

MPhil Econometrics – Limited Dependent Variables and Selection

Francis J. DiTraglia

University of Oxford

Compiled on 2020-12-16 at 11:58:30

Housekeeping

Lecturer: Francis J. DiTraglia
Email: francis.ditraglia@economics.ox.ac.uk
Course Materials: <https://economictricks.com>

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*.

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 .

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

You will show this on the problem set.

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

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

MLE for θ where $y_1, y_2, \dots, y_N \sim \text{iid Poisson}(\theta)$.

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

$$\hat{\theta} \equiv \arg \max_{\theta \in \Theta} \ell_N(\theta) = \bar{y}$$

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

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

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

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

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

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

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

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 $f(\mathbf{y}|\boldsymbol{\theta}) = 0$ & $p_o(\mathbf{y}) \neq 0$

$\text{KL}(p_o; f) \geq 0$ 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

$y_1, \dots, y_N \sim \text{iid } p_o$ where:

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

Mis-specified Model f_θ

$y_1, \dots, y_N \sim \text{iid Poisson}(\theta)$

KL Divergence

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

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

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

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

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

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

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

A more direct approach

Min KL \iff Max Expected Log-like.

$$\begin{aligned} \frac{d}{d\theta} \mathbb{E}[\log f(y|\theta)] &= \mathbb{E} \left[\frac{d}{d\theta} \{-\theta + y \log(\theta) - \log(y!)\} \right] \\ &= \mathbb{E}[-1 + y/\theta] = \mathbb{E}[y]/\theta - 1 = 0 \\ &\implies \boxed{\theta = \mathbb{E}[y]} \end{aligned}$$

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: $\theta_o = 1$

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 $\text{Poisson}(\theta)$ model for y_1, \dots, y_N , the MLE for θ is $\hat{\theta} = \bar{y}$
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 $y_1, \dots, y_N \sim \text{iid}$, 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 $\hat{\theta}$ is a consistent estimator of θ_o , the minimizer the KL divergence from the true distribution p_o to the $\text{Poisson}(\theta)$ model $f(y|\theta)$.

Maximum Likelihood Estimation Under Mis-specification

Note: expectations and variances are calculated using p_o

Theorem

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

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

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 θ_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: $\hat{\mathbf{J}}^{-1}\hat{\mathbf{K}}\hat{\mathbf{J}}^{-1}$

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

$$\boldsymbol{\theta}_o \equiv \arg \max_{\boldsymbol{\theta} \in \Theta} \mathbb{E} [\log f(\mathbf{y}_i | \boldsymbol{\theta})] \quad \hat{\boldsymbol{\theta}} \equiv \arg \max_{\boldsymbol{\theta} \in \Theta} \frac{1}{N} \sum_{i=1}^N \log f(\mathbf{y}_i | \boldsymbol{\theta})$$

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \rightarrow_d \mathcal{N}(\mathbf{0}, \mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1}) \quad \hat{\boldsymbol{\theta}} \approx \mathcal{N}(\boldsymbol{\theta}_o, \hat{\mathbf{J}}^{-1}\hat{\mathbf{K}}\hat{\mathbf{J}}^{-1}/N)$$

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

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

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, $\arg \max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}[\log f(\mathbf{y}_i|\boldsymbol{\theta})] = \arg \min_{\boldsymbol{\theta} \in \Theta} KL(p_o, f_{\boldsymbol{\theta}})$.

Isn't $\hat{\mathbf{K}}$ 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 $\hat{\mathbf{K}}$, the “ $\bar{\mathbf{x}} \bar{\mathbf{x}}'$ ” term appears to be missing, but it is in fact equal to zero, since $\hat{\boldsymbol{\theta}}$ is the solution to the MLE first-order condition.

Some Terminology

I will call $\hat{\mathbf{J}}^{-1} \hat{\mathbf{K}} \hat{\mathbf{J}}^{-1}$ 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 $\mathbf{y}_1, \dots, \mathbf{y}_N \sim \text{iid } f(\mathbf{y}|\boldsymbol{\theta}_o)$. Then, under mild regularity conditions:

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

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

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 $\mathbf{K} = \mathbf{J}$.

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

Step 1: Alternative Expression for \mathbf{K}

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

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

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

so it suffices to show that

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

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

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

Step 2: Chain Rule & Product Rule

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

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

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

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 $\mathbf{K} = \mathbf{J}$.

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

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

$$\begin{aligned} \mathbb{E} \left[\frac{1}{f(\mathbf{y}|\boldsymbol{\theta}_o)} \cdot \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(\mathbf{y}|\boldsymbol{\theta}_o) \right] &\equiv \int \left[\frac{1}{f(\mathbf{y}|\boldsymbol{\theta}_o)} \cdot \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(\mathbf{y}|\boldsymbol{\theta}_o) \right] p_o(\mathbf{y}) d\mathbf{y} \\ &= \int \left[\frac{1}{f(\mathbf{y}|\boldsymbol{\theta}_o)} \cdot \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(\mathbf{y}|\boldsymbol{\theta}_o) \right] f(\mathbf{y}|\boldsymbol{\theta}_o) d\mathbf{y} = \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(\mathbf{y}|\boldsymbol{\theta}_o) d\mathbf{y} \\ &= \frac{\partial^2}{\partial \theta_i \partial \theta_j} \int f(\mathbf{y}|\boldsymbol{\theta}_o) d\mathbf{y} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} (1) = 0 \end{aligned}$$

A Simple Example Continued Again: Asymptotic Variance Calculations

Poisson(θ) model, possibly mis-specified.

Ingredients

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

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

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

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

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

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

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

Correct Specification

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

Potential Mis-specification

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

A Simple Example Continued Again: Asymptotic Variance Calculations

Comparison of Asymptotic Distributions

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

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

Comparison of Asymptotic 95% CIs

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

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

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. $y \in \{0, 1, 2, \dots\}$
- ▶ $\mathbf{x} \equiv \{\text{years of schooling, age, married, etc.}\}$

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

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

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

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

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

Step 2: iterated expectations

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

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

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

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

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

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 $\mathbb{E}(y|\mathbf{x}) \approx \mathbf{x}'\beta$, OLS is a reasonable approach.
- ▶ **Problem:** y is a count so it *can't* be negative, but OLS prediction $\mathbf{x}'\beta$ 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 $\mathbb{E}(y|\mathbf{x}) = m(\mathbf{x}; \beta)$ where m is a known parametric function.
- ▶ Choose m so that it is always positive, regardless of \mathbf{x} and β .
- ▶ This means m *cannot* be linear.

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

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

How to estimate β_o ?

Assumption: $\mathbb{E}(y|\mathbf{x}) = \exp(\mathbf{x}'\beta_o)$

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

Nonlinear Least Squares (NLLS)

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

Poisson Regression (MLE)

$\hat{\beta}_{MLE}$ is the MLE for β_o under the model $y_i|\mathbf{x}_i \sim \text{indep. Poisson}(\exp(\mathbf{x}'_i\beta_o))$

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} : $f(y|\mathbf{x}, \boldsymbol{\theta})$.

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

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

Population

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

Important

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

Poisson Regression as a Conditional MLE

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

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

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

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

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

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

Estimated Partial Effect

$$\exp(\mathbf{x}'_i \hat{\boldsymbol{\beta}}) \hat{\beta}_j$$

Average Partial Effect

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

Estimated Average Partial Effect

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

Relative Effects

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

Problem Set

Poisson regression: $\text{APE} = \bar{y} \hat{\beta}_j$. 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 $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N \sim \text{iid } p_o$ and let $\hat{\boldsymbol{\theta}}$ 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) $\hat{\boldsymbol{\theta}}$ is consistent for the **pseudo-true** parameter value $\boldsymbol{\theta}_o$, defined as the *maximizer* of the expected log likelihood $\mathbb{E} [\log f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]$ over the parameter space Θ .
- (ii) $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \rightarrow_d \mathcal{N}(\mathbf{0}, \mathbf{J}^{-1} \mathbf{K} \mathbf{J}^{-1})$

where we define $\mathbf{J} \equiv -\mathbb{E} \left[\frac{\partial^2 \log f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]$ 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 $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N \sim \text{iid}$ 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) $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_o$

(ii) $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \rightarrow_d \mathcal{N}(\mathbf{0}, \mathbf{J}^{-1})$ 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 β maximizes $\mathbb{E} [\ell_i(\beta)]$?

Iterated Expectations

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

Simplify Inner Expectation

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

FOC for Inner Expectation

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

What value of β maximizes $\mathbb{E} [\ell_i(\beta)]$?

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

What does this mean?

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

In other words:

For any realization of \mathbf{x}_i and any β ,

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

so taking expectations of both sides:

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

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

We showed this for a particular choice of $m(\mathbf{x};\beta)$ 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 β_o .

Compare

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

Caveat

Strictly speaking we need to show that β_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

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

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

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

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

Asymptotic Variance Calculations for Poisson Regression

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

Notice

\mathbf{J} does not depend on y but \mathbf{K} does:

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

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

Possible Assumptions for $\text{Var}(y|\mathbf{x})$: 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

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

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

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

Asymptotic Variance Under Quasi-Poisson Assumption

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

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

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

How to estimate σ^2 under the Quasi-Poisson Assumption?

$$\text{Var}(y|\mathbf{x}) = \sigma^2 \mathbb{E}(y|\mathbf{x})$$

$$\sigma^2 = \text{Var}(y|\mathbf{x}) / \mathbb{E}(y|\mathbf{x})$$

$$\sigma^2 = \mathbb{E} \left[\{y - \mathbb{E}(y|\mathbf{x})\}^2 \middle| \mathbf{x} \right] / \mathbb{E}(y|\mathbf{x})$$

$$\sigma^2 = \mathbb{E} \left[\frac{\{y - \mathbb{E}(y|\mathbf{x})\}^2}{\mathbb{E}(y|\mathbf{x})} \middle| \mathbf{x} \right]$$

$$\sigma^2 = \mathbb{E} \left[\frac{\{y - \exp(\mathbf{x}'\beta_o)\}^2}{\exp(\mathbf{x}'\beta)} \middle| \mathbf{x} \right]$$

$$\mathbb{E}[\sigma^2] = \mathbb{E} \left(\mathbb{E} \left[\frac{\{y - \exp(\mathbf{x}'\beta_o)\}^2}{\exp(\mathbf{x}'\beta)} \middle| \mathbf{x} \right] \right)$$

$$\sigma^2 = \mathbb{E} \left[\frac{\{y - \exp(\mathbf{x}'\beta_o)\}^2}{\exp(\mathbf{x}'\beta)} \right]$$

$$\sigma^2 = \mathbb{E} \left[u^2(\beta_o) / \exp(\mathbf{x}'\beta_o) \right]$$

Consistent Estimator of σ^2

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

Robust Asymptotic Variance Matrix

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

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

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

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.

Lecture #3 – Models for Binary Outcomes

Properties of Binary Outcome Models

Linear Probability Model

Index Models (e.g. Logit & Probit)

Partial Effects

Conditional MLE for Index Models

Pseudo R-squared

Models for Binary Outcomes

Example

- ▶ Outcome: $y = 1$ if employed, 0 otherwise
- ▶ Predictors/Regressors: $\mathbf{x} = \{\text{age, sex, education, experience, ...}\}$

Question

How does x_j affect our prediction of y holding the other regressors constant?

We'll consider three models:

1. Linear Probability Model (LPM)
2. Logistic Regression (Logit)
3. Probit Regression (Probit)

Properties of Binary Outcome Models: $y \in \{0, 1\}$

Notation

$$p(\mathbf{x}) \equiv \mathbb{P}(y = 1|\mathbf{x})$$

Conditional Mean

$$\mathbb{E}(y|\mathbf{x}) = p(\mathbf{x})$$

Conditional Variance

$$\text{Var}(y|\mathbf{x}) = p(\mathbf{x}) [1 - p(\mathbf{x})]$$

$$\begin{aligned}\mathbb{E}(y|\mathbf{x}) &= 0 \times \mathbb{P}(y = 0|\mathbf{x}) + 1 \times \mathbb{P}(y = 1|\mathbf{x}) \\ &= \mathbb{P}(y = 1|\mathbf{x}) \equiv p(\mathbf{x})\end{aligned}$$

$$\begin{aligned}\mathbb{E}(y^2|\mathbf{x}) &= \{0^2 \times [1 - p(\mathbf{x})] + 1^2 \times p(\mathbf{x})\} \\ &= p(\mathbf{x})\end{aligned}$$

$$\begin{aligned}\text{Var}(y|\mathbf{x}) &= \mathbb{E}(y^2|\mathbf{x}) - \mathbb{E}(y|\mathbf{x})^2 \\ &= \{0^2 \times [1 - p(\mathbf{x})] + 1^2 \times p(\mathbf{x})\} - p(\mathbf{x})^2 \\ &= p(\mathbf{x}) [1 - p(\mathbf{x})]\end{aligned}$$

The Linear Probability Model: Assume $p(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$

Conditional Mean & Variance

- ▶ $\mathbb{E}(y|\mathbf{x}) = p(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$
- ▶ $\text{Var}(y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}(1 - \mathbf{x}'\boldsymbol{\beta})$

This is just Linear Regression!

$$y = \mathbf{x}'\boldsymbol{\beta} + u, \quad \mathbb{E}(u|\mathbf{x}) = 0$$

But u is Heteroskedastic

$$\text{Var}(u|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}(1 - \mathbf{x}'\boldsymbol{\beta})$$

$$\begin{aligned}\mathbb{E}(u|\mathbf{x}) &= \mathbb{E}(y - \mathbf{x}'\boldsymbol{\beta}|\mathbf{x}) = \mathbb{E}(y|\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta} \\ &= \mathbf{x}'\boldsymbol{\beta} - \mathbf{x}'\boldsymbol{\beta} = 0\end{aligned}$$

$$\begin{aligned}\text{Var}(u|\mathbf{x}) &= \mathbb{E} \left[\{u - \mathbb{E}(u|\mathbf{x})\}^2 | \mathbf{x} \right] = \mathbb{E} [u^2 | \mathbf{x}] \\ &= \mathbb{E} \left[(y - \mathbf{x}'\boldsymbol{\beta})^2 | \mathbf{x} \right] \\ &= \mathbb{E} (y^2 | \mathbf{x}) - 2\mathbb{E}(y|\mathbf{x})\mathbf{x}'\boldsymbol{\beta} + (\mathbf{x}'\boldsymbol{\beta})^2 \\ &= p(\mathbf{x}) - 2p(\mathbf{x})p(\mathbf{x}) + p(\mathbf{x})^2 \\ &= p(\mathbf{x}) [1 - p(\mathbf{x})]\end{aligned}$$

The Linear Probability Model: Assume $p(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$

Estimation

Since $\mathbb{E}(u|\mathbf{x}) = 0$ OLS estimation of $y = \mathbf{x}'\boldsymbol{\beta} + u$ is unbiased and consistent.

Inference

Since u is heteroskedastic, tests and CIs should use robust standard errors.

Is the LPM actually reasonable?

- ▶ Assumes $p(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} \implies$ changing x_j by Δ changes $p(\mathbf{x})$ by $\beta_j\Delta$
- ▶ If \mathbf{x} contains regressors without upper/lower bounds, $p(\mathbf{x})$ could be > 1 or < 0 !
- ▶ LPM could be a reasonable approximation but cannot be *literally* true.

Index Models: Constrain $p(\mathbf{x})$ to lie in $[0, 1]$

Index Model: $p(\mathbf{x}) = G(\mathbf{x}'\beta)$

Assume \mathbf{x} includes a constant, $0 \leq G(\cdot) \leq 1$, G is differentiable and strictly increasing, $\lim_{z \rightarrow \infty} G(z) = 1$, and $\lim_{z \rightarrow -\infty} G(z) = 0$.

Terminology

We call $\mathbf{x}'\beta$ the **linear index** and G the **index function**.

Partial Effects

Let $g(z) \equiv \frac{d}{dz} G(z)$. Then $\frac{\partial}{\partial x_j} p(\mathbf{x}) = g(\mathbf{x}'\beta)\beta_j$. Hence:

- ▶ The partial effect of x_j depends on the value of \mathbf{x} at which we evaluate g .
- ▶ G strictly increasing $\implies g(\cdot) > 0 \implies$ sign of partial effect determined by β_j .

Possible Choices of Index Function

CDFs as Index Functions

G satisfies the index model assumptions (prev. slide) iff it is a continuous CDF.

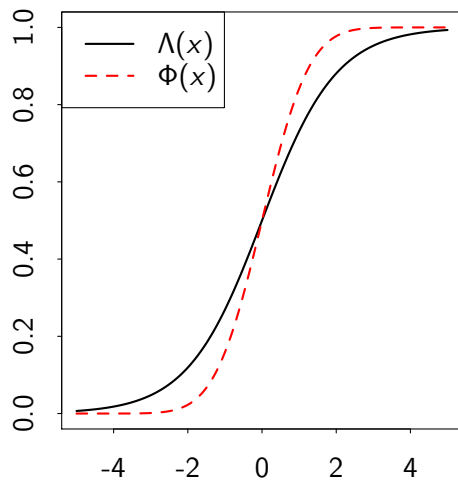
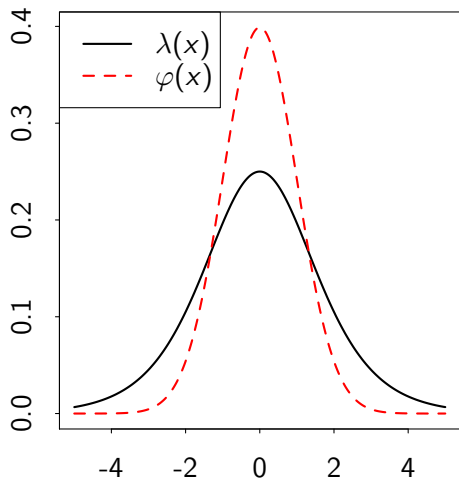
We focus on two examples:

1. Logit: $G(z) = \Lambda(z) \equiv \exp(z) / [1 + \exp(z)]$
2. Probit: $G(z) = \Phi(z) \equiv \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) dt$

Notation:

- ▶ Λ is the CDF of a “standard logistic” RV and Φ of a standard normal RV.
- ▶ λ is the density of a “standard logistic” RV and φ of a standard normal
- ▶ To treat Logit and Probit simultaneously, we'll write G as a placeholder.

Standard Logistic and Normal Densities and CDFs



Partial Effects: $\partial p(\mathbf{x})/\partial x_j$

LPM

$$\frac{\partial}{\partial x_j} \mathbf{x}'\boldsymbol{\beta} = \beta_j$$

Logit

$$\frac{\partial}{\partial x_j} \Lambda(\mathbf{x}'\boldsymbol{\beta}) = \frac{\beta_j \exp(\mathbf{x}'\boldsymbol{\beta})}{[1 + \exp(\mathbf{x}'\boldsymbol{\beta})]^2}$$

Probit

$$\frac{\partial}{\partial x_j} \Phi(\mathbf{x}'\boldsymbol{\beta}) = \frac{\beta_j \exp\{-(\mathbf{x}'\boldsymbol{\beta})^2/2\}}{\sqrt{2\pi}}$$

$$\frac{\partial}{\partial x_j} G(\mathbf{x}'\boldsymbol{\beta}) = g(\mathbf{x}'\boldsymbol{\beta})\beta_j$$

$$\begin{aligned} \frac{d}{dz} \Lambda(z) &\equiv \lambda(z) = \frac{d}{dz} \left(\frac{e^z}{1 + e^z} \right) = \frac{e^z(1 + e^z) - e^z e^z}{(1 + e^z)^2} \\ &= \frac{e^z}{(1 + e^z)^2} \end{aligned}$$

$$\frac{d}{dz} \Phi(z) = \varphi(z) = \frac{\exp\{-z^2/2\}}{\sqrt{2\pi}}$$

Comparing Logit, Probit, and LPM Partial Effects

$$\frac{\partial}{\partial x_j} G(\mathbf{x}'\beta) = g(\mathbf{x}'\beta)\beta_j, \quad \frac{d}{dz}\Lambda(z) \equiv \lambda(z) = \frac{e^z}{(1+e^z)^2}, \quad \frac{d}{dz}\Phi(z) \equiv \varphi(z) = \frac{\exp\{-z^2/2\}}{\sqrt{2\pi}}$$

Maximum Partial Effects

- ▶ λ and φ are unimodal with mode at 0

Logit $\lambda(0) = 0.25$

Probit $\varphi(0) = (2\pi)^{-1/2} \approx 0.4$

- ▶ *Maximum* partial effect when $\mathbf{x}'\beta = 0$

Logit $\beta_j\lambda(0) = 0.25\beta_j$

Probit $\beta_j\varphi(0) \approx 0.4\beta_j$

- ▶ LPM has constant partial effects β_j

Relative Effects

$$\frac{\frac{\partial}{\partial x_j} p(\mathbf{x})}{\frac{\partial}{\partial x_h} p(\mathbf{x})} = \frac{\beta_j g(\mathbf{x}'\beta)}{\beta_h g(\mathbf{x}'\beta)} = \frac{\beta_j}{\beta_h}$$

Relative effects do not depend on \mathbf{x} .

Average Partial Effects for Index Models

Partial Effect

$$\frac{\partial}{\partial x_j} G(\mathbf{x}'_i \boldsymbol{\beta}) = g(\mathbf{x}'_i \boldsymbol{\beta}) \beta_j$$

Average Partial Effect

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

Estimated Partial Effect

$$\frac{\partial}{\partial x_j} G(\mathbf{x}'_i \hat{\boldsymbol{\beta}}) = g(\mathbf{x}'_i \hat{\boldsymbol{\beta}}) \hat{\beta}_j$$

Estimated Average Partial Effect

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

Warning:

APE \neq partial effect evaluated at the average value of \mathbf{x} since $\mathbb{E}[f(Z)] \neq f(\mathbb{E}[Z])$.

Conditional MLE for Index Models: iid Observations

Conditional Likelihood

$$f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \begin{cases} 1 - G(\mathbf{x}'_i\boldsymbol{\beta}) & \text{if } y_i = 0 \\ G(\mathbf{x}'_i\boldsymbol{\beta}) & \text{if } y_i = 1 \end{cases} \iff f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = G(\mathbf{x}'_i\boldsymbol{\beta})^{y_i} [1 - G(\mathbf{x}'_i\boldsymbol{\beta})]^{1-y_i}$$

Conditional Log-Likelihood

$$\ell_i(\boldsymbol{\beta}) \equiv \log f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = y_i \log [G(\mathbf{x}'_i\boldsymbol{\beta})] + (1 - y_i) \log [1 - G(\mathbf{x}'_i\boldsymbol{\beta})]$$

Sample

$$\hat{\boldsymbol{\beta}} \equiv \arg \max_{\boldsymbol{\beta} \in \Theta} \frac{1}{N} \sum_{i=1}^N \ell_i(\boldsymbol{\beta})$$

Population

$$\boldsymbol{\beta}_o \equiv \arg \max_{\boldsymbol{\beta} \in \Theta} \mathbb{E} [\ell(\boldsymbol{\beta})]$$

Correct specification: $\mathbb{E}(y|\mathbf{x}) = p(\mathbf{x}) = G(\mathbf{x}'\boldsymbol{\beta}_o)$. Otherwise $\boldsymbol{\beta}_o = \text{KL-minimizer}$.

Asymptotic Variance Calculations for Index Models

Recall from last lecture.

Possibly Mis-specified Model

$\sqrt{N}(\hat{\beta} - \beta_o) \rightarrow_d \mathcal{N}(\mathbf{0}, \mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1})$ where $\mathbf{J} = -\mathbb{E} [\mathbf{H}_i(\beta_o)]$ and $\mathbf{K} = \mathbb{E} [\mathbf{s}_i(\beta_o)\mathbf{s}_i(\beta_o)']$

Correct Specification

$\sqrt{N}(\hat{\beta} - \beta_o) \rightarrow_d \mathcal{N}(\mathbf{0}, \mathbf{J}^{-1})$ where $\mathbf{J} = -\mathbb{E} [\mathbf{H}_i(\beta_o)]$

Asymptotic variance calculations for index models are complicated, but there's a clever trick for computing \mathbf{J} under correct specification.

Asymptotic Variance Calculation for Correctly Specified Index Models

$$\ell_i(\beta) = y_i \log \{ G(\mathbf{x}'_i \beta) \} + (1 - y_i) \log \{ 1 - G(\mathbf{x}'_i \beta) \}$$

Step 1: Calculate The Score Vector

$$\begin{aligned} \mathbf{s}_i &\equiv \frac{\partial}{\partial \beta} \ell_i(\beta) = \frac{y_i g(\mathbf{x}'_i \beta) \mathbf{x}_i}{G(\mathbf{x}'_i \beta)} - \frac{(1 - y_i) g(\mathbf{x}'_i \beta) \mathbf{x}_i}{1 - G(\mathbf{x}'_i \beta)} \\ &= \frac{g(\mathbf{x}'_i \beta) \mathbf{x}_i}{G(\mathbf{x}'_i \beta) [1 - G(\mathbf{x}'_i \beta)]} \{ [1 - G(\mathbf{x}'_i \beta)] y_i - G(\mathbf{x}'_i \beta) (1 - y_i) \} \\ &= \frac{g(\mathbf{x}'_i \beta) \mathbf{x}_i [y_i - G(\mathbf{x}'_i \beta)]}{G(\mathbf{x}'_i \beta) [1 - G(\mathbf{x}'_i \beta)]} \end{aligned}$$

Asymptotic Variance Calculation for Correctly Specified Index Models

$$\mathbf{s}_i = \frac{g(\mathbf{x}'_i\beta)\mathbf{x}_i \{y_i - G(\mathbf{x}'_i\beta)\}}{G(\mathbf{x}'_i\beta) \{1 - G(\mathbf{x}'_i\beta)\}}$$

Step 2: Start Calculating the Hessian but give up because it's a nightmare.

$$\begin{aligned}\mathbf{H}_i(\beta) &\equiv \frac{\partial \mathbf{s}_i}{\partial \beta'} = \frac{\partial}{\partial \beta'} \left([y_i - G(\mathbf{x}'_i\beta)] \left[\frac{g(\mathbf{x}'_i\beta)\mathbf{x}_i}{G(\mathbf{x}'_i\beta) \{1 - G(\mathbf{x}'_i\beta)\}} \right] \right) \\ &= \frac{-g(\mathbf{x}'_i\beta)^2 \mathbf{x}_i \mathbf{x}'_i}{G(\mathbf{x}'_i\beta) \{1 - G(\mathbf{x}'_i\beta)\}} + [y_i - G(\mathbf{x}'_i\beta)] \underbrace{\frac{\partial}{\partial \beta'} \left\{ \frac{g(\mathbf{x}'_i\beta)\mathbf{x}_i}{G(\mathbf{x}'_i\beta) [1 - G(\mathbf{x}'_i\beta)]} \right\}}_{\text{a nasty awful mess: call it } \mathbf{M}(\mathbf{x}_i, \beta)}\end{aligned}$$

Asymptotic Variance Calculation for Correctly Specified Index Models

$$\mathbf{H}_i(\beta) = \frac{-g(\mathbf{x}'_i\beta)^2 \mathbf{x}_i \mathbf{x}'_i}{G(\mathbf{x}'_i\beta) \{1 - G(\mathbf{x}'_i\beta)\}} + [y_i - G(\mathbf{x}'_i\beta)] \mathbf{M}(\mathbf{x}_i, \beta)$$

Step 3: Calculate the *Conditional Expectation* instead...

$$\begin{aligned} -\mathbb{E} [\mathbf{H}_i(\beta) | \mathbf{x}_i] &= \frac{g(\mathbf{x}'_i\beta)^2 \mathbf{x}_i \mathbf{x}'_i}{G(\mathbf{x}'_i\beta) \{1 - G(\mathbf{x}'_i\beta)\}} + \underbrace{\mathbb{E} [y_i - G(\mathbf{x}'_i\beta) | \mathbf{x}_i]}_{\text{equals zero under correct spec.}} \mathbf{M}(\mathbf{x}_i, \beta) \\ &= \frac{g(\mathbf{x}'_i\beta)^2 \mathbf{x}_i \mathbf{x}'_i}{G(\mathbf{x}'_i\beta) \{1 - G(\mathbf{x}'_i\beta)\}} \end{aligned}$$

Step 4: Iterated Expectations

$$\mathbf{J} = -\mathbb{E} [\mathbf{H}_i(\beta_o)] = \mathbb{E} \{ \mathbb{E} [\mathbf{H}_i(\beta_o) | \mathbf{x}_i] \} = \mathbb{E} \left\{ \frac{g(\mathbf{x}'_i\beta_o)^2 \mathbf{x}_i \mathbf{x}'_i}{G(\mathbf{x}'_i\beta_o) \{1 - G(\mathbf{x}'_i\beta_o)\}} \right\}$$

Asymptotic Variance Calculation for Correctly Specified Index Models

Asymptotic Distribution

$$\sqrt{N}(\hat{\beta} - \beta_o) \rightarrow_d \mathcal{N}(\mathbf{0}, \mathbf{J}^{-1}), \quad \mathbf{J}^{-1} = \mathbb{E} \left\{ \frac{g(\mathbf{x}'_i \beta_o)^2 \mathbf{x}_i \mathbf{x}'_i}{G(\mathbf{x}'_i \beta_o) \{1 - G(\mathbf{x}'_i \beta_o)\}} \right\}^{-1}$$

Consistent Estimator

$$\hat{\mathbf{J}}^{-1} \equiv \left\{ \frac{1}{N} \sum_{i=1}^N \frac{g(\mathbf{x}'_i \hat{\beta})^2 \mathbf{x}_i \mathbf{x}'_i}{G(\mathbf{x}'_i \hat{\beta}) [1 - G(\mathbf{x}'_i \hat{\beta})]} \right\}^{-1}$$

Notes

- ▶ Assumes correct specification, i.e. $p(\mathbf{x}) = \mathbb{E}(y|\mathbf{x}) = G(\mathbf{x}'\beta_o)$
- ▶ In contrast, *robust* variance matrix $\mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1}$ is complicated, but R can do it.

McFadden (1974) – “Pseudo R-squared”

Model with Intercept Only

$L(\bar{y}) \equiv$ maximized sample Likelihood

$\ell(\bar{y}) \equiv$ maximized sample log-likelihood

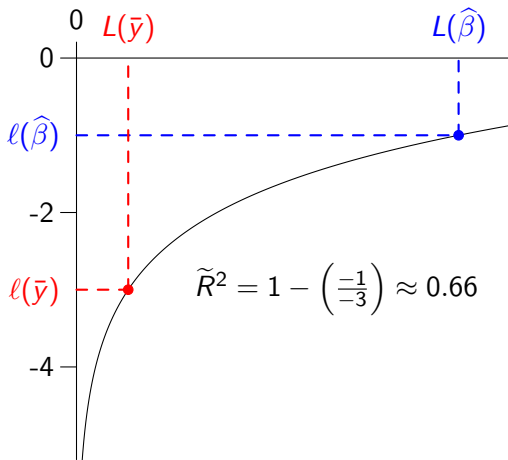
Full Model

$L(\hat{\beta}) \equiv$ maximized sample Likelihood

$\ell(\hat{\beta}) \equiv$ maximized sample log-likelihood

Pseudo R-squared

$$\tilde{R}^2 \equiv 1 - \ell(\hat{\beta})/\ell(\bar{y})$$



McFadden (1974) – “Pseudo R-squared”

Pseudo R-squared

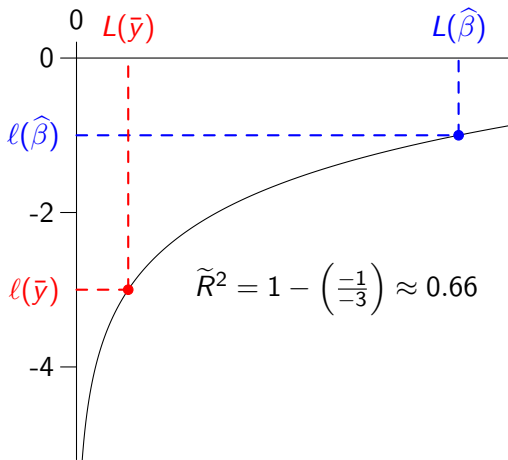
$$\tilde{R}^2 \equiv 1 - \ell(\hat{\beta}) / \ell(\bar{y})$$

Always between 0 and 1

Show this on the problem set!

Health Warning

I don't recommend using pseudo- R^2 : it's arbitrary and can be misleading. Other people use it so I'm telling you what it is.



Lecture #4 – Random Utility Models

Overview of Random Utility Models

Identification of Choice Models

Index Models as Special Cases (e.g. Logit & Probit)

Multinomial and Conditional Logit

Discrete Choice – Basic Terminology

Decision-maker

Household, person, firm, etc.

Alternatives

Products, courses of action, etc.

Choice Set

The collection of all alternatives available to the decision-maker.

Restrictions on the Choice Set

We assume that:

1. Choices are mutually exclusive: choose only *one* alternative.
2. Choice set is *exhaustive*: contains every alternative (always choose something)
3. The number of alternatives is finite.

We can always redefine the choice set to satisfy 1 and 2

$$\underbrace{\{\text{Beer}, \text{Pizza}\}}_{\text{not mutually exclusive}} \rightarrow \underbrace{\{\text{Beer only}, \text{Pizza only}, \text{Beer and Pizza}\}}_{\text{mutually exclusive}}$$

$$\underbrace{\{\text{Beer only}, \text{Pizza only}, \text{Beer and Pizza}\}}_{\text{not exhaustive}} \rightarrow \underbrace{\{\text{Beer only}, \text{Pizza only}, \text{Beer and Pizza}, \text{Something Else}\}}_{\text{exhaustive}}$$

Random Utility Models (RUMs)

Following Train (2009), use n to index individuals!

Notation

- ▶ N decision-makers $n = 1, \dots, N$
- ▶ J alternatives $j = 1, \dots, J$.

Utility and Decision Rule

- ▶ Decision-maker n obtains utility U_{nj} from choosing alternative j
- ▶ Maximize utility: decision-maker n chooses alternative i iff $U_{ni} > U_{nj}$ for any $j \neq i$

Random Utility Models

Researcher Observes

- ▶ Attributes x_{nj} of each alternative (e.g. product characteristics)
- ▶ Attributes s_n of the decision-maker (e.g. demographics)
- ▶ Choices but **not utilities**

Representative Utility V_{nj}

Assume researcher can specify a function $V_{nj}(x_{nj}, s_n)$ relating attributes x_{nj} of each alternative j and attributes s_n of each decision-maker n to her utilities U_{nj} .

Error Terms ε_{nj}

$\varepsilon_{nj} \equiv U_{nj} - V_{nj}$ is the difference between *true* utility U_{nj} and representative utility V_{nj}

Random Utility Models (RUMs)

What are the error terms?

ε_{nj} for $j = 1, \dots, J$ represent unobserved factors that affect choices but are not captured by representative utilities (i.e. our model)

Treat Errors as Random

Let $\varepsilon' \equiv [\varepsilon_{n1} \dots \varepsilon_{nJ}]$ have density function $f(\varepsilon_n)$. Characterizes unobserved heterogeneity across decision-makers.

Choice Probabilities

$$P_{ni} \equiv \mathbb{P}(U_{ni} > U_{nj} \quad \forall j \neq i) = \int_{\mathbb{R}^J} \mathbb{1} \{ \varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \quad \forall j \neq i \} f(\varepsilon_n) d\varepsilon_n$$

This all sounds a bit abstract...

Basic Idea

1. Write down a parametric model for $V_{nj}(x_{nj}, s_n)$ with unknown parameters θ .
2. Choose a distribution f for the errors (heterogeneity) ε_n .
3. Back out choice probabilities as a function of parameters θ .
4. Use observed choices and attributes to find the MLE $\hat{\theta}$.

Looking Back; Looking Ahead

- ▶ Logit and Probit are special cases of RUMs: choice between two alternatives.
- ▶ RUMs provide a framework to estimate more complicated discrete choice models.

Some Complications

Computation

- ▶ Integral linking choice probabilities to parameters θ rarely has a closed form.
- ▶ Logit-type models are a well-known *exception*.
- ▶ More generally: use *Monte Carlo Simulation* to approximate the integral.

Identification

- ▶ Roughly speaking, we say that a parameter is *identified* if it could be uniquely determined by observing the *whole population*.
- ▶ What parameters of RUMs are identified from choices and attributes?

A Very Simple Example

Transport Decision

- ▶ Exactly two ways to get to work: by **car** or by **bus**.
- ▶ Observe two attributes: cost in **time** T and **money** M of each mode of transport.

Econometrician's Model: (β, γ) unknown

$$V_{\text{car}} = \beta T_{\text{car}} + \gamma M_{\text{car}}$$

$$U_{\text{car}} = V_{\text{car}} + \varepsilon_{\text{car}}$$

$$V_{\text{bus}} = \beta T_{\text{bus}} + \gamma M_{\text{bus}}$$

$$U_{\text{bus}} = V_{\text{bus}} + \varepsilon_{\text{bus}}$$

Choice Probabilities

$$P_{\text{car}} = \mathbb{P}(\varepsilon_{\text{bus}} - \varepsilon_{\text{car}} < V_{\text{car}} - V_{\text{bus}})$$

$$P_{\text{bus}} = \mathbb{P}(\varepsilon_{\text{car}} - \varepsilon_{\text{bus}} < V_{\text{bus}} - V_{\text{car}}) = 1 - P_{\text{car}}$$

A Very Simple Example: Who drives to work?

What is common?

Parameters: (β, γ) . Our goal is to identify and estimate these.

Observed Heterogeneity

- ▶ Alice lives next to the bus stop: her T_{bus} is low.
- ▶ Bob is 70 and gets a discount on public transport: his M_{bus} is low.
- ▶ Clara and her roommates work at the same office and can carpool: her M_{car} is low.

Unobserved Heterogeneity

James hates to drive ($\varepsilon_{\text{car}} - \varepsilon_{\text{bus}} < 0$) but Steve loves driving ($\varepsilon_{\text{car}} - \varepsilon_{\text{bus}} > 0$).

Identification – What can we learn from data?

Only differences in utility matter

- ▶ $\mathbb{P}(U_{ni} > U_{nj} \quad \forall j \neq i) = \mathbb{P}(U_{ni} - U_{nj} > 0 \quad \forall j \neq i)$
- ▶ All that matters is how much better or worse a given alternative is than the others.

Consequences

1. We cannot identify a different intercept for each alternative.
2. We can only identify differences of effects for decision-maker attributes.

If there are J alternatives, we can identify only $(J - 1)$ intercepts.

Equivalently: normalize one intercept to zero.

$$\text{Intercept} \Rightarrow \mathbb{E}[\varepsilon_{nj}] = 0$$

- ▶ Suppose $U_{nj} = \mathbf{x}'_{nj}\beta + \varepsilon_{nj}^*$ where \mathbf{x}_{nj} *excludes* a constant and $\mathbb{E}[\varepsilon_{nj}^*] \neq 0$.
- ▶ Equivalent model: $U_{nj} = \alpha + \mathbf{x}'_{nj}\beta + \varepsilon_{nj}$ where $\mathbb{E}[\varepsilon_{nj}] = 0$ by construction.

Why not a different intercept for each alternative?

$$U_{\text{car}} = \alpha_{\text{car}} + \beta T_{\text{car}} + \gamma M_{\text{car}} + \varepsilon_{\text{car}}$$

$$U_{\text{bus}} = \alpha_{\text{bus}} + \beta T_{\text{bus}} + \gamma M_{\text{bus}} + \varepsilon_{\text{bus}}$$

$$U_{\text{bus}} - U_{\text{car}} = (\alpha_{\text{bus}} - \alpha_{\text{car}}) + \beta (T_{\text{bus}} - T_{\text{car}}) + \gamma (M_{\text{bus}} - M_{\text{car}}) + (\varepsilon_{\text{bus}} - \varepsilon_{\text{car}})$$

Only differences of effects for decision-maker attributes are identified.

Can we identify the effects of income Y separately for Bus and Car?

$$U_{\text{car}} = \theta_{\text{car}} Y + \beta T_{\text{car}} + \gamma M_{\text{car}} + \varepsilon_{\text{car}}$$

$$U_{\text{bus}} = \theta_{\text{bus}} Y + \beta T_{\text{bus}} + \gamma M_{\text{bus}} + \varepsilon_{\text{bus}}$$

$$U_{\text{bus}} - U_{\text{car}} = (\theta_{\text{bus}} - \theta_{\text{car}}) Y + \beta (T_{\text{bus}} - T_{\text{car}}) + \gamma (M_{\text{bus}} - M_{\text{car}}) + (\varepsilon_{\text{bus}} - \varepsilon_{\text{car}})$$

Equivalent to normalizing one of the θ s to zero.

More on Identification – What can we learn from data?

The scale of utility is irrelevant

- ▶ Let λ be an arbitrary positive constant.
- ▶ Original Model: $U_{nj} = V_{nj} + \varepsilon_{nj}$, $\text{Var}(\varepsilon_{nj}) = \sigma^2$
- ▶ Re-scaled Model: $\lambda U_{nj} = \lambda V_{nj} + \lambda \varepsilon_{nj} \iff U_{nj}^* = V_{nj}^* + \varepsilon_{nj}^*$, $\text{Var}(\varepsilon_{nj}^*) = \lambda^2 \sigma^2$

$\text{Var}(\varepsilon_{nj})$ determines the scale of β

- ▶ $U_{nj} = \mathbf{x}_{nj}'\beta + \varepsilon_{nj}$, $\text{Var}(\varepsilon_{nj}) = \sigma^2 \iff U_{nj}^* = \mathbf{x}_{nj}'(\beta/\sigma) + \varepsilon_{nj}^*$, $\text{Var}(\varepsilon_{nj}^*) = 1$
- ▶ Can't directly compare coefs. across models with different normalizations for ε_{nj} .
- ▶ Recall: we had to re-scale Logit and Probit coefs. to compare them.

Only differences in utility matter \implies only differences in errors matter.

Notation

- ▶ $\tilde{\varepsilon}_{nj} \equiv \varepsilon_{nj} - \varepsilon_{ni}$ be the *difference* of errors ε_{nj} and ε_{ni} .
- ▶ $\tilde{\varepsilon}_{ni} \equiv$ vector of all unique differences, taking ε_{ni} as the “base case”
 - ▶ E.g. $\varepsilon'_n = (\varepsilon_{n1}, \varepsilon_{n2}, \varepsilon_{n3}) \implies \tilde{\varepsilon}'_{n1} = (\varepsilon_{n2} - \varepsilon_{n1}, \varepsilon_{n3} - \varepsilon_{n1})$
 - ▶ Note: J errors $\Rightarrow (J - 1)$ unique *differences*
- ▶ Let g be the joint density of $\tilde{\varepsilon}_{ni}$.

Choice Probabilities

$$\begin{aligned} P_{ni} &\equiv \mathbb{P}(U_{ni} > U_{nj} \quad \forall j \neq i) = \mathbb{P}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \quad \forall j \neq i) \\ &= \mathbb{P}(\tilde{\varepsilon}_{nji} < V_{ni} - V_{nj} \quad \forall j \neq i) = \int_{\mathbb{R}^{J-1}} \mathbb{1}\{\tilde{\varepsilon}_{nji} < V_{ni} - V_{nj} \quad \forall j \neq i\} g(\tilde{\varepsilon}_{ni}) d\tilde{\varepsilon}_{ni} \end{aligned}$$

How to obtain the index models from last lecture? (E.g. Probit and Logit)

1. Two alternatives, e.g. Bus or Something Else
2. Let $y_n = 1$ if decision-maker n chooses alternative 1; zero otherwise.
3. $V_{nj} = \mathbf{s}'_n \gamma_j$ (representative utility depends only on attributes of decision-maker)
4. $(\varepsilon_{n2} - \varepsilon_{n1}) \sim G$ independently of \mathbf{s}_n .

$$\begin{aligned} U_{n1} - U_{n2} &= (\mathbf{s}'_n \gamma_1 - \mathbf{s}'_n \gamma_2) + (\varepsilon_{n1} - \varepsilon_{n2}) = \mathbf{s}'_n (\gamma_1 - \gamma_2) + (\varepsilon_{n1} - \varepsilon_{n2}) \\ &= \mathbf{s}'_n \gamma + (\varepsilon_{n1} - \varepsilon_{n2}) \end{aligned}$$

$$\mathbb{P}(y_n = 1 | \mathbf{s}_n) = \mathbb{P}(U_{n1} - U_{n2} > 0 | \mathbf{s}_n) = \mathbb{P}(\varepsilon_{n2} - \varepsilon_{n1} < \mathbf{s}'_n \gamma | \mathbf{s}_n) = G(\mathbf{s}'_n \gamma)$$

The Logit Family of Choice Models

Theorem

Suppose that $\varepsilon_{n1}, \dots, \varepsilon_{nJ} \sim \text{iid } F$ where $F(z) = \exp\{-\exp(-z)\}$. Then,

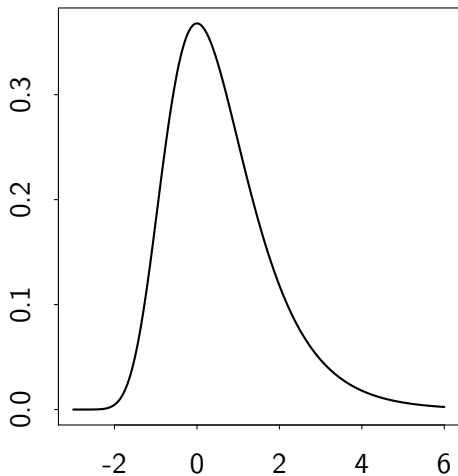
$$P_{ni} = \mathbb{P}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \quad \forall j \neq i) = \frac{\exp(V_{ni})}{\sum_{j=1}^J \exp(V_{nj})}$$

Notes

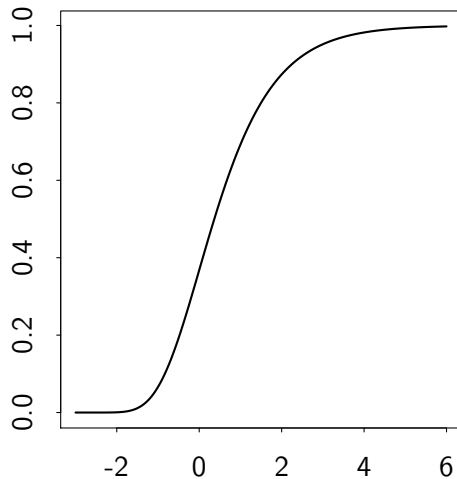
- ▶ This is a special case where the choice probabilities have a closed-form solution!
- ▶ $F(z) = \exp\{-\exp(-z)\}$ is the Gumbel aka Type I Extreme Value CDF
- ▶ Corollary: the *difference* of independent Gumbel RVs is a standard Logistic RV

The Gumbel Distribution (aka Type I Extreme Value)

Gumbel Density



Gumbel CDF



Different specifications of V_{nj} yield different models.

Multinomial Logit

- ▶ $V_{nj} = \mathbf{s}'_n \gamma_j$ ← only attributes that are fixed across alternatives (e.g. n 's income)
- ▶ Can only identify differences $(\gamma_j - \gamma_i)$. Typical to normalize $\gamma_1 = \mathbf{0}$.

Conditional Logit

- ▶ $V_{nj} = \mathbf{x}'_{nj} \beta$ ← only attributes that vary across alternatives (e.g. price)
- ▶ Note that β is fixed across alternatives.

Mixed Logit

- ▶ $V_{nj} = \mathbf{s}'_n \gamma_j + \mathbf{x}'_{nj} \beta$ ← a combination of the two

The Likelihood for Random Utility Models

Notation

- ▶ $y_n \in \{1, \dots, J\} \equiv n$'s choice.
- ▶ \mathbf{z}_n vector of all regressors for n
- ▶ $\boldsymbol{\theta}$ vector of all unknown parameters
- ▶ Choice Probs. $P_{ni} \equiv \mathbb{P}(y_n = i | \mathbf{z}_n, \boldsymbol{\theta})$

Note

Likelihood is easy, but choice probabilities are usually hard (logit is an exception).

Likelihood

$$f(y_n | \mathbf{z}_n, \boldsymbol{\theta}) = \prod_{j=1}^J P_{nj}^{\mathbb{1}\{y_n=j\}}$$

Log Likelihood

$$\ell_N(\boldsymbol{\theta}) = \sum_{n=1}^N \sum_{j=1}^J \mathbb{1}\{y_n = j\} \log P_{nj}$$

Logit Choice Probabilities

$$P_{ni} = \exp(V_{ni}) / \sum_{j=1}^J \exp(V_{nj})$$

Interpreting Multinomial Logit Coefficients

- ▶ Partial effects tricky to derive and interpret.
- ▶ Better approach: examine **log-odds ratios**
- ▶ Normalizing $\gamma_1 = \mathbf{0}$, we have $\exp(\mathbf{s}_n \gamma_1) = \exp(0) = 1$. Hence,

$$\frac{P_{ni}}{P_{n1}} = \frac{\exp(\mathbf{s}_n \gamma_i)}{\sum_{j=1}^J \exp(\mathbf{s}_n \gamma_j)} \times \frac{\sum_{j=1}^J \exp(\mathbf{s}_n \gamma_j)}{\exp(\mathbf{s}_n \gamma_1)} = \frac{\exp(\mathbf{s}_n \gamma_i)}{\exp(\mathbf{s}_n \gamma_1)} = \exp(\mathbf{s}_n \gamma_i)$$

- ▶ Taking logs: $\log(P_{ni}/P_{n1}) = \log[\exp(\mathbf{s}_n \gamma_i)] = \mathbf{s}_n' \gamma_i$.

Punchline

$\gamma_i^{(k)}$ is the marginal effect of $s_n^{(k)}$ on the **relative probability** that $y = i$ compared to $y = 1$ **measured on the log scale** – e.g. taking the bus relative to driving.

Interpreting Conditional Logit Coefficients

You'll derive these on the problem set!

Partial Effects

- ▶ The attributes \mathbf{x}_{nj} are *specific* to a particular alternative j .
- ▶ Hence: partial effects are much simpler for conditional logit than multinomial.

Own Attribute

$$\frac{\partial P_{nj}}{\partial \mathbf{x}_{nj}} = P_{nj}(1 - P_{nj})\beta$$

Cross-Attribute ($j \neq i$)

$$\frac{\partial P_{nj}}{\partial \mathbf{x}_{ni}} = -P_{nj}P_{ni}\beta$$

If increasing $\mathbf{x}_{nj}^{(k)}$ makes $y = j$ *more likely*, it must make $y = i$ *less likely*

The Independence of Irrelevant Alternatives (IIA)

Or why people don't like logit models...

Logit Choice Probabilities

$$P_{ni} = \frac{\exp(V_{ni})}{\sum_{j=1}^J \exp(V_{nj})} \implies \frac{P_{ni}}{P_{nj}} = \exp(V_{ni} - V_{nj})$$

In Words

The relative probability of choosing i versus j only depends on the representative utilities for i and j . This is called the **independence of irrelevant alternatives (IIA)**.

Why is this a problem

IIA arises in logit models because $\varepsilon_{n1}, \dots, \varepsilon_{nJ}$ are *independent*. In reality “some alternatives are more similar than others,” i.e. errors may be correlated.

An Example where IIA is Unreasonable – Choosing Presidential Candidates

Model

- ▶ $V_{nj} = (\text{Demographics}_n)' \gamma_j + (\text{Ideology}_{nj})' \beta$
- ▶ (Ideology_{nj}) = similarity between voter n 's ideology and candidate j 's.
- ▶ Candidates = {Trump, Sanders, Warren}

Consider a group of voters who all have the *same* demographics and ideology

E.g. white, centrist, female, mid-westerners between the age of 45 and 50 with an average household income between \$50 and \$55 thousand USD.

Same regressors \Rightarrow same V_{nj}

V_{nj} doesn't vary over n within the group: $\{V_{\text{Trump}}, V_{\text{Sanders}}, V_{\text{Warren}}\}$

An Example where IIA is Unreasonable – Choosing Presidential Candidates

Two-way Race

Suppose 2/3 of this group of voters chooses Sanders over Trump: $P_{\text{Sanders}}/P_{\text{Trump}} = 2$

Assumption

Sanders and Warren are ideologically similar $\implies V_{\text{Warren}} \approx V_{\text{Sanders}}$

Implications of Logit

- ▶ Relative choice probabilities are the *same* in a two-way race or a three-way race.
- ▶ $P_{\text{Warren}}/P_{\text{Sanders}} = \exp(V_{\text{Warren}} - V_{\text{Sanders}}) \approx 1$

An Example where IIA is Unreasonable – Choosing Presidential Candidates

Logit Implication for Three-way Race

$$P_{\text{Sanders}} = 2P_{\text{Trump}}, \quad P_{\text{Sanders}} \approx P_{\text{Warren}}, \quad P_{\text{Trump}} + P_{\text{Sanders}} + P_{\text{Warren}} = 1$$

$$\implies P_{\text{Trump}} + 2P_{\text{Trump}} + 2P_{\text{Trump}} = 1$$

$$P_{\text{Trump}} = 1/5$$

$$P_{\text{Warren}} = P_{\text{Sanders}} = 2/5$$

What we'd actually expect in a Three-way Race

1/3 Trump, 1/3 Sanders and 1/3 Warren – i.e. Warren “splits” the Sanders vote.

What's going wrong?

Logit assumes $\varepsilon_{\text{Warren}}$ and $\varepsilon_{\text{Sanders}}$ are independent but in reality they're not.

Lecture #5 – Sample Selection

Examples of Sample Selection

The Heckman Selection Model

What is sample selection?

Question

Thus far we have always assumed that $(y_1, \mathbf{x}_1), \dots, (y_N, \mathbf{x}_N)$ are a random sample from the population of interest. What if they aren't?

Example 1: MPhil Admissions

- ▶ Suppose we want to improve admissions decisions at Oxford.
- ▶ $y \equiv$ overall marks in 1st year of Oxford Economics MPhil
- ▶ $\mathbf{x} \equiv \{\text{undergrad grades, letters of reference, } \dots\}$
- ▶ What we observe: \mathbf{x} for all applicants; y for applicants who were **admitted**.
- ▶ What we want: $\mathbb{E}(y|\mathbf{x})$ for **all applicants**.

Example 2: A Model of Wage Offers

Gronau (1974; JPE)

Question

How do wage offers w_i^o vary with \mathbf{x}_i for all people in the population.

Problem

Only observe w_i^o for people who *accept* their offer, i.e. those who are employed.

Mathematically

$$\mathbb{E}(w_i^o | \mathbf{x}_i) \neq \mathbb{E}(w_i^o | \mathbf{x}_i, \text{Employed})$$

The Heckman Selection Model (Heckit) — Is β_1 identified?

Outcome Equation

$$y_1 = \mathbf{x}'_1 \beta_1 + u_1$$

Assumptions

- (a) Observe $y_2, \mathbf{x}' = (\mathbf{x}'_1, \mathbf{x}'_2)$; only observe y_1 if $y_2 = 1$.
- (b) (u_1, v_2) are mean zero and jointly independent of \mathbf{x} .
- (c) $v_2 \sim \text{Normal}(0, 1)$
- (d) $\mathbb{E}(u_1 | v_2) = \gamma_1 v_2$ where γ_1 is an unknown constant.

Participation Equation

$$y_2 = \mathbb{1} \{ \mathbf{x}' \boldsymbol{\delta}_2 + v_2 > 0 \}$$

Notes

- ▶ $\mathbb{E}(u_1) = \mathbb{E}(v_2) = 0$ is not restrictive: just include intercepts in both equations.
- ▶ Assumption (d) would be *implied* by assuming that (u_1, v_2) are jointly normal.
- ▶ These assumptions are strong. They can be weakened a bit, but not too much.

Step 1: Show that u_1 and \mathbf{x} are conditionally independent given v_2 .

Assumption (b)

(u_1, v_2) are jointly independent of \mathbf{x} .

Equivalently

$$f_{1,2|\mathbf{x}}(u_1, v_2|\mathbf{x}) = f_{1,2}(u_1, v_2), \quad \text{and} \quad f_{1|\mathbf{x}}(u_1|\mathbf{x}) = f_1(u_1), \quad \text{and} \quad f_{2|\mathbf{x}}(v_2|\mathbf{x}) = f_2(v_2)$$

Therefore

$$f_{1|2,\mathbf{x}}(u_1|v_2, \mathbf{x}) = \frac{f_{1,2|\mathbf{x}}(u_1, v_2|\mathbf{x})}{f_{2|\mathbf{x}}(v_2|\mathbf{x})} = \frac{f_{1,2}(u_1, v_2)}{f_2(v_2)} = f_{1|2}(u_1|v_2)$$

In Words

Conditioning on (v_2, \mathbf{x}) gives the same information about u_1 as conditioning on v_2 only.

Step 2: Calculate $\mathbb{E}(y_1|\mathbf{x}, v_2)$; show that if v_2 were observed we'd be done.

$$\begin{aligned}\mathbb{E}(y_1|\mathbf{x}, v_2) &= \mathbb{E}(\mathbf{x}'_1\boldsymbol{\beta}_1 + u_1|\mathbf{x}, v_2) && \text{(Substitute Outcome Eq.)} \\ &= \mathbf{x}'_1\boldsymbol{\beta}_1 + \mathbb{E}(u_1|\mathbf{x}, v_2) && (\mathbf{x}_1 \text{ is a subset of } \mathbf{x}) \\ &= \mathbf{x}'_1\boldsymbol{\beta}_1 + \mathbb{E}(u_1|v_2) && \text{(apply result of Step 1)} \\ &= \mathbf{x}'_1\boldsymbol{\beta}_1 + \gamma_1 v_2 && \text{(apply Assumption (d))}\end{aligned}$$

Step 3: Relate v_2 (unobserved) to \mathbf{x} and y_2 (both observed).

$$\begin{aligned}\mathbb{E}(y_1|\mathbf{x}, y_2) &= \mathbb{E}_{v_2|(\mathbf{x}, y_2)} [\mathbb{E}(y_1|\mathbf{x}, y_2, v_2)] && \text{(Law of Iterated Expectations)} \\ &= \mathbb{E}_{v_2|(\mathbf{x}, y_2)} [\mathbb{E}(y_1|\mathbf{x}, v_2)] && \text{(Participation Eq: } y_2 = g(\mathbf{x}, v_2)) \\ &= \mathbb{E} [\mathbf{x}'_1 \boldsymbol{\beta}_1 + \gamma_1 v_2 | \mathbf{x}, y_2] && \text{(apply result of Step 2)} \\ &= \mathbf{x}'_1 \boldsymbol{\beta}_1 + \gamma_1 \mathbb{E}(v_2 | \mathbf{x}, y_2) && (\mathbf{x}_1 \text{ is a subset of } \mathbf{x})\end{aligned}$$

Therefore

$$\mathbb{E}(y_1|\mathbf{x}, y_2 = 1) = \mathbf{x}'_1 \boldsymbol{\beta}_1 + \gamma_1 \mathbb{E}(v_2 | \mathbf{x}, y_2 = 1).$$

What is the significance of Step 3?

- ▶ Define $h(\mathbf{x}) \equiv \mathbb{E}(v_2|\mathbf{x}, y_2 = 1)$. Then: $\mathbb{E}(y_1|\mathbf{x}, y_2 = 1) = \mathbf{x}'_1\beta_1 + \gamma_1 h(\mathbf{x})$
- ▶ Note that $h(\mathbf{x})$ is a random variable: a function of \mathbf{x} .
- ▶ Step 3 shows that a linear regression of y_1 on \mathbf{x}_1 and $h(\mathbf{x})$ for the *selected* sample, those with $y_2 = 1$, identifies β_1 and γ_1 !
- ▶ All that remains is to figure out what function h is...

Note: Selection Bias Enters Through γ_1

Assumption (d)

$\mathbb{E}(u_1|v_2) = \gamma_1 v_2$ allows *dependence* between errors in participation and outcome eqs.

Step 3

$$\mathbb{E}(y_1|\mathbf{x}, y_2 = 1) = \mathbf{x}'_1 \beta_1 + \gamma_1 \mathbb{E}(v_2|\mathbf{x}, y_2 = 1)$$

Therefore

If $\gamma_1 = 0$ there is no selection bias: in this case $\mathbb{E}(y_1|\mathbf{x}, y_2 = 1) = \mathbf{x}'_1 \beta$ so regressing y_1 on \mathbf{x}_1 for the subset of individuals with $y_2 = 1$ identifies β_1 .

Step 4: Determine the distribution of v_2 given $(\mathbf{x}, y_2 = 1)$.

$$\mathbb{P}(v_2 \leq t | \mathbf{x}, y_2 = 1) = \mathbb{P}(v_2 \leq t | \mathbf{x}, v_2 > -\mathbf{x}'\delta_2) \quad (\text{participation eq.})$$

$$= \frac{\mathbb{P}(\{v_2 \leq t\} \cap \{v_2 > -\mathbf{x}'\delta_2\} | \mathbf{x})}{\mathbb{P}(v_2 > -\mathbf{x}'\delta_2 | \mathbf{x})} \quad (\text{defn. of cond. prob.})$$

$$= \frac{\mathbb{P}\{v_2 \in (-\mathbf{x}'\delta_2, t]\}}{\mathbb{P}(v_2 > -\mathbf{x}'\delta_2)} \quad (v_2 \text{ and } \mathbf{x} \text{ are indep.})$$

$$= \frac{\mathbb{P}\{z \in (c, t]\}}{\mathbb{P}(z > c)} \quad (v_2 \text{ is standard normal})$$

where $z \sim \text{Normal}(0, 1)$ and we define the shorthand $c \equiv -\mathbf{x}'\delta_2$.

Step 5: Calculate the Expectation of a Truncated Normal

Recall: $z \sim \text{Normal}(0, 1)$ and $c \equiv -\mathbf{x}'\boldsymbol{\delta}_2$

CDF

$$\mathbb{P}(z \leq t | z > c) = \frac{\mathbb{P}\{z \in (c, t]\}}{\mathbb{P}(z > c)} = \mathbb{1}\{c \leq t\} \left[\frac{\Phi(t) - \Phi(c)}{1 - \Phi(c)} \right]$$

Density

$$f(t | z > c) = \frac{d}{dt} \mathbb{P}(z \leq t | z > c) = \begin{cases} 0, & t \leq c \\ \varphi(t) / [1 - \Phi(c)], & t > c \end{cases}$$

Step 5: Calculate the Expectation of a Truncated Normal

Recall: $z \sim \text{Normal}(0, 1)$ and $c \equiv -\mathbf{x}'\boldsymbol{\delta}_2$

$$\begin{aligned}\mathbb{E}(z|z > c) &= \int_{-\infty}^{\infty} tf(t|z > c) dt = \frac{1}{1 - \Phi(c)} \int_c^{\infty} t\varphi(t) dt \\&= \left[\frac{1}{1 - \Phi(c)} \right] \left(\frac{1}{\sqrt{2\pi}} \right) \int_c^{\infty} t \exp \{ -t^2/2 \} dt \\&= \left[\frac{1}{1 - \Phi(c)} \right] \left(\frac{1}{\sqrt{2\pi}} \right) \left[-\exp \{ -t^2/2 \} \right]_c^{\infty} \\&= \left[\frac{1}{1 - \Phi(c)} \right] \left(\frac{\exp \{ -c^2/2 \}}{\sqrt{2\pi}} \right) = \frac{\varphi(c)}{1 - \Phi(c)}\end{aligned}$$

Step 6: Put everything together.

Recall: Step 3

$$y_1 = \mathbf{x}'_1 \beta_1 + \gamma_1 h(\mathbf{x}) + \eta, \quad \mathbb{E}[\eta | \mathbf{x}_1, h(\mathbf{x})] = 0, \quad h(\mathbf{x}) \equiv \mathbb{E}(v_2 | \mathbf{x}, y_2 = 1)$$

Using Steps 4–5

$$h(\mathbf{x}) = \frac{\varphi(-\mathbf{x}'\delta_2)}{1 - \Phi(-\mathbf{x}'\delta_2)} = \frac{\varphi(\mathbf{x}'\delta_2)}{\Phi(\mathbf{x}'\delta_2)} \quad \text{since } \varphi(-c) = \varphi(c) \text{ and } 1 - \Phi(c) = \Phi(-c).$$

Inverse Mills Ratio

$\varphi(c)/\Phi(c)$ is the inverse Mills Ratio, traditionally denoted by $\lambda \implies h(\mathbf{x}) = \lambda(\mathbf{x}'\delta_2)$.

Careful!

In an earlier lecture λ denoted the standard logistic density. Here it's something else!

The Heckman Two-step Estimator aka “Heckit”

Observables

Observe (y_{2i}, \mathbf{x}_i) for a random sample of size N ; only observe y_{1i} for those with $y_{2i} = 1$.

First Step – Estimate δ_2 from Full Sample

- ▶ Run Probit on the Participation Eq. $\mathbb{P}(y_{2i} = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_i'\delta_2)$ for the full sample.
- ▶ Define $\hat{\lambda}_i \equiv \lambda(\mathbf{x}_i'\hat{\delta}_2)$ where $\hat{\delta}_2$ is the MLE for δ_2 .

Second Step – Estimate (β_1, γ_1) from Selected Sample

Using the observations for which y_{1i} is observed, regress y_{1i} on $(\mathbf{x}_{1i}, \hat{\lambda}_i)$ by OLS to obtain estimates $(\hat{\beta}_1, \hat{\gamma}_1)$.

The Heckman Two-step Estimator aka “Heckit”

Theorem

Under the assumptions from above, the 2-step “Heckit” estimators satisfy

$$\begin{bmatrix} \hat{\delta}_2 \\ \hat{\beta}_1 \\ \hat{\gamma}_1 \end{bmatrix} \rightarrow_p \begin{bmatrix} \delta_2 \\ \beta_1 \\ \gamma_1 \end{bmatrix} \quad \text{and} \quad \sqrt{N} \begin{bmatrix} \hat{\delta}_2 - \delta_2 \\ \hat{\beta}_1 - \beta_1 \\ \hat{\gamma}_1 - \gamma_1 \end{bmatrix} \rightarrow_d \text{Normal}(\mathbf{0}, \mathbf{\Omega}) \quad \text{as } N \rightarrow \infty.$$

Standard Errors

The asymptotic variance matrix $\mathbf{\Omega}$ is complicated: the usual OLS standard errors from step two are incorrect as they do not account for the estimation of δ_2 in step one.

The Big Picture: How does Heckit solve the selection problem?

- ▶ If we regress y_{1i} on \mathbf{x}_{1i} for the selected sample, there is an omitted variable.
- ▶ Under the Heckit assumptions, the omitted variable is precisely $\lambda(\mathbf{x}_i'\boldsymbol{\delta}_2)$.
- ▶ Hence: a regression of y_{1i} on \mathbf{x}_{1i} and $\lambda(\mathbf{x}_i'\boldsymbol{\delta}_2)$ is correctly specified.

Why is the second step regression identified?

- ▶ If \mathbf{x}_i contains some variables that are *not* in \mathbf{x}_{1i} , we have an **exclusion restriction**: i.e. there are variables that affect participation but not outcomes.
- ▶ Even if there are no exclusion restrictions, λ is nonlinear so $\lambda(\mathbf{x}'_{1i}\boldsymbol{\delta}_2)$ will not be perfectly co-linear with \mathbf{x}_{1i} .
- ▶ Without exclusion restrictions identification comes *solely* from nonlinearity in λ .
- ▶ Depending on the values where it is evaluated, λ can be *close* to linear, leading to very imprecise estimates unless you have an exclusion restriction.