Microeconometrics

Nonlinear Models

by Vanessa Berenguer-Rico

University of Oxford

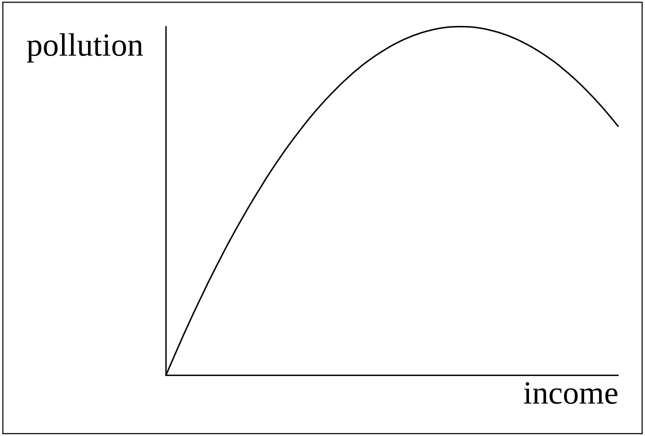
Michaelmas Term 2016

Outline

- Nonlinear Models: Preliminaries
  - Nonlinearities in variables
  - Nonlinearities in Parameters
- Estimation Methods
  - Nonlinear Least Squares (NLLS)
  - Maximum Likelihood (ML)

Nonlinear Models: Preliminaries

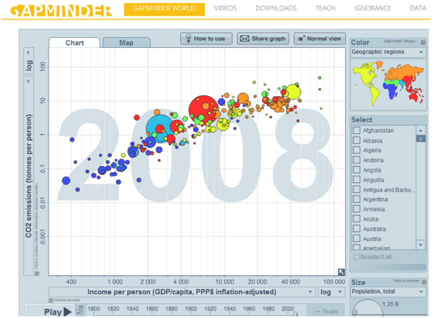
Environmental Kuznets Curve Hypothesis



Nonlinear Models: Preliminaries

Environmental Kuznets Curve Hypothesis

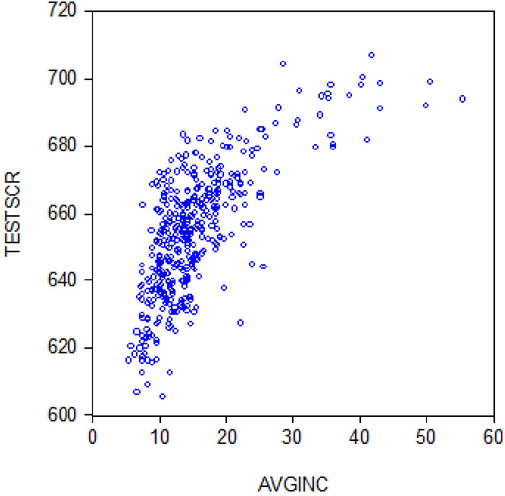
- GapMinder Data: http://www.gapminder.org/



- BTW: The Joy of Stats (Poisson min 22:02) (http://www.gapminder.org/videos/the-joy-of-stats/)

Nonlinear Models: Preliminaries

Test Score vs District Income (thousands of dollars)
Stock and Watson, p. 297



Nonlinear Models: Preliminaries

Types Nonlinear Regression Models

- In this course: **Parametric Models**. Example:

$$y_i = g\left(x_i, \theta\right) + \varepsilon_i$$

- Nonparametric Models. Example:

$$y_i = h\left(x_i\right) + u_i$$

- Semiparametric Models. Example:

$$y_i = m\left(x_i\right) + \beta z_i + e_i$$

Nonlinear Models: Preliminaries

Nonlinear Parametric Regression Models

- Nonlinear in variables

$$y_i = \theta g\left(x_i\right) + u_i$$

  Examples: Piecewise linear, polynomials, logarithmic

- Nonlinear in parameters

$$y_i = g\left(x_i, \theta\right) + u_i$$

  Examples: Exponential, Logistic, CES production function, Box-Cox

Nonlinear Models: Preliminaries

Nonlinear Parametric Regression Models

- Intrinsically linear models. Example

$$y_i = Ak_i^{\alpha} l_i^{\beta} e^{u_i}$$

- Intrinsically non-linear. Example:

$$y_i = Ak_i^{\alpha} l_i^{\beta} + u_i$$

Nonlinear Models: Nonlinear in Variables

Nonlinearity in Variables: Examples

- Polynomials (SW p. 297)

$$testScore_i = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2 + \beta_3 Income_i^3 + u_i$$

- Logarithms (SW p. 312)

$$\ln(testScore_i) = \beta_0 + \beta_1 \ln(Income_i) + u_i$$

Nonlinear Models: Nonlinear in Variables

Nonlinearity in Variables: Examples

- Dummies: Piecewise linear (SW p. 322)

$$testScore_i = \beta_0 + \beta_1 STR_i + \beta_2 HiEL_i + \beta_3 STR_i \times HiEL_i + u_i$$

- Cross-Products: (SW p. 328)

$$testScore_i = \beta_0 + \beta_1 STR_i + \beta_2 PctEL_i + \beta_3 STR_i \times PctEL_i + u_i$$

Nonlinear Models: Nonlinear in Variables

Nonlinearity in Variables: Polynomials

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_2 x_i^r + u_i$$

- Linear in parameters: OLS
- Testing the null that the model is linear: $H_o : \beta_2 = \beta_3 = ... = \beta_r = 0$ vs H_a : at least one $\beta_j \neq 0$, $j = 2, ..., r$ ($q = r - 1$ restrictions. Wald test)
- Which degree polynomial should we use? General to Specific vs Specific to General

Nonlinear Models: Nonlinear in Variables

Nonlinearity in Variables: Polynomials

Example (Stock and Watson)

$$\widehat{testScore}_i = \underset{(117.66)}{600.1} + \underset{(7.07)}{5.02} Income_i - \underset{(-3.31)}{0.096} Income_i^2 + \underset{(1.971)}{0.00069} Income_i^3$$

- Which degree polynomial should we use?
- General to Specific vs Specific to General

Nonlinear Models: Nonlinear in Variables

Nonlinearity in Variables: Logarithms (SW p. 314)

- Case I
  $$y_i = \beta_1 + \beta_2 \ln(x_i) + u_i$$
  A 1% change in x is associated with a change in y of $0.01\beta_2$
- Case II
  $$\ln(y_i) = \beta_1 + \beta_2 (x_i) + u_i$$
  A change in x by 1 unit is associated with a $100\beta_2\%$ change in y
- Case III
  $$\ln(y_i) = \beta_1 + \beta_2 \ln(x_i) + u_i$$
  A 1% change in x is associated with a $\beta_2\%$ change in y, i.e. β_2 is the elasticity of y with respect to x

Nonlinear Models: Nonlinear in Variables

Recall: Elasticities
Stock and Watson, p. 352

- Let
$$y_i = f(x_i, \theta) + u_i \quad \text{with} \quad E(y_i | x_i) = f(x_i, \theta)$$
- Slope of $f(x)$ evaluated at at point $\tilde{x}$ is
$$slope(\tilde{x}) = \left. \frac{df(x)}{dx} \right|_{x=\tilde{x}}$$
- Elasticity of y with respect to x:
$$\varepsilon_{yx} = \frac{\frac{dy}{y}}{\frac{dx}{x}} = \frac{dy}{dx}\frac{x}{y} = \frac{d \ln y}{d \ln x}$$

Nonlinear Models: Nonlinear in Variables

Recall: Elasticities
Stock and Watson, p. 352

- Let
$$y_i = f\left(x_i, \theta\right) + u_i \quad \text{with} \quad E\left(y_i | x_i\right) = f\left(x_i, \theta\right)$$
- Elasticity of $E\left(y|x\right)$ with respect to x:
$$\varepsilon_{E(y|x)x} = \frac{\frac{dE(y|x)}{E(y|x)}}{\frac{dx}{x}} = \frac{dE\left(y|x\right)}{dx}\frac{x}{E\left(y|x\right)} = \frac{d\ln E\left(y|x\right)}{d\ln x}$$

Nonlinear Models: Nonlinear in Variables

Recall: Elasticities

- Linear: $y = \beta_1 + \beta_2 x + u$

$$\varepsilon_{E(y|x)x} = \beta_2 x / \left(\beta_1 + \beta_2 x\right)$$

- Linear-log: $y = \beta_1 + \beta_2 \ln(x) + u$

$$\varepsilon_{E(y|x)x} = \beta_2 / \left(\beta_1 + \beta_2 \ln(x)\right)$$

- Log-linear: $\ln(y) = \beta_1 + \beta_2 x + u$

$$\varepsilon_{E(y|x)x} = \beta_2 x$$

- Log-Log: $\ln(y) = \beta_1 + \beta_2 \ln(x) + u$

$$\varepsilon_{E(y|x)x} = \beta_2$$

Nonlinear Models: Nonlinear in Variables

Nonlinearity in Variables: Logarithms
(Stock and Watson)

Case I

$$\widehat{testScore}_i = \underset{(3.8)}{557.8} + \underset{(1.40)}{36.42} \ln\left(Income_i\right)$$

A 1% increase in income is associated with an increase in test scores of $0.01 \times 36.42 = 0.36$ points

Nonlinear Models: Nonlinear in Variables

Nonlinearity in Variables: Logarithms
(Stock and Watson)

Case II

$$\widehat{\ln\left(testScore_i\right)} = \underset{(0.003)}{6.439} + \underset{(0.00018)}{0.00284} Income_i$$

A change in income by 1 unit is associated with a 0.28% change in test scores

Nonlinear Models: Nonlinear in Variables

Nonlinearity in Variables: Logarithms
(Stock and Watson)

Case III

$$\ln\left(\widehat{testScore_i}\right) = \underset{(0.006)}{6.336} + \underset{(0.0021)}{0.0554} \ln\left(Income_i\right)$$

A 1% change in income is associated with a 0.0554% change in test scores, i.e. the elasticity of test scores with respect to income is 0.0554

Nonlinear Models: Nonlinear in Variables

Nonlinearity in Variables: Dummies

- Constant

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + u_i$$

- Slope

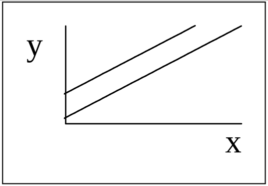
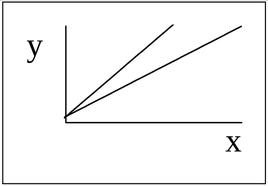
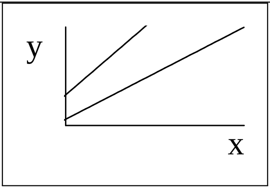
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i D_i + u_i$$

- Constant & Slope

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 x_i D_i + u_i$$

Nonlinear Models: Nonlinear in Variables

Nonlinearity in Variables: Dummies

Constant	Slope	Constant&Slope
		

Nonlinear Models: Nonlinear in Variables

Dummies: Student-teacher ratio and % of English learners
(Stock and Watson)

$$\widehat{testScore}_i = \underset{(11.9)}{682.2} - \underset{(0.59)}{0.97} STR_i + \underset{(19.5)}{5.6}\, HiEL_i + \underset{(0.97)}{1.28} STR_i \times HiEL_i$$

- Districts with a low fraction of English learners ($HiEL_i = 0$) estimated regression: $682.2 - 0.97 STR_i$
- Districts with a high fraction of English learners ($HiEL_i = 1$) estimated regression: $682.2 + 5.6 - 0.97 STR_i - 1.28 STR_i = 687.8 - 2.25 STR_i$
- Reduction student-ratio by 1 increases test scores by 0.97 point in districts with low fractions of English learners but 2.25 points in districts with high fractions of English learners

Nonlinear Models: Nonlinear in Variables

Nonlinearity in Variables: Cross-Products
(Stock and Watson)

$$\widehat{testScore}_i = \underset{(11.9)}{686.3} - \underset{(0.59)}{1.12} STR_i - \underset{(0.37)}{0.67} PctEL_i + \underset{(0.019)}{0.0012} STR_i \times PctEL_i + u_i$$

- Percentage of English learners at the median ($PctEL_i = 8.85$), the slope of the line relating test scores and student-teacher ratio is: $-1.11 = -1.12 + 0.0012 \times 8.85$
- For a district with 8.85% English learners, the estimated effect of a one-unit reduction in the student-teacher ratio is to increase test scores by 1.11 point

Nonlinear Models: Nonlinear in Variables

Nonlinearity in Variables: Cross-Products
(Stock and Watson)

$$\widehat{testScore}_i = \underset{(11.9)}{686.3} - \underset{(0.59)}{1.12} STR_i - \underset{(0.37)}{0.67} PctEL_i + \underset{(0.019)}{0.0012} STR_i \times PctEL_i + u_i$$

- Percentage of English learners at the 75th percentile $(PctEL_i = 23.0)$, the slope of the line relating test scores and student-teacher ratio is: $-1.09 = -1.12 + 0.0012 \times 23$
- For a district with 8.85% English learners, the estimated effect of a one-unit reduction in the student-teacher ratio is to increase test scores by 1.11 point
- Notice, however that $t = 0.0012/0.019 = 0.06$

Nonlinear Models: Nonlinear in Parameters

Nonlinearity in Parameters: Examples

- Logistic Curve (Stock and Watson p. 348)

$$y_i = \frac{1}{1 + e^{-(\beta_1 + \beta_2 x_i)}} + u_i$$

- Negative Exponential Growth (Stock and Watson p. 349)

$$y_i = \beta_1 \left(1 - e^{\beta_2 \left(x_i - \beta_3\right)}\right) + u_i$$

Nonlinear Models: Nonlinear in Parameters

Nonlinearity in Parameters: Examples
(see Greene p.233)

- Box Cox transformation (Box and Cox, 1964)
- Generalization of the linear model

$$y_i = \alpha + \beta \frac{x_i^{\lambda} - 1}{\lambda} + u_i$$

  - If $\lambda = 1$: linear model: $y_i = \alpha^* + \beta x_i + u_i$
  - If $\lambda = 0$: by L'Hôpital rule

$$\lim_{\lambda \to 0} \frac{x_i^{\lambda} - 1}{\lambda} = \lim_{\lambda \to 0} \frac{d\left(x_i^{\lambda} - 1\right)/d\lambda}{1} = \lim_{\lambda \to 0} x^{\lambda} \times \ln x = \ln x$$

  - If $\lambda = -1$: linear model: $y_i = \tilde{\alpha} + \tilde{\beta} \frac{1}{x_i} + u_i$

Nonlinear Models: Nonlinear in Parameters

Nonlinearity in Parameters: Examples
(see Greene p.222)

- CES production function: Constant elasticity of substitution

$$y = \gamma \left[\delta k^{-\rho} + (1-\delta)\, l^{-\rho} \right]^{-\nu/\rho} e^{\varepsilon}$$

- Take logs:

$$\ln y = \ln \gamma - \frac{\nu}{\rho} \ln \left[\delta k^{-\rho} + (1-\delta)\, l^{-\rho} \right] + \varepsilon$$

still nonlinear in parameters

Nonlinear Models: Nonlinear in Parameters

Coefficient Interpretation in Nonlinear Regression
(Cameron and Trivedi p. 122)

$$y = g\left(x, \theta\right) + u \quad \text{with} \quad E\left[y|x\right] = g\left(x, \theta\right)$$

Marginal effects

$$\frac{\partial E\left[y|x\right]}{\partial x}$$

- Linear model: $y = \beta x + u$

$$\frac{\partial E\left[y|x\right]}{\partial x} = \beta$$

- Exponential model: $y = \exp\left(\beta x\right) + u$

$$\frac{\partial E\left[y|x\right]}{\partial x} = \exp\left(\beta x\right)\beta$$

Nonlinear Models: Nonlinear in Parameters

Marginal effects: Three different estimates
(Cameron and Trivedi p. 122)

- Average response of all individuals

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial E\left[y_i|x_i\right]}{\partial x_i}$$

- Response of the average individual

$$\left.\frac{\partial E\left[y|x\right]}{\partial x}\right|_{\bar{x}} = \exp\left(\beta\bar{x}\right)\beta$$

- Response of the representative individual with $x = x^*$ (e.g. person who is female with 12 years of schooling)

$$\left.\frac{\partial E\left[y|x\right]}{\partial x}\right|_{x^*} = \exp\left(\beta x^*\right)\beta$$

Nonlinear Models: Nonlinear in Parameters

Single Index Models
(Cameron and Trivedi p. 123)

- Let

$$E[y|x] = g(x'\beta)$$

where x is a vector of regressors ($x' = [x_1\ x_2\ ...x_k]$)

- Marginal effect of x_j:

$$\frac{\partial E[y|x]}{\partial x_j} = \dot{g}(x'\beta)\,\beta_j$$

where $\dot{g}(z) = \partial g(z)/\partial z$

- Therefore, relative effects

$$\frac{\partial E[y|x]/\partial x_j}{\partial E[y|x]/\partial x_k} = \frac{\beta_j}{\beta_k}$$

Nonlinear Regression Models: Estimation

- Nonlinear in Variables: OLS

$$y_i = \theta g\left(x_i\right) + \varepsilon_i$$

OLS Estimator: $g\left(x_i\right)$ regressor (Interpretation)

- Nonlinear in parameters:

$$y_i = g\left(x_i, \theta\right) + \varepsilon_i$$

Extremum Estimators: NLLS, Maximum Likelihood, GMM

Nonlinear Models: Nonlinear in Parameters

Nonlinearity in Parameters: Examples

- Logistic Curve

$$y_i = \frac{1}{1 + e^{-(\beta_1 + \beta_2 x_i)}} + u_i$$

- Negative Exponential Growth

$$y_i = \beta_1 \left(1 - e^{\beta_2 \left(x_i - \beta_3\right)}\right) + u_i$$

- Box-Cox type

$$y_i = \alpha + \beta x_i^{\gamma} + u_i$$

- CES production function

$$\ln y = \ln \gamma - \frac{\nu}{\rho} \ln \left[\delta k^{-\rho} + (1 - \delta)\, l^{-\rho}\right] + \varepsilon$$

Nonlinear Models: Estimation

Extremum Estimators
(Cameron and Trivedi p. 124)

Extremum estimators are a very general class of estimators that minimize or maximize an objective function. In particular, an estimator $\hat{\theta}$ is called an extremum estimator if there is an objective function $Q_n(\theta)$ such that

$$\hat{\theta} = \arg\max_{\theta \in \Theta} Q_n(\theta)$$

where Θ is the parameter space i.e. the set of possible parameter values (ex. $\Theta \subset \mathbb{R}^K$)

Nonlinear Models: Estimation

Extremum Estimators: Two Examples

- NLLS: Nonlinear Least Squares
- ML: Maximum Likelihood

Nonlinear Least Squares

Let

$$y_i = g\left(x_i, \theta_0\right) + u_i$$

where θ_0 a $K \times 1$ vector of unknown parameters.

The Nonlinear Least Squares (NLLS) estimator is defined as

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} Q_n\left(\theta\right)$$

where

$$Q_n\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n} u_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - g\left(x_i, \theta\right)\right)^2$$

Nonlinear Least Squares

The Nonlinear Least Squares (NLLS) estimator is defined as

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} Q_n(\theta) = \underset{\theta \in \Theta}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - g(x_i, \theta)\right)^2$$

First order conditions:

$$\dot{Q}_n(\theta) = \frac{\partial Q_n(\theta)}{\partial \theta} = -2 \frac{1}{n} \sum_{i=1}^{n} \left(y_i - g(x_i, \theta)\right) \dot{g}(x_i, \theta) = 0$$

where $\dot{g}(x_i, \theta) = \partial g(x_i, \theta) / \partial \theta$.

Nonlinear Least Squares

Example:

- The model

$$y_i = \theta_{00} + \theta_{01} x_{1i} + \theta_{02} x_{2i} + \theta_{01}\theta_{02} x_{3i} + u_i$$

- NLLS

$$\begin{aligned}
\hat{\theta} &= \arg\min_{\theta \in \Theta} Q_n(\theta) \\
&= \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \theta_0 - \theta_1 x_{1i} - \theta_2 x_{2i} - \theta_1 \theta_2 x_{3i} \right)^2 \\
&= \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} u_i^2(\theta)
\end{aligned}$$

Nonlinear Least Squares

- NLLS

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \theta_0 - \theta_1 x_{1i} - \theta_2 x_{2i} - \theta_1 \theta_2 x_{3i} \right)^2$$

- First order conditions:

$$\begin{aligned}
\frac{\partial Q_n(\theta)}{\partial \theta_0} &= -2 \frac{1}{n} \sum_{i=1}^{n} u_i(\theta) = 0 \\
\frac{\partial Q_n(\theta)}{\partial \theta_1} &= -2 \frac{1}{n} \sum_{i=1}^{n} u_i(\theta) \left(x_{1i} + \theta_2 x_{3i} \right) = 0 \\
\frac{\partial Q_n(\theta)}{\partial \theta_2} &= -2 \frac{1}{n} \sum_{i=1}^{n} u_i(\theta) \left(x_{2i} + \theta_1 x_{3i} \right) = 0
\end{aligned}$$

- **System of Nonlinear Equations !!!**

Nonlinear Least Squares

Example:

- The model

$$y_i = \theta_{00} + \theta_{01} e^{\theta_{02} x_i} + u_i$$

- NLLS

$$\begin{aligned}
\hat{\theta} &= \arg\min_{\theta \in \Theta} Q_n(\theta) \\
&= \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \theta_0 - \theta_1 e^{\theta_2 x_i} \right)^2 \\
&= \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} u_i^2(\theta)
\end{aligned}$$

Nonlinear Least Squares

- NLLS
$$\hat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \theta_0 - \theta_1 e^{\theta_2 x_i}\right)^2$$

- First order conditions:

$$\begin{aligned}
\frac{\partial Q_n(\theta)}{\partial \theta_0} &= -2\frac{1}{n}\sum_{i=1}^{n} u_i(\theta) = 0 \\
\frac{\partial Q_n(\theta)}{\partial \theta_1} &= -2\frac{1}{n}\sum_{i=1}^{n} u_i(\theta)\, e^{\theta_2 x_i} = 0 \\
\frac{\partial Q_n(\theta)}{\partial \theta_2} &= -2\frac{1}{n}\sum_{i=1}^{n} u_i(\theta)\, \theta_1 x_i e^{\theta_2 x_i} = 0
\end{aligned}$$

- **System of Nonlinear Equations !!!**

Nonlinear Least Squares

- The Nonlinear Least Squares (NLLS) estimator is defined as

$$\hat{\theta} = \arg\min_{\theta \in \Theta} Q_n(\theta) = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i, \theta))^2$$

- First order conditions:

$$\dot{Q}_n(\theta) = \frac{\partial Q_n(\theta)}{\partial \theta} = -2\frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i, \theta))\, \dot{g}(x_i, \theta) = 0$$

where $\dot{g}(x_i, \theta) = \partial g(x_i, \theta) / \partial \theta$.

- **These systems of equations do not have an explicit (closed-form) solution... What to do?**

Nonlinear Least Squares

Numerical Methods:

- Grid Search: (Trial and error principle). Construct a grid

  $$[\theta_{\min}, \dots, \theta_{\max}]$$

  and search for the value that solves the least squares problem

- Iterative algorithms: Newton-Raphson, Gauss-Newton, etc.......

Nonlinear Least Squares

Iterative algorithms: Newton's Methods

- Based on Taylor's Theorem (Taylor approximations)
- (From Bartle's Introduction to Real Analysis): *"A very useful technique in the analysis of real functions is the approximation of functions by polynomials.[...] a fundamental theorem in this area which goes back to Brook Taylor (1685-1731), although the remainder term was not provided until much later by Joseph-Louis Lagrange (1736-1813). Taylor's Theorem is a powerful result that has many applications. [...] some of its applications to numerical estimation, inequalities, extreme values of a function, and convex functions."*

Iterative algorithms: Newton's Methods

Taylor's theorem: (As in Bartle p. 184): Let $m \in \mathbb{N}$, let $I := [a, b]$, and let $f : I \to \mathbb{R}$ be such that f and its derivatives $\dot{f}, \ddot{f}, ..., f^{(m)}$ are continuous on I and that $f^{(m+1)}$ exists on (a, b). If $x_0 \in I$, then for any x in I there exists a point c between x and x_0 such that

$$\begin{aligned} f(x) &= f(x_0) + \frac{\dot{f}(x_0)}{1!}(x - x_0) + \frac{\ddot{f}(x_0)}{2!}(x - x_0)^2 \\ &\quad + ... + \frac{f^{(m)}(x_0)}{m!}(x - x_0)^n + \frac{f^{(m+1)}(c)}{(m+1)!}(x - x_0)^{m+1}. \end{aligned}$$

Taylor's Theorem

Taylor theorem conclusion:

$$f(x) = P_m(x) + R_m(x)$$

with polynomial approximation

$$\begin{aligned} P_m(x) &= f(x_0) + \frac{\dot{f}(x_0)}{1!}(x - x_0) + \frac{\ddot{f}(x_0)}{2!}(x - x_0)^2 \\ &\quad + \ldots + \frac{f^{(m)}(x_0)}{m!}(x - x_0)^n \end{aligned}$$

and remainder

$$R_m(x) = \frac{f^{(m+1)}(c)}{(m+1)!}(x - x_0)^{m+1}$$

Taylor theorem: Example
(Bartle p. 185)

Taylor Theorem with $m = 2$ to approximate $\sqrt[3]{1+x}$ $(x > -1)$

- Take the function $f(x) = (1+x)^{1/3}$, the point $x_0 = 0$, and $m = 2$
- First derivative

$$\dot{f}(x) = \frac{1}{3}(1+x)^{-2/3} \quad \text{so that} \quad \dot{f}(0) = \frac{1}{3}$$

- Secon derivative

$$\ddot{f}(x) = \frac{1}{3}\left(-\frac{2}{3}\right)(1+x)^{-5/3} \quad \text{so that} \quad \ddot{f}(0) = -\frac{2}{9}$$

Taylor theorem: Example
(Bartle p. 185)

Taylor Theorem with $m = 2$ to approximate $\sqrt[3]{1+x}$ $(x > -1)$

- Hence,

$$f(x) = P_m(x) + R_m(x) = 1 + \frac{1}{3}x - \frac{2}{9}x^2 + R_2$$

with

$$R_2 = \frac{1}{3!}\dddot{f}(c)\,x^3 = \frac{5}{81}(1+c)^{-8/3}x^3$$

for some point c between 0 and x

Taylor theorem: Example
(Bartle p. 185)

- Remainder

$$R_2 = \frac{1}{3!}\dddot{f}(c)\,x^3 = \frac{5}{81}(1+c)^{-8/3}x^3$$

  for some point c between 0 and x
- Let $x = 0.3$ for instance, so $P_2(0.3) = 1.09$ for $\sqrt[3]{1.3}$. In this case, $c > 0$, then $(1+c)^{-8/3} < 1$ and the error is at most

$$R_2(0.3) \leq \frac{5}{81}\left(\frac{3}{10}\right)^3 = \frac{1}{600} < 0.17 \times 10^{-2}.$$

- So we get: $\left|\sqrt[3]{1.3} - 1.09\right| < 0.5 \times 10^{-2}$: Two decimal place accuracy is assured

Nonlinear Least Squares

Recall: Our objective is to solve:

- The Nonlinear Least Squares (NLLS) estimator is defined as

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} Q_n(\theta) = \underset{\theta \in \Theta}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - g(x_i, \theta)\right)^2$$

- First order conditions:

$$\dot{Q}_n(\theta) = \frac{\partial Q_n(\theta)}{\partial \theta} = -2\frac{1}{n} \sum_{i=1}^{n} \left(y_i - g(x_i, \theta)\right) \dot{g}(x_i, \theta) = 0$$

  where $\dot{g}(x_i, \theta) = \partial g(x_i, \theta) / \partial \theta$.

- **These systems of equations do not have an explicit (closed-form) solution... What to do?**

Taylor theorem application: Newton's Method
(Bartle p. 189)

"It is often desirable to estimate a solution of an equation with a high degree of accuracy. [...] A method that often results in much more rapid convergence is based on the geometric idea of successively approximating a curve by tangent lines. The method is named after its discoverer, Isaac Newton."

Iterative algorithms: Newton's Method

Taylor theorem application: Newton's Method
(Bartle p. 189)

- Let f be a differentiable function that has a zero at r and let x_1 be an initial estimate of r
- The line tangent to the graph at $(x_1, f(x_1))$ has the equation

$$y = f(x_1) + \dot{f}(x_1)(x - x_1)$$

and crosses the x-axis at the point

$$x_2 = x_1 - \frac{f(x_1)}{\dot{f}(x_1)}$$

Iterative algorithms: Newton's Method

Taylor theorem application: Newton's Method
(Bartle p. 189)

- If we replace x_1 by the second estimate x_2, a new point x_3 is obtained (so on and so for)... At the p-th iteration we get

$$x_{p+1} = x_p - \frac{f\left(x_p\right)}{\dot{f}\left(x_p\right)}$$

- Newton's Method is based on the fact that x_p will converge rapidly to a root of the equation $f\left(x\right) = 0$. Key tool to show this is Taylor's theorem

Nonlinear Least Squares

- The Nonlinear Least Squares (NLLS) estimator is defined as

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} Q_n(\theta) = \underset{\theta \in \Theta}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i, \theta))^2$$

- First order conditions:

$$\dot{Q}_n(\theta) = \frac{\partial Q_n(\theta)}{\partial \theta} = -2 \frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i, \theta)) \dot{g}(x_i, \theta) = 0$$

  where $\dot{g}(x_i, \theta) = \partial g(x_i, \theta) / \partial \theta$.

- **These systems of equations do not have an explicit (closed-form) solution... What to do?**

Nonlinear Least Squares

Iterative algorithms: Newton Methods

- Based on Taylor's Theorem (Taylor approximations)
- (From Bartle's Introduction to Real Analysis): *"A very useful technique in the analysis of real functions is the approximation of functions by polynomials.[...] a fundamental theorem in this area which goes back to Brook Taylor (1685-173 1), although the remainder term was not provided until much later by Joseph-Louis Lagrange (1736-1813). Taylor's Theorem is a powerful result that has many applications. [...] some of its applications to numerical estimation, inequalities, extreme values of a function, and convex functions."*

Nonlinear Least Squares

Gauss-Newton Method
(Amemiya, 1985 p. 139)

- Gauss-Newton Method specially designed to calculate NLLS
- Consider the simple model:

$$y_i = g\left(x_i, \theta\right) + u_i$$

  where both x_i and θ are scalars.
- First order Taylor approximation around and initial estimate $\hat{\theta}_{(1)}$:

$$g\left(x_i, \theta\right) \approx g\left(x_i, \hat{\theta}_{(1)}\right) + \dot{g}\left(x_i, \hat{\theta}_{(1)}\right)\left(\theta - \hat{\theta}_{(1)}\right)$$

- Linearized model:

$$y_i \approx g\left(x_i, \hat{\theta}_{(1)}\right) + \dot{g}\left(x_i, \hat{\theta}_{(1)}\right)\left(\theta - \hat{\theta}_{(1)}\right) + u_i$$

Nonlinear Least Squares

Gauss-Newton Method
(Amemiya, 1985 p. 139)

Linearized model:

$$y_i \approx g\left(x_i, \hat{\theta}_{(1)}\right) + \dot{g}\left(x_i, \hat{\theta}_{(1)}\right)\left(\theta - \hat{\theta}_{(1)}\right) + u_i$$

This is a linear model with unknown θ that we can estimate by least squares again

$$\begin{aligned}
\hat{\theta}_{(2)} &= \arg\min_{\theta\in\Theta} \tilde{Q}_n(\theta) \\
&= \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - g\left(x_i, \hat{\theta}_{(1)}\right) + \dot{g}\left(x_i, \hat{\theta}_{(1)}\right)\left(\theta - \hat{\theta}_{(1)}\right)\right)^2
\end{aligned}$$

Nonlinear Least Squares

Gauss-Newton Method
(Amemiya, 1985 p. 139)

Linear model with unknown θ that we can estimate by least squares again

$$\begin{aligned}\hat{\theta}_{(2)} &= \arg\min_{\theta\in\Theta} \tilde{Q}_n(\theta) \\ &= \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - g\left(x_i, \hat{\theta}_{(1)}\right) - \dot{g}\left(x_i, \hat{\theta}_{(1)}\right)\left(\theta - \hat{\theta}_{(1)}\right)\right)^2\end{aligned}$$

FOC

$$\begin{aligned}&\frac{\partial \tilde{Q}_n(\theta)}{\partial\theta} \\ &= -2\frac{1}{n}\sum_{i=1}^{n}\left(y_i - g\left(x_i, \hat{\theta}_{(1)}\right) - \dot{g}\left(x_i, \hat{\theta}_{(1)}\right)\left(\theta - \hat{\theta}_{(1)}\right)\right)\dot{g}\left(x_i, \hat{\theta}_{(1)}\right) \\ &= 0\end{aligned}$$

Gauss-Newton Method
(Amemiya, 1985 p. 139)

The second-round estimator is

$$\hat{\theta}_{(2)} = \hat{\theta}_{(1)} + \left[\sum_{i=1}^{n} \dot{g}\left(x_i, \hat{\theta}_{(1)}\right)^2\right]^{-1} \left[\sum_{i=1}^{n} \left(y_i - g\left(x_i, \hat{\theta}_{(1)}\right)\right) \dot{g}\left(x_i, \hat{\theta}_{(1)}\right)\right]$$

Nonlinear Least Squares

Gauss-Newton Method
(Amemiya, 1985 p. 139)

At the p-th iteration we get

$$\hat{\theta}_{(p+1)} = \hat{\theta}_{(p)} + \left[\sum_{i=1}^{n} \dot{g}\left(x_i, \hat{\theta}_{(p)}\right)^2\right]^{-1} \left[\sum_{i=1}^{n} \left(y_i - g\left(x_i, \hat{\theta}_{(p)}\right)\right) \dot{g}\left(x_i, \hat{\theta}_{(p)}\right)\right]$$

- The iteration continues until convergence is achieved
- So: starting with an initial value $\hat{\theta}_{(0)}$, the process can be iterated until $\left|\hat{\theta}_{(p+1)} - \hat{\theta}_{(p)}\right|$ is small enough

Nonlinear Models: Estimation

Extremum Estimators
(Cameron and Trivedi p. 124)

Extremum estimators are a very general class of estimators that minimize or maximize an objective function. In particular, an estimator $\hat{\theta}$ is called an extremum estimator if there is an objective function $Q_n(\theta)$ such that

$$\hat{\theta} = \arg\max_{\theta \in \Theta} Q_n(\theta)$$

where Θ is the parameter space i.e. the set of possible parameter values (ex. $\Theta \subset \mathbb{R}^K$)

Maximum Likelihood

Extremum Estimators Examples:
Maximum Likelihood

$$y_i = g\left(x_i, \theta\right) + \varepsilon_i \quad \text{with} \quad \varepsilon_i \sim iidN\left(0, \sigma_\varepsilon^2\right)$$

The Maximum Likelihood (ML) estimator is defined as

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\max} Q_n\left(\theta\right)$$

where

$$Q_n\left(\theta\right) = \frac{1}{n} \sum_{i=1}^{n} \ln f\left(y_i | x_i, \theta\right)$$

and where $f\left(y_i | x_i, \theta\right)$ is the conditional likelihood for observation i.

Maximum Likelihood

What is a likelihood function?
(Stock and Watson p. 438)

Definition: The likelihood function is the joint probability distribution of the data, treated as a function of the unknown coefficients

Maximum Likelihood

What is a likelihood function?
(Unconditional)

- Let $(y_1, ..., y_n)$ be *i.i.d.* and let $f(y_i, \theta)$ be probability density function (or probability mass) for each i. Example: $y_i \sim B(1, p)$ so that (Bernoulli) $f(y_i, p) = p^{y_i}(1-p)^{1-y_i}$
- Since $y_i \sim i.i.d.$ the joint density function is

$$f(y_1, ..., y_n, \theta) = f(y_1, \theta) \times f(y_2, \theta) \times ... \times f(y_n, \theta) = \prod_{i=1}^{n} f(y_i, \theta)$$

- The likelihood function is

$$L_n(\theta, y_1, ..., y_n) = \prod_{i=1}^{n} f(\theta, y_i)$$

Maximum Likelihood

- The likelihood function is

$$L_n\left(\theta, y_1, ..., y_n\right) = \prod_{i=1}^{n} f\left(\theta, y_i\right)$$

- The log-likelihood function is

$$\mathcal{L}_n\left(\theta\right) = \ln\left(L_n\left(\theta, y_1, ..., y_n\right)\right) = \ln\left(\prod_{i=1}^{n} f\left(\theta, y_i\right)\right) = \sum_{i=1}^{n} \ln f\left(\theta, y_i\right)$$

Maximum Likelihood

(Stock and Watson p. 438)

Definition: The maximum likelihood estimator (MLE) of the unknown coefficients are the value of the coefficients that maximize the likelihood function. In effect, the MLE chooses the values of the parameters to maximize the probability of drawing the data that are actually observed. In this sense, the MLEs are the parameter values "most likely" to have produced the data.

Maximum Likelihood Example

Example: Poisson distribution (Ross, p. 136)

- A random variable Y that takes on one of the values 0,1,2,... is said to be Poisson random variable with parameter λ if, for some $\lambda > 0$,

$$P\left(Y = y\right) = f\left(y, \lambda\right) = \frac{e^{-\lambda}\lambda^{y}}{y!} \quad y = 0, 1, 2, \ldots$$

- Introduced by Simeón Denis Poisson in a book published in 1837 titled (translated from French) "Investigations into the Probability of Verdicts in Criminal and Civil Matters."
- A book that applies probability theory to lawsuits, criminal trials, etc...

Maximum Likelihood

Example: Poisson distribution (Ross, p. 136)

- Used in situations in which a certain kind of occurrence happens at random over a period of time. For instance:
  - Number of misprints on a page (or group of pages) of a book (or my lecture notes...)
  - Number of people in a community who survive to age 100
  - Number of customers entering a store on a given day
  - etc... etc...

Maximum Likelihood Example

Example: Poisson distribution (Ross, p. 136)
For a Poisson random variable Y

- Probability mass:

$$f(y, \lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \quad y = 0, 1, 2, \ldots$$

- $E[Y] = V[Y] = \lambda$
- We would like to estimate λ by Maximum Likelihood

Maximum Likelihood Example

Example: Poisson distribution

- Density function

$$f\left(y_i, \lambda\right) = \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \quad y_i = 0, 1, 2, ...$$

  tell us, for a fixed λ, the the probability of occurrence of $Y = y$.

- Likelihood function for individual i:

$$L\left(\lambda, y_i\right) = f\left(\lambda, y_i\right) = \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \quad y_i = 0, 1, 2, ...$$

  is the density as a function of the parameters, λ in this case

Maximum Likelihood Example

Example: Poisson distribution for a random sample (iid) y_i, $i = 1, ..., n$

- Joint Density function: By iid

$$f(y_1, ..., y_n, \lambda) = \prod_{i=1}^{n} f(y_i, \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

- Likelihood function:

$$L_n(\lambda, y_1, ..., y_n) = \prod_{i=1}^{n} L(\lambda, y_i) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

Maximum Likelihood

Maximum Likelihood Example

Example: Poisson distribution for a random sample (iid) y_i, $i = 1, ..., n$

- Likelihood function:

$$L_n(\lambda, y_1, ..., y_n) = \prod_{i=1}^{n} L(\lambda, y_i) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

- Log-Likelihood function:

$$\mathcal{L}_n(\lambda) = \ln(L_n(\lambda, y_1, ..., y_n)) = \ln\left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\right)$$

Maximum Likelihood Example

Example: Poisson distribution for a random sample (iid) y_i, $i = 1, ..., n$

- Log-Likelihood function:

$$\begin{aligned} \mathcal{L}_n(\lambda) &= \ln\left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\right) \\ &= \sum_{i=1}^{n} \ln\left(\frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\right) \\ &= \sum_{i=1}^{n} \left(-\lambda \ln(e) + y_i \ln(\lambda) - \ln(y_i!)\right) \\ &= -N\lambda + \ln(\lambda)\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln(y_i!) \end{aligned}$$

Extremum Estimators Examples: Maximum Likelihood

The Maximum Likelihood (ML) estimator is defined as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} Q_n(\theta)$$

where

$$Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln f(\theta, y_i)$$

and where $f(\theta, y_i)$ is the likelihood for observation i.

Maximum Likelihood Example

Example: Poisson distribution for a random sample (iid) y_i, $i = 1, ..., n$. The Maximum Likelihood (ML) estimator is defined as

$$\hat{\lambda} = \arg\max_{\lambda} Q_n(\lambda)$$

where

$$\begin{aligned} Q_n(\lambda) &= \frac{1}{n}\sum_{i=1}^{n} \ln f(\lambda, y_i) \\ &= -\lambda + \ln(\lambda)\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\ln(y_i!) \end{aligned}$$

Maximum Likelihood Example

Example: Poisson distribution for a random sample (iid) y_i, $i = 1, ..., n$. The Maximum Likelihood (ML) estimator is defined as

$$\begin{aligned}\hat{\lambda} &= \arg\max_{\lambda} Q_n(\lambda) \\ &= \arg\max_{\lambda} \left(-\lambda + \ln(\lambda) \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{1}{n} \sum_{i=1}^{n} \ln(y_i!) \right)\end{aligned}$$

Maximum Likelihood Example

Example: Poisson distribution for a random sample (iid) y_i, $i = 1, ..., n$

- MLE

$$\hat{\lambda} = \arg\max_{\lambda} Q_n(\lambda)$$
$$= \arg\max_{\lambda} \left(-\lambda + \ln(\lambda) \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{1}{n} \sum_{i=1}^{n} \ln(y_i!) \right)$$

- FOC:

$$\frac{\partial Q_n(\lambda)}{\partial \lambda} = -1 + \frac{1}{\lambda} \frac{1}{n} \sum_{i=1}^{n} y_i = 0$$

- Therefore, the MLE is:

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Maximum Likelihood

Some notation etc...

Cameron and Trivedi, p. 139

- Joint probability mass or density: $f(y, x|\theta)$
- Likelihood function: $f(y, x|\theta)$ as a function of θ given the data (y, x) is denoted by: $L_n(\theta|y, x)$
- Maximizing $L_n(\theta|y, x)$ is equivalent to maximizing the log-likelihood function $\mathcal{L}_n(\theta|y, x)$

Maximum Likelihood (Conditional)

Some notation etc...

Cameron and Trivedi, p. 139

- Likelihood function: $L_n(\theta|y,x) = f(y,x|\theta) = f(y|x,\theta)f(x|\theta)$ requires specification of both: conditional and marginal
- Instead, estimation is usually based on the conditional likelihod function $L_n(\theta) = f(y|x,\theta)$
- Goal of regression is to model the behavior of y given x

Maximum Likelihood (Conditional)

Example: **The Poisson regression model** (Cameron and Trivedi p. 117 or Greene p. 843)

- Each y_i is drawn from a Poisson population with parameter λ_i, which is related to the regressors x_i
- Primary equation of the model

$$P\left(Y = y_i | \lambda_i\left(x_i\right)\right) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \quad y = 0, 1, 2, ...$$

- The most common formulation for λ_i is the loglinear model:

$$\ln \lambda_i = x_i'\beta$$

- In this case:

$$E\left[y_i | x_i\right] = V\left[y_i | x_i\right] = \lambda_i = e^{x_i'\beta}$$

- So

$$\frac{\partial E\left[y_i | x_i\right]}{\partial x_i} = \lambda_i \beta$$

Maximum Likelihood
(Conditional)

Example: **The Poisson regression model** (Cameron and Trivedi p. 117 or Greene p. 843)

$$y_i = E\left[y_i | x_i\right] + u_i$$

in this case is

$$y_i = e^{x_i'\beta} + u_i$$

and our objective is to estimate β by ML

Maximum Likelihood Example

Example: **The Poisson regression model**: Now y_i and $\lambda_i = e^{x_i'\beta}$ vary for each individual:

- Density: By iid

$$f\left(y_1, ..., y_n | x_i, \beta\right) = \prod_{i=1}^{n} f\left(y_i | x_i, \beta\right) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

- Conditional Likelihood function:

$$L_n\left(y_1, ..., y_n | x_1, ... x_n, \beta\right) = \prod_{i=1}^{n} L\left(y_i | x_i, \beta\right) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

Maximum Likelihood Example

Example: **The Poisson regression model**:

- Conditional Likelihood function:

$$L_n\left(y_1,\ldots,y_n|x_1,\ldots x_n,\beta\right)=\prod_{i=1}^{n}L\left(y_i|x_i,\beta\right)=\prod_{i=1}^{n}\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

- Conditional Log-Likelihood function:

$$\mathcal{L}_n\left(\beta\right)=\ln\left(L_n\left(y_1,\ldots,y_n|x_1,\ldots x_n,\beta\right)\right)=\ln\left(\prod_{i=1}^{n}\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right)$$

Maximum Likelihood Example

Example: **The Poisson regression model**: Log-Likelihood function: $\ln(\lambda_i) = x_i'\beta$ or equivalently $\lambda_i = e^{x_i'\beta}$

$$
\begin{aligned}
\mathcal{L}_n(\beta) &= \ln\left(\prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right) \\
&= \sum_{i=1}^{n} \ln\left(\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right) \\
&= \sum_{i=1}^{n} \left(-\lambda_i \ln(e) + y_i \ln(\lambda_i) - \ln(y_i!)\right) \\
&= \sum_{i=1}^{n} \left(-e^{x_i'\beta} + y_i x_i'\beta - \ln(y_i!)\right)
\end{aligned}
$$

Extremum Estimators Examples: Maximum Likelihood

The Maximum Likelihood (ML) estimator is defined as

$$\hat{\beta} = \arg\max_{\beta} Q_n\left(\beta\right)$$

where

$$Q_n\left(\beta\right) = \frac{1}{n}\sum_{i=1}^{n} \ln f\left(y_i | x_i, \beta\right)$$

and where $f\left(y_i | x_i, \beta\right)$ is the conditional likelihood for observation i.

Maximum Likelihood Example

Example: **The Poisson regression model**: The Maximum Likelihood (ML) estimator is defined as

$$\hat{\beta} = \arg\max_{\beta} Q_n(\beta)$$

where

$$\begin{aligned} Q_n(\beta) &= \frac{1}{n}\sum_{i=1}^{n} \ln f(y_i|x_i,\beta) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(-e^{x_i'\beta} + y_i x_i'\beta - \ln(y_i!)\right) \end{aligned}$$

Maximum Likelihood Example

Example: **The Poisson regression model**:

- MLE

$$\hat{\beta} = \arg\max_{\beta} Q_n(\beta) = \arg\max_{\beta} \left(\frac{1}{n} \sum_{i=1}^{n} \left(-e^{x_i'\beta} + y_i x_i'\beta - \ln(y_i!) \right) \right)$$

- FOC:

$$\frac{\partial Q_n(\beta)}{\partial \beta} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - e^{x_i'\beta} \right) x_i = 0$$

- **System of Nonlinear equations!!! Numerical Methods**

Nonlinear Models: Estimation

Extremum Estimators: Asymptotics
Consistency and Asymptotic Normality

- The asymptotic analysis of extremum estimators is a technical matter
- Amemiya (1985) is a good reference for this material
- General treatment for Extremum Estimators
- Application to particular cases: NLLS and ML

Nonlinear Models: Estimation

Recall: Extremum Estimators

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} Q_n(\theta)$$

Nonlinear Models: Estimation

Extremum Estimators: Consistency

Heuristics (for a formal argument: Amemiya, p. 106)

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} Q_n(\theta)$$

Let the following assumptions hold:
(A) The parameter space Θ is compact (θ_0 is in Θ)
(B) $Q_n(\theta)$ is continuous in $\theta \in \Theta$
(C) For a nonstochastic function $Q(\theta)$, uniformly in θ,

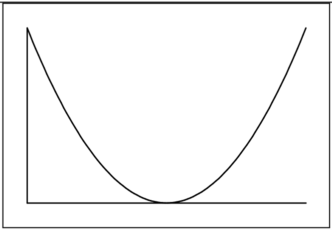
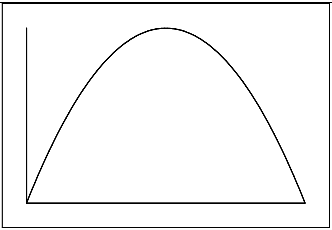
$$Q_n(\theta) \xrightarrow{p} Q(\theta)$$

and $Q(\theta)$ attains a unique global maximum at θ_0.
Under (A), (B), and (C):

$$\hat{\theta}_n \xrightarrow{p} \theta_0.$$

Nonlinear Models: Estimation

Extremum Estimators: Consistency

$Q_n(\theta)$	
NLLS	ML
	

Nonlinear Models: Estimation

Extremum Estimators: Asymptotic Normality
(Heuristics)

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} Q_n(\theta)$$

The first order conditions are

$$\dot{Q}_n(\theta) = \frac{\partial Q_n(\theta)}{\partial \theta} = 0$$

Then the estimator $\hat{\theta}_n$ is the value that solves this system of equations. Therefore,

$$\dot{Q}_n\left(\hat{\theta}_n\right) = 0$$

Extremum Estimators: Asymptotic Normality
(Heuristics)

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} Q_n(\theta)$$

From the FOC:

$$\dot{Q}_n\left(\hat{\theta}_n\right) = 0.$$

Let us apply a first order Taylor expansion to $\dot{Q}_n\left(\hat{\theta}_n\right)$ around the true parameter θ_0:

$$\dot{Q}_n\left(\hat{\theta}_n\right) \approx \dot{Q}_n(\theta_0) + \ddot{Q}_n(\theta_0)\left(\hat{\theta}_n - \theta_0\right)$$

Extremum Estimators: Asymptotic Normality
(Heuristics)

Let us apply a first order Taylor expansion to $\dot{Q}_n\left(\hat{\theta}_n\right)$ around the true parameter θ_0:

$$\dot{Q}_n\left(\hat{\theta}_n\right) \approx \dot{Q}_n\left(\theta_0\right) + \ddot{Q}_n\left(\theta_0\right)\left(\hat{\theta}_n - \theta_0\right)$$

Hence,

$$\dot{Q}_n\left(\hat{\theta}_n\right) = 0 \approx \dot{Q}_n\left(\theta_0\right) + \ddot{Q}_n\left(\theta_0\right)\left(\hat{\theta}_n - \theta_0\right)$$

and

$$\left(\hat{\theta}_n - \theta_0\right) \approx \ddot{Q}_n\left(\theta_0\right)^{-1}\dot{Q}_n\left(\theta_0\right)$$

Nonlinear Models: Estimation

Extremum Estimators: Asymptotic Normality
(Heuristics)

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \approx \ddot{Q}_n\left(\theta_0\right)^{-1}\sqrt{n}\dot{Q}_n\left(\theta_0\right)$$

Under appropriate regularity conditions (see Amemiya p. 111)

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, A\left(\theta_0\right)^{-1} B\left(\theta_0\right) A\left(\theta_0\right)^{-1}\right)$$

where $A\left(\theta_0\right) = p\lim \ddot{Q}_n\left(\theta_0\right)$ and $B\left(\theta_0\right)$ is the asymptotic variance of $\sqrt{n}\dot{Q}_n\left(\theta_0\right)$.
Consistent estimates of $A\left(\theta_0\right)$ and $B\left(\theta_0\right)$ can be obtained.
Inferences can be performed.

Maximum Likelihood
(Conditional)

Example: **The Poisson regression model** (Cameron and Trivedi p. 117 or Greene p. 843)

$$y_i = E\left[y_i | x_i\right] + u_i$$

in this case is

$$y_i = e^{x_i'\beta} + u_i$$

and our objective is to estimate β by ML

Motivating Regressions: Conditional Expectation

- **Main focus of this course**: Conditional Expectation: $E[y|x]$

$$y = E[y|x] + u$$

- First part of the course $E[y|x]$ linear: Example: Linear wage equation

$$E[wage|x] = \beta_0 + \beta_1 educ + \beta_2 \text{exper} + \beta_3 female$$

- Second part of the course $E[y|x]$ is nonlinear. Example: Poisson Regression for number of arrests

$$E[narr86|x] = \exp(\beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime)$$

Motivating Regressions: Conditional Expectation

- **Main focus of this course**: Conditional Expectation: $E\left[y|x\right]$

$$y = E\left[y|x\right] + u$$

- Second part of the course $E\left[y|x\right]$ is nonlinear. Example: Poisson Regression for number of arrests

$$E\left[narr86|x\right] = \exp\left(\beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime\right)$$

- Binary choice models. Example: Labor force participation

$$E\left[inlf_i|x\right] = P\left(inlf_i = 1|x\right) = g\left(nwifeinc_i, educ_i, \exp er_i kidslt6_i, \theta\right)$$

Cross-sectional Data: California Test Score

- **California Test Score**: Data: Stock and Watson (p. 51)
  - $tscore_i$: average of the math and science test scores for all fifth grades in 1999 in district i
  - str_i : average student-teacher ratio in district i
  - $\exp en_i$: average expenditure per pupil
  - eng_i: percentage of students still learning English

Cross-sectional Data: Wage Equations

- **Wage Equations**: Data: Wooldrige (p. 218)
  - w_i : hourly wage
  - $educ_i$: years of formal education
  - $\exp er_i$: years of workforce experience
  - $female_i$: 1 if person i is female, otherwise
  - $married_i$: 1 if person i is married, otherwise

Cross-sectional Data: Labor Force Participation

- **Labor Force Participation**: Data: Wooldrige (p. 239)
  - $inlf_i$: 1 if woman i reports working for a wage outside the home, 0 otherwise
  - $nwifeinc_i$: husband's earnings
  - $educ_i$: years of education
  - $\exp er_i$: past years of labor market experience
  - $kidslt6_i$: number of children less than six years old
  - $kidsge6_i$: number of kids between 6 and 18 years of age

Cross-sectional Data: Crime Data

- **Crime**: Data: Wooldrige (p. 4, 78,172, 295, 583)
  - $crime_i$: some measure of the frequency of criminal activity
  - Ex: $narr86_i$: number of times a man was arrested
  - $pcnv_i$: proportion of prior arrests leading to conviction
  - $tottime_i$: total time the man has spent in prison prior to 1986 since reaching the age of 18
  - $ptime86_i$: months spent in prison in 1986
  - $qemp86_i$: number of quarters in 1986 during which the man was legally employed