \in \Theta$, then

$\mathfrak{G}_n = \nabla_\beta L_n(U, \tilde{\theta}_n)' \mathbf{A}_n^{f,\beta\beta}(\tilde{\theta}_n)^{-1} \mathbf{S}_n(\tilde{\theta}_n, \tilde{\gamma}_n)^{-1} \mathbf{A}_n^{f,\beta\beta}(\tilde{\theta}_n)^{-1} \nabla_\beta L_n(U, \tilde{\theta}_n) \rightarrow_d \chi_k^2$

Moreover $\mathfrak{H}_n - \mathfrak{G}_n \rightarrow_p 0$