M-Estimation

• An extremum estimator is one obtained as the optimizer of a criterion function, \( q(z, b) \).

Examples:

- OLS: \( b = \arg \max \left ( -\mathbf{e}'\mathbf{e}/T \right ) \)
- MLE: \( b_{\text{MLE}} = \arg \max \ln L = \sum_{i=1, \ldots, T} \ln f(y_i, x_i, b) \)
- GMM: \( b_{\text{GMM}} = \arg \max -g(y_i, x_i, b)' W g(y_i, x_i, b) \)

• There are two classes of extremum estimators:
  - M-estimators: The objective function is a sample average or a sum.
  - Minimum distance estimators: The objective function is a measure of a distance.

• "M" stands for a maximum or minimum estimators -- Huber (1967).
**M-Estimation**

- The objective function is a sample average or a sum.
- We want to minimize a population (first) moment:
  \[ \min_{\beta} E[q(z,\beta)] \]

  - Using the LLN, we move from the population first moment to the sample average:
    \[ \sum q(z_i,\beta) / T \rightarrow E[q(z,\beta)] \]
  - We want to obtain: \( \beta = \text{argmin} \sum q(z_i,\beta) \) (or divided by \( T \))
  - In general, we solve the f.o.c. (or zero-score condition):
    \[ \text{Zero-Score: } \sum \frac{\partial q(z_i,\beta)}{\partial \beta'} = 0 \]
  - To check the s.o.c., we define the (pd) Hessian:
    \[ H = \sum \frac{\partial^2 q(z_i,\beta)}{\partial \beta \partial \beta'} \]

**M-Estimation**

- If \( s(z,\beta) = \frac{\partial q(z,\beta)}{\partial \beta'} \) exists (almost everywhere), we solve
  \[ \sum s(z_i,\beta_M) / T = 0 \quad (*) \]

- If, in addition, \( E_X[s(z,\beta)] = \frac{\partial}{\partial \beta'} E_X[q(z,\beta)] \) - i.e., differentiation and integration are exchangeable-, then
  \[ E_X[\partial q(z,\beta) / \partial \beta'] = 0. \]

- Under this assumptions the M-estimator is said to be of \( \phi \)-type (\( \phi = s(z,\beta) \)=score). Often, \( \beta_M \) is taken to be the solution of (*) without checking whether it is indeed a minimum).

- Otherwise, the M-estimator is of \( \rho \)-type. (\( \rho = q(z,\beta) \)).
M-Estimation: LS & ML

- Least Squares
  - DGP: \( y = f(x; \beta) + \varepsilon, \quad z = [y, x] \)
  - \( q(z; \beta) = S(\beta) = \varepsilon' \varepsilon = \sum_{i=1,...,T} (y_i - f(x_i; \beta))^2 \)
  - Now, we move from population to sample moments
    - \( q(z; b) = S(b) = e' e = \sum_{i=1,...,T} (y_i - f(x_i; b))^2 \)
    - \( \hat{b}_{NLLS} = \text{argmin } S(b) \)

- Maximum Likelihood
  - Let \( f(x; \beta) \) be the pdf of the data.
  - \( L(x; \beta) = \prod_{i=1,...,T} f(x_i; \beta) \)
  - \( \log L(x; \beta) = \sum_{i=1,...,T} \ln f(x_i; \beta) \)
  - Now, we move from population to sample moments
    - \( q(z; b) = -\log L(x; b) \)
    - \( \hat{b}_{MLE} = \text{argmin } -\log L(x; b) \)

M-Estimation: Minimum \( L_p \)-estimators

- Minimum \( L_p \)-estimators
  - \( q(z; \beta) = (1/p) |x - \beta|^p \) for \( 1 \leq p \leq 2 \)
  - \( s(z; \beta) = |x - \beta|^{p-1} \)
  - \( s(z; \beta) = -|x - \beta|^{p-1} \)

- Special cases:
  - \( p = 2 \) : We get the sample mean (LS estimator for \( \beta \)).
    - \( s(z; \beta) = \sum_i (x_i - \hat{b}_M) = 0 \) \( \Rightarrow \hat{b}_M = \sum_i x_i / T \)
  - \( p = 1 \) : We get the sample median as the estimator with the least absolute deviation (LAD) for the median \( \beta \). (There is no unique solution if \( T \) is even.)

Note: Unlike LS, LAD does not have an analytical solving method. Numerical optimization is not feasible. Linear programming is used.
The Score Vector

- Let \( \{X = X_i; X_2; \ldots \} \) be i.i.d.
- If \( s(z,b) = \partial q(z,b)/\partial b' \) exists, we solve
  \[
  \sum_i s(z_i,b_M)/T = 0 \quad (s(z,b) \text{ is a } k \times 1 \text{ vector}).
  \]
  - \( E[s(z,b_0)] = E[\partial q(z,b)/\partial b'] = 0 \)
  - Using the LLN: \( \sum s(z_i,b)/T \xrightarrow{p} E[s(z,b_0)] = 0 \)
  - \( V = \text{Var}[s(z,b_0)] = E[s(z,b)s(z,b)'] \quad (V \text{ is a } k \times k \text{ matrix}). \)
    - \( E[\partial q(z,b)/\partial b'] (\partial q(z,b)/\partial b) \)
    - Using the LLN: \( \sum [s(z_i,b)s(z_i,b)']/T \xrightarrow{p} \text{Var}[s(z,b_0)] \)

- Using the Lindeberg-Levy CLT: \( \sum s(z_i,b)/\sqrt{T} \xrightarrow{d} N(0,V) \)

Note: We have already shown these results for the ML case.

The Hessian Matrix

- \( H(z,b) = E[\partial s(z,b)/\partial b] = E[\partial^2 q(z,b)/\partial b\partial b'] \)
- Using the LLN: \( \sum [\partial s(z,b)/\partial b]/T \xrightarrow{p} H(z,b_0) \)

- In general, the Information (Matrix) Equality does not hold. That is, \( H \neq V \). The equality only holds if the model is correctly specified.

  - Recall the Mean Value Theorem: \( f(x) = f(a) + f'(b) (x-a) \quad a<b<x \)
  - Apply MVT to the score:
    \[
    \sum_i s(z_i,b_M) = \sum_i s(z_i,b_0) + \sum_i H(z_i,b^*) (b_M-b_0) \quad b_0<b^*<b_M
    \]
    - \( \sum_i s(z_i,b_0) + \sum_i H(z_i,b^*) (b_M-b_0) \)
    - \( \Rightarrow (b_M-b_0) = [\sum_i H(z_i,b^*)]^{-1} \sum_i s(z_i,b_0) \)
    - \( \Rightarrow \sqrt{T} (b_M-b_0) = [\sum_i H(z_i,b^*)/T]^{-1} \sum_i s(z_i,b_0)/\sqrt{T} \)
The Asymptotic Theory

- **Theorem:** Consistency of M-estimators

Let \( \{X = X_1; X_2; \ldots \} \) be i.i.d. and assume

1. \( b \in B \), where \( B \) is compact. ("compact")
2. \( \sum q(X_i, b)/T \xrightarrow{p} g(b) \) uniformly in \( b \) for some continuous function \( g: B \rightarrow \mathbb{R} \) ("continuity")
3. \( g(b) \) has a unique global minimum at \( b_0 \). ("identification")

Then, \( b_M \xrightarrow{p} b_0 \)

**Remark:** a) Since \( X \) are i.i.d. by the LLN (without uniformity) it must hold \( g(b) = E_X[q(Xb)] \), thus \( E_X[q(z, b_0)] = \min_{b \in B} E_X[q(z, \beta)] \).

b) If \( B \) is not compact, find a compact subset \( B_0 \), with \( b_0 \in B_0 \) and \( P[b_M \in B_0] \rightarrow 1 \).

The Asymptotic Theory

- **Theorem:** Asymptotic Normality of M-estimators

Assumptions:

1. \( b_M \xrightarrow{p} b_0 \) for some \( b_0 \in B \)
2. \( b_M \) is of \( \psi \)-type and \( s \) is continuously (for almost all \( x \)) differentiable w.r.t. \( b \).
3. \( \sum [\partial s(z, b)/\partial b]/T \xrightarrow{p} H(z, b_0) \) for \( b^* \xrightarrow{p} b_0 \)
4. \( \sum s(z, b)/\sqrt{T} \xrightarrow{d} N(0, V_0) \) \( V_0 = \text{Var}[s(z, b_0)] < \infty \)

Then, \( \sqrt{T} (b_M - b_0) = \left( \sum H(z, b^*)/T \right)^{-1} \left[ -\sum s(z, b_0)/\sqrt{T} \right] \)

\[
\Rightarrow \sqrt{T} (b_M - b_0) \xrightarrow{d} N(0, H_0^{-1}V_0H_0^{-1})
\]

- \( V = E[s(z, b)s(z, b)^*] = E[(\partial q(z, b)/\partial b)(\partial q(z, b)/\partial b)] \)
- \( H = \partial s(z, b)/\partial b = E[\partial^2 q(z, b)/\partial b \partial b^*] \)
Asymptotic Normality

- Summary
  - $b_M \xrightarrow{d} b_0$
  - $b_M \xrightarrow{d} N(b_0, \text{Var}[b_0])$
  - $\text{Var}[b_M] = (1/T) H_0^{-1} V_0 H_0^{-1}$
  - If the model is correctly specified: $-H = V$

Then, $\text{Var}[b] = V_0$

- $H$ and $V$ are evaluated at $b_0$;
  - $H = \sum_i [\partial^2 q(z, b)/\partial b \partial b']$
  - $V = \sum_i [\partial q(z, b)/\partial b][\partial q(z, b)/\partial b']$

M-Estimation: Example

- DGP: $y = f(x, \beta) + \epsilon = \exp(x \beta) + \epsilon$
- Objective function:
  $q(X, \beta) = \frac{1}{2} \epsilon' \epsilon = \frac{1}{2} [y - \exp(X \beta)]' [y - \exp(X \beta)]$

- Score: $s(z, \beta) = \partial q(z, \beta)/\partial \beta = \partial f(x, \beta)/\partial \beta' \epsilon$
  $= - [\exp(X \beta)X]' [y - \exp(X \beta)]$
  $= - [\exp(X \beta)X]' \epsilon = -X' \exp(X \beta)' y + X' \exp(2\sum x_i \beta)$

- $V = \text{Var}[s(z, \beta)] = E[[\exp(X \beta)X]' \epsilon \epsilon' [\exp(X \beta)X]]$

- $H = E[\partial^2 q(z, \beta)/\partial \beta \partial \beta'] = E[\partial f(x, \beta)/\partial \beta' \partial f(x, \beta)/\partial \beta']$
  - $\partial^2 f(x, \beta)/\partial \beta \partial \beta' \epsilon = E[\exp(X \beta)X'X \exp(X \beta)' - \exp(X \beta)X' \epsilon]$

- $\text{Var}[b_M] = (1/T) H_0^{-1} V_0 H_0^{-1}$
M-Estimation: Example

- \( \text{Var}[b_M] = (1/T) \, H_0^{-1} \, V_0 \, H_0^{-1} \)

- We approximate
  \[
  \text{Var}[b_M] = (1/T) \left\{ \sum \left[ \frac{\partial s(z_i, b_M)}{\partial b_M} \right]^2 \right\}^{-1} \left[ \sum s(z_i, b_M) \, s(z_i, b_M)' \right]
  \times \left[ \sum \left[ \frac{\partial s(z_i, b_M)}{\partial b_M} \right]^2 \right]^{-1}
  \]

  \[s(z_i, b_M) = - \left[ \exp(\mathbf{x}_i' b_M) \mathbf{x}_i \right]' \left[ \mathbf{y}_i - \exp(\mathbf{x}_i' b_M) \right] = - \mathbf{x}_i' \exp(\mathbf{x}_i' b_M)' \mathbf{e}_i\]

Two-Step M-Estimation

- Sometimes, nonlinear models depend not only on our parameter of interest \( \beta \), but nuisance parameters or unobserved variables in some way. It is common to estimate \( \beta \) using a “two-step” procedure:
  1st-stage: \( y_2 = g(w_i; \gamma) + \nu \Rightarrow \) we estimate \( \gamma \), say \( c \)
  2nd-stage \( y = f(x; \beta, c) + \varepsilon \Rightarrow \) we estimate \( \beta \), given \( c \).

- The objective function: \( \min_{\beta} \{ \sum q(x; \beta, c) = \varepsilon' \varepsilon \} \)

- Examples:
  1. DHW Test for endogeneity
  2. Weighted NLLS: \( \min_{\beta} \{ \sum (y - f(x; \beta))^2 / g(z; c) \} \)
  3. Selection Bias Model: \( y = X\beta + \delta \hat{h} + \varepsilon \), \( \hat{h} = G(z; c) \).
Two-Step M-Estimation

- Properties --Pagan (1984, 1986), generated regressors:
  - Consistency. We need to apply a uniform weak LLN.
  - Asymptotic normality: We need to apply CLT.

- Two interesting results:
  - The 2S estimator can be consistent even in some cases where \( g(z;\gamma) \) is not correctly specified --i.e., situations where \( c \) may be inconsistent.
  - The S.E. --i.e., \( \text{Var}[b_{2S}] \) needs to be adjusted by the 1st stage estimation, in most cases.

Two-Step M-Estimation

Recall

\[ \sqrt{T}(b_M - b_0) = H_0^{-1}[-\sum s(z_i, b_{0i}, c)/\sqrt{T}] + o(1) \] (*)

The question is whether the following equation holds:

\[ \sum s(z_i, b_{0i}, c)/\sqrt{T} = \sum s(z_i, b_{0i}, c_0)/\sqrt{T} + o(1) \] (**)

where \( c_0 \) is the true value of \( \gamma \).

If this equality holds, \( b_M \) would be consistent.

- Let’s do a 1st order Taylor expansion:

\[ \sum s(z_i, b_{0i}, c)/\sqrt{T} \approx \sum s(z_i, b_{0i}, c_0)/\sqrt{T} + F_0(c - c_0)/\sqrt{T} \] (***)

where \( F_0 = \partial s(z, b_{0i}, c)/\partial \gamma \)

Note: If \( c = c_0 \) or \( F_0 = 0 \), then (***) holds.
Two-Step M-Estimation
• We can also write
\[
\sqrt{T} (c - c_0) = H_{c0}^{-1} \left[ \sum_i s(w_i, c) / \sqrt{T} \right] + o(1)
= \sum_i h(w_i, c) / \sqrt{T} + o(1)
\]

• Then, substituting back in (***) and then in (*), we have
\[
\sqrt{T} (b_M - b_0) = H_0^{-1} \left[ \sum_i r(z_i, b_0, c_0) / \sqrt{T} \right] + o(1), \quad (***)
\]

where \( r(z_i, b_0, c_0) = s(z_i, b_0, c_0) + F_0 h(w_i, c_0) \)

**Note**: Difference between (*) and (***): \( r(z_i, b_0, c_0) \) replaces \( s(z_i, b_0, c_0) \). The second term in \( r(z_i, b_0, c_0) \) reflects the 1st-stage adjustment.

• \( \text{Var}[b_M] = (1 / T) H_0^{-1} \text{Var}[r(z_i, b_0, c_0)] H_0^{-1} \)

Applications
• Heteroscedasticity Autocorrelation Consistent (HAC) Variance-Covariance Matrix
  – Non-spherical disturbances in NLLS

• Quasi Maximum Likelihood (QML)
  – Misspecified density assumption in ML
  – Information Equality may not hold
Special case of M-estimation: NL Regression

• We start with a regression model: \( y_i = f(x_i, \beta) + \varepsilon_i \)

• Q: What makes a regression model nonlinear?

• Recall that OLS can be applied to nonlinear functional forms. But, for OLS to work, we need intrinsic linearity –i.e., the model linear in the parameters.

Example: A nonlinear functional form, but intrinsic linear:
\[ y_i = \exp(\beta_1) + \beta_2 x_i + \beta_2 x_i^2 + \varepsilon_i \]

Example: A non intrinsic linear model:
\[ y_i = \beta_0 + \beta_1 x_i^{\beta_2} + \varepsilon_i. \]

Nonlinear Least Squares

• Least squares: Min _\beta_ \( S(\beta) = \{ \frac{1}{2} \sum_i [y_i - f(x_i, \beta)]^2 \} = \frac{1}{2} \sum_i \varepsilon_i^2 \)

F.o.c.:
\[
\frac{\partial}{\partial \beta} \left( \frac{1}{2} \sum_i [y_i - f(x_i, \beta)]^2 \right) = \frac{1}{2} \sum_i (-2) [y_i - f(x_i, \beta)] \frac{\partial f(x_i, \beta)}{\partial \beta} = -\sum_i c_i x_i^0
\]

\Rightarrow -\sum_i c_i x_i^0 = 0 \quad \text{we solve for } b_{\text{NLLS}}

In general, there is no explicit solution, like in the OLS case:
\[ b = g(X, y) = (XX)'X'y \]

• In this case, we have a nonlinear model: the f.o.c. cannot be solved explicitly for \( b_{\text{NLLS}} \). That is, the nonlinearity of the f.o.c. defines a nonlinear model.
Nonlinear Least Squares: Example

• Q: How to solve this kind of set of equations?

Example: Min $\beta$ $S(\beta) = \{\frac{1}{2} \sum_i [y_i - f(x_i, \beta)]^2 \} = \frac{1}{2} \sum_i e_i^2$

$f_i = f(x, \beta) + e_i = \beta_0 + \beta_1 x_i \beta_2 + e_i$.

f.o.c.:

$\frac{\partial}{\partial \beta_0} [\frac{1}{2} \sum_i e_i^2] = \sum_i (-1) (y_i - \beta_0 - \beta_1 x_i \beta_2) = 0$

$\frac{\partial}{\partial \beta_1} [\frac{1}{2} \sum_i e_i^2] = \sum_i (-1) (y_i - \beta_0 - \beta_1 x_i \beta_2) x_i \beta_2 = 0$

$\frac{\partial}{\partial \beta_2} [\frac{1}{2} \sum_i e_i^2] = \sum_i (-1) (y_i - \beta_0 - \beta_1 x_i \beta_2) \beta_1 x_i \beta_2 \ln(x_i) = 0$

• Nonlinear equations require a nonlinear solution. This defines a nonlinear regression model: the f.o.c. are not linear in $\beta$.

Note: If $\beta_2 = 1$, we have a linear model. We would get the normal equations from the f.o.c.

Nonlinear Least Squares: Example

Example: Min $\beta$ $S(\beta) = \{\frac{1}{2} \sum_i [y_i - (\beta_0 + \beta_1 x_i \beta_2)]^2 \}$

• From the f.o.c., we cannot solve for $\beta$ explicitly. But, using some steps, we can still minimize RSS to obtain estimates of $\beta$.

• Nonlinear regression algorithm:
1. Start by guessing a plausible values for $\beta$, say $\beta^0$.
2. Calculate RSS for $\beta^0$, => get $RSS(\beta^0)$.
3. Make small changes to $\beta^0$, => get $\beta^1$.
4. Calculate RSS for $\beta^1$, => get $RSS(\beta^1)$.
5. If $RSS(\beta^1) < RSS(\beta^0)$, => $\beta^1$ becomes your new starting point.
6. Repeat steps 3-5 until you $RSS(\beta)$ cannot be lowered. => get $\beta^*$.

$\Rightarrow \beta^*$ is the (nonlinear) least squares estimates.
**NLLS: Linearization**

- We start with a nonlinear model: $y_i = f(x_i \beta) + \epsilon_i$

- We expand the regression around some point, $\beta^0$:
  
  $f(x_i \beta) = f(x_i \beta^0) + \sum_k \left[ \left( \frac{\partial f(x_i \beta^0)}{\partial \beta_k} \right) \beta_k - \beta_k^0 \right]$
  
  $= f(x_i \beta^0) + \sum_k x_{i0} \beta_k - \beta_k^0$
  
  $= f(x_i \beta^0) - \sum_k x_{i0} \beta_k^0 + \sum_k x_{i0} \beta_k$
  
  where
  
  $f_i^0 = f(x_i \beta^0) - x_{i0} \beta^0$  (\(f_i^0\) does not depend on unknowns)

Now, $f(x_i \beta)$ is (approximately) linear in the parameters! That is,

$y_i = f_i^0 + x_{i0} \beta + \epsilon_i^0$  ($\epsilon_i^0 = \epsilon_i + \text{linearization error } i$)

$\Rightarrow y_i = y_i - f_i^0 = x_{i0} \beta + \epsilon_i^0$

---

- **NLLS: Linearization**

  - We linearized $f(x_i \beta)$ to get:
    
    $y = f^0 + X^0 \beta + \epsilon^0$  ($\epsilon^0 = \epsilon + \text{linearization error}$)
    
    $\Rightarrow y^0 = y^0 = y - f^0 = X^0 \beta + \epsilon^0$

  - Now, we can do OLS:
    
    $b_{\text{NLLS}} = (X^0 X^0)^{-1} X^0 y^0$

  - Note: $X^0$ are called pseudo-regressors.

  - In general, we get different $b_{\text{NLLS}}$ for different $\beta^0$. An algorithm can be used to get the best $b_{\text{NLLS}}$.

  - We will resort to numerical optimization to find the $b_{\text{NLLS}}$.  

---

Review of Probability and Statistics in Simulation
NLLS: Linearization

• We can also compute the asymptotic covariance matrix for the NLLS estimator as usual, using the pseudo regressors and the RSS:

\[
\text{Est. Var}[\beta_{\text{NLLS}} | X^0] = s_{\text{NLLS}}^2 (X^0 X^0)^{-1}
\]

\[
\sigma_{\text{NLLS}} = [y - f(x_i, \beta_{\text{NLLS}})]' [y - f(x_i, \beta_{\text{NLLS}})] / (T-k).
\]

• Since the results are asymptotic, we do not need a degrees of freedom correction. However, a \(df\) correction is usually included.

Note: To calculate \(s_{\text{NLLS}}^2\), we calculate the residuals from the nonlinear model, not from the linearized model (linearized regression).

NLLS: Linearization - Example

• Nonlinear model: \(y_i = f(x_i, \beta^0) + \epsilon_i = \beta_0 + \beta_1 x_i \beta_2 + \epsilon_i\)
• Linearize the model to get:

\[
y^0 = y - f^0 = X^0 \hat{\beta} + \epsilon^0,
\]

where \(f^0 = f(x_i, \beta^0) - x_i^0 \beta^0\)

Get \(x_i^0 = \frac{\partial f(x_i, \beta)}{\partial \beta} \bigg|_{\beta = \beta^0}\)

\[
\frac{\partial f(x_i, \beta)}{\partial \beta_0} = 1
\]

\[
\frac{\partial f(x_i, \beta)}{\partial \beta_1} = x_i^2
\]

\[
\frac{\partial f(x_i, \beta)}{\partial \beta_2} = \beta_1 x_i \beta_2 \ln(x_i)
\]

\[
f_{i}^0 = \beta_{0}^0 + \beta_{1}^0 x_{i}^0 \beta_{2}^0 - \{\beta_{0}^0 + \beta_{0}^1 x_{i} \beta_{0}^2 + \beta_{0}^2 \beta_{1}^0 x_{i} \beta_{0}^2 \ln(x_i)\}
\]

\[
y_{i}^0 = \beta_0 + \beta_1 x_i \beta_2 + \beta_2 \beta_1 x_i \beta_2 \ln(x) + \epsilon_{i}^0
\]

To get \(\beta_{\text{NLLS}}\), regress \(y^0\) on a constant, \(x^0 \beta_2\), and \(\beta_0' x^0 \beta_{0}^2 \ln(x)\).
### Gauss-Newton Algorithm

- Recall that $\mathbf{b}_{NLLS}$ depends on $\mathbf{b}^0$. That is, 
  
  $$
  \mathbf{b}_{NLLS}(\mathbf{b}^0) = (\mathbf{X}^0 \mathbf{X}^0)^{-1} \mathbf{X}^0 \mathbf{y}^0
  $$

- We use a Gauss-Newton algorithm to find $\mathbf{b}_{NLLS}$. Recall GN:
  
  $$
  \beta_{k+1} = \beta_k + (J^T J)^{-1} J^T \mathbf{e} \\
  \text{Jacobian} = \frac{\delta f(x_i; \beta)}{\delta \beta}.
  $$

- Given a $\mathbf{b}_{NLLS}$ at step $m$, $\mathbf{b}(m)$, we find the $\mathbf{b}_{NLLS}$ for step $j+1$ by:
  
  $$
  \mathbf{b}(j+1) = \mathbf{b}(j) + [\mathbf{X}^0(j)^T \mathbf{X}^0(j)]^{-1} \mathbf{X}^0(j)^T \mathbf{e}(j)
  $$

Columns of $\mathbf{X}^0(j)$ are the derivatives:

$$
\frac{\partial f(x_i, \mathbf{b}(j))}{\partial \mathbf{b}(j)^T}
$$

$$
\mathbf{e}(j) = \mathbf{y} - f(x, \mathbf{b}(j))
$$

- The update vector is the slopes in the regression of the residuals on $\mathbf{X}^0$. The update is zero when they are orthogonal. (Just like OLS)

### Box-Cox Transformation

- It’s a simple transformation that allows non-linearities in the CLM.

  $$
  \begin{align*}
  \mathbf{y} &= f(\mathbf{x}, \mathbf{b}) + \mathbf{e} = \sum_k \mathbf{x}_k(\lambda) \beta_k + \mathbf{e} \\
  \mathbf{x}_k(\lambda) &= (\mathbf{x}_k^\lambda - 1)/\lambda \\
  \lim_{\lambda \to 0} (\mathbf{x}_k^\lambda - 1)/\lambda &= \ln \mathbf{x}_k
  \end{align*}
  $$

- For a given $\lambda$, OLS can be used. An iterative process can be used to estimate $\lambda$. OLS standard errors have to be corrected. Probably, not a very efficient method.

- NLLS or MLE will work fine.

- We can have a more general Box-Cox transformation model:

  $$
  \mathbf{y}^{(x1)} = \sum_k \mathbf{x}_k^{(x2)} \beta_k + \mathbf{e}
  $$
Testing non-linear restrictions

- Testing linear restrictions as before.
- Non-linear restrictions introduce slight modification to the usual tests. We want to test:
  \[ H_0: R(\beta) = 0 \]
  where \( R(\beta) \) is a non-linear function, with rank \( \frac{\partial R(\beta)}{\partial \beta} = G(\beta) \) = J.

- A Wald test can be based on \( m = R(b_{NLLS}) - 0 \):
  \[ W = m'(\text{Var}[m | X])^{-1}m = R(b_{NLLS})'(\text{Var}[R(b_{NLLS}) | X])^{-1}R(b_{NLLS}) \]

Problem: We do not know the distribution of \( R(b_{NLLS}) \), but we know the distribution of \( b_{NLLS} \).

Solution: Linearize \( R(b_{NLLS}) \) around \( \beta \)
\[ R(b_{NLLS}) \approx R(\beta) + G(b_{NLLS})(b_{NLLS} - \beta) \]

Testing non-linear restrictions

- Linearize \( R(b_{NLLS}) \) around \( \beta (=b_0) \)
  \[ R(b_{NLLS}) \approx R(\beta) + G(b_{NLLS})(b_{NLLS} - \beta) \]

- Recall \( \sqrt{T} (b_M - b_0) \rightarrow N(0, \text{Var}[b_0]) \)
  where \( \text{Var}[b_0] = H(\beta)^{-1}V(\beta)H(\beta)^{-1} \)
  \[ => \sqrt{T} [R(b_{NLLS}) - R(\beta)] \rightarrow N(0, G(\beta) \text{Var}[b_0] G(\beta)') \]
  \[ => \text{Var}[R(b_{NLLS})] = (1/T) G(\beta) \text{Var}[b_0] G(\beta)' \]

- Then,
  \[ W = TR(b_{NLLS})'(G(b_{NLLS}) \text{Var}[b_{NLLS}] G(b_{NLLS})')^{-1}R(b_{NLLS}) \]
  \[ => W \rightarrow \chi^2_J \]
### NLLS - Application: A NIST Application (Greene)

<table>
<thead>
<tr>
<th>Y</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.138</td>
<td>1.309</td>
</tr>
<tr>
<td>3.421</td>
<td>1.471</td>
</tr>
<tr>
<td>3.597</td>
<td>1.490</td>
</tr>
<tr>
<td>4.340</td>
<td>1.565</td>
</tr>
<tr>
<td>4.882</td>
<td>1.611</td>
</tr>
<tr>
<td>5.660</td>
<td>1.680</td>
</tr>
</tbody>
</table>

The equation is:

\[
y = \beta_0 + \beta_1 x^{\beta_2} + \epsilon.
\]

### NLLS - Application: Iterations (Greene)

```
NLSQ;LHS=Y ;FCN=b0+B1*X^B2 ;LABELS=b0,B1,B2 ;MAXIT=500;TLF;TLB;OUTPUT=1;DFC;START=0,1,5  
BEGIN NLSQ iterations. Linearized regression.
Iteration= 1; Sum of squares= 149.719219 ; Gradient= 149.718223
Iteration= 2; Sum of squares= 5.04072877 ; Gradient= 5.03960538
Iteration= 3; Sum of squares= .137768222E-01; Gradient= .125711747E-01
Iteration= 4; Sum of squares= .186786786E-01; Gradient= .174668584E-01
Iteration= 5; Sum of squares= .121182327E-02; Gradient= .301702148E-08
Iteration= 6; Sum of squares= .121182025E-02; Gradient= .134513256E-15
Iteration= 7; Sum of squares= .121182025E-02; Gradient= .644990175E-20
Convergence achieved
```

Gradient = \[e^0 'X^0 \cdot [X^0 'X^0]^{-1} \cdot X^0 ' \cdot e^0\]
NLLS - Application: Results (Greene)

User Defined Optimization
Nonlinear least squares regression
LHS=Y
Mean = 4.00633
Standard deviation = 1.23398
Number of observs. = 6
Model size
Parameters = 3
Degrees of freedom = 3
Residuals
Sum of squares = .00121
Standard error of e = .02010
Fit
R-squared = .99984

| Variable | Coefficient | Standard Error | b/St.Er. | P[|Z|>z] |
|----------|-------------|----------------|----------|--------|
| B0       | -.54559**   | .22460         | -2.429   | .0151  |
| B1       | 1.08072***  | .13698         | 7.890    | .0000  |
| B2       | 3.37287***  | .17847         | 18.899   | .0000  |

NLLS – Application: Solution (Greene)

The pseudo regressors and residuals at the solution are:

X10   X20   X30
1  xβ2  β1xβ2lnx  e0
1  2.47983  0.721624  .0036
1  3.67566  1.5331  -.0058
1  3.83826  1.65415  -.0055
1  4.52972  2.19255  -.0097
1  4.99466  2.57397  .0298
1  5.75358  3.22585  -.0124

X0’e0 = .3375078D-13
.3167466D-12
.1283528D-10
Application 2: Doctor Visits (Greene)

- German Individual Health Care data: N=27,236
- Model for number of visits to the doctor

Application 2: Conditional Mean and Projection

Notice the problem with the linear approach. Negative predictions.
Application 2: NL Model Specification (Greene)

- Nonlinear Regression Model \( y = \exp(\mathbf{X}\beta) + \varepsilon \)

\( \mathbf{X} = \text{one, age, health\_status, married, educ., household\_income, nkids} \)

- \( \text{nlsq;lhs=docvis;start=0,0,0,0,0,0,0;labels=k\_b;fcn=exp(b1'x);maxit=25;out...} \)

Begin NLSQ iterations. Linearized regression.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Sum of squares</th>
<th>Gradient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1014865.00</td>
<td>257025.070</td>
</tr>
<tr>
<td>2</td>
<td>.130154610E+11</td>
<td>.130145942E+11</td>
</tr>
<tr>
<td>3</td>
<td>.175441482E+10</td>
<td>.175354986E+10</td>
</tr>
<tr>
<td>4</td>
<td>235369144.</td>
<td>234509185.</td>
</tr>
<tr>
<td>5</td>
<td>3161046.6</td>
<td>30763872.3</td>
</tr>
<tr>
<td>6</td>
<td>4684627.59</td>
<td>3871393.70</td>
</tr>
<tr>
<td>7</td>
<td>1224759.31</td>
<td>467169.410</td>
</tr>
<tr>
<td>8</td>
<td>778596.192</td>
<td>33500.289</td>
</tr>
<tr>
<td>9</td>
<td>746343.830</td>
<td>450.321350</td>
</tr>
<tr>
<td>10</td>
<td>745898.272</td>
<td>.287180441</td>
</tr>
<tr>
<td>11</td>
<td>745897.985</td>
<td>.929822308E-03</td>
</tr>
<tr>
<td>15</td>
<td>745897.984</td>
<td>.188041512E-10</td>
</tr>
</tbody>
</table>

Application 2: NL Regression Results (Greene)

<table>
<thead>
<tr>
<th>Nonlinear least squares regression</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LHS=DOCVIS Mean</td>
<td>3.183525</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>5.689690</td>
</tr>
<tr>
<td>WTS=none Number of observs.</td>
<td>27326</td>
</tr>
<tr>
<td>Model size Parameters</td>
<td>7</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>27319</td>
</tr>
<tr>
<td>Residuals Sum of squares</td>
<td>745898.0</td>
</tr>
<tr>
<td>Standard error of e</td>
<td>5.224584</td>
</tr>
<tr>
<td>Fit R-squared</td>
<td>.1567778</td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>.1568087</td>
</tr>
<tr>
<td>Info crit. LogAmemiya Prd. Crt.</td>
<td>3.307006</td>
</tr>
<tr>
<td>Akaike Info. Criter.</td>
<td>3.307263</td>
</tr>
<tr>
<td>Not using OLS or no constant.</td>
<td>R^2 &amp; P may be &lt; 0.</td>
</tr>
</tbody>
</table>

| Variable  | Coefficient  | Standard Error  | b/St.Er. | P[|Z|>z] |
|-----------|--------------|-----------------|----------|----------|
| B1        | 2.37667859   | .06972582       | 34.086   | .0000    |
| B2        | .00809310    | .00088490       | 9.146    | .0000    |
| B3        | -.21721398   | .00313992       | -69.178  | .0000    |
| B4        | .00371129    | .02051147       | .181     | .8564    |
| B5        | -.01096227   | .00435601       | -2.517   | .0118    |
| B6        | -.26584001   | .05664473       | -4.693   | .0000    |
| B7        | -.09152326   | .02128053       | -4.301   | .0000    |
Partial Effects in the Nonlinear Model (Greene)

What are the slopes?
Conditional Mean Function = \( E[y|x] = \exp(\mathbf{x}'\beta) \)
Derivatives of the conditional mean are the partial effects
\[
\frac{\partial E[y|x]}{\partial \mathbf{x}} = \exp(\mathbf{x}'\beta) \times \beta
\]
= a scaling of the coefficients that depends on the data
Usually computed using the sample means of the data.

Asymptotic Variance of the Slope Estimator (Greene)

\[ \hat{\delta} = \text{estimated partial effects} = \frac{\partial \hat{E}[y|x]}{\partial \mathbf{x}} \big| (\mathbf{x} = \bar{x}) \]

To estimate \( \text{Asy.Var}[\hat{\delta}] \), we use the delta method:
\[ \hat{\delta} = \exp(\bar{x}'\hat{\beta}) \hat{\beta} \]
\[ \hat{\mathbf{G}} = \frac{\partial \hat{\delta}}{\partial \beta} = \exp(\bar{x}'\hat{\beta}) \mathbf{I} + \hat{\beta} \exp(\bar{x}'\hat{\beta})\bar{x}' \]
\[ \text{Est.Asy.Var}[\hat{\delta}] = \hat{\mathbf{G}} \text{Est.Asy.Var}[\hat{\beta}] \hat{\mathbf{G}}' \]
Computing the Slopes (Greene)

calc;k=col(x)$
nlsq;lhs=docvis;start=0,0,0,0,0,0,0
          ;labels=k_b;fcn=exp(b1'x);
matr;xbar=mean(x)$
calc;mean=exp(xbar'b)$
matr;me=b*mean$
matr;g=mean*iden(k)+mean*b*xbar$
matr;vme=g*varb*g'$
matr;stat(me,vme)$

Partial Effects at the Means of X (Greene)

| Variable | Coefficient | Standard Error | b/St. Er. | P[|Z|>z] |
|----------|-------------|----------------|-----------|---------|
| Constant | 6.48148***  | .20680         | 31.342    | .0000   |
| AGE      | .02207***   | .00239         | 9.216     | .0000   |
| HSAT     | -.59241***  | .00660         | -89.740   | .0000   |
| MARRIED  | .01005      | .05593         | .180      | .8574   |
| EDUC     | -.02988**   | .01186         | -2.519    | .0118   |
| HHNINC   | -.72495***  | .15450         | -4.692    | .0000   |
| HHKIDS   | -.24958***  | .05796         | -4.306    | .0000   |
What About Just Using LS? (Greene)

| Variable | Coefficient | Standard Error | b/St.Er. | P(|Z|>z) | Mean of X |
|----------|-------------|----------------|----------|----------|-----------|
| Constant | .02385640   | .00327769      | 7.278    | .0000    | 43.5256898 |
| AGE      | -.02458941  | .01441043      | -.294    | .7688    | .75861817  |
| NEWHSAT  | -.86828751  | .01455653      | -60.254  | .0000    | 6.78566201 |
| MARRIED  | -.02458941  | .01441043      | -.294    | .7688    | .75861817  |
| EDUC     | -.04909154  | .01455653      | -3.372   | .0000    | 11.3206310 |
| HHINCOME | -.02174923  | .01455653      | -5.353   | .0000    | 11.3206310 |
| HHKIDS   | -.38033746  | .07513138      | -5.062   | .0000    | 40.073000  |

**Least Squares Coefficient Estimates**

| Variable | Coefficient | Standard Error | t-Value | P(|t|>t) |
|----------|-------------|----------------|---------|---------|
| Constant | 9.12437987  | .25731934      | 35.459  | .0000   |
| AGE      | .02385640   | .00327769      | 7.278   | .0000   |
| NEWHSAT  | -.86828751  | .01441043      | -60.254 | .0000   |
| MARRIED  | -.02458941  | .01441043      | -.294   | .7688   |
| EDUC     | -.04909154  | .01455653      | -3.372  | .0000   |
| HHINCOME | -.02174923  | .01455653      | -5.353  | .0000   |
| HHKIDS   | -.38033746  | .07513138      | -5.062  | .0000   |

**Estimated Partial Effects**

<table>
<thead>
<tr>
<th>ME_1</th>
<th>Constant term, marginal effect not computed</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME_2</td>
<td>.02207102</td>
</tr>
<tr>
<td>ME_3</td>
<td>-.59237330</td>
</tr>
<tr>
<td>ME_4</td>
<td>.01012122</td>
</tr>
<tr>
<td>ME_5</td>
<td>-.02989567</td>
</tr>
<tr>
<td>ME_6</td>
<td>-.72498339</td>
</tr>
<tr>
<td>ME_7</td>
<td>-.24959690</td>
</tr>
</tbody>
</table>