# MPhil Econometrics – Limited Dependent Variables and Selection


Francis J. DiTraglia

University of Oxford



Complied on 2020-01-30 at 12:34:30

# Housekeeping

| | |
|---|---|
| **Lecturer:** | Francis J. DiTraglia |
| **Email:** | francis.ditraglia@ox.ac.uk |
| **Course Materials:** | `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 $\theta$ equals the mean of $y$.

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

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

You will show this on the problem set.

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

### 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 $\varphi$ is convex then $\varphi(\mathbb{E}[y]) \leq \mathbb{E}[\varphi(y)]$, with equality iff $\varphi$ 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 $f_\theta$

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

## 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$(\theta)$; True Dist. $p_o(0) = p_o(2) = \frac{2}{5}$ and $p_o(1) = \frac{1}{5}$

## Best Approximation

What parameter value $\theta_o$ makes the Poisson$(\theta)$ 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($\theta$); True Dist. $p_o(0) = p_o(2) = \frac{2}{5}$ and $p_o(1) = \frac{1}{5}$

## Best Approximation

What parameter value $\theta_o$ makes the Poisson($\theta$) 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($\theta$) model for $y_1, \ldots, y_N$, the MLE for $\theta$ 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 $\theta_o$, the minimizer the KL divergence from the true distribution $p_o$ to the Poisson($\theta$) 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 $\Theta$.

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

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

# 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) $\widehat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_o$.

(ii) $\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \to_d \mathcal{N}(\mathbf{0}, \mathbf{J}^{-1})$ 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 $p_o(\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta}_o)$

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

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

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

# A Simple Example Continued Again: Asymptotic Variance Calculations

Poisson($\theta$) model, possibly mis-specified.

## Ingredients

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

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

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

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

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

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

# Lecture #2 – Poisson Regression

Review: Minimum MSE Predictor / Minimum MSE Linear Predictor

Why not just use OLS?

Conditional Maximum Likelihood Estimation

Poisson Regression: A Robust Model for Count Data

Asymptotic Variance Calculations for Poisson Regression

# How to predict a count variable?

## Example

Suppose we want to predict $y$ using $\mathbf{x}$, where:

- $y \equiv \#$ of children a woman has: a count variable, i.e. $y \in \{0, 1, 2, \ldots\}$
- $\mathbf{x} \equiv$\{years of schooling, age, married, etc.\}

## Minimum MSE Predictor

$\mu(\mathbf{x}) \equiv \mathbb{E}(y|\mathbf{x})$ minimizes $\mathbb{E}\left[\{y - \varphi(\mathbf{x})\}^2\right]$ over all possible predictors $\varphi(\cdot)$.

## Minimum MSE Linear Predictor

$\boldsymbol{\beta} \equiv \mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]^{-1} \mathbb{E}[\mathbf{x}y]$ minimizes $\mathbb{E}\left[(y - \mathbf{x}'\boldsymbol{\theta})^2\right]$ over all linear predictors $\mathbf{x}'\boldsymbol{\theta}$.

# Proof: $\mathbb{E}(y|\mathbf{x})$ is the minimum MSE predictor

## Step 1: add and subtract $\mu(\mathbf{x}) \equiv \mathbb{E}(y|\mathbf{x})$

$$\begin{aligned}
\mathbb{E}\left[\{y - \varphi(\mathbf{x})\}^2\right] &= \mathbb{E}\left[\left\{\left(y - \mu(\mathbf{x})\right) - \left(\varphi(\mathbf{x}) - \mu(\mathbf{x})\right)\right\}^2\right] \\
&= \mathbb{E}\left[\{y - \mu(\mathbf{x})\}^2\right] - 2\mathbb{E}\left[\{y - \mu(\mathbf{x})\}\{\varphi(\mathbf{x}) - \mu(\mathbf{x})\}\right] + \mathbb{E}\left[\{\varphi(\mathbf{x}) - \mu(\mathbf{x})\}^2\right]
\end{aligned}$$

## Step 2: iterated expectations

$$\begin{aligned}
\mathbb{E}\left[\{y - \mu(\mathbf{x})\}\{\varphi(\mathbf{x}) - \mu(\mathbf{x})\}\right] &= \mathbb{E}\Big(\mathbb{E}\left[\{y - \mu(\mathbf{x})\}\{\varphi(\mathbf{x}) - \mu(\mathbf{x})\}|\mathbf{x}\right]\Big) \\
&= \mathbb{E}\Big(\left[\varphi(\mathbf{x}) - \mu(\mathbf{x})\right]\left[\mathbb{E}(y|\mathbf{x}) - \mu(\mathbf{x})\right]\Big) = 0
\end{aligned}$$

## Step 3: combine steps 1 & 2

$$\mathbb{E}\left[\{y - \varphi(\mathbf{x})\}^2\right] = \underbrace{\mathbb{E}\left[\{y - \mu(\mathbf{x})\}^2\right]}_{\text{constant wrt } \varphi} + \underbrace{\mathbb{E}\left[\{\varphi(\mathbf{x}) - \mu(\mathbf{x})\}^2\right]}_{\text{cannot be negative; zero if } \varphi=\mu}$$

# Proof: OLS is the Minimum MSE Linear Predictor

## Objective Function

$$\mathbb{E}\left[\left(y - \mathbf{x}'\boldsymbol{\theta}\right)^2\right] = \mathbb{E}[y^2] - 2\mathbb{E}[y\mathbf{x}']\boldsymbol{\theta} + \boldsymbol{\theta}'\mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]\boldsymbol{\theta}$$

## Recall: Matrix Differentiation

$$\frac{\partial(\mathbf{a}'\mathbf{z})}{\partial\mathbf{z}} = \mathbf{a}, \quad \frac{\partial(\mathbf{z}'\mathbf{A}\mathbf{z})}{\partial\mathbf{z}} = (\mathbf{A} + \mathbf{A}')\mathbf{z}$$

## First-Order Condition

$$-2\mathbb{E}\left[\mathbf{x}'y\right] + 2\mathbb{E}[\mathbf{x}\mathbf{x}']\boldsymbol{\beta} = 0 \implies \boldsymbol{\beta} = \mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]^{-1}\mathbb{E}\left[\mathbf{x}'y\right]$$

# Problems with linear-in-parameters models for count data

Best predictor is $\mathbb{E}(y|\mathbf{x})$ but how can we estimate this?

Plain-vanilla OLS?

- If $\mathbb{E}(y|\mathbf{x}) \approx \mathbf{x}'\boldsymbol{\beta}$, OLS is a reasonable approach.
- Problem: $y$ is a count so it *can't* be negative, but OLS prediction $\mathbf{x}'\boldsymbol{\beta}$ could be.

OLS for $\log(y)$?

- Log-linear model $\log(y) = \mathbf{x}'\beta + \varepsilon$
- Solves the problem of negative predictions: $\log(y)$ *can* be negative.
- Problem: if $y$ is a count it could equal zero but $\log(0) = -\infty$!

# A realistic model for count data *must* be nonlinear in parameters.

## General Approach

- Assume that $\mathbb{E}(y|\mathbf{x}) = m(\mathbf{x};\boldsymbol{\beta})$ where $m$ is a known parametric function.
- Choose $m$ so that it is always positive, regardless of $\mathbf{x}$ and $\boldsymbol{\beta}$.
- This means $m$ *cannot* be linear.

## This Lecture: $m(\mathbf{x};\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$

- Always strictly positive
- Common choice in practice
- Everything I'll discuss works with other choices of $m$, making appropriate changes.

# How to estimate $\beta_o$?

Assumption: $\mathbb{E}(y|\mathbf{x}) = \exp\left(\mathbf{x}'\boldsymbol{\beta}_o\right)$

Using our argument from above, $\boldsymbol{\beta}_o$ minimizes $\mathbb{E}\left[\{y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})\}^2\right]$ over all $\boldsymbol{\beta}$.

Nonlinear Least Squares (NLLS)

$\widehat{\beta}_{NLLS}$ is the minimizer of $\sum_{i=1}^N \{y_i - \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\}^2$

Poisson Regression (MLE)

$\widehat{\beta}_{MLE}$ is the MLE for $\beta_o$ under the model $y_i|\mathbf{x}_i \sim$ indep. Poisson$\big(\exp(\mathbf{x}_i'\boldsymbol{\beta}_o)\big)$

# Conditional versus Unconditional MLE

## Last Lecture: Unconditional MLE

Model *unconditional* dist. of a random vector $\mathbf{y}$: $f(\mathbf{y}|\boldsymbol{\theta})$.

## This Lecture: Conditional MLE

Model *conditional* dist. of a random variable $y$ *given* a random vector $\mathbf{x}$: $f(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})$.

## Why Conditional MLE?

- Unconditional MLE requires joint distribution: $f(y,\mathbf{x}|\boldsymbol{\theta}) = f(y|\mathbf{x},\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})$
- $\mathbb{E}(y|\mathbf{x})$ only depends on $f(y|\mathbf{x},\boldsymbol{\theta})$ not $f(\mathbf{x}|\boldsymbol{\theta})$.
- Not interested in $f(\mathbf{x}|\boldsymbol{\theta})$; coming up with a good model for it is challenging.
- Caveat: unconditional MLE is more efficient provided the model for $\mathbf{x}$ is correct.

# The Conditional Maximum Likelihood Estimator

Assuming iid data.

Sample

$$\boldsymbol{\theta}_o \equiv \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{N} \sum_{i=1}^{N} \log f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$$

Population

$$\boldsymbol{\theta}_o \equiv \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}\left[\log f(y_i|\mathbf{x}_i, \boldsymbol{\theta})\right]$$

Important

- We only model the conditional distribution $y|\mathbf{x}$, but...
- ...the expectation $\mathbb{E}[\log f(y_i|\mathbf{x}_i, \boldsymbol{\theta})]$ is taken over the *joint distribution* of $(y, \mathbf{x})$.
- $f(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ is merely a *function* of the RVs $(y_i, \mathbf{x}_i)$.

# Poisson Regression as a Conditional MLE

Model: $y_i|\mathbf{x}_i \sim \text{Poisson}\left(\exp\{\mathbf{x}_i'\beta\}\right)$

$$\ell_i(\boldsymbol{\beta}) \equiv \log f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = y_i\mathbf{x}_i'\boldsymbol{\beta} - \exp(\mathbf{x}_i'\boldsymbol{\beta}) - \log(y_i!)$$

$$\underbrace{\mathbf{s}_i(\boldsymbol{\beta})}_{\text{score vector}} \equiv \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{x}_i\left[y_i - \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right]$$

$$\widehat{\boldsymbol{\beta}} \text{ solves } \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \underbrace{\left[y_i - \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right]}_{\text{residual: } u_i} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i u_i(\boldsymbol{\beta}) = \mathbf{0}$$

# Average Partial Effects

## Partial Effects

For continuous $x_j$, we call $\frac{\partial}{\partial x_j}\mathbb{E}(y|\mathbf{x})$ the partial effect of $x_j$. For discrete $x_j$ the partial effect is the difference of $\mathbb{E}(y|\mathbf{x})$ at two different values of $x_j$

## Average Partial Effects (APE)

In nonlinear models, partial effects typically vary with $\mathbf{x}$. The average partial effect is the expectation of the partial effect over the distribution of $\mathbf{x}$.

# Average Partial Effects for Poisson Regression

## Partial Effect

$$\frac{\partial}{\partial x_j}\mathbb{E}(y|\mathbf{x}) = \frac{\partial}{\partial x_j}\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) = \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\beta_j$$

## Average Partial Effect

$$\mathbb{E}\left[\frac{\partial}{\partial x_j}\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right] = \mathbb{E}\left[\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right]\beta_j$$

## Estimated Partial Effect

$$\exp\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)\widehat{\beta}_j$$

## Estimated Average Partial Effect

$$\left[\frac{1}{N}\sum_{i=1}^{N}\exp\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)\right]\widehat{\beta}_j$$

## Relative Effects

The *ratio* of partial effects does not depend on $\mathbf{x}$: relative effects are constant.

## Problem Set

Poisson regression: APE$=\bar{y}\widehat{\beta}_j$. Multiply by $\bar{y}$ to put coefficients on the scale of OLS.

# Conditional MLE Under Mis-specification

Basically identical to the unconditional version.

**Theorem**

Suppose that $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N \sim$ iid $p_o$ and let $\widehat{\boldsymbol{\theta}}$ denote the Conditional MLE for $\boldsymbol{\theta}$ under the possibly mis-specified model $f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$. Then, under mild regularity conditions:

(i) $\widehat{\boldsymbol{\theta}}$ is consistent for the pseudo-true parameter value $\boldsymbol{\theta}_o$, defined as the *maximizer* of the expected log likelihood $\mathbb{E}[\log f(y|\mathbf{x}, \boldsymbol{\theta})]$ over the parameter space $\Theta$.

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

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

# Conditional MLE Under Correct Specification

Basically identical to the unconditional version.

**Theorem**

Suppose that $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N \sim$ iid where the conditional distribution of $y_i|\mathbf{x}_i$ is given by $f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_o)$. Then, under mild regularity conditions,

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

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

# What value of $\boldsymbol{\beta}$ maximizes $\mathbb{E}\left[\ell_i(\boldsymbol{\beta})\right]$?

## Iterated Expectations

$$\mathbb{E}[\ell_i(\boldsymbol{\beta})] = \mathbb{E}\left\{\mathbb{E}\left[\ell_i(\boldsymbol{\beta})|\mathbf{x}_i\right]\right\} = \mathbb{E}\left\{\mathbb{E}\left[y_i \mathbf{x}_i'\boldsymbol{\beta} - \exp(\mathbf{x}_i'\boldsymbol{\beta}) - \log\left(y_i!\right)|\mathbf{x}_i\right]\right\}$$

## Simplify Inner Expectation

$$\mathbb{E}\left[\ell_i(\boldsymbol{\beta})|\mathbf{x}_i\right] = \mathbf{x}_i'\boldsymbol{\beta}\mathbb{E}\left[y_i|\mathbf{x}_i\right] - \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) - \underbrace{\mathbb{E}\left[\log\left(y_i!\right)|\mathbf{x}_i\right]}_{\text{constant wrt } \mathbf{x}_i}$$

## FOC for Inner Expectation

$$\frac{\partial}{\partial\boldsymbol{\beta}}\mathbb{E}\left[\ell_i(\boldsymbol{\beta})|\mathbf{x}_i\right] = \left\{\mathbb{E}\left[y_i|\mathbf{x}_i\right] - \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right\}\mathbf{x}_i = \mathbf{0}$$

# What value of $\beta$ maximizes $\mathbb{E}\left[\ell_i(\boldsymbol{\beta})\right]$?

$$\boxed{\frac{\partial}{\partial \boldsymbol{\beta}}\mathbb{E}\left[\ell_i(\boldsymbol{\beta})|\mathbf{x}_i\right] = \left\{\mathbb{E}\left[y_i|\mathbf{x}_i\right] - \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right\}\mathbf{x}_i = \mathbf{0}}$$

## What does this mean?

Since $\mathbb{E}\left[y_i|\mathbf{x}_i\right] = \exp\left(\mathbf{x}_i'\boldsymbol{\beta}_o\right)$, setting $\boldsymbol{\beta} = \boldsymbol{\beta}_o$ solves the FOC for the inner expectation!

## In other words:

For any realization of $\mathbf{x}_i$ and any $\boldsymbol{\beta}$,

$$\mathbb{E}[\ell_i(\boldsymbol{\beta})|\mathbf{x}_i] \leq \mathbb{E}[\ell_i(\boldsymbol{\beta}_o)|\mathbf{x}_i]$$

so taking expectations of both sides:

$$\mathbb{E}\left[\ell_i(\boldsymbol{\beta})\right] = \mathbb{E}\left\{\mathbb{E}[\ell_i(\boldsymbol{\beta})|\mathbf{x}_i]\right\} \leq \mathbb{E}\left\{\mathbb{E}[\ell_i(\boldsymbol{\beta}_o)|\mathbf{x}_i]\right\} = \mathbb{E}\left[\ell_i(\boldsymbol{\beta}_o)\right]$$

# Poisson Regression is consistent if $\mathbb{E}(y|\mathbf{x})$ is correctly specified.

We showed this for a particular choice of $m(\mathbf{x}; \boldsymbol{\beta})$ but the result is general.

### Result

Provided that we have correctly specified $\mathbb{E}(y_i|\mathbf{x}_i)$, it *doesn't matter* if $y_i|\mathbf{x}_i$ actually follows a Poisson distribution: Poisson regression is *still consistent* for $\boldsymbol{\beta}_o$.

### Compare

This is very similar to our result for the Poisson($\theta$) model from last lecture.

### Caveat

Strictly speaking we need to show that $\boldsymbol{\beta}_o$ is the *unique* maximizer of the expected log likelihood. *Multiple solutions* if $\mathbf{x}_i$ perfectly co-linear (compare to OLS regression).

# Asymptotic Variance Calculations for Poisson Regression

$$\underbrace{\mathbf{s}_i(\boldsymbol{\beta})}_{\text{score vector}} \equiv \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{x}_i \left[ y_i - \exp\left(\mathbf{x}_i' \boldsymbol{\beta}\right)\right] = \mathbf{x}_i u_i(\boldsymbol{\beta})$$

$$\underbrace{\mathbf{H}_i(\boldsymbol{\beta})}_{\text{Hessian matrix}} \equiv \frac{\partial \mathbf{s}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = -\exp\left(\mathbf{x}_i' \boldsymbol{\beta}\right) \mathbf{x}_i \mathbf{x}_i'$$

$$\mathbf{J} \equiv -\mathbb{E}\left[\mathbf{H}_i(\boldsymbol{\beta}_o)\right] = \mathbb{E}\left[\exp\left(\mathbf{x}_i' \boldsymbol{\beta}_o\right) \mathbf{x}_i \mathbf{x}_i'\right]$$

$$\mathbf{K} \equiv \mathrm{Var}\left[\mathbf{s}_i(\boldsymbol{\beta}_o)\right] = \mathbb{E}\left[\mathbf{s}_i(\boldsymbol{\beta}_o)\mathbf{s}_i(\boldsymbol{\beta}_o)'\right] = \mathbb{E}\left[u_i^2(\boldsymbol{\beta}_o)\mathbf{x}_i \mathbf{x}_i'\right]$$

# Asymptotic Variance Calculations for Poisson Regression

$$\boxed{\mathbf{J} = \mathbb{E}\left[\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_o\right)\mathbf{x}_i\mathbf{x}_i'\right], \quad \mathbf{K} = \mathbb{E}\left[u_i^2(\boldsymbol{\beta}_o)\mathbf{x}_i\mathbf{x}_i'\right]}$$

## Notice

**J** does not depend on $y$ but **K** does:

$$\mathbf{K} = \mathbb{E}\left[u_i^2(\boldsymbol{\beta}_o)\mathbf{x}_i\mathbf{x}_i'\right] = \mathbb{E}\left\{\mathbb{E}\left[u_i^2(\boldsymbol{\beta}_o)|\mathbf{x}_i\right]\mathbf{x}_i\mathbf{x}_i'\right\} = \mathbb{E}\left(\mathbb{E}\left[\left\{y_i - \mathbb{E}(y_i|\mathbf{x}_i)\right\}^2|\mathbf{x}_i\right]\right)$$

$$= \mathbb{E}\left[\mathrm{Var}(y_i|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'\right]$$

Assumptions about $\mathrm{Var}(y|\mathbf{x})$ affect the asymptotic variance through **K**.

# Possible Assumptions for Var($y|\mathbf{x}$): Strongest to Weakest

1. Poisson Assumption: $\text{Var}(y|\mathbf{x}) = \mathbb{E}(y|\mathbf{x})$
   - holds if Poisson model is correct.
2. Quasi-Poisson Assumption: $\text{Var}(y|\mathbf{x}) = \sigma^2\mathbb{E}(y|\mathbf{x})$
   - Allows for possibility that $y|\mathbf{x}$ is *not* Poisson
   - Overdispersion: $\sigma^2 > 1 \implies \text{Var}(y|\mathbf{x}) > \mathbb{E}(y|\mathbf{x})$
   - Underdispersion $\sigma^2 > 1 \implies \text{Var}(y|\mathbf{x}) < \mathbb{E}(y|\mathbf{x})$
   - If $\sigma^2 = 1$ we're back to the Poisson Assumption.
3. No Assumption: $\text{Var}(y|\mathbf{x})$ unspecified

# Asymptotic Variance Under Poisson Assumption

$$\boxed{\mathbf{J} = \mathbb{E}\left[\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_o\right)\mathbf{x}_i\mathbf{x}_i'\right], \quad \mathbf{K} = \mathbb{E}\left[\text{Var}(y_i|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'\right]}$$

Assumption: $\text{Var}(y_i|\mathbf{x}_i) = \mathbb{E}(y_i|\mathbf{x}_i) = \exp\left(\mathbf{x}_i'\boldsymbol{\beta}_o\right)$

- Implies $\mathbf{K} = \mathbb{E}\left[\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_o\right)\mathbf{x}_i\mathbf{x}_i'\right]$
- Hence $\mathbf{K} = \mathbf{J}$ (Information Matrix Equality)
- Therefore: $\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) \to_d \mathcal{N}(\mathbf{0}, \mathbf{J}^{-1})$
- Consistent Estimator: $\widehat{\mathbf{J}}^{-1} = \left[\frac{1}{N}\sum_{i=1}^{N}\exp\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)\mathbf{x}_i\mathbf{x}_i'\right]^{-1}$

# Asymptotic Variance Under Quasi-Poisson Assumption

$$\boxed{\mathbf{J} = \mathbb{E}\left[\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_o\right)\mathbf{x}_i\mathbf{x}_i'\right], \quad \mathbf{K} = \mathbb{E}\left[\text{Var}(y_i|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'\right]}$$

Assumption: $\text{Var}(y_i|\mathbf{x}_i) = \sigma^2\mathbb{E}(y_i|\mathbf{x}_i) = \sigma^2 \exp\left(\mathbf{x}_i'\boldsymbol{\beta}_o\right)$

- Implies $\mathbf{K} = \sigma^2\mathbb{E}\left[\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_o\right)\mathbf{x}_i\mathbf{x}_i'\right] = \sigma^2\mathbf{J}$
- Hence $\mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1} = \sigma^2\mathbf{J}^{-1}$
- Therefore: $\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) \rightarrow_d \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{J}^{-1})$
- Consistent estimator of $\mathbf{J}^{-1}$ on prev. slide but how can we estimate $\sigma^2$?

# How to estimate $\sigma^2$ under the Quasi-Poisson Assumption?

$$
\begin{aligned}
\text{Var}(y|\mathbf{x}) &= \sigma^2 \mathbb{E}(y|\mathbf{x}) \\
\sigma^2 &= \text{Var}(y|\mathbf{x}) / \mathbb{E}(y|\mathbf{x}) \\
\sigma^2 &= \mathbb{E}\left[ \{y - \mathbb{E}(y|\mathbf{x})\}^2 | \mathbf{x} \right] / \mathbb{E}(y|\mathbf{x}) \\
\sigma^2 &= \mathbb{E}\left[ \left. \frac{\{y - \mathbb{E}(y|\mathbf{x})\}^2}{\mathbb{E}(y|\mathbf{x})} \right| \mathbf{x} \right] \\
\sigma^2 &= \mathbb{E}\left[ \left. \frac{\{y - \exp(\mathbf{x}'\boldsymbol{\beta}_o)\}^2}{\exp(\mathbf{x}'\boldsymbol{\beta})} \right| \mathbf{x} \right] \\
\mathbb{E}[\sigma^2] &= \mathbb{E}\left( \mathbb{E}\left[ \left. \frac{\{y - \exp(\mathbf{x}'\boldsymbol{\beta}_o)\}^2}{\exp(\mathbf{x}'\boldsymbol{\beta})} \right| \mathbf{x} \right] \right) \\
\sigma^2 &= \mathbb{E}\left[ \frac{\{y - \exp(\mathbf{x}'\boldsymbol{\beta}_o)\}^2}{\exp(\mathbf{x}'\boldsymbol{\beta})} \right] \\
\sigma^2 &= \mathbb{E}\left[ u^2(\boldsymbol{\beta}_o) / \exp(\mathbf{x}'\boldsymbol{\beta}_o) \right]
\end{aligned}
$$

## Consistent Estimator of $\sigma^2$

$$
\widehat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \frac{[y_i - \exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})]^2}{\exp(\mathbf{x}_i\widehat{\boldsymbol{\beta}})} = \frac{1}{N} \sum_{i=1}^{N} \frac{\widehat{u}_i^2}{\exp(\mathbf{x}_i\widehat{\boldsymbol{\beta}})}
$$

# Robust Asymptotic Variance Matrix

$$\boxed{\mathbf{J} = \mathbb{E}\left[\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_o\right)\mathbf{x}_i\mathbf{x}_i'\right], \quad \mathbf{K} = \mathbb{E}\left[u_i^2(\boldsymbol{\beta}_o)\mathbf{x}_i\mathbf{x}_i'\right]}$$

## No Assumption on $\text{Var}(y_i|\mathbf{x}_i)$

- $\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) \to_d \mathcal{N}(\mathbf{0}, \mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1})$
- Consistent Estimator: $\widehat{\mathbf{J}}^{-1} = \left[\frac{1}{N}\sum_{i=1}^{N}\exp\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)\mathbf{x}_i\mathbf{x}_i'\right]^{-1}$
- Consistent Estimator: $\widehat{\mathbf{K}} = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - \exp(\mathbf{x}_i\widehat{\boldsymbol{\beta}})\right]^2\mathbf{x}_i\mathbf{x}_i' = \frac{1}{N}\sum_{i=1}^{N}\widehat{u}_i^2\mathbf{x}_i\mathbf{x}_i'$

# Why Poisson Regression rather than NLLS?

## Assume that $\mathbb{E}(y|\mathbf{x}) = \exp(\mathbf{x}'\beta_o)$

Both Poisson Reg. & NLLS are consistent if the conditional mean is correctly specified.

## Count data are typically heteroskedastic.

If $\text{Var}(y|\mathbf{x})$ varies with $\mathbf{x}$, NLLS will be relatively inefficient.

## Efficiency of Poisson Regression

- Correct model $\implies$ lowest variance among all estimators that leave the distribution of $\mathbf{x}$ unspecified.
- $\text{Var}(y|\mathbf{x}) = \sigma^2\mathbb{E}(y|\mathbf{x}) \implies$ Poisson regression is more efficient than NLLS and various other count data models.