Day 4B and 5A
Nonlinear methods

© A. Colin Cameron
Univ. of Calif.- Davis

Anvanced Econometrics
Bavarian Graduate Program in Economics

Based on A. Colin Cameron and Pravin K. Trivedi (2009, 2010), Microeconometrics using Stata (MUS), Stata Press. and A. Colin Cameron and Pravin K. Trivedi (2005), Microeconometrics: Methods and Applications (MMA), C.U.P.

Jully 22-26, 2013

1. Introduction

- Consider nonlinear estimator, one for which there is no explicit solution for $\widehat{\boldsymbol{\theta}}$.
  - example we work with is Poisson MLE
- Topics include
  - interpreting coefficients
  - general estimation theory
  - key nonlinear estimators: MLE and NLS
  - calculating marginal effects
  - statistical inference
  - computational methods

Overview

2. Poisson Regression Model

- Poisson regression is a leading example of a nonlinear model.
- Consider count data with $y = 0, 1, 2, ...$
  - OLS has problem that $\mathrm{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta} < 0$ is possible
  - And OLS is inefficient (based on homoskedasticity, normality).
  - So what do we do?
- Starting point from statistics is Poisson:
  - Poisson density (or more precisely probability mass function):

$$\Pr[Y = y|\mu] = e^{-\mu}\mu^y / y!$$

where $\mu = \mathrm{E}[y] > 0$, $\mathrm{V}[y] = \mu$, $y! = y \times (y-1) \times \cdots \times 1$.

- For regression the mean $\mu > 0$ varies with regressors $\mathbf{x}$
  - Conditional mean for Poisson regression model:

$$\mu_i = \mathrm{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta}), \quad \mu_i > 0$$

where $\mathbf{x}_i$ and $\boldsymbol{\beta}$ are $K \times 1$ vectors.

Maximum likelihood

- Likelihood principle
  - Likelihood principle says choose that value of $\boldsymbol{\beta}$ which makes the probability of observing our data $(\mathbf{y}, \mathbf{X})$ as high as possible.
  - Joint density $f(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})$ gives probability of observing $\mathbf{y}$ given $\boldsymbol{\beta}$ (and $\mathbf{X}$).
  - Likelihood $\mathrm{L}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = f(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})$ reinterprets as probability of $\boldsymbol{\beta}$ given $\mathbf{y}$ (and $\mathbf{X}$).
  - Maximize the likelihood, or equivalently the log-likelihood.
- In general
  - $f(y_i|\mathbf{x}_i, \boldsymbol{\beta})$ is conditional (on $\mathbf{x}_i$) density for one observation.
  - $f(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = f(y_1|\mathbf{x}_1, \boldsymbol{\theta}) \times \cdots f(y_N|\mathbf{x}_N, \boldsymbol{\theta}) = \prod_{i=1}^{N} f(y_i|\mathbf{x}_i, \boldsymbol{\beta})$ is joint conditional density for N independent observations.
  - $\mathrm{L}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} f(y_i|\mathbf{x}_i, \boldsymbol{\beta})$ is the likelihood function.
  - $\ln \mathrm{L}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \ln(\prod_{i=1}^{N} f(y_i|\mathbf{x}_i, \boldsymbol{\beta})) = \sum_{i=1}^{N} \ln f(y_i|\mathbf{x}_i, \boldsymbol{\beta})$ is the log-likelihood function.

MLE for Poisson

- Conditional density for one observation is $e^{-\mu}\mu^y/y!$ with $\mu = e^{\mathbf{x}'\boldsymbol{\beta}}$

$$\begin{aligned} f(y_i|\mathbf{x}_i,\boldsymbol{\beta}) &= \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))\exp(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}/y_i! \\ \ln f(y_i|\mathbf{x}_i,\boldsymbol{\beta}) &= -\exp(\mathbf{x}_i'\boldsymbol{\beta}) + y_i\mathbf{x}_i'\boldsymbol{\beta} - \ln(y_i!) \end{aligned}$$

- Log-likelihood for N independent observations:

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^N \ln f(y_i|\mathbf{x}_i,\boldsymbol{\beta}) = \sum_{i=1}^N \{-\exp(\mathbf{x}_i'\boldsymbol{\beta}) + y_i\mathbf{x}_i'\boldsymbol{\beta} - \ln(y_i!)\}.$$

- Differentiate w.r.t. $\boldsymbol{\beta}$:

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^N \{-\exp(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i + y_i\mathbf{x}_i\} = \sum_{i=1}^N (y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i$$

- ML first-order conditions

$$\sum_{i=1}^N (y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{0}.$$

The Complications

- We cannot solve the nonlinear first-order conditions in $\boldsymbol{\beta}$.
  - How do we compute $\widehat{\boldsymbol{\beta}}$?
    - ★ Use an iterative gradient method such as Newton-Raphson.
  - How do we do asymptotic theory for $\widehat{\boldsymbol{\beta}}$?
    - ★ Linearize the nonlinear first-order conditions.
- The conditional mean $\mathrm{E}[y|\mathbf{x}]$ is nonlinear in $\mathbf{x}$ and $\boldsymbol{\beta}$.
  - How do we interpret $\boldsymbol{\beta}$?
    - ★ Use the marginal effect $\partial \mathrm{E}[y|\mathbf{x}]/\partial \mathbf{x} = \exp(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}$.
    - ★ This varies with the evaluation point $\mathbf{x}$.
- Similar complications hold for all nonlinear models.

3. Data Example: Poisson for doctor visits

- Data from MUS chapter 10.
  - Use Poisson regression as dependent variable docvis is a count.

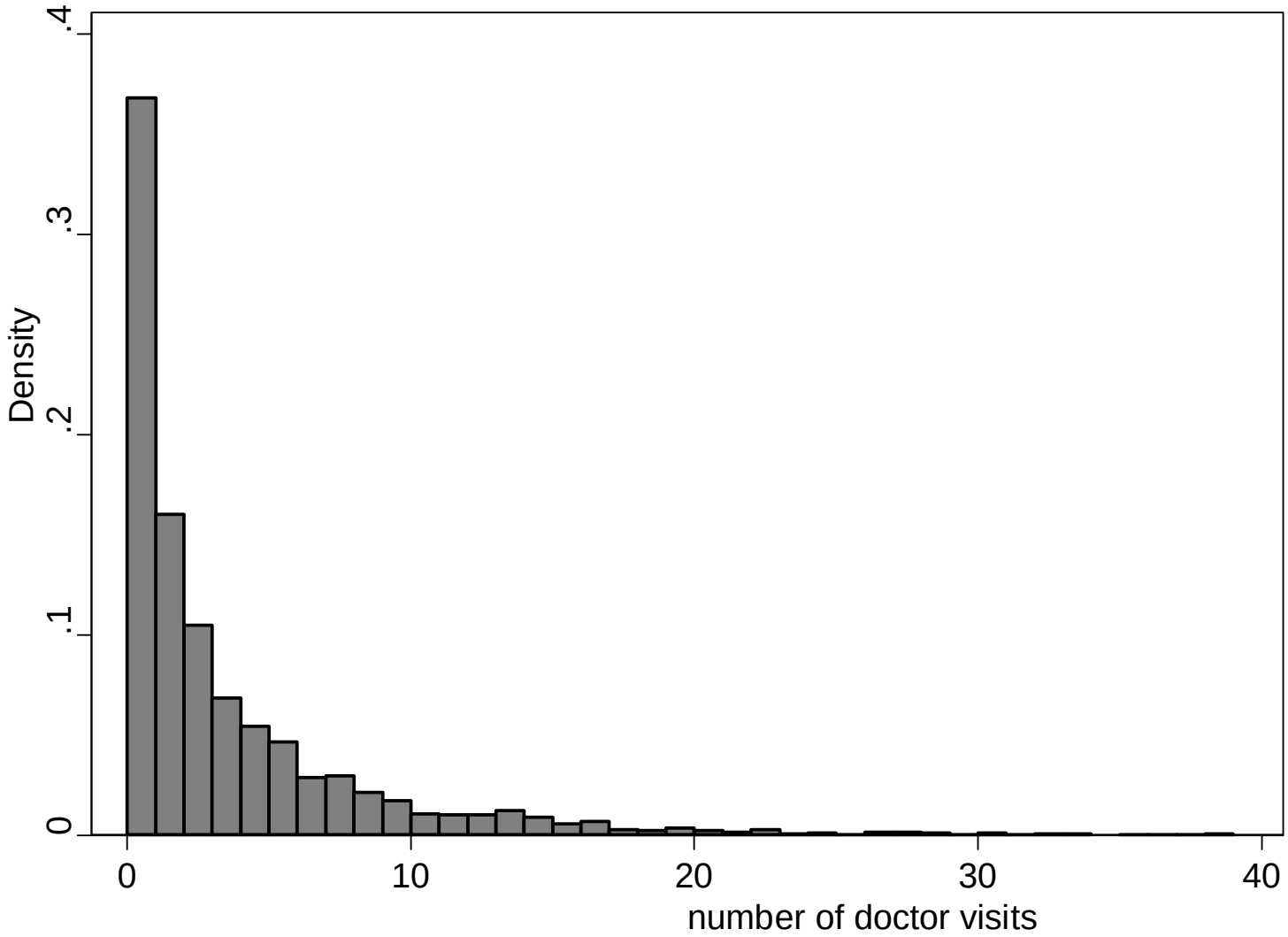
```
. use mus10data.dta, clear

. quietly keep if year02==1

. describe docvis private chronic female income
```

variable name	storage type	display format	value label	variable label
docvis	int	%8.0g		number of doctor visits
private	byte	%8.0g		= 1 if private insurance
chronic	byte	%8.0g		= 1 if a chronic condition
female	byte	%8.0g		= 1 if female
income	float	%9.0g		Income in $ / 1000

```
. summarize docvis private chronic female income
```

Variable	Obs	Mean	Std. Dev.	Min	Max
docvis	4412	3.957389	7.947601	0	134
private	4412	.7853581	.4106202	0	1
chronic	4412	.3263826	.4689423	0	1
female	4412	.4718948	.4992661	0	1
income	4412	34.34018	29.03987	-49.999	280.777

```
. histogram docvis if docvis < 40, width(1)
```



- Poisson MLE: default standard errors (do not use)
  - Converges quickly (2 iterations) and coefficients highly statistically significant.

```
. poisson docvis private chronic female income

Iteration 0:   log likelihood = -18504.413
Iteration 1:   log likelihood = -18503.549
Iteration 2:   log likelihood = -18503.549

Poisson regression                              Number of obs   =       4412
                                                LR chi2(4)      =    8852.71
                                                Prob > chi2     =     0.0000
Log likelihood = -18503.549                     Pseudo R2       =     0.1930

------------------------------------------------------------------------------
      docvis |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     private |   .7986652    .027719    28.81   0.000      .744337    .8529934
     chronic |   1.091865   .0157985    69.11   0.000     1.060901     1.12283
      female |   .4925481   .0160073    30.77   0.000     .4611744    .5239218
      income |    .003557   .0002412    14.75   0.000     .0030844    .0040297
       _cons |  -.2297262   .0287022    -8.00   0.000    -.2859814    -.173471
------------------------------------------------------------------------------
```

- Poisson MLE: robust standard errors (use these)
  - Same coefficient estimates, different s.e.'s, still highly statistically significant.

```
. poisson docvis private chronic female income, vce(robust)

Iteration 0:   log pseudolikelihood = -18504.413
Iteration 1:   log pseudolikelihood = -18503.549
Iteration 2:   log pseudolikelihood = -18503.549

Poisson regression                              Number of obs   =       4412
                                                Wald chi2(4)    =     594.72
                                                Prob > chi2     =     0.0000
Log pseudolikelihood = -18503.549               Pseudo R2       =     0.1930
```

docvis	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
private	.7986652	.1090014	7.33	0.000	.5850263	1.012304
chronic	1.091865	.0559951	19.50	0.000	.9821167	1.201614
female	.4925481	.0585365	8.41	0.000	.3778187	.6072774
income	.003557	.0010825	3.29	0.001	.0014354	.0056787
_cons	-.2297262	.1108732	-2.07	0.038	-.4470338	-.0124186

- For Poisson the default standard errors can be much too small.
  - Reason: Algebra reveals that the default s.e.'s are based on Poisson assumption that $V[y] = \mu$ and instead for these data $V[y] >> \mu$.

```
. * Comparison of standard errors
. quietly poisson docvis private chronic female income

. estimates store DEFAULT

. quietly poisson docvis private chronic female income, vce(robust)

. estimates store ROBUST

. estimates table DEFAULT ROBUST, b(%9.4f) se(%9.3f) stats(N r2 F)
```

Variable	DEFAULT	ROBUST
private	0.7987	0.7987
	0.028	0.109
chronic	1.0919	1.0919
	0.016	0.056
female	0.4925	0.4925
	0.016	0.059
income	0.0036	0.0036
	0.000	0.001
_cons	-0.2297	-0.2297
	0.029	0.111
N	4412.0000	4412.0000
r2		
F		

legend: b/se

4. Marginal effects

- Interpret coefficients using $\text{ME}_j = \frac{\partial \text{E}[y|\mathbf{x}]}{\partial x_j}$, the marginal effect on the conditional mean of a one unit change in the j^{th} regressor.
- For Poisson

$$\text{ME}_j = \frac{\partial \text{E}[y|\mathbf{x}]}{\partial x_j} = \frac{\partial \exp(\mathbf{x}'\boldsymbol{\beta})}{\partial x_j} = \exp(\mathbf{x}'\boldsymbol{\beta})\beta_j.$$

- 1. For Poisson the sign of β_j equals the sign of ME_j
  - Reason: $\exp(\mathbf{x}'\boldsymbol{\beta}) > 0$.
- 2. For Poisson if β_j is twice β_k then ME_j is twice ME_k
  - Reason: $\frac{\text{ME}_j}{\text{ME}k} = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})\beta_j}{\exp(\mathbf{x}'\boldsymbol{\beta})\beta_k} = \frac{\beta_j}{\beta_k}$.
  - This is the case for any single-index model with $\text{E}[y|\mathbf{x}] = g(\mathbf{x}'\boldsymbol{\beta})$.
- 3. ME_j differs with the point of evaluation $\mathbf{x}$
- 4. For Poisson: β_j is a semi-elasticity since $\text{ME}_j/\text{E}[y|\mathbf{x}] = \beta_j$.

AME, MEM and MER: Three estimates of ME

- **1.** AME: Average marginal effect $= \frac{1}{N}\sum_{i=1}^{N} \text{ME}_{ij}$.
  - Stata 11: `margins, dydx(*)` or Stata 10 add-on `margeff`.
- **2.** MEM: Marginal effect at mean = ME at $\mathbf{x} = \mathbf{x}^*$.
  - Stata 11: `margins, dydx(*) atmean` or Stata 10 `mfx`.
- **3.** MER: Marginal effect at a representative value = ME at $\mathbf{x} = \mathbf{x}^*$.
  - Stata 11: `margins, dydx(*) at()` or Stata 10 `mfx`.
- For Poisson these are
  - (1) $\frac{1}{N}\sum_{i=1}^{N} \exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})\widehat{\beta}_j$; (2) $\exp(\overline{\mathbf{x}}'\widehat{\boldsymbol{\beta}})\widehat{\beta}_j$; (3) $\exp(\mathbf{x}^{*\prime}\widehat{\boldsymbol{\beta}})\widehat{\beta}_j$.
- Preceding method uses derivatives.
  - For discrete regressors use finite difference method
    - ★ $\text{ME} = \text{E}[y|d=1,\mathbf{x}] - \text{E}[y|d=0,\mathbf{x}]$

- Average Marginal effect (AME)
  - Large effects: e.g. Doctor visits 50% higher for `female`
  - Here AME is 10%-40% higher than MEM

```
. * Marginal effects AME and MEM
. margins, dydx(*)
Warning: cannot perform check for estimable functions.

Average marginal effects                        Number of obs   =       4412
Model VCE    : Robust

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : private chronic female income
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
private	3.160629	.4352572	7.26	0.000	2.307541	4.013717
chronic	4.320935	.2757872	15.67	0.000	3.780402	4.861468
female	1.949204	.2313726	8.42	0.000	1.495722	2.402686
income	.0140765	.0043457	3.24	0.001	.0055591	.0225939

- Marginal effect at mean (MEM)
  - Here AME was about 10%-40% higher than MEM

```
. margins, dydx(*)

Average marginal effects                        Number of obs   =       4412
Model VCE    : Robust

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : 1.private 1.chronic 1.female income
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.private	2.404721	.2438573	9.86	0.000	1.926769	2.882672
1.chronic	4.599174	.2886176	15.94	0.000	4.033494	5.164854
1.female	1.900212	.2156694	8.81	0.000	1.477508	2.322917
income	.0140765	.0043457	3.24	0.001	.0055591	.0225939

```
Note: dy/dx for factor levels is the discrete change from the base level.
```

- Finite difference method used to compute AME
  - Use prefix `i.` to declare regressors as indicator variables
  - Compared to calculus AME:
    - ★ lower for private, higher for chronic, similar for female.

```
. * Marginal effects using finite difference for binary regressors
. quietly poisson docvis i.private i.chronic i.female income, vce(robust)

. margins, dydx(*)

Average marginal effects                        Number of obs   =       4412
Model VCE    : Robust

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : 1.private 1.chronic 1.female income
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.private	2.404721	.2438573	9.86	0.000	1.926769	2.882672
1.chronic	4.599174	.2886176	15.94	0.000	4.033494	5.164854
1.female	1.900212	.2156694	8.81	0.000	1.477508	2.322917
income	.0140765	.0043457	3.24	0.001	.0055591	.0225939

Note: dy/dx for factor levels is the discrete change from the base level.

- More generally prefixes `i.` and `c.` create factor variables to get marginal effects with interactions.

5. Estimation Theory: m-Estimator

- An m-estimator $\widehat{\boldsymbol{\theta}}$ of the $q \times 1$ parameter vector $\boldsymbol{\theta}$ maximizes

$$Q_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} q(y_i, \mathbf{x}_i, \boldsymbol{\theta}).$$

- $Q_N(\boldsymbol{\theta})$ is the average or sum of N scalar sub-functions $q_i(\cdot)$.
  - MLE: $q(y_i, \mathbf{x}_i, \boldsymbol{\theta}) = \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$.
  - OLS: $q(y_i, \mathbf{x}_i, \boldsymbol{\beta}) = -\frac{1}{2}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$.
  - NLS: $q(y_i, \mathbf{x}_i, \boldsymbol{\beta}) = -\frac{1}{2}(y_i - g(\mathbf{x}_i, \boldsymbol{\beta}))^2$ for specified $g(\mathbf{x}_i, \boldsymbol{\beta})$.
- Consider the local maximum of $Q_N(\boldsymbol{\theta})$ that solves the estimating equation

$$\frac{1}{N}\sum_{i=1}^{N} \left.\frac{\partial q(y_i, \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\widehat{\boldsymbol{\theta}}} = \mathbf{0}.$$

Consistency

- Informally $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ if the sample condition $\frac{1}{N}\sum_{i=1}^{N} \frac{\partial q(y_i,\mathbf{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\widehat{\boldsymbol{\theta}}} = \mathbf{0}$ holds in the population:

$$\mathrm{E}\Big[\frac{\partial q(y_i,\mathbf{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_0}\Big] = \mathbf{0}.$$

  - Here $\boldsymbol{\theta}_0$ is the "true value" in the data generating process (d.g.p.).
  - Formal condition: $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ if plim $Q_N(\boldsymbol{\theta})$ is maximized at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.
- For Poisson with $\frac{1}{N}\sum_{i=1}^{N}(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{0}$ we require

$$\mathrm{E}[(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}_0))\mathbf{x}_i] = \mathbf{0},$$

  which is the case if

$$\mathrm{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta}_0).$$

- Poisson MLE consistent if the conditional mean is correctly specified.
  - For other models correct mean may be insufficient for consistency.

Asymptotic normality

- Define the following:

$$\begin{aligned} \text{gradient term:} \quad & \mathbf{g}_i(\boldsymbol{\theta}) = \partial q_i(\boldsymbol{\theta})/\partial \boldsymbol{\theta}. \\ \text{Hessian term:} \quad & \mathbf{H}_i(\boldsymbol{\theta}) = \partial^2 q_i(\boldsymbol{\theta})/\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}' = \partial \mathbf{g}_i(\boldsymbol{\theta})/\partial \boldsymbol{\theta}' \end{aligned}$$

- Asymptotic normality:

$$\widehat{\boldsymbol{\theta}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\theta}_0,\ \mathrm{V}[\widehat{\boldsymbol{\theta}}]],$$

- The variance-covariance matrix of the estimator (VCE) is

$$\mathrm{V}[\widehat{\boldsymbol{\theta}}] = \left(\mathrm{E}\left[\textstyle\sum_i \mathbf{H}_i(\boldsymbol{\theta})\right]\right)^{-1} \mathrm{E}\left[\textstyle\sum_i \sum_j \mathbf{g}_i(\boldsymbol{\theta})\mathbf{g}_j(\boldsymbol{\theta})'\right] \left(\mathrm{E}\left[\textstyle\sum_i \mathbf{H}_i(\boldsymbol{\theta})'\right]\right)^{-1}$$

- If data are independent over i we use the robust sandwich estimate

$$\mathrm{V}[\widehat{\boldsymbol{\theta}}] = \left(\textstyle\sum_i \mathbf{H}_i(\widehat{\boldsymbol{\theta}})\right)^{-1} \left(\textstyle\sum_i \mathbf{g}_i(\widehat{\boldsymbol{\theta}})\mathbf{g}_i(\widehat{\boldsymbol{\theta}})'\right) \left(\textstyle\sum_i \mathbf{H}_i(\widehat{\boldsymbol{\theta}})'\right)^{-1}$$

Limit normal derivation

- The first-order conditions are $\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$.
- An exact-first order Taylor series expansion yields

$$\begin{aligned}\frac{1}{N}\sum_i \mathbf{g}_i(\widehat{\boldsymbol{\theta}}) &= \frac{1}{N}\sum_i \mathbf{g}_i(\boldsymbol{\theta}_0) + \frac{1}{N}\sum_i \left.\frac{\partial g_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}^+} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &= \frac{1}{N}\sum_i \mathbf{g}_i(\boldsymbol{\theta}_0) + \frac{1}{N}\sum_i \mathbf{H}_i(\boldsymbol{\theta}^+)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).\end{aligned}$$

- Set this to zero (first-order conditions) and solve

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \left(\frac{1}{N}\sum_i \mathbf{H}_i(\boldsymbol{\theta}^+)\right)^{-1} \frac{1}{N}\sum_i \mathbf{g}_i(\boldsymbol{\theta}_0)$$

  - Linearized so like $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \left(\frac{1}{N}\sum_i \mathbf{x}_i\mathbf{x}_i\right)^{-1} \frac{1}{N}\sum_i \mathbf{x}_i\mathbf{u}_i$.

- We have

$$\begin{aligned}\sqrt{N}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) &= -\left(\tfrac{1}{N}\textstyle\sum_i \mathbf{H}_i(\boldsymbol{\theta}^+)\right)^{-1}\tfrac{1}{\sqrt{N}}\textstyle\sum_i \mathbf{g}_i(\boldsymbol{\theta}_0)\\ &\xrightarrow{d} \mathbf{A}_0^{-1}\times\mathcal{N}[\mathbf{0},\ \mathbf{B}_0]\\ &\xrightarrow{d} \mathcal{N}[\mathbf{0},\ \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}]\end{aligned}$$

by a LLN and a CLT, where

$$\begin{aligned}\mathbf{A}_0 &= \operatorname{plim}\tfrac{1}{N}\textstyle\sum_i \mathbf{H}_i(\boldsymbol{\theta}_0)\\ &= \lim \mathrm{E}[\tfrac{1}{N}\textstyle\sum_i \mathbf{H}_i(\boldsymbol{\theta}_0)]\\ \mathbf{B}_0 &= \operatorname{plim}\tfrac{1}{N}\textstyle\sum_i\sum_j \mathbf{g}_i(\boldsymbol{\theta}_0)\mathbf{g}_j(\boldsymbol{\theta}_0)'\\ &= \lim \mathrm{E}[\tfrac{1}{N}\textstyle\sum_i\sum_j \mathbf{g}_i(\boldsymbol{\theta}_0)\mathbf{g}_j(\boldsymbol{\theta}_0)']\end{aligned}$$

Different Estimates of the VCE

- Need to get estimate $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}]$ of $\mathrm{V}[\widehat{\boldsymbol{\theta}}]$

$$\mathrm{V}[\widehat{\boldsymbol{\theta}}] = \left(\mathrm{E}\left[\textstyle\sum_i \mathbf{H}_i(\boldsymbol{\theta})\right]\right)^{-1} \mathrm{E}\left[\textstyle\sum_i \sum_j \mathbf{g}_i(\boldsymbol{\theta})\mathbf{g}_j(\boldsymbol{\theta})'\right] \left(\mathrm{E}\left[\textstyle\sum_i \mathbf{H}_i(\boldsymbol{\theta})'\right]\right)^{-1}$$

- Leading estimates
  - Robust for independent:
    $\widehat{\mathrm{V}}_{\text{rob}}[\widehat{\boldsymbol{\theta}}] = \left(\sum_i \widehat{\mathbf{H}}_i\right)^{-1} \left(\frac{N}{N-q} \sum_i \widehat{\mathbf{g}}_i \widehat{\mathbf{g}}_i'\right) \left(\sum_i \widehat{\mathbf{H}}_i\right)^{-1}$
  - Cluster-robust:
    $\widehat{\mathrm{V}}_{\text{clus}}[\widehat{\boldsymbol{\theta}}] = \left(\sum_c \widehat{\mathbf{H}}_{cc}\right)^{-1} \left(\frac{C}{C-1} \sum_c \widehat{\mathbf{g}}_c \widehat{\mathbf{g}}_c'\right) \left(\sum_c \widehat{\mathbf{H}}_{cc}\right)^{-1}$
  - Default for MLE:
    $\widehat{\mathrm{V}}_{\text{def}}[\widehat{\boldsymbol{\theta}}] = -\left(\sum_i \mathrm{E}[\mathbf{H}_i(\boldsymbol{\theta})]|_{\widehat{\boldsymbol{\theta}}}\right)^{-1}$
- Clusters rewrite f.o.c. as $\sum_{c=1}^{C} \mathbf{g}_c(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$ for the c^{th} of C clusters, where $\mathbf{g}_c(\boldsymbol{\theta}) = \sum_{i:i\in c} \mathbf{g}_i(\boldsymbol{\theta})$.

Poisson Example

- For Poisson example

$$\begin{aligned} \sum_{i=1}^{N} \mathbf{g}_i(\boldsymbol{\beta}) &= \sum_{i=1}^{N} (y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i \\ \sum_{i=1}^{N} \mathbf{H}_i(\boldsymbol{\beta}) &= \sum_{i=1}^{N} -\exp(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i' \end{aligned}$$

- Three standard estimates are

$$\begin{aligned} \widehat{\mathrm{V}}_{\mathrm{rob}}[\widehat{\boldsymbol{\beta}}] &= \left(\textstyle\sum_i e^{\mathbf{x}_i'\widehat{\boldsymbol{\beta}}}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \left(\textstyle\sum_i (y_i - e^{\mathbf{x}_i'\widehat{\boldsymbol{\beta}}})^2\mathbf{x}_i\mathbf{x}_i'\right) \left(\textstyle\sum_i e^{\mathbf{x}_i'\widehat{\boldsymbol{\beta}}}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \\ \widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}] &= \left(\textstyle\sum_i e^{\mathbf{x}_i'\widehat{\boldsymbol{\beta}}}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \left(\textstyle\sum_c \frac{C}{C-1} \sum_c \widehat{\mathbf{g}}_c\widehat{\mathbf{g}}_c'\right) \left(\textstyle\sum_i e^{\mathbf{x}_i'\widehat{\boldsymbol{\beta}}}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \\ \widehat{\mathrm{V}}_{\mathrm{def}}[\widehat{\boldsymbol{\beta}}] &= \left(\textstyle\sum_i e^{\mathbf{x}_i'\widehat{\boldsymbol{\beta}}}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \end{aligned}$$

  - $\widehat{\mathrm{V}}_{\mathrm{clu}}[\widehat{\boldsymbol{\beta}}]$ used $\widehat{\mathbf{g}}_c = \sum_{i:i\in c}(y_i - e^{\mathbf{x}_i'\widehat{\boldsymbol{\beta}}})\mathbf{x}_i$.
  - $\widehat{\mathrm{V}}_{\mathrm{def}}[\widehat{\boldsymbol{\beta}}]$ relies on $\mathrm{E}[(y_i - e^{\mathbf{x}_i'\boldsymbol{\beta}})^2] = e^{\mathbf{x}_i'\boldsymbol{\beta}}$ (so imposes $\mathrm{V}[y|\mathbf{x}] = \mathrm{E}[y|\mathbf{x}]$).

6. Maximum Likelihood Estimator (MLE)

- Special case of the preceding m-estimator theory.
- With N independent observations $\widehat{\boldsymbol{\theta}}_{\text{ML}}$ maximizes the log-likelihood function:

$$\ln \mathsf{L}(\boldsymbol{\theta}) = \ln \left(\prod_{i=1}^{N} f(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \right) = \sum_{i=1}^{N} \ln f(y_i|\mathbf{x}_i, \boldsymbol{\theta}).$$

- The following properties hold provided essentially that
  - **1.** the true density is $f(y_i|\mathbf{x}_i, \boldsymbol{\theta}_0)$; and
  - **2.** the range of y does not depend on $\boldsymbol{\theta}$ (regularity conditions)
- The MLE is consistent: $\widehat{\boldsymbol{\theta}}_{\text{ML}} \xrightarrow{p} \boldsymbol{\theta}_0$.
- The MLE is asymptotically normal: $\widehat{\boldsymbol{\theta}}_{\text{ML}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\theta}, \mathsf{V}[\widehat{\boldsymbol{\theta}}_{\text{ML}}]]$ with

$$\mathsf{V}[\widehat{\boldsymbol{\theta}}_{\text{ML}}] = \left(\sum_{i=1}^{N} \mathsf{E} \left[\frac{\partial \ln f_i}{\partial \boldsymbol{\theta}} \frac{\partial \ln f_i}{\partial \boldsymbol{\theta}'} \right]_{\widehat{\boldsymbol{\theta}}} \right)^{-1} = - \left(\sum_{i=1}^{N} \mathsf{E} \left[\frac{\partial^2 \ln f_i}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]_{\widehat{\boldsymbol{\theta}}} \right)^{-1}.$$

ML Terminology

- Score: $\partial \ln \mathsf{L}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$.
  - For correctly specified model $\mathsf{E}[\partial \ln \mathsf{L}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}] = \mathbf{0}$.
- Information matrix: $\mathsf{V}\left[\frac{\partial \ln \mathsf{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathsf{E}\left[\frac{\partial \ln \mathsf{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln \mathsf{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right]$.
- Information matrix equality:

$$\mathsf{E}\left[\frac{\partial \ln \mathsf{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln \mathsf{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right] = -\mathsf{E}\left[\frac{\partial^2 \ln \mathsf{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]$$

  - Holds if density correctly specified
  - The asymptotic variance of the MLE is minus the inverse of the information matrix.
- The MLE is fully efficient as its variance matrix is the Cramer-Rao lower bound.

Quasi-MLE: Misspecified density

- What if $f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ is misspecified?
  - The MLE is then called the quasi-MLE or pseudo-MLE
- The quasi-MLE is inconsistent in general.
  - $\widehat{\boldsymbol{\theta}}_{\text{QML}} \xrightarrow{p} \boldsymbol{\theta}^*$ that maximizes $\text{plim}\, N^{-1} \ln \text{L}(\boldsymbol{\theta})$.
  - In general $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$, though there are some notable exceptions.
  - $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ if the specified density is in the linear exponential family (Poisson, binomial, one-parameter gamma, normal) and $\text{E}[y|\mathbf{x}]$ is correctly specified.
- The quasi-MLE is still asymptotically normal with
  - $\widehat{\boldsymbol{\theta}}_{\text{QML}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\theta}^*,\ \text{V}[\widehat{\boldsymbol{\theta}}_{\text{QML}}]]$.
- Given independence over i, "robust" standard errors are based on
  - $\widehat{\text{V}}[\widehat{\boldsymbol{\theta}}_{\text{QML}}] = \left(\sum_i \frac{\partial^2 \ln f_i}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\Big|_{\widehat{\boldsymbol{\theta}}}\right)^{-1} \times \left(\sum_i \frac{\partial \ln f_i}{\partial \boldsymbol{\theta}} \frac{\partial \ln f_i}{\partial \boldsymbol{\theta}'}\Big|_{\widehat{\boldsymbol{\theta}}}\right)^{-1} \times \left(\sum_i \frac{\partial^2 \ln f_i}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\Big|_{\widehat{\boldsymbol{\theta}}}\right)^{-1}$.

7. Nonlinear Least Squares: Definition

- Nonlinear regression model:

$$y_i = g(\mathbf{x}_i, \boldsymbol{\beta}) + u_i,$$

  where the function $g(\cdot)$ is specified.
  - Exponential mean example: $g(\mathbf{x}_i, \boldsymbol{\beta}) = \exp(\mathbf{x}_i'\boldsymbol{\beta})$.
- Nonlinear least squares (NLS) estimator $\widehat{\boldsymbol{\beta}}_{\text{NLS}}$ minimizes the sum of squared residuals

$$S_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - g(\mathbf{x}_i, \boldsymbol{\beta}))^2.$$

- There is no explicit solution for $\widehat{\boldsymbol{\beta}}_{\text{NLS}}$ in general.
  - So compute using iterative gradient methods.

Nonlinear Least Squares: Properties

- Assume that $\mathrm{E}[y_i|\mathbf{x}_i] = g(\mathbf{x}_i, \boldsymbol{\beta})$ so conditional mean correct.
- NLS is consistent and asymptotically normal with

$$\widehat{\boldsymbol{\beta}}_{\text{NLS}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\beta},\ \mathrm{V}[\widehat{\boldsymbol{\beta}}_{\text{NLS}}]].$$

- Given independence over i robust estimate of VCE is

$$\begin{aligned}\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] \quad = \quad & \left(\sum_{i=1}^{N} \left.\frac{\partial g_i}{\partial \boldsymbol{\beta}}\right|_{\widehat{\boldsymbol{\beta}}} \left.\frac{\partial g_i}{\partial \boldsymbol{\beta}'}\right|_{\widehat{\boldsymbol{\beta}}}\right)^{-1} \times \\ & \left(\sum_{i=1}^{N} \widehat{u}_i^2 \left.\frac{\partial g_i}{\partial \boldsymbol{\beta}}\right|_{\widehat{\boldsymbol{\beta}}} \left.\frac{\partial g_i}{\partial \boldsymbol{\beta}'}\right|_{\widehat{\boldsymbol{\beta}}}\right) \left(\sum_{i=1}^{N} \left.\frac{\partial g_i}{\partial \boldsymbol{\beta}}\right|_{\widehat{\boldsymbol{\beta}}} \left.\frac{\partial g_i}{\partial \boldsymbol{\beta}'}\right|_{\widehat{\boldsymbol{\beta}}}\right)^{-1}\end{aligned}$$

  - where $\widehat{u}_i = (y_i - g(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}))$ is the NLS residual.
- Exactly the same as OLS except $\partial g_i/\partial \boldsymbol{\beta}|_{\widehat{\boldsymbol{\beta}}}$ replaces $\mathbf{x}_i$!

- NLS regression for the model $y = \exp(\mathbf{x}'\boldsymbol{\beta}) + u$.

```
. * Nonlinear least-squares regression (command nl)
. generate one = 1

. nl (docvis = exp({xb: private chronic female income one})), vce(robust) nolog
(obs = 4412)

Nonlinear regression                            Number of obs =      4412
                                                R-squared     =    0.3046
                                                Adj R-squared =    0.3038
                                                Root MSE      =  7.407479
                                                Res. dev.     =  30185.68
```

docvis	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
/xb_private	.7105104	.1086194	6.54	0.000	.4975618	.923459
/xb_chronic	1.057318	.0558352	18.94	0.000	.947853	1.166783
/xb_female	.4320225	.0694662	6.22	0.000	.2958338	.5682112
/xb_income	.002558	.0012544	2.04	0.041	.0000988	.0050173
/xb_one	-.040563	.1126216	-0.36	0.719	-.2613578	.1802319

- NLS slope coeffs are: 0.71, 1.06, 0.43, 0.0026.
  Poisson slope coeffs: 0.79, 1.09, 0.49, 0.0036.
  NLS robust standard errors are: 0.109, 0.056, 0.069, 0.0012.
  Poisson robust standard errors: 0.109, 0.056, 0.058, 0.0011.

8. Statistical inference: Wald test

- Wald test of linear restrictions on θ: same method as for OLS.
- Wald test of h nonlinear restrictions:

$$H_0 : \mathbf{h}(\theta) = \mathbf{0} \text{ against } H_a : \mathbf{h}(\theta) \neq \mathbf{0}.$$

- Base test on does $\mathbf{h}(\widehat{\theta}) \simeq \mathbf{0}$?
- By Taylor series with $\mathbf{R} = \partial \mathbf{h}(\theta)/\partial \theta'$

$$\begin{aligned} \mathbf{h}(\widehat{\theta}) &\simeq \mathbf{h}(\theta) + \widehat{\mathbf{R}}(\widehat{\theta} - \theta) \\ &= \widehat{\mathbf{R}}(\widehat{\theta} - \theta) \text{ under } H_0 : \mathbf{h}(\theta) = \mathbf{0} \\ &\stackrel{a}{\sim} \mathcal{N}[\mathbf{0}, \widehat{\mathbf{R}}\mathrm{V}[\widehat{\theta}]\widehat{\mathbf{R}}'] \end{aligned}$$

- The Wald test statistics are

$$\begin{aligned} \mathrm{W} &= \mathbf{h}(\widehat{\theta})'[\widehat{\mathbf{R}}\widehat{\mathrm{V}}[\widehat{\theta}]\widehat{\mathbf{R}}']^{-1}\mathbf{h}(\widehat{\theta}) \stackrel{a}{\sim} \chi^2(h) \text{ under } H_0 \\ \mathrm{F} &= (\mathrm{W}/h) \stackrel{a}{\sim} F(h, N-q) \text{ under } H_0. \end{aligned}$$

Confidence interval (delta method)

- Consider scalar $\gamma = g(\boldsymbol{\theta})$ for specified function $g(\cdot)$.
- A 95% confidence interval for γ is then

$$\widehat{\gamma} \pm 1.96 \times s^2_{\widehat{\gamma}}.$$

  - where $s^2_{\widehat{\gamma}} = \left.\frac{\partial \gamma}{\partial \boldsymbol{\theta}'}\right|_{\widehat{\boldsymbol{\theta}}} \times \widehat{\mathrm{V}}[\widehat{\boldsymbol{\theta}}] \times \left.\frac{\partial \gamma}{\partial \boldsymbol{\theta}}\right|_{\widehat{\boldsymbol{\theta}}}$.

- Derivation:

$$\begin{aligned} g(\widehat{\boldsymbol{\theta}}) &\simeq g(\boldsymbol{\theta}) + g'(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\ \widehat{\gamma} &\simeq \gamma + \left.\frac{\partial \gamma}{\partial \boldsymbol{\theta}'}\right|_{\widehat{\boldsymbol{\theta}}} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\ \mathrm{Var}[\widehat{\gamma}] &= \left.\frac{\partial \gamma}{\partial \boldsymbol{\theta}'}\right|_{\widehat{\boldsymbol{\theta}}} \mathrm{Var}[\widehat{\boldsymbol{\theta}}] \left.\frac{\partial \gamma}{\partial \boldsymbol{\theta}}\right|_{\widehat{\boldsymbol{\theta}}} \end{aligned}$$

  - called the delta method as derivative taken.

Data example

- Wald test of $H_0 : \beta_2/\beta_2 = 1$ against $H_a : \beta_2/\beta_2 = 1$.
  95% confidence interval for $\gamma = \beta_2/\beta_2 - 1$.

```
. * Wald test of nonlinear hypothesis
. quietly poisson docvis private chronic female income, vce(robust)

. testnl _b[female]/_b[private]=1

  (1)  _b[female]/_b[private] = 1

               chi2(1) =       13.51
           Prob > chi2 =        0.0002

.
. * Delta method confidence interval
. nlcom _b[female]/_b[private] - 1

       _nl_1:  _b[female]/_b[private] - 1
```

docvis	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	-.383286	.1042734	-3.68	0.000	-.587658	-.1789139

- Reject H_0 as $p = 0.0002$; 95% conf. interval is $(-0.59, -0.18)$.

Likelihood-based tests

- Test $H_0 : \mathbf{h}(\boldsymbol{\theta}_0) = \mathbf{0}$ which imposes h restrictions.
- Three tests Wald, likelihood ratio, and Lagrange multiplier (or score) test are
  - asymptotically equivalent under H_0 (all $\chi^2(h)$)
  - asymptotically equivalent under local alternatives $\mathbf{h}(\boldsymbol{\theta}_0) = \mathbf{c}/\sqrt{N}$.
  - Wald most often used as can robustify.
- Define
  - $L(\boldsymbol{\theta})$ is the likelihood function.
  - $\widehat{\boldsymbol{\theta}}_u$ = unrestricted MLE that maximizes $\ln L(\boldsymbol{\theta})$.
  - $\widetilde{\boldsymbol{\theta}}_r$ = restricted MLE that maximizes $\ln L(\boldsymbol{\theta}) - \boldsymbol{\lambda}'\mathbf{h}(\boldsymbol{\theta})$. (where $\boldsymbol{\lambda}$ is vector of Lagrangian multipliers.)

1. Wald test: Does $\mathbf{h}(\widehat{\boldsymbol{\theta}}_u) \simeq \mathbf{0}$?

$$W = \mathbf{h}(\widehat{\boldsymbol{\theta}}_u)'[\widehat{\mathbf{R}}\mathrm{V}[\widehat{\boldsymbol{\beta}}]\widehat{\mathbf{R}}']^{-1}\mathbf{h}(\widehat{\boldsymbol{\theta}}_u).$$

2. Likelihood Ratio Test: Does $\ln L(\widetilde{\boldsymbol{\theta}}_r) \simeq \ln L(\widehat{\boldsymbol{\theta}}_u)$?

$$\mathrm{LR} = -2[\ln L(\widetilde{\boldsymbol{\theta}}_r) - \ln L(\widehat{\boldsymbol{\theta}}_u)].$$

3. Score test: Does $\partial \ln L(\boldsymbol{\theta})/\partial\boldsymbol{\theta} \simeq \mathbf{0}$ when evaluated at $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}_r$?
$[L(\boldsymbol{\theta})$ is for unrestricted model so $\partial \ln L(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$ at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_u]$.

$$\text{Score} = \mathrm{LM} = \left.\frac{\partial \ln L}{\partial \boldsymbol{\theta}'}\right|_{\widetilde{\boldsymbol{\theta}}_r} (N\widetilde{\mathbf{A}}^{-1}) \left.\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right|_{\widetilde{\boldsymbol{\theta}}_r}$$

4. LM test: does Lagrange multiplier $\widetilde{\boldsymbol{\lambda}}_r \simeq \mathbf{0}$?
Equals 3. as maximizing $\ln L(\boldsymbol{\theta}) - \boldsymbol{\lambda}'\mathbf{h}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ implies

$$\left.\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right|_{\widetilde{\boldsymbol{\theta}}_r} = \left.\frac{\partial \mathbf{h}(\boldsymbol{\theta})'}{\partial \boldsymbol{\theta}}\right|_{\widetilde{\boldsymbol{\theta}}_r} \times \widetilde{\boldsymbol{\lambda}}_r.$$

Score vector is a full rank matrix multiple of the Lagrange multipliers.

11. Gradient methods

- Consider estimator $\widehat{\boldsymbol{\theta}}$ that is a local maximum, solving

$$\mathbf{g}(\widehat{\boldsymbol{\theta}}) = \mathbf{0}, \text{ where } \mathbf{g}(\boldsymbol{\theta}) = \frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

- Iterative gradient methods update the s^{th} round estimate $\widehat{\boldsymbol{\theta}}_s$ by a matrix weighted average of the gradient

$$\widehat{\boldsymbol{\theta}}_{s+1} = \widehat{\boldsymbol{\theta}}_s + \mathbf{A}_s \mathbf{g}(\widehat{\boldsymbol{\theta}}_s), \qquad s = 1, ..., S.$$

  - motivation: change $\widehat{\boldsymbol{\theta}}$ in direction with greatest impact on $Q(\widehat{\boldsymbol{\theta}})$ i.e. where the gradient is largest, though need to scale the gradient.
  - the weighting matrix $\mathbf{A}_s$ is a $q \times q$ matrix that depends on $\widehat{\boldsymbol{\theta}}_s$

- Newton-Raphson is leading example with $\mathbf{A}_s = -\mathbf{H}(\widehat{\boldsymbol{\theta}}_s)^{-1}$, where

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \frac{\partial^2 Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

is the Hessian matrix for the optimization problem.

Newton-Raphson

- Motivate NR: second-order Taylor series expansion around $\widehat{\boldsymbol{\theta}}_s$

$$Q(\boldsymbol{\theta}) = Q(\widehat{\boldsymbol{\theta}}_s) + \mathbf{g}(\widehat{\boldsymbol{\theta}}_s)'(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_s) + \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_s)'\mathbf{H}(\widehat{\boldsymbol{\theta}}_s)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_s) + R,$$

  where R is a remainder that we now ignore.
- Maximize the approximation $Q^*(\boldsymbol{\theta})$ w.r.t. respect to $\boldsymbol{\theta}$.

$$\frac{\partial Q^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{g}(\widehat{\boldsymbol{\theta}}_s) + \mathbf{H}(\widehat{\boldsymbol{\theta}}_s)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_s) = \mathbf{0}.$$

- Solving yields the Newton-Raphson iteration

$$(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_s) = -\mathbf{H}(\widehat{\boldsymbol{\theta}}_s)^{-1}\mathbf{g}(\widehat{\boldsymbol{\theta}}_s).$$

  - This increases $Q(\boldsymbol{\theta})$ if $\mathbf{H}(\widehat{\boldsymbol{\theta}}_s)$ is negative definite (and we ignore R)
  - This works especially well if $Q(\boldsymbol{\theta})$ is globally concave.

Gradient methods: Stopping criteria

- Stop iterations when
  - (1) small change in the coefficient vector (`tolerance`)
  - (2) small change in the objective function (`ltolerance`)
  - (3) small gradient relative to the Hessian (`nrtolerance`)
  - (4) small gradient relative to the coefficients (gtolerance)
  - (5) maximum number of iterations reached (i.e. NOT CONVERGED)
- Stata default values for these criteria can be changed
  See `help maximize`.
- Stata built-in commands use (1), (2) and (5) only.

Poisson Example

- For the Poisson MLE

$$\begin{aligned} Q(\beta) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\{-\exp(\mathbf{x}_i'\beta) + y_i\mathbf{x}_i'\beta - \ln y_i!\} \\ g(\beta) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(y_i - \exp(\mathbf{x}_i'\beta))\mathbf{x}_i \\ H(\beta) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} -\exp(\mathbf{x}_i'\beta)\mathbf{x}_i\mathbf{x}_i'. \end{aligned}$$

- The Newton-Raphson iterations are

$$\widehat{\beta}_{s+1} = \widehat{\beta}_s + \left[\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\exp(\mathbf{x}_i'\widehat{\beta}_s)\mathbf{x}_i\mathbf{x}_i'\right]^{-1} \times \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(y_i - \exp(\mathbf{x}_i'\widehat{\beta}_s))\mathbf{x}_i.$$

  - $H(\beta) = -\mathbf{X}'\mathbf{D}\mathbf{X}$ where $\mathbf{X}$ is $N \times k$ regressors and $\mathbf{D} = \text{Diag}[\exp(\mathbf{x}_i'\beta)]$ is an $N \times N$ diagonal matrix with positive entries.
  - So $H(\beta)$ is negative definite for all β and objective function is globally concave.
  - NR will work very well here unless regressors are highly multicollinear.

Application

- Apply Newton-Raphson to Poisson MLE using Mata.

```
. * Newton-Raphson in Mata for Poisson MLE
. * Set up data and local macros for dependent variable and regressors
. generate cons = 1

. local y docvis

. local xlist private chronic female income cons

. * Mata commands for Poisson MLE NR iterations
. mata:
------------------------------------------------- mata (type end to exit) -----------------
:   st_view(y=., ., "`y'")              // read in stata data to y and X

:   st_view(X=., ., tokens("`xlist'"))

:   b = J(cols(X),1,0)                  // compute starting values

:   n = rows(X)

:   iter = 1                            // initialize number of iterations

:   cha = 1                             // initialize stopping criterion
```

```
:    do {
>       mu = exp(X*b)
>       grad = X'(y-mu)                  // k x 1 gradient vector
>       hes = cross(X, mu, X)            // negative of the k x k Hessian matrix
>       bold = b
>       b = bold + cholinv(hes)*grad
>       cha = (bold-b)'(bold-b)/(bold'bold)
>       iter = iter + 1
>    } while (cha > 1e-16)               // end of iteration loops

:    mu = exp(X*b)

:    hes = cross(X, mu, X)

:    vgrad = cross(X, (y-mu):^2, X)

:    vb = cholinv(hes)*vgrad*cholinv(hes)*n/(n-cols(X))

:    iter                               // num iterations
  13

:    cha                                // stopping criterion
  1.11462e-24

:    st_matrix("b",b')                  // pass results from Mata to Stata

:    st_matrix("V",vb)                  // pass results from Mata to Stata

: end
```

- Mata computes $\widehat{\boldsymbol{\beta}}$ in b and $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$ in V.
  - Use Stata command ereturn to nicely display results.

```
. * Present results, nicely formatted using Stata command ereturn
. matrix colnames b = `xlist'

. matrix colnames V = `xlist'

. matrix rownames V = `xlist'

. ereturn post b V

. ereturn display
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
private	.7986654	.1090509	7.32	0.000	.5849295	1.012401
chronic	1.091865	.0560205	19.49	0.000	.9820669	1.201663
female	.4925481	.058563	8.41	0.000	.3777666	.6073295
income	.003557	.001083	3.28	0.001	.0014344	.0056796
cons	-.2297263	.1109236	-2.07	0.038	-.4471325	-.0123202

- Results are the same as from Stata command poisson.

10. Appendix: Theory for extremum estimator

- More general framework than m-estimators.
  Additionally includes generalized method of moments (GMM).
- Extremum estimator $\widehat{\boldsymbol{\theta}}$ maximizes

$$Q_N(\boldsymbol{\theta}) = Q_N(\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}).$$

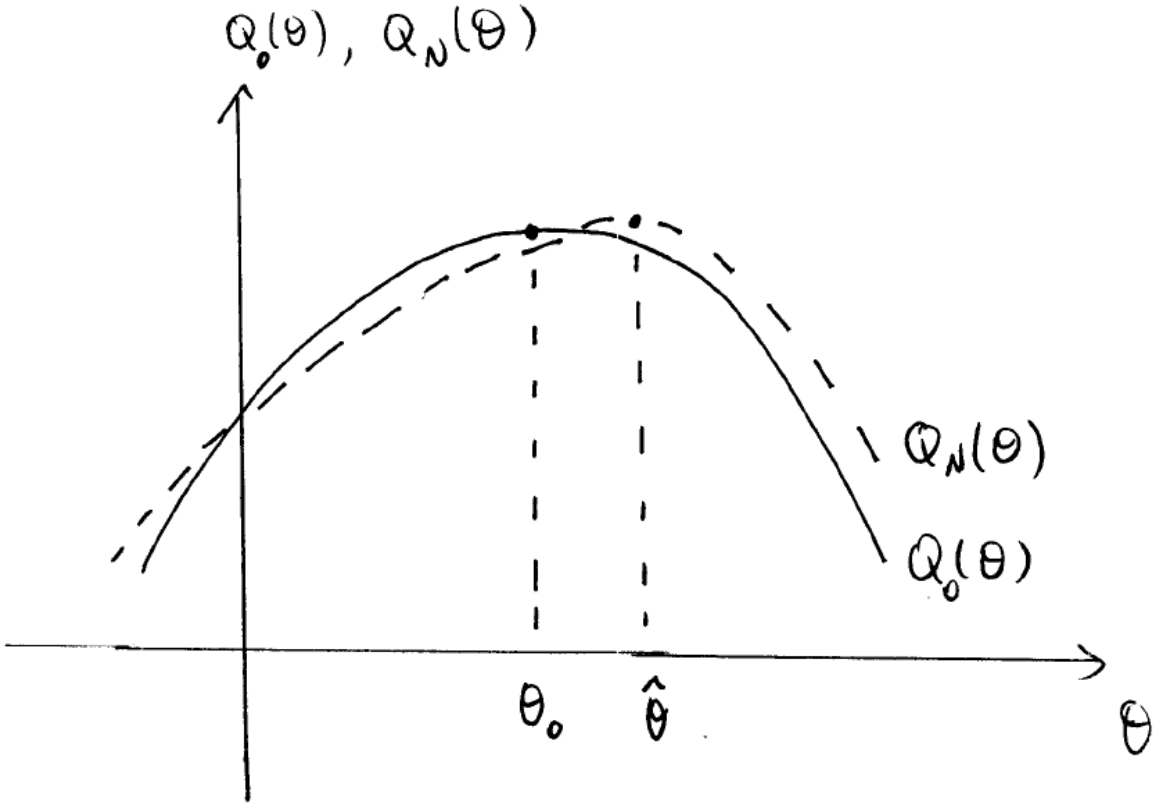
- For a global maximum, so $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} Q_N(\boldsymbol{\theta})$.
- Usually a local maximum, solving

$$\left. \frac{\partial Q_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\widehat{\boldsymbol{\theta}}} = \mathbf{0}.$$

- Leading examples: MLE, NLS, and GMM.

Consistency

- If $Q_N(\boldsymbol{\theta}) \xrightarrow{p} Q_0(\boldsymbol{\theta})$ (where $Q_0(\boldsymbol{\theta})$ is nonstochastic) then the local maximum $\xrightarrow{p}$ to each other
- So $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is the local maximum of $Q_0(\boldsymbol{\theta})$.



Consistency of Local Maximum *(Amemiya (1985, Theorem 4.1.2)).*
Make the assumptions:

(i) *The parameter space* Θ *is an open subset of* R^q;

(ii) $Q_N(\boldsymbol{\theta})$ *is a measurable function of the data for all* $\boldsymbol{\theta} \in \Theta$, *and* $\partial Q_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ *exists and is continuous in an open neighborhood of* $\boldsymbol{\theta}_0$;

(iii) *The objective function* $Q_N(\boldsymbol{\theta})$ *converges uniformly in probability to* $Q_0(\boldsymbol{\theta})$ *in an open neighborhood of* $\boldsymbol{\theta}_0$, *and* $Q_0(\boldsymbol{\theta})$ *attains a unique local maximum at* $\boldsymbol{\theta}_0$.

Then one of the solutions to $\partial Q_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$ *is* **consistent** *for* $\boldsymbol{\theta}_0$.

Limit normal distribution

- Based on exact first-order Taylor series (mean-value theorem).
- For $f(\cdot)$ differentiable there is always x^+ between x and x_0 such that

$$f(x) = f(x_0) + f'(x^+)(x - x_0).$$

- Extend to $\mathbf{f}(\cdot)$ a vector function of the vector $\boldsymbol{\theta}$

$$\mathbf{f}(\widehat{\boldsymbol{\theta}}) = \mathbf{f}(\boldsymbol{\theta}_0) + \left.\frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}^+} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

- Here $\mathbf{f}(\boldsymbol{\theta}) = \partial Q_N(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is already a first derivative. Then

$$\left.\frac{\partial Q_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\widehat{\boldsymbol{\theta}}} = \left.\frac{\partial Q_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_0} + \left.\frac{\partial^2 Q_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}^+} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

- Set the right-hand side to $\mathbf{0}$ and solve for

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\ \sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\left(\left.\frac{\partial^2 Q_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}^+}\right)^{-1} \sqrt{N}\left.\frac{\partial Q_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}_0}.$$

Asymptotic Normality of Local Maximum *(Amemiya (1985, Theorem 4.1.3)). In addition to preceding assumptions for consistency assume:*

(i) $\partial^2 Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ *exists and is continuous in an open convex neighborhood of* $\boldsymbol{\theta}_0$;

(ii) $\partial^2 Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\big|_{\boldsymbol{\theta}^+}$ *converges in prob. to finite nonsingular matrix*

$$\mathbf{A}_0 = \text{plim}\, \partial^2 Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\big|_{\boldsymbol{\theta}_0}$$

for any sequence $\boldsymbol{\theta}^+$ *such that* $\boldsymbol{\theta}^+ \xrightarrow{p} \boldsymbol{\theta}_0$;

(iii) $\sqrt{N}\,\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}_0} \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{B}_0]$, *where*

$$\mathbf{B}_0 = \text{plim}\left[N\,\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta} \times \partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'\big|_{\boldsymbol{\theta}_0}\right].$$

Then

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}]$$

where $\widehat{\boldsymbol{\theta}}$ *is the consistent solution to* $\partial Q_N(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$.

11. Appendix: Theory for Poisson MLE

- The Poisson quasi-MLE maximizes

$$Q_N(\boldsymbol{\beta}) = N^{-1}\sum_i \left\{-e^{\mathbf{x}_i'\boldsymbol{\beta}} + y_i\mathbf{x}_i'\boldsymbol{\beta} - \ln y_i!\right\}$$

- Then

$$\begin{aligned} Q_0(\boldsymbol{\beta}) &= \text{plim} N^{-1}\sum_i \left\{-e^{\mathbf{x}_i'\boldsymbol{\beta}} + y_i\mathbf{x}_i'\boldsymbol{\beta} - \ln y_i!\right\} \\ &= \lim N^{-1}\sum_i \left\{-\mathsf{E}\left[e^{\mathbf{x}_i'\boldsymbol{\beta}}\right] + \mathsf{E}[y_i\mathbf{x}_i'\boldsymbol{\beta}] - \mathsf{E}\left[\ln y_i!\right]\right\} \text{ by LLN} \\ &= \lim N^{-1}\sum_i \left\{-\mathsf{E}\left[e^{\mathbf{x}_i'\boldsymbol{\beta}}\right] + \mathsf{E}\left[e^{\mathbf{x}_i'\boldsymbol{\beta}_0}\mathbf{x}_i'\boldsymbol{\beta}\right] - \mathsf{E}\left[\ln y_i!\right]\right\}. \end{aligned}$$

- The key step is

$$\begin{aligned} &\text{plim}\, N^{-1}\sum_i y_i\mathbf{x}_i'\boldsymbol{\beta} \\ &= \lim N^{-1}\sum_i \mathsf{E}\left[y_i\mathbf{x}_i'\boldsymbol{\beta}\right] \text{ by LLN} \\ &= \lim N^{-1}\sum_i \mathsf{E}\left[e^{\mathbf{x}_i'\boldsymbol{\beta}_0}\mathbf{x}_i'\boldsymbol{\beta}\right] \text{ if } \mathsf{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta}_0). \end{aligned}$$

Poisson Consistency

- Then

$$Q_0(\boldsymbol{\beta}) = \lim N^{-1} \sum_i \left\{ -\mathrm{E}\left[e^{\mathbf{x}_i'\boldsymbol{\beta}}\right] + \mathrm{E}\left[e^{\mathbf{x}_i'\boldsymbol{\beta}_0}\mathbf{x}_i'\boldsymbol{\beta}\right] - \mathrm{E}\left[\ln y_i!\right] \right\}$$

  so (noting that $\mathrm{E}[\ln y_i!]$ depends on $\boldsymbol{\beta}_0$ and not $\boldsymbol{\beta}$)

$$\begin{aligned} \frac{\partial Q_0(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -\lim N^{-1} \sum_i \mathrm{E}\left[e^{\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x}_i\right] + \lim N^{-1} \sum_i \mathrm{E}\left[e^{\mathbf{x}_i'\boldsymbol{\beta}_0}\mathbf{x}_i\right] \\ &= \mathbf{0} \text{ at } \boldsymbol{\beta} = \boldsymbol{\beta}_0. \end{aligned}$$

- So $\widehat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ as
  - $Q_N(\boldsymbol{\beta}) \xrightarrow{p} Q_0(\boldsymbol{\beta})$ and
  - $Q_0(\boldsymbol{\beta})$ is maximized at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

- A simpler heuristic "proof" is that $\widehat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ if

$$\begin{aligned} \mathrm{E}\left[\left.\frac{\partial Q_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_0}\right] &= N^{-1}\textstyle\sum_i (y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}_0))\mathbf{x}_i \\ &= \mathbf{0}. \end{aligned}$$

- This is the case if the d.g.p. such that

$$\mathrm{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta}_0)$$

  i.e. that the conditional mean is correctly specified

  - this result holds more generally if the specified density is in linear exponential family.

Poisson MLE: Limit normal distribution

- For Poisson

$$\begin{aligned}\sqrt{N}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) &= -\left(N^{-1}\textstyle\sum_i \mathbf{H}_i(\boldsymbol{\beta}^+)\right)^{-1} \times \sqrt{N}\textstyle\sum_i \mathbf{g}_i(\boldsymbol{\beta}_0) \\ &= -\left(N^{-1}\textstyle\sum_i -e^{\mathbf{x}_i'\boldsymbol{\beta}^+}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} \times N^{-1/2}\textstyle\sum_i (y_i - e^{\mathbf{x}_i'\boldsymbol{\beta}_0})\mathbf{x}\end{aligned}$$

- First term by a LLN

$$N^{-1}\textstyle\sum_i -e^{\mathbf{x}_i'\boldsymbol{\beta}^+}\mathbf{x}_i\mathbf{x}_i' \xrightarrow{p} \mathbf{A}_0 = -\lim \textstyle\sum_i \mathrm{E}[e^{\mathbf{x}_i'\boldsymbol{\beta}_0}\mathbf{x}_i\mathbf{x}_i']$$

- Second term by a CLT and with independence over i

$$\begin{aligned}&N^{-1/2}\textstyle\sum_i (y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}_0))\mathbf{x}_i \\ &\xrightarrow{p} \mathcal{N}\left[0,\ \mathbf{B}_0 = \lim N^{-1}\textstyle\sum_i \mathrm{E}[\mathrm{E}[(y_i - e^{\mathbf{x}_i'\boldsymbol{\beta}_0})^2|\mathbf{x}_i]\mathbf{x}_i\mathbf{x}_i']\right]\end{aligned}$$

- Combining

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}].$$

- Note that if y_i is Poisson distributed then

$$\mathsf{E}[(y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}_0))^2|\mathbf{x}_i] = \mathsf{V}[y_i|\mathbf{x}_i] = \mathsf{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta}_0)$$

  - so $\mathbf{B}_0 = -\mathbf{A}_0$ and
  - $\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, -\mathbf{A}_0^{-1}]$.