MPhil Econometrics – Limited Dependent Variables and Selection

Francis J. DiTraglia

University of Oxford

Complied on 2020-01-27 at 19:13:28

Housekeeping

Lecturer:	Francis J. DiTraglia
Email:	francis.ditraglia@ox.ac.uk
Slides & Errata:	`http://ditraglia.com/teaching`
Office:	2132 Manor Road Building
Meetings:	10:30 on Wednesdays (after lecture) or by appointment

References

- **Wooldridge (2010) – *Econometric Analysis of Cross Section & Panel Data***
- Cameron & Trivedi (2005) – *Microeconometrics: Methods and Applications*
- Train (2009) – *Discrete Choice Methods with Simulation*

Lecture $\#1$ – Maximum Likelihood Estimation Under Mis-specification

Review: the Poisson Distribution

The Kullback-Leibler Divergence

Example: Consistency of Poisson MLE

Asymptotic Theory for MLE Under Mis-specification

The Information Matrix Equality

Example: Asymptotic Variance Calculations for Poisson MLE

“All models are wrong; some are useful.”

Question

What happens if we carry out maximum likelihood estimation, but our model is *wrong*?

This Lecture

Examine a simple example in excruciating detail; present the general theory.

Next Lecture

Apply what we’ve learned to study Poisson Regression, a model for count data.

Suppose that $y \sim \text{Poisson}(\theta)$

Support Set: $\{0, 1, 2, \dots\}$

A Poisson Random Variable is a *count*.

$$\sum_{y=0}^{\infty} \frac{e^{-\theta}\theta^y}{y!} = e^{-\theta} \sum_{y=0}^{\infty} \frac{\theta^y}{y!} = e^{-\theta}\left(e^{\theta}\right) = 1$$

Probability Mass Function

$$f(y|\theta) = \frac{e^{-\theta}\theta^y}{y!}$$

Expected Value: $\mathbb{E}(y) = \theta$

Poisson parameter θ equals the mean of y.

$$\mathbb{E}(y) = \sum_{y=0}^{\infty} y \frac{e^{-\theta}\theta^y}{y!} = \sum_{y=1}^{\infty} y \frac{e^{-\theta}\theta^y}{y!}$$

$$= \theta \sum_{y=1}^{\infty} \frac{e^{-\theta}\theta^{y-1}}{(y-1)!} = \theta \sum_{y=0}^{\infty} \frac{e^{-\theta}\theta^y}{y!} = \theta$$

Variance: $\text{Var}(y) = \theta$

You will show this on the problem set.

MLE for θ where $y_1, y_2, \ldots, y_N \sim$ iid Poisson(θ).

The Likelihood (iid data)

$L_N(\theta) \equiv \prod_{i=1}^{N} \frac{e^{-\theta}\theta^{y_i}}{y_i!}$

The Log-Likelihood

$\ell_N(\theta) = \sum_{i=1}^{N} [y_i \log(\theta) - \theta - \log(y_i!)]$

Maximum Likelihood Estimator

$\widehat{\theta} \equiv \underset{\theta \in \Theta}{\arg\max}\ \ell_N(\theta) = \bar{y}$

$$\frac{d}{d\theta}\ell_N(\theta) = \sum_{i=1}^{N}\left[\frac{y_i}{\theta} - 1\right]$$

$$\frac{d}{d\theta}\ell_N(\widehat{\theta}) = 0$$

$$\sum_{i=1}^{N}\left[y_i/\widehat{\theta} - 1\right] = 0$$

$$\left(\sum_{i=1}^{N} y_i\right)/\widehat{\theta} = N$$

$$\frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y} = \widehat{\theta}$$

The Kullback-Leibler (KL) Divergence

Motivation

How well does a parametric model $f(\mathbf{y}|\boldsymbol{\theta})$ approximate a *true* density/pmf $p_o(\mathbf{y})$?

Definition

$$KL(p_o; f_{\boldsymbol{\theta}}) \equiv \mathbb{E}\left[\log\left\{\frac{p_o(\mathbf{y})}{f(\mathbf{y}|\boldsymbol{\theta})}\right\}\right]$$

KL Properties

1. *Asymmetric*: $KL(p_o; f_{\boldsymbol{\theta}}) \neq KL(f_{\boldsymbol{\theta}}; p_o)$
2. $KL(p_o; f_{\boldsymbol{\theta}}) \geq 0$; zero iff $p_o = f_{\boldsymbol{\theta}}$
3. Min KL iff max expected log-likelihood

Alternative Expression

$$\boxed{\mathbb{E}\left[\log\left\{\frac{p_o(\mathbf{y})}{f(\mathbf{y}|\boldsymbol{\theta})}\right\}\right] = \underbrace{\mathbb{E}\left[\log p_o(\mathbf{y})\right]}_{\text{Constant wrt } \boldsymbol{\theta}} - \underbrace{\mathbb{E}\left[\log f(\mathbf{y}|\boldsymbol{\theta})\right]}_{\text{Expected Log-like.}}}$$

All expectations are wrt p_o

$p_o(\mathbf{y})$ and $f(\mathbf{y}|\boldsymbol{\theta})$ are merely *functions* of the RV $\mathbf{y}$

$$\mathbb{E}[\log p_o(\mathbf{y})] = \int \log p_o(\mathbf{y}) p_o(\mathbf{y})\, d\mathbf{y}$$

$$\mathbb{E}[\log f(\mathbf{y}|\boldsymbol{\theta})] = \int \log f(\mathbf{y}|\boldsymbol{\theta}) p_o(\mathbf{y})\, d\mathbf{y}$$

Watch Out!

$KL = \infty$ if $\exists \mathbf{y}$ with $f(\mathbf{y}|\boldsymbol{\theta}) = 0$ & $p_o(\mathbf{y}) \neq 0$

$\mathrm{KL}(p_o; f) \geq 0$ with equality iff $p_o = f$

Jensen's Inequality

If φ is convex then $\varphi(\mathbb{E}[y]) \leq \mathbb{E}[\varphi(y)]$, with equality iff φ is linear or y is constant.

log is concave so $(-\log)$ is convex

$$
\begin{aligned}
\mathbb{E}\left[\log\left\{\frac{p_o(y)}{f(y)}\right\}\right] = \mathbb{E}\left[-\log\left\{\frac{f(y)}{p_o(y)}\right\}\right] &\geq -\log\left\{\mathbb{E}\left[\frac{f(y)}{p_o(y)}\right]\right\} \\
&= -\log\left\{\int_{-\infty}^{\infty} \frac{f(y)}{p_o(y)} \cdot p_o(y)\, dy\right\} \\
&= -\log\left\{\int_{-\infty}^{\infty} f(y)\, dy\right\} \\
&= -\log(1) = 0
\end{aligned}
$$

A Simple Example: Calculating the KL Divergence

Remember: all expectations are calculated using p_o.

True Distribution p_o

$y_1, \ldots, y_N \sim$ iid p_o where:

$p_o(0) = \frac{2}{5}, p_o(1) = \frac{1}{5}, p_o(2) = \frac{2}{5}$.

Mis-specified Model f_θ

$y_1, \ldots, y_N \sim$ iid Poisson(θ)

KL Divergence

$KL(p_o; f_\theta) = \theta - \log\theta + (\text{Constant})$

$$\boxed{KL(p_o; f_\theta) = \mathbb{E}[\log p_o(y)] - \mathbb{E}[\log f(y|\theta)]}$$

$$\begin{aligned}\mathbb{E}[\log p_o(y)] &= \sum_{\text{all } y} \log\left[p_o(y)\right] p_o(y) \\ &= \log\left(\frac{2}{5}\right) \times \frac{2}{5} + \log\left(\frac{1}{5}\right) \times \frac{1}{5} + \log\left(\frac{2}{5}\right) \times \frac{2}{5}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\log f(y|\theta)] &= \sum_{\text{all } y} \log\left[\frac{e^{-\theta}\theta^y}{y!}\right] p_o(y) \\ &= \log\left(e^{-\theta}\right) \times \frac{2}{5} + \log\left(e^{-\theta}\theta\right) \times \frac{1}{5} + \log\left(\frac{e^{-\theta}\theta^2}{2}\right) \times \frac{2}{5} \\ &= -\left[\theta - \log(\theta) + \log(2) \times \frac{2}{5}\right]\end{aligned}$$

A Simple Example Continued: Minimizing the KL Divergence

Model = Poisson(θ); True Dist. $p_o(0) = p_o(2) = \frac{2}{5}$ and $p_o(1) = \frac{1}{5}$

Best Approximation

What parameter value θ_o makes the Poisson(θ) model *as close as possible* to the true distribution p_o, where we measure "closeness" using the KL-divergence?

Using the previous slide

$$KL(p_o; f_\theta) = \theta - \log\theta + (\text{Const.})$$

$$\text{FOC:} \quad 0 = 1 - \frac{1}{\theta} \implies \boxed{\theta = 1}$$

A more direct approach

Min KL $\iff$ Max Expected Log-like.

$$\frac{d}{d\theta}\mathbb{E}[\log f(y|\theta)] = \mathbb{E}\left[\frac{d}{d\theta}\{-\theta + y\log(\theta) - \log(y!)\}\right]$$

$$= \mathbb{E}[-1 + y/\theta] = \mathbb{E}[y]/\theta - 1 = 0$$

$$\implies \boxed{\theta = \mathbb{E}[y]}$$

A Simple Example Continued: Minimizing the KL Divergence

Model = Poisson(θ); True Dist. $p_o(0) = p_o(2) = \frac{2}{5}$ and $p_o(1) = \frac{1}{5}$

Best Approximation

What parameter value θ_o makes the Poisson(θ) model *as close as possible* to the true distribution p_o, where we measure "closeness" using the KL-divergence?

Using the previous slide: $\theta_o = 1$

A more direct approach: $\theta_o = \mathbb{E}[y]$

Both Methods Agree

- For the specified p_o we have: $\mathbb{E}[y] = 0 \times \frac{1}{5} + 1 \times \frac{2}{5} + 2 \times \frac{2}{5} = 1$.
- The "Direct approach" is general: works for *any* p_o (under regularity conditions)

Is this just a coincidence?

We have shown that:

1. Under an iid Poisson(θ) model for $y_1, \ldots, y_N$, the MLE for θ is $\widehat{\theta} = \bar{y}$
2. For *any* (reasonable) p_o, setting $\theta_o = \mathbb{E}[y_i]$ minimizes $KL(p_o; f_\theta)$.

By the (weak) law of large numbers:

If $y_1, \ldots, y_N \sim$ iid, then $\bar{y}$ is a consistent estimator of $\mathbb{E}[y_i]$ as N approaches infinity.

So at least in this example...

The maximum likelihood estimator $\widehat{\theta}$ is a consistent estimator of θ_o, the minimizer the KL divergence from the true distribution p_o to the Poisson(θ) model $f(y|\theta)$.

Maximum Likelihood Estimation Under Mis-specification

Note: expectations and variances are calculated using p_o

Theorem

Suppose that $\mathbf{y}_1, \ldots, \mathbf{y}_N \sim$ iid p_o and let $\widehat{\boldsymbol{\theta}}$ denote the MLE for $\boldsymbol{\theta}$ under the possibly mis-specified model $f(\mathbf{y}|\boldsymbol{\theta})$. Then, under mild regularity conditions:

(i) $\widehat{\boldsymbol{\theta}}$ is consistent for the pseudo-true parameter value $\boldsymbol{\theta}_o$, defined as the minimizer of $KL(p_o, f_{\boldsymbol{\theta}})$ over the parameter space Θ.

(ii) $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \rightarrow_d \mathcal{N}(\mathbf{0}, \mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1})$

where we define $\mathbf{J} \equiv -\mathbb{E}\left[\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]$ and $\mathbf{K} \equiv \text{Var}\left[\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial\boldsymbol{\theta}}\right]$.

Why is this result such a big deal?

1. Provides an interpretation of MLE when we acknowledge that our models are only an *approximation* or reality: MLE recovers the pseudo-true parameter $\boldsymbol{\theta}_o$.
2. Yields a formula for standard errors that is robust to mis-specification of our model: compare to Heteroskedasticity consistent SEs for regression.
3. If the model is correctly specified, we recover the "classical" MLE result.

A Consistent Asymptotic Variance Matrix Estimator: $\widehat{\mathbf{J}}^{-1}\widehat{\mathbf{K}}\widehat{\mathbf{J}}^{-1}$

$\widehat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}_o$ plus Uniform Weak Law of Large Numbers: Newey & McFadden (1994)

$$\boldsymbol{\theta}_o \equiv \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\ \mathbb{E}\left[\log f(\mathbf{y}_i|\boldsymbol{\theta})\right] \qquad \widehat{\theta} \equiv \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\ \frac{1}{N}\sum_{i=1}^{N}\log f(\mathbf{y}|\boldsymbol{\theta})$$

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \to_d \mathcal{N}(\mathbf{0}, \mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1}) \qquad \widehat{\boldsymbol{\theta}} \approx \mathcal{N}(\boldsymbol{\theta}_o, \widehat{\mathbf{J}}^{-1}\widehat{\mathbf{K}}\widehat{\mathbf{J}}^{-1}/N)$$

$$\mathbf{J} \equiv -\mathbb{E}\left[\frac{\partial^2 \log f(\mathbf{y}_i|\boldsymbol{\theta}_o)}{\partial^2\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] \qquad \widehat{\mathbf{J}} \equiv -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 \log f(\mathbf{y}_i|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$$

$$\mathbf{K} \equiv \mathrm{Var}\left[\frac{\partial \log f(\mathbf{y}_i|\boldsymbol{\theta}_o)}{\partial\boldsymbol{\theta}}\right] \qquad \widehat{\mathbf{K}} \equiv \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial \log f(\mathbf{y}_i|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}}\right]\left[\frac{\partial \log f(\mathbf{y}_i|\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}}\right]'$$

Some Notes on the Preceding Slide

What happened to the KL divergence?

$\mathbb{E}[\log p_o(\mathbf{y})]$ does not involve $\boldsymbol{\theta}$. Hence, $\underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\ \mathbb{E}\left[\log f(\mathbf{y}_i|\boldsymbol{\theta})\right] = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\min}\ KL(p_o, f_{\boldsymbol{\theta}})$.

Isn't $\widehat{\mathbf{K}}$ missing a term?

The sample variance of $\mathbf{x}$ is given by $\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right) - (\bar{\mathbf{x}}\bar{\mathbf{x}}')$ where $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$. In our formula for $\widehat{\mathbf{K}}$, the "$\bar{\mathbf{x}}\bar{\mathbf{x}}'$" term appears to be missing, but it is in fact equal to zero, since $\widehat{\boldsymbol{\theta}}$ is the solution to the MLE first-order condition.

Some Terminology

I will call $\widehat{\mathbf{J}}^{-1}\widehat{\mathbf{K}}\widehat{\mathbf{J}}^{-1}$ the robust asymptotic variance matrix estimator, since it is correct regardless of whether the model is correctly specified.

Maximum Likelihood Estimation Under Correct Specification

“Classical” large-sample theory for MLE

Theorem

Suppose that $\mathbf{y}_1, \ldots, \mathbf{y}_N \sim$ iid $f(\mathbf{y}|\boldsymbol{\theta}_o)$. Then, under mild regularity conditions:

(i) $\boldsymbol{\theta}_o$ is consistent for $\boldsymbol{\theta}_o$.

(ii) $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \rightarrow_d \mathcal{N}(\mathbf{0}, \mathbf{J}^{-1})$ where $\mathbf{J} \equiv -\mathbb{E}\left[\dfrac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial^2\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]$.

Why? If $p_o(\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta}_o)$, then:

1. $KL(p_o; f_{\boldsymbol{\theta}})$ equals *zero* at $\boldsymbol{\theta} = \boldsymbol{\theta}_o$.
2. The *information matrix equality* gives $\mathbf{K} = \mathbf{J}$ which implies $\mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1} = \mathbf{J}^{-1}$.

The Information Matrix Equality: if $p_o(\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta}_o)$, then $\mathbf{K} = \mathbf{J}$.

$$\boxed{\mathbf{J} \equiv -\mathbb{E}\left[\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial^2 \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right], \quad \mathbf{K} \equiv \mathrm{Var}\left[\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}}\right]}$$

Step 1: Alternative Expression for $\mathbf{K}$

$$\mathrm{Var}\left[\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}}\right] = \mathbb{E}\left[\left\{\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}}\right\}\left\{\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}}\right\}'\right] - \mathbb{E}\left[\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}}\right]\mathbb{E}\left[\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}}\right]'$$

but since $\boldsymbol{\theta}_o$ minimizes $\mathbb{E}\left[\log f(\mathbf{y}|\boldsymbol{\theta})\right]$,

$$\mathbb{E}\left[\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}}\right] = \frac{\partial}{\partial \boldsymbol{\theta}}\mathbb{E}\left[\log f(\mathbf{y}|\boldsymbol{\theta}_o)\right] = \mathbf{0}$$

so it suffices to show that

$$-\mathbb{E}\left[\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial^2 \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] = \mathbb{E}\left[\left\{\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}}\right\}\left\{\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}}\right\}'\right]$$

The Information Matrix Equality: if $p_o(\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta}_o)$, then $\mathbf{K} = \mathbf{J}$.

$$\boxed{\text{suffices to show} \quad -\mathbb{E}\left[\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = \mathbb{E}\left[\left\{\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial\boldsymbol{\theta}}\right\}\left\{\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial\boldsymbol{\theta}}\right\}'\right]}$$

Step 2: Chain Rule & Product Rule

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\mathbf{y}|\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_i}\left[\frac{\partial}{\partial\theta_j}\log f(\mathbf{y}|\boldsymbol{\theta})\right] = \frac{\partial}{\partial\theta_i}\left[\frac{1}{f(\mathbf{y}|\boldsymbol{\theta})}\cdot\frac{\partial}{\partial\theta_j}f(\mathbf{y}|\boldsymbol{\theta})\right]$$

$$= \left[-\frac{1}{f^2(\mathbf{y}|\boldsymbol{\theta})}\cdot\frac{\partial}{\partial\theta_i}f(\mathbf{y}|\boldsymbol{\theta})\right]\left[\frac{\partial}{\partial\theta_j}f(\mathbf{y}|\boldsymbol{\theta})\right] + \frac{1}{f(\mathbf{y}|\boldsymbol{\theta})}\cdot\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(\mathbf{y}|\boldsymbol{\theta})$$

$$= -\left[\frac{1}{f(\mathbf{y}|\boldsymbol{\theta})}\cdot\frac{\partial}{\partial\theta_i}f(\mathbf{y}|\boldsymbol{\theta})\right]\left[\frac{1}{f(\mathbf{y}|\boldsymbol{\theta})}\cdot\frac{\partial}{\partial\theta_j}f(\mathbf{y}|\boldsymbol{\theta})\right] + \frac{1}{f(\mathbf{y}|\boldsymbol{\theta})}\cdot\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(\mathbf{y}|\boldsymbol{\theta})$$

$$= -\frac{\partial}{\partial\theta_i}\log f(\mathbf{y}|\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\log f(\mathbf{y}|\boldsymbol{\theta}) + \frac{1}{f(\mathbf{y}|\boldsymbol{\theta})}\cdot\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(\mathbf{y}|\boldsymbol{\theta})$$

The Information Matrix Equality: if $p_o(\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta}_o)$, then $\mathbf{K} = \mathbf{J}$.

$$\boxed{\text{suffices to show } \; -\mathbb{E}\left[\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = \mathbb{E}\left[\left\{\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial\boldsymbol{\theta}}\right\}\left\{\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}_o)}{\partial\boldsymbol{\theta}}\right\}'\right]}$$

Step 3: Multiply by -1, Evaluate at $\boldsymbol{\theta}_o$, and Take Expectations

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\mathbf{y}|\boldsymbol{\theta}) = -\frac{\partial}{\partial\theta_i}\log f(\mathbf{y}|\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\log f(\mathbf{y}|\boldsymbol{\theta}) + \frac{1}{f(\mathbf{y}|\boldsymbol{\theta})}\cdot\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(\mathbf{y}|\boldsymbol{\theta})$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\mathbf{y}|\boldsymbol{\theta}_o)\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta_i}\log f(\mathbf{y}|\boldsymbol{\theta}_o)\frac{\partial}{\partial\theta_j}\log f(\mathbf{y}|\boldsymbol{\theta}_o)\right] - \underbrace{\mathbb{E}\left[\frac{1}{f(\mathbf{y}|\boldsymbol{\theta}_o)}\cdot\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(\mathbf{y}|\boldsymbol{\theta}_o)\right]}_{\text{suffices to show this is zero!}}$$

The Information Matrix Equality: if $p_o(\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta}_o)$, then $\mathbf{K} = \mathbf{J}$.

$$\boxed{\text{suffices to show } \mathbb{E}\left[\frac{1}{f(\mathbf{y}|\boldsymbol{\theta}_o)} \cdot \frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\mathbf{y}|\boldsymbol{\theta}_o)\right] = 0}$$

Step 4: Use $p_o(\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta}_o)$

$$\mathbb{E}\left[\frac{1}{f(\mathbf{y}|\boldsymbol{\theta}_o)} \cdot \frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\mathbf{y}|\boldsymbol{\theta}_o)\right] \equiv \int \left[\frac{1}{f(\mathbf{y}|\boldsymbol{\theta}_o)} \cdot \frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\mathbf{y}|\boldsymbol{\theta}_o)\right] p_o(\mathbf{y})\, d\mathbf{y}$$

$$= \int \left[\frac{1}{\color{red}{f(\mathbf{y}|\boldsymbol{\theta}_o)}} \cdot \frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\mathbf{y}|\boldsymbol{\theta}_o)\right] \color{red}{f(\mathbf{y}|\boldsymbol{\theta}_o)}\, d\mathbf{y} = \int \frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\mathbf{y}|\boldsymbol{\theta}_o)\, d\mathbf{y}$$

$$= \frac{\partial^2}{\partial\theta_i\partial\theta_j} \int f(\mathbf{y}|\boldsymbol{\theta}_o)\, d\mathbf{y} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}(1) = 0$$

A Simple Example Continued Again: Asymptotic Variance Calculations

Poisson(θ) model, possibly mis-specified.

Ingredients

$$\boxed{\begin{aligned} \log f(y|\theta) &= -\theta + y\log(\theta) - \log(y!) \\ \frac{d}{d\theta}\log f(y|\theta) &= -1 + y/\theta \\ \frac{d^2}{d\theta^2}\log f(y|\theta) &= -y/\theta^2 \\ \theta_o = \mathbb{E}[y], &\quad \widehat{\theta} = \bar{y} \end{aligned}}$$

$$J = -\mathbb{E}\left[\frac{d^2}{d\theta^2}\log f(y|\theta_o)\right] = 1/\mathbb{E}[y]$$

$$\widehat{J} = -\frac{1}{N}\sum_{i=1}^{N}\frac{d^2}{d\theta^2}\log f(y_i|\widehat{\theta}) = 1/\bar{y}$$

$$K = \mathrm{Var}\left[\frac{d}{d\theta}\log f(y|\theta_o)\right] = \mathrm{Var}(y)/\mathbb{E}[y]^2$$

$$\widehat{K} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{d}{d\theta}\log f(y_i|\widehat{\theta})\right]^2 = s_y^2/(\bar{y})^2$$

where $s_y^2 \equiv \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})$ and $\bar{y} \equiv \frac{1}{N}\sum_{i=1}^{n} y_i$

A Simple Example Continued Again: Asymptotic Variance Calculations

From Previous Slide

$$\theta_0 = \mathbb{E}[y], \quad J = 1/\mathbb{E}[y], \quad \widehat{J} = 1/\bar{y}, \quad K = \mathrm{Var}(y)/\mathbb{E}[y]^2, \quad \widehat{K} = s_y^2/(\bar{y})^2$$

Correct Specification

$$\boxed{y_1, \ldots, y_N \sim \text{ iid Poisson}(\theta_o)} \implies \boxed{J = K = 1/\theta_o} \implies \boxed{J^{-1}KJ^{-1} = \theta_o = \mathbb{E}[y]}$$

Potential Mis-specification

$$\boxed{y_1, \ldots, y_N \sim \text{ iid}} \implies \boxed{J = 1/\mathbb{E}[y], \quad K = \mathrm{Var}(y)/\mathbb{E}[y]^2} \implies \boxed{J^{-1}KJ^{-1} = \mathrm{Var}(y)}$$

A Simple Example Continued Again: Asymptotic Variance Calculations

Comparison of Asymptotic Distributions

$$\boxed{y_1, \ldots, y_N \sim \text{ iid Poisson}(\theta_o)} \implies \sqrt{N}(\widehat{\theta} - \theta_o) = \sqrt{N}(\bar{y} - \mathbb{E}[y]) \to_d \mathcal{N}(0, \mathbb{E}[y])$$

$$\boxed{y_1, \ldots, y_N \sim \text{ iid}} \implies \sqrt{N}(\widehat{\theta} - \theta_o) = \sqrt{N}(\bar{y} - \mathbb{E}[y]) \to_d \mathcal{N}(0, \text{Var}[y])$$

Comparison of Asymptotic 95% CIs

$$\boxed{y_1, \ldots, y_N \sim \text{ iid Poisson}(\theta_o)} \implies \bar{y} \pm 1.96 \times \sqrt{\bar{y}/N}$$

$$\boxed{y_1, \ldots, y_N \sim \text{ iid}} \implies \bar{y} \pm 1.96 \times s_y/\sqrt{N}$$

Punch Line

Unless $\text{Var}(y) = \mathbb{E}[y]$, CIs/tests that assume the Poisson model is true are wrong!