Introductory Graduate Econometrics

Francis J. DiTraglia

University of Oxford

This Version: 2021-01-17 21:38
Latest Version: `https://economictricks.com`

Abstract

These are lecture notes to accompany my lectures as part of the first-year MPhil course in econometrics at Oxford. For more details, see the course website: economictricks.com. If you spot any typos, send me a Github pull request or an email: francis.ditraglia@economics.ox.ac.uk.

Contents

Chapter 1

Maximum Likelihood Estimation Under Mis-specification

In your introductory lectures on probability and statistics, you studied the method of **maximum likelihood estimation** (MLE). As you may recall, MLE posits a distribution $f_{\boldsymbol{\theta}}$ for a random vector $\mathbf{y}$ in terms of an unknown parameter value $\boldsymbol{\theta}$ and uses an observed sample $\mathbf{y}_1, \ldots, \mathbf{y}_N$ to estimate $\boldsymbol{\theta}$ by maximizing the likelihood function $L(\boldsymbol{\theta})$ or equivalently the log-likelihood function $\ell(\boldsymbol{\theta})$. Most introductory treatments of MLE assume that $f_{\boldsymbol{\theta}}$ is **correctly specified**, in other words that each observation $\mathbf{y}_i$ is in fact a draw from this distribution. But what if this assumption is *wrong* and our model is in fact mis-specified? In this chapter we examine the general theory for MLE under mis-specification and apply it to a simple example.

1.1 Review: MLE for a Poisson Distribution

We'll begin by introducing the running example for the chapter: MLE for the parameter of a Poisson distribution. This material should be familiar from your lectures on introductory probability earlier in the academic year, but a little review never hurts! If you are confident that you already understand this material, skip to the next section. We say that y follows a **Poisson distribution** with parameter θ, written $y \sim \text{Poisson}(\theta)$, if y has probability mass function (pmf)

$$f(y; \theta) = \frac{e^{-\theta}\theta^y}{y!}, \quad y \in \{0, 1, 2, \ldots\}. \tag{1.1}$$

The Poisson distribution is a common model for *count data*, examples of which include the number of patents a firm has obtained, or the number of people who live in a household.[1]

[1] To be clear, I am not claiming that the Poisson distribution is necessarily a good model for these examples: merely that both are instances of *count data*.

To verify that (1.1) is indeed a valid pmf, we need to check that $f(y;\theta)$ is strictly positive for every value in the support set of y and that the pmf sums to one over this set. The first requirement clearly holds for any $y \in \{0, 1, 2, \dots\}$. For the second,

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

using the Taylor series $e^x = \sum_{k=0}^{\infty} x^k/k!$. If $y \sim$ Poisson(θ) then both the mean and variance of y equal θ. For this reason, the parameter θ of a Poisson distribution *must* be non-negative.[2] To help refresh your memory on basic probability calculations, you will prove the result for the variance on your first problem set.

Lemma 1.1. *Suppose that $y \sim$ Poisson(θ). Then, $\mathbb{E}(y) = Var(y) = \theta$.*

Proof of Lemma 1.1. Here I prove $\mathbb{E}(y) = \theta$ only. To prove that $\text{Var}(y) = \theta$, use a similar argument to calculate $\mathbb{E}(y^2)$ and then write $\text{Var}(y) = \mathbb{E}(y^2) - \mathbb{E}(y)^2$. By definition, $\mathbb{E}(y) = \sum_{y=0}^{\infty} y f(y;\theta)$. But since the first term in this sum equals zero, we can write

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

Now, since $y/y! = 1/(y-1)!$, pulling a factor of θ in front of the sum gives

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

Finally, Re-indexing the sum to start at zero rather than one, it follows that

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

since $\sum_{y=0}^{\infty} e^{-\theta}\theta^y/y! = 1$ as shown in our discussion above. □

Suppose we observe a random sample $y_1, y_2, \dots, y_N \sim$ iid Poisson(θ) and wish to estimate the unknown parameter θ. Because the data are iid, the sample likelihood $L_N(\theta)$ and log-likelihood $\ell_N(\theta)$ are given by

$$L_N(\theta) \equiv \prod_{i=1}^{N} \frac{e^{-\theta}\theta^{y_i}}{y_i!}, \quad \ell_N(\theta) = \sum_{i=1}^{N} \left[y_i \log(\theta) - \theta - \log(y_i!)\right]. \tag{1.2}$$

To find the maximum likelihood estimator $\widehat{\theta}$ we optimize either of these functions over θ, the unknown parameter.[3] In general it's much easier to work with the log-likelihood, so

[2] If $\theta = 0$, then we have a *degenerate* Poisson distribution with $\mathbb{P}(y = 0) = 1$.

[3] Because log is a monotonic function, the solutions to both problems are the same.

we define the solution to the maximum likelihood problem as

$$\widehat{\theta} \equiv \underset{\theta \in [0,+\infty)}{\arg\max}\, \ell_N(\theta).$$

Differentiating the log-likelihood from (1.2),

$$\frac{d}{d\theta}\ell_N(\theta) = \sum_{i=1}^{N} \frac{d}{d\theta}\left[y_i \log(\theta) - \theta - \log(y_i!)\right] = \sum_{i=1}^{N}\left(\frac{y_i}{\theta} - 1\right) \tag{1.3}$$

$$\frac{d^2}{d\theta^2}\ell_N(\theta) = \sum_{i=1}^{N} \frac{d}{d\theta}\left(\frac{y_i}{\theta} - 1\right) = \sum_{i=1}^{N}\left(-\frac{y_i}{\theta^2}\right) \tag{1.4}$$

we see that $\ell_N(\theta)$ is strictly concave.[4] Thus, the first order condition identifies the unique global maximum, in particular:

$$\sum_{i=1}^{N}\left(y_i/\widehat{\theta} - 1\right) = 0 \quad \Longleftrightarrow \quad \widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y} \tag{1.5}$$

so the maximum likelihood estimator for θ turns out to be the sample mean $\bar{y}$. Given that $\theta = \mathbb{E}(y)$ this seems reasonable enough! Notice that, because $\bar{y} \geq 0$ our MLE $\widehat{\theta}$ automatically respects the constraint $\theta \in [0, +\infty)$. We will return to this example and the associated calculations below.

1.2 The Kullback-Leibler (KL) Divergence

Our goal in this chapter is to answer the following question: how can we interpret MLE if our model is incorrect? For example, what if $y_1, \ldots, y_N$ do *not* really come from a Poisson distribution, but we apply Poisson MLE nonetheless? Answering this question will require us to think about models as *approximations* rather than reality. In this section we will describe a way to quantify how well a parametric model $f(\mathbf{y}; \boldsymbol{\theta})$ approximates $p_o(\mathbf{y})$, the unknown true density or pmf of a random vector $\mathbf{y}$—the **Kullback-Leibler divergence**.

Definition 1.1 (Kullback Leibler Divergence)**.** Let $\mathbf{y}$ be a RV with true density $p_o(\mathbf{y})$ and let $f_{\boldsymbol{\theta}}$ be a possibly mis-specified parametric model for $\mathbf{y}$. The quantity

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

is called the Kullback-Leibler divergence from p_o to $f_{\boldsymbol{\theta}}$. If $\mathbf{y}$ is discrete, the integral is replaced with a sum and p_o is a pmf rather than a density.

[4] Remember that y_i cannot be negative. Unless all observations y_i are equal to zero, the second derivative of the log likelihood function is strictly negative.

Notice that the expectation in the definition of the KL divergence is taken with respect to the true density or pmf $p_o(\mathbf{y})$. To make this clearer, notice that we can express it as

$$\mathrm{KL}(p_o; f_{\boldsymbol{\theta}}) = \mathbb{E}[h(\mathbf{y})] = \int h(\mathbf{y}) p_o(\mathbf{y})\, d\mathbf{y}, \quad h(\mathbf{y}) \equiv \log\left[\frac{p_o(\mathbf{y})}{f(\mathbf{y};\boldsymbol{\theta})}\right].$$

In other words, the KL divergence is simply an expectation of a *function* h of the random vector $\mathbf{y}$. It just so happens that h involves $f(\mathbf{y};\boldsymbol{\theta})$ and $p_o(\mathbf{y})$, which is perfectly fine because both of these are themselves functions of $\mathbf{y}$. The KL divergence has three important properties. First, it is **asymmetric**:

$$\mathrm{KL}(p_o; f_{\boldsymbol{\theta}}) \neq \mathrm{KL}(f_{\boldsymbol{\theta}}; p_o).$$

This is apparent from Definition 1.1: we end up with a totally different integral if we interchange the roles of p_o and $f_{\boldsymbol{\theta}}$.[5] Second, the KL divergence is non-negative and equals zero if and only if the model and the true distribution are identical.

Lemma 1.2. *$KL(p_o; f_{\boldsymbol{\theta}}) \geq 0$ with equality if and only if $p_o = f_{\boldsymbol{\theta}}$.*

Proof of Lemma 1.2. This proof relies on Jensen's Inequality: if φ is a convex function then $\varphi(\mathbb{E}[y]) \leq \mathbb{E}[\varphi(y)]$, with equality if and only if φ is linear or y is constant. Now, since log is concave, $(-\log)$ is convex. It follows that

$$\begin{aligned}
\mathbb{E}\left[\log\left\{\frac{p_o(\mathbf{y})}{f(\mathbf{y};\boldsymbol{\theta})}\right\}\right] &= \mathbb{E}\left[-\log\left\{\frac{f(\mathbf{y};\boldsymbol{\theta})}{p_o(\mathbf{y})}\right\}\right] \geq -\log\left\{\mathbb{E}\left[\frac{f(\mathbf{y};\boldsymbol{\theta})}{p_o(\mathbf{y})}\right]\right\} \\
&= -\log\left\{\int \frac{f(\mathbf{y};\boldsymbol{\theta})}{p_o(\mathbf{y})} \cdot p_o(\mathbf{y})\, d\mathbf{y}\right\} = -\log\left\{\int f(\mathbf{y};\boldsymbol{\theta})\, d\mathbf{y}\right\} \\
&= -\log(1) = 0.
\end{aligned}$$

□

As seen from 1.2, the KL divergence cannot be negative. In fact it can be **infinite**. The integral in 1.1 is only taken over values in the support set of $\mathbf{y}$, i.e. values for which $p_o(\mathbf{y}) \neq 0$. If $f_{\boldsymbol{\theta}}$ equals zero when evaluated at one of these values, the KL divergence is infinitely large. In other words: if our model $f_{\boldsymbol{\theta}}$ *rules out* values of $\mathbf{y}$ that can actually occur in reality, the KL divergence is no longer well-defined. This problem is easy to avoid. We simply need to ensure that our model has the same support set as the real data. So if y is a count that could possibly take on any value in $\{0, 1, 2, \dots\}$, our model should not rule out any of these values. With this caveat out of the way, we will assume throughout the rest of our discussion that the KL divergence is finite.

The third important property of the KL divergence concerns its relationship to maxi-

[5] The reason that we call it the KL *divergence* rather than the KL *distance* is because a distance function, aka a metric, must be symmetric, whereas the KL divergence is not.

mum likelihood estimation. Since the log of a ratio equals the difference of the logarithms,

$$\mathbb{E}\left[\log\left\{\frac{p_o(\mathbf{y})}{f(\mathbf{y};\boldsymbol{\theta})}\right\}\right] = \mathbb{E}\left[\log p_o(\mathbf{y})\right] - \mathbb{E}\left[\log f(\mathbf{y};\boldsymbol{\theta})\right]. \tag{1.6}$$

The first term in (1.6), $\mathbb{E}[\log p_o(\mathbf{y})]$ does not depend on f or $\boldsymbol{\theta}$. Regardless of which parametric model we consider, this term is fixed. The second term in (1.6), $\mathbb{E}[\log f(\mathbf{y};\boldsymbol{\theta})]$, is called the **expected log-likelihood**. This term depends both on the choice of parametric model f and the value of the parameter vector $\boldsymbol{\theta}$. Because (1.6) takes the form

$$\text{KL} = \text{Constant} - \mathbb{E}[\text{log-likelihood}]$$

to minimize the KL divergence, it suffices to *maximize* the expected log-likelihood.[6] Thus the value of $\boldsymbol{\theta}$ that maximizes the expected log-likelihood is the value of $\boldsymbol{\theta}$ that minimizes the KL divergence. We call this the **pseudo-true** parameter value, denoted $\boldsymbol{\theta}_o$.

Definition 1.2 (Pseudo-true Parameter Value)**.** Let $p_o(\mathbf{y})$ be the true density or pmf of a random vector $\mathbf{y}$ and $f_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, be a possibly mis-specified parametric model for $\mathbf{y}$. Then we call $\boldsymbol{\theta}_o$ defined by

$$\boldsymbol{\theta}_o \equiv \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \text{KL}(p_o, f_{\boldsymbol{\theta}}) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}[\log f(\mathbf{y};\boldsymbol{\theta})]$$

the pseudo-true parameter value under the model $f_{\boldsymbol{\theta}}$.

The intuition behind the idea of a pseudo-true parameter values is as follows. If our model were correctly specified, then the value of $\boldsymbol{\theta}$ that maximized the expected log-likelihood would be the *true* value of $\boldsymbol{\theta}$: the parameter value of the distribution from which the data were actually drawn. When our model is mis-specified, however, there is no choice of $\boldsymbol{\theta}$ in the parameter space $\boldsymbol{\Theta}$ such that f_θ coincides with the true distribution. There is, however, a choice of $\boldsymbol{\theta}$ that makes our model $f_{\boldsymbol{\theta}}$ *as close as possible* to the truth, p_o, where we define "closeness" as minimum KL divergence. We call this parameter value the pseudo-true value.

The best way to understand this is with an example. Suppose that y is a discrete RV with support set $\{0, 1, 2\}$ and probability mass function p_o where $p_o(0) = 2/5$, $p_o(1) = 1/5$, and $p_o(2) = 2/5$. Since a Poisson random variable can take on any value in $\{0, 1, 2, \dots\}$ we can tell immediately that y does *not* follow a Poisson distribution.[7] Suppose that we nevertheless choose as our parametric model for y a Poisson(θ) distribution. Our question is: which choice of θ makes the Poisson(θ) distribution as close as possible to the true pmf p_o of y? To really drive home the point about KL divergence

[6] A very poor choice of the parameter $\boldsymbol{\theta}$ will result in a very large *negative* value for the expected log-likelihood, and hence a very large *positive* value for the KL divergence.

[7] Here the support set of our model is *larger* than the true support set of y. This is fine: the problem of an infinite KL divergence can only arise when our model *excludes* part of the true support set of $\mathbf{y}$.

and expected log-likelihood, we will carry out this calculation two different ways: first by working out the full expression for the KL divergence, and then more simply by directly maximizing the expected log-likelihood.

We can calculate the KL divergence for this example as follows. For $\mathbb{E}[\log p_o(y)]$,

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

which, as expected, does not depend on the parameter θ. For $\mathbb{E}[f(y;\theta)]$,

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

which, as expected, *does* depend on θ. Combining these, we find that the KL divergence from the true pmf p_o to our Poisson(θ) model is given by

$$\mathrm{KL}(p_o, f_\theta) = \mathbb{E}[\log p_o(y)] - \mathbb{E}[\log f(y;\theta)] = (\text{Constant}) + \theta - \log(\theta).$$

This is a strictly convex function of θ, so the first-order condition $1 - 1/\theta = 0$ is necessary and sufficient for a global minimum. Hence we obtain $\theta_o = 1$. The Poisson distribution that best approximates p_o, in the sense of minimizing the KL divergence, is a Poisson(1).

At the risk of beating a dead horse, let's try this calculation again. This time, rather than minimizing the KL divergence, we'll maximize the expected log-likelihood, using the equivalence from Definition 1.2. The expected log-likelihood is given by

$$\mathbb{E}[\log f(y;\theta)] = \mathbb{E}[y\log(\theta) - \theta - \log(y!)] = \log(\theta)\mathbb{E}[y] - \theta - \mathbb{E}[\log(y!)]$$

but since $\mathbb{E}[\log(y!)]$ does not depend on θ, we can write this as

$$\mathbb{E}[\log f(y;\theta)] = (\text{Constant}) + \log(\theta)\mathbb{E}[y] - \theta.$$

Since $\mathbb{E}[y]$ is non-negative, this is a strictly *concave* function of θ and hence the first-order condition $\mathbb{E}[y]/\theta - 1 = 0$ is necessary and sufficient for a global maximum. Thus we obtain $\theta_o = \mathbb{E}[y]$. For the specified p_o we have: $\mathbb{E}[y] = 0 \cdot \frac{1}{5} + 1 \cdot \frac{2}{5} + 2 \cdot \frac{2}{5} = 1$ and thus the two approaches give the same result. The advantage of the second approach over the first is that it shows us that *no matter* what the true distribution p_o actually is, the pseudo-true parameter value under a Poisson model is $\theta_o = \mathbb{E}[y]$.

1.3 Asymptotic Theory for Mis-specified MLE

As shown in section 1.1, under an iid Poisson(θ) model, the MLE is $\widehat{\theta} = \bar{y}$. And as shown in section 1.2, when this model is mis-specified, the value of θ that minimizes the KL divergence is $\theta_o = \mathbb{E}[y]$, the population mean of y.[8] Now, since $y_1, \ldots, y_N$ are iid draws from a distribution with finite mean and variance, we can apply the weak law of large numbers and the central limit theorem to obtain

$$\bar{y} \to_p \mathbb{E}[y], \quad \text{and} \quad \sqrt{N}(\bar{y} - \mathbb{E}[y]) \to_d \mathcal{N}(0, \text{Var}(y)) \tag{1.7}$$

or writing the same thing using more suggestive notation,

$$\widehat{\theta} \to_p \theta_o, \quad \text{and} \quad \sqrt{N}(\widehat{\theta} - \theta_o) \to_d \mathcal{N}(0, \text{Var}(y)).$$

Crucially, neither of these asymptotic results relies upon the assumption that y_i is actually a draw from a Poisson distribution! Indeed, they continue to hold *regardless* of the true distribution p_o from which the y_i were drawn, subject to mild regularity conditions.[9] So, at least in this example, the maximum likelihood estimator $\widehat{\theta}$ turns out to be a consistent and asymptotically normal estimator of θ_o, the pseudo-true parameter value defined in Definition 1.2. It turns out that this is *not a coincidence*. Under mild regularity conditions, maximum likelihood estimators are consistent for the pseudo-true parameter value, and asymptotically normal.

Theorem 1.1. *Suppose that* $\mathbf{y}_1, \ldots, \mathbf{y}_N \sim$ *iid* p_o *and let* $\widehat{\boldsymbol{\theta}}$ *denote the MLE for* $\boldsymbol{\theta}$ *under the possibly mis-specified model* $f(\mathbf{y}; \boldsymbol{\theta})$*. Then, under mild regularity conditions:*

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

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

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* $\mathbf{K} \equiv Var\left[\dfrac{\partial \log f(\mathbf{y}; \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}}\right]$.

Theorem 1.1 provides an interpretation of MLE when we acknowledge that our models are only an *approximation* of reality. It also provides a way of computing standard errors that are *robust* to mis-specification of our model. In particular, let

$$\widehat{\mathbf{J}} \equiv -\frac{1}{N}\sum_{i=1}^{N} \frac{\partial^2 \log f(\mathbf{y}_i; \widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}, \quad \widehat{\mathbf{K}} \equiv \frac{1}{N}\sum_{i=1}^{N} \left[\frac{\partial \log f(\mathbf{y}_i; \widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\right]\left[\frac{\partial \log f(\mathbf{y}_i; \widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\right]'. \tag{1.8}$$

[8] It bears repeating that all expectations in this chapter are taken with respect to the *true* distribution p_o, e.g. $\mathbb{E}[y] = \sum_{\text{all } y} y\, p_o(y)$ for a discrete RV.

[9] There are two conditions. First, we require that the support of y is a subset of the support set of a Poisson RV, so that the KL divergence is finite. Without this condition, the pseudo-true value is undefined. Second, to apply the law of large numbers and central limit theorem, we require that the first and second moments of y exist and are finite.

Under regularity conditions, the **robust asymptotic variance matrix estimator** $\widehat{\mathbf{J}}^{-1}\widehat{\mathbf{K}}\widehat{\mathbf{J}}^{-1}$ is consistent for $\mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1}$. But isn't $\widehat{\mathbf{K}}$ from (1.8) missing a term? The sample variance of $\mathbf{x}$ is given by $\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right) - (\bar{\mathbf{x}}\bar{\mathbf{x}}')$ where $\bar{\mathbf{x}} = N^{-1}\sum_{i=1}^{N}\mathbf{x}_i$ but the formula for $\widehat{\mathbf{K}}$ doesn't contain the "$\bar{\mathbf{x}}\bar{\mathbf{x}}'$" component. This isn't a mistake: since $\widehat{\boldsymbol{\theta}}$ is the solution to the MLE first-order condition, the missing term is precisely equal to zero. Theorem 1.1 combined with the robust asymptotic variance matrix estimator justifies the approximation

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

from which we can construct tests and confidence intervals for $\boldsymbol{\theta}_o$. If, miraculously, our model turns out to be *correctly specified*, we obtain the following corollary to Theorem 1.1.

Corollary 1.1. *If $p_o(\mathbf{y}) = f(\mathbf{y};\boldsymbol{\theta}_o)$, then under the conditions of Theorem 1.1,*

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

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

where $\mathbf{J}$ is as defined in Theorem 1.1.

You are likely to have encountered some form of Corollary 1.1 in your earlier exposure to MLE. Its proof relies on two lemmas. First, by Lemma 1.2, if $p_o(\mathbf{y}) = f(\mathbf{y};\boldsymbol{\theta}_o)$ then $KL(p_o; f_{\boldsymbol{\theta}})$ equals *zero* at $\boldsymbol{\theta} = \boldsymbol{\theta}_o$. In other words, under correct specification, the pseudo-true parameter value is simply the *true* parameter value. Second, by the **information matrix equality**, in a correctly specified maximum likelihood model we have $\mathbf{K} = \mathbf{J}$. A proof of this result appears in section 1.4 below. Using this fact, we obtain $\mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1} = \mathbf{J}^{-1}$. Since $\widehat{\mathbf{J}}^{-1}$, from (1.8), is a consistent estimator of $\mathbf{J}^{-1}$, under correct specification we can carry out inference using $\widehat{\mathbf{J}}^{-1}$ in place of the more complicated robust asymptotic variance matrix estimator.

To clarify the distinction between Theorem 1.1 and Corollary 1.1, we will return one final time to the Poisson MLE example from above. Recall that $\theta_o = \mathbb{E}[y]$, $\widehat{\theta} = \bar{y}$, and

$$\log f(y;\theta) = y\log(\theta) - \theta - \log(y!), \quad \frac{d}{d\theta}\log f(y;\theta) = \frac{y}{\theta} - 1, \quad \frac{d^2}{d\theta^2}\log f(y;\theta) = -\frac{y}{\theta^2}.$$

Substituting these into the definitions of $\mathbf{J}$ and $\mathbf{K}$ from Theorem 1.1, we obtain

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

both of which are *scalars* rather than matrices since θ is one-dimensional. Similarly,

specializing the estimators $\widehat{\mathbf{J}}$ and $\widehat{\mathbf{K}}$ from (1.8) to the Poisson MLE example, we obtain

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

where $s_y^2 \equiv N^{-1}\sum_{i=1}^{N}(y_i-\bar{y})^2$ and $\bar{y} \equiv N^{-1}\sum_{i=1}^{n} y_i$. Now consider two cases. Suppose first that the Poisson models is correct. Then, since

$$J^{-1} = (1/\mathbb{E}[y])^{-1} = \mathbb{E}[y]$$

we obtain $\sqrt{N}(\widehat{\theta}-\theta_o) \to_d \mathcal{N}(0,\mathbb{E}[y])$ justifying the approximation $\widehat{\theta} \approx \mathcal{N}(\theta_o,\bar{y})$. Recall from above that the mean of a Poisson random variable equals the variance, so this result makes perfect sense *if our model is correctly specified*. Now suppose that the Poisson model is incorrect. Since

$$J^{-1}KJ^{-1} = \left(\frac{1}{\mathbb{E}[y]}\right)^{-1}\left[\frac{\text{Var}(y)}{\mathbb{E}[y]^2}\right]\left(\frac{1}{\mathbb{E}[y]}\right)^{-1} = \text{Var}(y)$$

we obtain $\sqrt{N}(\widehat{\theta}-\theta_o) \to_d \mathcal{N}(0,\text{Var}(y))$ justifying the approximation $\widehat{\theta} \approx \mathcal{N}(\theta_o, s_y^2)$. This is exactly the same result that we obtained by proceeding "from first principles" in (1.7), i.e. without appealing to Theorem 1.1. If the Poisson model is incorrect, then there is no reason to suppose that $\text{Var}(y) = \mathbb{E}(y)$. If we make this assumption when it is false, our tests and confidence intervals will be incorrect. The robust asymptotic variance calculation avoids this problem: regardless of whether the model is correct or incorrect, it gives the right answer.

An interesting feature of the Poisson MLE example is that, regardless of whether the model is correct or incorrect, we are fundamentally estimating the *same parameter*: θ_o is the mean of y regardless of whether y is a Poisson(θ_o) random variable or not. We will see a very similar phenomenon emerge in our next chapter, when we extend the Poisson MLE example to consider **Poisson regression**: a regression model for count data.

1.4 Appendix: The Information Matrix Equality

Lemma 1.3 (Information Matrix Equality)**.** *If* $p_o(\mathbf{y}) = f(\mathbf{y};\boldsymbol{\theta}_o)$*, then* $\mathbf{K} = \mathbf{J}$ *where*

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

Proof of Lemma 1.3. This proof assumes that $\mathbf{y}$ is a continuous random vector. For

a discrete random vector, simply replace the integrals in the final step with sums.

We begin by finding a simpler expression for the matrix $\mathbf{K}$. Using the fact that $\text{Var}(\mathbf{x}) = \mathbb{E}[\mathbf{x}\mathbf{x}'] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]'$, we can write

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

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

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

so it suffices to show that

$$-\mathbb{E}\left[\frac{\partial^2 \log f(\mathbf{y};\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right] = \mathbb{E}\left[\left\{\frac{\partial \log f(\mathbf{y};\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}}\right\}\left\{\frac{\partial \log f(\mathbf{y};\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}}\right\}'\right]. \tag{1.9}$$

The remainder of the proof is devoted to establishing (1.9). We do this element-by-element, showing the equality of the (i,j) elements of the two matrices. By the chain and product rules, along with some algebra, we see that

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

Multiplying this result by -1, gives

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

Evaluating the result at $\boldsymbol{\theta}_o$, and taking expectations of both sides of the equality,

$$-\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\mathbf{y};\boldsymbol{\theta}_o)\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta_i}\log f(\mathbf{y};\boldsymbol{\theta}_o)\frac{\partial}{\partial\theta_j}\log f(\mathbf{y};\boldsymbol{\theta}_o)\right] - \mathbb{E}\left[\frac{1}{f(\mathbf{y};\boldsymbol{\theta}_o)}\cdot\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(\mathbf{y};\boldsymbol{\theta}_o)\right].$$

Accordingly, it remains only to show that

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

We have not yet used the assumption that the model is correctly specified. Now is the

time to do so. Since, $p_o(\mathbf{y}) = f(\mathbf{y}; \boldsymbol{\theta}_o)$, we can re-write the expectation as an integral

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{f(\mathbf{y};\boldsymbol{\theta}_o)} \cdot \frac{\partial^2}{\partial\theta_i \partial\theta_j} f(\mathbf{y};\boldsymbol{\theta}_o)\right] &\equiv \int_{\mathcal{Y}} \left[\frac{1}{f(\mathbf{y};\boldsymbol{\theta}_o)} \cdot \frac{\partial^2}{\partial\theta_i \partial\theta_j} f(\mathbf{y};\boldsymbol{\theta}_o)\right] p_o(\mathbf{y})\, d\mathbf{y} \\
&= \int_{\mathcal{Y}} \left[\frac{1}{f(\mathbf{y};\boldsymbol{\theta}_o)} \cdot \frac{\partial^2}{\partial\theta_i \partial\theta_j} f(\mathbf{y};\boldsymbol{\theta}_o)\right] f(\mathbf{y};\boldsymbol{\theta}_o)\, d\mathbf{y} \\
&= \int_{\mathcal{Y}} \frac{\partial^2}{\partial\theta_i \partial\theta_j} f(\mathbf{y};\boldsymbol{\theta}_o)\, d\mathbf{y} \\
&= \frac{\partial^2}{\partial\theta_i \partial\theta_j} \int_{\mathcal{Y}} f(\mathbf{y};\boldsymbol{\theta}_o)\, d\mathbf{y}.
\end{aligned}$$

where $\mathcal{Y}$ denotes the support set of $\mathbf{y}$. But since $f(\mathbf{y};\boldsymbol{\theta}_o)$ its integral over $\mathcal{Y}$ with respect to $\mathbf{y}$ equals one. The result follows because the second cross-partial derivative of 1 is indeed zero! □

Chapter 2

Poisson Regression

In this chapter we begin by expanding the discussion from chapter 1 to consider *conditional* maximum likelihood estimation based on a model of the form $f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$ where $\mathbf{x}$ is a vector of observed covariates. This extension is important because it allows us to study *regression models* under mis-specification. We apply these results to a popular model for count data: the **Poisson regression** model.

2.1 How to Predict Count Data?

Suppose our goal is to build a model that predicts the number of children a woman has, y, using a set of covariates $\mathbf{x} \equiv$\{years of schooling, age, married, etc.\}. In this example, y is a **count variable**, i.e. $y \in \{0, 1, 2, \dots\}$. How could we go about making our predictions?

Before discussing what makes count data special, let's think about this problem in general. Suppose we want to predict y using $\mathbf{x}$. What function $\varphi(\mathbf{x})$ should we use to construct our prediction $\widehat{y}$ of the outcome y? To answer this question, we first need to be clear about what counts as a "good" prediction and what counts as a bad one. The most common way of doing this is by specifying a **loss function** $\mathscr{L}(y, \widehat{y})$ that tells us the *loss* we incur from predicting $\widehat{y}$ when the truth is y.[1] There are many possible loss functions that we could choose. Ideally we would choose one that's specific to the application we have in mind. For example, if the head of Ocado wants to predict demand for home grocery delivery, the loss function should weigh the costs of having idle trucks and drivers against the opposing costs of being unable to respond to an unexpected surge in demand. Without a particular application in mind, we need to choose something, and a very common choice is **squared error loss**, namely $\mathcal{L}(y, \widehat{y}) = (y - \widehat{y})^2$. Given a choice of loss function, we can choose a predictor to minimize **expected loss**.[2] The function φ that minimizes expected squared error loss, $\mathbb{E}\left[\{y - \varphi(\mathbf{x})\}^2\right]$, is called the

[1] A loss function is simply the negative of a utility function: $\mathscr{L}(y, \widehat{y}) = -U(y, \widehat{y})$. In words: a utility function tells us the *gain* from predicting $\widehat{y}$ when the truth is y.

[2] This is just like maximizing expected utility, but with the sign reversed.

minimum **mean squared error** (MSE) predictor. As you may recall from your lectures on probability and statistics, this function turns out to be the conditional expectation of y given $\mathbf{x}$, as shown in the following lemma.

Lemma 2.1 (Minimum MSE Predictor)**.** *The function* $\mu(\mathbf{x}) \equiv \mathbb{E}(y|\mathbf{x})$ *minimizes the mean squared error* $\mathbb{E}\left[\{y - \varphi(\mathbf{x})\}^2\right]$ *over all possible predictors* $\varphi(\cdot)$*.*

Proof of Lemma 2.1. At first glance, this appears to be a very challenging problem because φ is a *function* rather than a scalar or a vector. But in fact, there's a simple way to prove this result without using any advanced mathematics. First we use the oldest trick in the book. To obtain an expression that contains $\mu(\mathbf{x}) \equiv \mathbb{E}(y|\mathbf{x})$, we add and subtract this quantity, yielding:

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

The second term in the preceding expression is a bit unwieldy. Let's see if we can simplify it. By iterated expectations,

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

since $\mu(\mathbf{x}) \equiv \mathbb{E}(y|\mathbf{x})$. Because this middle term turns out to be zero, we are left with only the first and third terms from our expression for MSE from above, namely

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

Consider first $\mathbb{E}\left[\{y - \varphi(\mathbf{x})\}^2\right]$. Because this term does not involve φ, it is a constant from the perspective of our optimization problem and hence can be ignored. All that remains is to minimize $\mathbb{E}\left[\{\varphi(\mathbf{x}) - \mu(\mathbf{x})\}^2\right]$ over φ. Notice that if we set $\varphi(\mathbf{x}) = \mu(\mathbf{x})$, this term becomes *exactly zero*. This is the smallest possible value we can achieve, since $\mathbb{E}[(\text{something})^2]$ cannot be negative. Therefore, $\mu(\mathbf{x})$ is the unique solution to our original optimization problem. □

The minimum mean squared error predictor of y is $\mathbb{E}(y|\mathbf{x})$, but in practice it is usually very difficult to estimate this function. In principle, it depends on $\mathbf{x}$ in a completely arbitrary way. If $\mathbf{x}$ contains more than a small number of regressors, we would need a truly gargantuan dataset to be able to learn the conditional mean function.[3] The

[3]Here I allude to what is called the **curse of dimensionality** in the machine learning and non-parametric econometrics literatures.

typical response to this difficulty is to *restrict* the set of functions φ that we consider for predicting y. For example, we might ask: what is the best function of the form $\mathbf{x}'\boldsymbol{\theta}$ for predicting y from $\mathbf{x}$? Under squared error loss, the answer to this question is called the **minimum MSE linear predictor**. As you have likely seen in your econometrics lectures from earlier in the year, the solution to this constrained problem is the linear regression predictor.

Lemma 2.2 (Minimum MSE Linear Predictor)**.** *Provided that* $\mathbb{E}[\mathbf{x}\mathbf{x}']$ *is invertible, the parameter vector that uniquely minimizes* $\mathbb{E}\left[(y - \mathbf{x}'\boldsymbol{\theta})^2\right]$ *is* $\boldsymbol{\beta} \equiv \mathbb{E}\left[\mathbf{x}\mathbf{x}'\right]^{-1}\mathbb{E}[\mathbf{x}y]$.

Proof. Recall the following facts from matrix differentiation:

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

Taking the first order condition and re-arranging,

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

If the conditional mean function $\mathbb{E}(y|\mathbf{x})$ is a linear function, then the solutions from Lemmas 2.1 and 2.1 coincide. While it is unlikely that the conditional mean function is *exactly* linear, it may be at least *approximately* linear, $\mathbb{E}(y|\mathbf{x}) \approx \mathbf{x}'\boldsymbol{\beta}$, in which case linear regression should provide reasonably accurate predictions.

So what's special about count data? If $y \in \{0, 1, 2, \dots\}$, then we know in advance that we should never predict a negative value. This presents a problem for linear regression: depending on the value of $\mathbf{x}$ that we observe, $\mathbf{x}'\boldsymbol{\beta}$ could be negative. How could we avoid this problem? One idea would be to consider a **log-linear model** of the form $\log(y) = \mathbf{x}'\beta + \varepsilon$, i.e. to carry out linear regression with y on the *log scale*. This solves the problem of negative predictions since it is perfectly fine for $\log(y)$ to be negative. At the same time, it introduces a new one: if $y = 0$, which is entirely possible for a count variable, then $\log(y) = -\infty$. To avoid these difficulties, we will consider parametric models for count data that are **nonlinear in parameters**. In particular, we will assume that $\mathbb{E}(y|\mathbf{x}) = m(\mathbf{x};\boldsymbol{\beta})$ where m is a known function of and unknown parameter vector β, and m cannot take on negative values. This means that m *cannot* be a linear function of $\boldsymbol{\beta}$.[4] In the discussion that follows, we will focus on a choice of m that is common in practice, namely $m(\mathbf{x};\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$. As required, this is strictly positive and hence nonlinear in parameters. Making appropriate changes to the notation, everything that we discuss below goes through for alternative choices of m.

Assumption 2.1. $\mathbb{E}(y|\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta}_o)$

[4] We say that a function m is **linear** if and only if $m(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2) = m(\boldsymbol{\beta}_1)+m(\boldsymbol{\beta}_2)$ and $m(a\boldsymbol{\beta}_1) = am(\boldsymbol{\beta}_1)$ for any vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and any scalar a. Suppose that $m(\boldsymbol{\beta}) > 0$. Then, taking $a = -1$, $m(-\boldsymbol{\beta})$ must be negative. Thus, a function that only takes on positive values cannot be linear.

So how can we estimate $\boldsymbol{\beta}_o$? Under Assumption 2.1, $\boldsymbol{\beta}_o$ minimizes $\mathbb{E}\left[\{y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})\}^2\right]$ over all parameter vectors $\boldsymbol{\beta}$ by Lemma 2.1. Converting this to an analogous problem for our sample dataset, one idea might be to use the **nonlinear least squares estimator**

$$\widehat{\beta}_{NLLS} \equiv \arg\min_{\beta} \sum_{i=1}^{N} \{y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})\}^2.$$

Another approach is **Poisson regression**, in which $\widehat{\boldsymbol{\beta}}$ is defined as the **conditional ML estimator** for $\boldsymbol{\beta}_o$ under the model $y_i|\mathbf{x}_i \sim$ independent Poisson$\big(\exp(\mathbf{x}_i'\boldsymbol{\beta}_o)\big)$. While either approach is reasonable, we will mainly concern ourselves with Poisson regression in this chapter.[5] Before beginning our discussion of Poisson regression, we explain how the results on misspecified maximum likelihood estimation can be extended to models that condition on covariates $\mathbf{x}$.

2.2 Conditional Maximum Likelihood Estimation

In chapter 1, we considered models of the form $f(\mathbf{y};\boldsymbol{\theta})$ for the unconditional distribution of a random vector $\mathbf{y}$ in terms of an unknown parameter vector $\boldsymbol{\theta}$. If we were willing to specify a model of the form $f(y,\mathbf{x};\boldsymbol{\theta})$, we could immediately apply our earlier results to study Poisson regression by taking "$\mathbf{y}$" to be the vector $(y,\mathbf{x})$. Notice, however, that this would require us to specify a model for the *joint* distribution of $(y,\mathbf{x})$. By the definition of a conditional density/pmf,

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

Thus, modeling $(y,\mathbf{x})$ jointly would require us to model the *marginal* distribution of the covariates $\mathbf{x}$. In regression applications, however, it is typically much more natural to model the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$: we're not interested in the marginal distribution of $\mathbf{x}$ and coming up with a good model for it could be challenging. In **conditional MLE**, we ignore the marginal distribution of $\mathbf{x}$, and specify a parametric model $f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$ for the unknown true conditional distribution $p_o(\mathbf{y}|\mathbf{x})$. Given a sample of iid observations $\{\mathbf{x}_i,\mathbf{y}_i\}_{i=1}^N$, the conditional maximum likelihood estimator is given by

$$\widehat{\boldsymbol{\theta}} \equiv \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{N}\sum_{i=1}^{N} \log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}).$$

In the conditional MLE case, the **pseudo-true parameter value** is

$$\boldsymbol{\theta}_o \equiv \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}\left[\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})\right] \tag{2.1}$$

[5] The chapter concludes with some further thoughts on the relative merits of the NLLS approach.

where the expectation is taken over the *joint* distribution $p_o(\mathbf{x}, \mathbf{y})$ of $(\mathbf{x}, \mathbf{y})$. The following are conditional MLE analogues of Theorem 1.1 and Corollary 1.1 from chapter 1.

Theorem 2.1. *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* $f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})$*. 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* $\mathbb{E}\left[\log f(y|\mathbf{x}, \boldsymbol{\theta})\right]$ *over the parameter space* Θ*, and*

(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

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

and all expectations are taken with respect to the true joint distribution $p_o(\mathbf{x}, \mathbf{y})$ *of* $(\mathbf{x}_i, \mathbf{y}_i)$*.*

Corollary 2.1. *Suppose that* $f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_o)$ *is the true conditional distribution of* $\mathbf{y}_i|\mathbf{x}_i$*. Then, under the conditions of Theorem 2.1*

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

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

where $\mathbf{J}$ *is as defined in Theorem 2.1.*

Theorem 2.1 and Corollary 2.1 are a bit more difficult to understand than their unconditional counterparts from chapter 1, so let's take them apart. By iterated expectations,

$$\mathbb{E}\left[\log f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})\right] = \mathbb{E}_{\mathbf{x}}\left\{\mathbb{E}\left[\log f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})|\mathbf{x}\right]\right\}. \tag{2.2}$$

The outer expectation in (2.2) is taken with respect to the marginal distribution of $\mathbf{x}$, while the inner expectation is taken with respect to the true conditional distribution of $\mathbf{y}$ given $\mathbf{x}$, namely $p_o(\mathbf{y}|\mathbf{x})$. Thus, we can write

$$\mathbb{E}[\log f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})|\mathbf{x}] = \int \log f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}) p_o(\mathbf{y}|\mathbf{x})\, d\mathbf{y} \equiv h(\mathbf{x}; \boldsymbol{\theta}) \tag{2.3}$$

introducing the shorthand $h(\mathbf{x}; \boldsymbol{\theta})$ to emphasize the point that the inner expectation is a *function* of $\mathbf{x}$ and $\boldsymbol{\theta}$. Using (2.2) and (2.3), we can re-express the pseudo-true parameter value defined in (2.1) as

$$\boldsymbol{\theta}_o = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_{\mathbf{x}}\left[h(\mathbf{x}; \boldsymbol{\theta})\right].$$

This is a slightly less intimidating way of writing the optimization problem. How could we go about solving it? To keep things simple, suppose that our problem is sufficiently

well-behaved that the first-order condition identifies a unique global maximum, and we can exchange expectation and differentiation. Then $\boldsymbol{\theta}_o$ satisfies

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}} \left[h(\mathbf{x}; \boldsymbol{\theta}_o) \right] = \mathbb{E}_{\mathbf{x}} \left[\frac{\partial}{\partial \boldsymbol{\theta}} h(\mathbf{x}; \boldsymbol{\theta}_o) \right] = \mathbf{0}.$$

In words, $\boldsymbol{\theta}_o$ makes the partial derivatives of $h(\mathbf{x}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ *average out* to zero over the distribution of $\mathbf{x}$. For a particular realization $\boldsymbol{x}$ of the random vector $\mathbf{x}$, we may very well have $\frac{\partial}{\partial \theta_j} h(\boldsymbol{x}; \boldsymbol{\theta}_o) \neq 0$. Any realizations at which this derivative is positive, however, are counterbalanced by other realizations at which it is negative. When realizations are weighted by their probabilities, the average derivative equals zero. Here's another way of thinking about it. The maximizer of $h(\mathbf{x}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ may well depend on the value of $\mathbf{x}$ at which we evaluate h. But we can't use a different value of $\boldsymbol{\theta}$ for every realization $\boldsymbol{x}$ of $\mathbf{x}$: we have been asked to find a *single* value of the parameter vector to solve (2.1). This means we have to choose a parameter vector that works well *on average* across all the different "sub-problems" defined by realizations of $\mathbf{x}$, even though it may not be the optimal choice for any them.

But what exactly *are* these sub-problems? Returning to the definition of h from (2.3), we see that they are instances of the *unconditional* maximum likelihood problem from chapter 1. For any fixed realization $\boldsymbol{x}$ of $\mathbf{x}$, $\log f(\mathbf{y}|\mathbf{x} = \boldsymbol{x}; \boldsymbol{\theta})$ is a possibly mis-specified parametric model for $\mathbf{y}$, and $p_o(\mathbf{y}|\mathbf{x} = \boldsymbol{x})$ is the true distribution. This shows us that the pseudo-true parameter value from (2.1) is the value of $\boldsymbol{\theta}$ that minimizes the KL divergence from $p_o(\mathbf{y}|\mathbf{x} = \boldsymbol{x})$ to $f(\mathbf{y}|\mathbf{x} = \boldsymbol{x}; \boldsymbol{\theta})$ *on average* across all realizations of $\mathbf{x}$.[6] Now, suppose that our model is correctly specified, that is $f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}) = p_o(\mathbf{y}|\mathbf{x})$. For a given realization of $\mathbf{x}$, what is the parameter value that minimizes the KL divergence from $p_o(\mathbf{y}|\mathbf{x} = \boldsymbol{x})$ to $f(\mathbf{y}|\mathbf{x} = \boldsymbol{x}; \boldsymbol{\theta})$? Again, for a given value of $\boldsymbol{x}$, we're back to the setting we studied in chapter 1. As shown in Lemma 1.2, the smallest possible value that the KL divergence can take is zero, and this value is attained when the model and true distribution are identical. It follows that, under correct specification, $\boldsymbol{\theta}_o$ is the solution to *all* of the sub-problems. To put it another way, under correct specification we have $\frac{\partial}{\partial \boldsymbol{\theta}} h(\mathbf{x}; \boldsymbol{\theta}_o) = \mathbf{0}$ not just on average, but for *every* value of $\mathbf{x}$. Thus, for conditional maximum likelihood, the pseudo-true parameter equals the *true* parameter under correct specification, just as it did for unconditional maximum likelihood in chapter 1 above.

[6] Recall that minimizing the KL divergence is *equivalent* to maximizing the expected log-likelihood: see section 1.2.

2.3 Poisson Regression: A Model for Count Data

The **Poisson regression model** posits that

$$y_i|\mathbf{x}_i \sim \text{independent Poisson}\big(\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\big), \quad i = 1, \ldots, N \tag{2.4}$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters.[7] Using our results for the Poisson distribution from section 1.1, (2.4) implies that $\mathbb{E}[y_i|\mathbf{x}_i] = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, which is precisely Assumption 2.1. We will maintain this assumption throughout. We will *not*, however, assume that the true conditional distribution of $y_i|\mathbf{x}_i$ is Poisson. Among other things, doing so would require us to assume that $\text{Var}(y_i|\mathbf{x}_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta})$. This is a very unrealistic assumption in most applications, as we will discuss further below. Thus, while maintaining Assumption 2.1, we will treat the Poisson regression model as a possibly mis-specified maximum likelihood model, following the theory from section 2.2.

Substituting $\exp(\mathbf{x}_i\boldsymbol{\beta})$ for θ in (1.1), the log-likelihood for a single observation from the Poisson regression model is given by

$$\ell_i(\boldsymbol{\beta}) \equiv \log f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = y_i(\mathbf{x}_i'\boldsymbol{\beta}) - \exp(\mathbf{x}_i'\boldsymbol{\beta}) - \log(y_i!)\,. \tag{2.5}$$

Because we refer to it so frequently below, both in this chapter and those that follows, it is helpful to have a convenient name and notation for the vector of partial derivatives of $\ell_i(\boldsymbol{\beta})$ with respect to β. We call it the **score vector**, denoted $\mathbf{s}_i(\boldsymbol{\beta})$. Differentiating (2.5), the score vector for the Poisson regression model is given by

$$\mathbf{s}_i(\boldsymbol{\beta}) \equiv \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{x}_i\left[y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})\right] = \mathbf{x}_i u_i(\boldsymbol{\beta}) \tag{2.6}$$

where we define the Poisson regression **residual** $u_i(\boldsymbol{\beta}) \equiv y_i - \exp(\mathbf{x}_i'\boldsymbol{\beta})$. It will similarly be helpful to have a convenient name and notation for the matrix of second derivatives of $\ell_i(\boldsymbol{\beta})$. We call this the **Hessian matrix**, denoted $\mathbf{H}_i(\boldsymbol{\beta})$. Differentiating (2.6),

$$\mathbf{H}_i(\boldsymbol{\beta}) \equiv \frac{\partial \mathbf{s}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = -\exp(\mathbf{x}_i'\boldsymbol{\beta})\,\mathbf{x}_i\mathbf{x}_i'. \tag{2.7}$$

By definition, the conditional MLE for $\boldsymbol{\beta}$ under (2.4) is given by

$$\widehat{\boldsymbol{\beta}} \equiv \underset{\boldsymbol{\beta}\in\mathbb{R}^J}{\arg\max}\ \ell_N(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^J}{\arg\max}\ \frac{1}{N}\sum_{i=1}^{N}\left[y_i(\mathbf{x}_i'\boldsymbol{\beta}) - \exp(\mathbf{x}_i'\boldsymbol{\beta}) - \log(y_i!)\right].$$

[7] Throughout this chapter, we maintain Assumption 2.1. As pointed out above, however, we could replace $\exp(\mathbf{x}'\boldsymbol{\beta})$ with any known parametric function $m(\mathbf{x};\boldsymbol{\beta})$ of $\boldsymbol{\beta}$ that is strictly positive. Subject to minor changes of notation, the key results from this section continue to apply.

By (2.6) and (2.7), however, notice that

$$\frac{\partial \ell_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{H}_i(\boldsymbol{\beta}) = - \left[\frac{1}{N} \sum_{i=1}^{N} \exp(\mathbf{x}_i' \boldsymbol{\beta}) \mathbf{x}_i \mathbf{x}_i' \right].$$

Since the expression on the right hand side is a negative semi-definite matrix regardless of the value of $\boldsymbol{\beta}$, it follows that the sample log-likelihood for a Poisson regression model is a *concave* function.[8] Because any local maximum of a concave function is also a global maximum, the MLE for a Poisson regression model satisfies the first order condition

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

Although no closed-form solution exists, the concavity of the log-likelihood makes it easy to compute $\widehat{\boldsymbol{\beta}}$ numerically, e.g. by using a variant of the Newton-Raphson algorithm.

2.4 Partial Effects in the Poisson Regression Model

For continuous x_j, we call $\frac{\partial}{\partial x_j} \mathbb{E}(y|\mathbf{x})$ the **partial effect** of x_j on the conditional mean function.[9] In a linear model, i.e. $\mathbb{E}(y|\mathbf{x}) = \mathbf{x}'\beta$, the partial effect of x_j is simply β_j. In nonlinear models, however, partial effects typically *vary* with $\mathbf{x}$. Under Assumption 2.1, for example, we have

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

Since $\exp(\cdot) > 0$, the *sign* of the partial effect of x_j is completely determined by that of β_j. The *magnitude*, however, depends on the value of $\mathbf{x}$ at which we evaluate the derivative. In contrast, **relative effects**—ratios of partial effects—do not depend on $\mathbf{x}$, since

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

for this model. When partial effects vary with $\mathbf{x}$, a reasonable way to summarize them is by *averaging* over the distribution of $\mathbf{x}$ in the population. This gives rise to what is called an **average partial effect** or APE for short. For our Poisson regression model

[8] To see why, note that we can write the second derivative matrix of ℓ_N as $-\sum_{i=1}^{N} \mathbf{v}_i \mathbf{v}_i'$ by defining $\mathbf{v}_i \equiv \mathbf{x}_i \exp(\mathbf{x}_i'\boldsymbol{\beta}/2)/\sqrt{N}$. Since each of the matrices $\mathbf{v}_i \mathbf{v}_i'$ is positive semi-definite, so is their sum. Multiplying by -1 converts the result to a negative semi-definite matrix.

[9] For discrete x_j, the partial effect is the difference of $\mathbb{E}(y|\mathbf{x})$ at two different values of x_j.

with an exponential conditional mean function, the APE is

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

To estimate partial effects and average partial effects, we simply substitute our maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ for the unknown parameter vector $\boldsymbol{\beta}$, yielding $\exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})\widehat{\beta}_j$ as an estimated partial effect evaluated at $\mathbf{x}_i$ and

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

as an estimated average partial effect. On the problem set, you'll show that for Poisson regression model the estimated APE in fact equals $\bar{y}\widehat{\beta}_j$. This implies that multiplying Poisson coefficients by $\bar{y}$ puts them roughly on the same scale as OLS coefficients estimated from the same data.

2.5 What if the Poisson Assumption Fails?

As explained in section 2.3, we maintain Assumption 2.1 throughout this chapter but treat the Poisson regression model as a possibly mis-specified conditional likelihood model. So what is the pseudo-true parameter value for Poisson regression? In other words, what value of $\boldsymbol{\beta}$ maximizes the expected log-likelihood $\mathbb{E}\left[\ell_i(\boldsymbol{\beta})\right]$? By (2.5) and iterated expectations,

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

Simplifying the inner expectation, we see that,

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

Now, suppose that we ignore our original problem of maximizing $\mathbb{E}[\ell_i(\boldsymbol{\beta})]$ and decide instead to maximize $\mathbb{E}[\ell_i(\boldsymbol{\beta})|\mathbf{x}_i]$. Note that this is one of the "sub-problems" discussed in section 2.2 above, in which $\mathbf{x}$ is held fixed. The first order condition for this problem is

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

Because the objective function is concave, the solution to this first-order condition is a global optimum. So what is the solution? Under Assumption 2.1, $\mathbb{E}(y_i|\mathbf{x}_i) = \exp(\mathbf{x}_i'\beta_o)$. It follows that setting $\boldsymbol{\beta} = \boldsymbol{\beta}_o$ makes $\mathbb{E}[y_i|\mathbf{x}_i] - \exp(\mathbf{x}_i'\boldsymbol{\beta})$ exactly zero. Notice that this

solution *does not depend* on $\mathbf{x}_i$. Thus, for any realization of $\mathbf{x}_i$ and any $\boldsymbol{\beta}$,

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

Taking expectations of both sides, it follows that

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

Since $\mathbb{E}[\ell_i(\boldsymbol{\beta}_o)] \geq \mathbb{E}[\ell_i(\boldsymbol{\beta})]$ for any value of $\boldsymbol{\beta}$, we have shown that the pseudo-true parameter value under the Poisson regression model from (2.4) *coincides* with the parameter $\boldsymbol{\beta}_o$ of the conditional mean function from Assumption 2.1.[10] In other words, as long as $\mathbb{E}(y|\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta}_o)$, Poisson regression will recover $\boldsymbol{\beta}_o$ *regardless* of whether the true conditional distribution of $y_i|\mathbf{x}_i$ is a Poisson. While we established this fact for a particular choice of $m(\mathbf{x};\boldsymbol{\beta})$ the result is general: Poisson Regression is consistent for the true parameters of the conditional mean function as long as we have correctly specified $\mathbb{E}(y|\mathbf{x})$, i.e. as long as $\mathbb{E}(y|\mathbf{x}) = m(\mathbf{x};\boldsymbol{\beta})$ for some value of $\boldsymbol{\beta}$. This is similar to the result that we obtained for the simple Poisson MLE example from chapter 1 above: regardless of whether $y_i \sim \text{Poisson}(\theta)$, $\widehat{\theta}$ is still consistent for $\mathbb{E}(y_i)$.

2.6 Alternative Asymptotic Variance Matrices

Using the notation defined in section 2.3, the matrices $\mathbf{J}$ and $\mathbf{K}$ from Theorem 2.1 are

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

in the Poisson regression case.[11] Notice that $\mathbf{J}$ only depends on the marginal distribution of $\mathbf{x}$, while $\mathbf{K}$ depends on the *joint* distribution of $(\mathbf{x}, y)$. By iterated expectations,

$$\mathbf{K} = \mathbb{E}\left(\mathbb{E}\left[\left\{y_i - \mathbb{E}(y_i|\mathbf{x}_i)\right\}^2|\mathbf{x}_i\right]\mathbf{x}_i\mathbf{x}_i'\right) = \mathbb{E}\left[\text{Var}(y_i|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'\right] \tag{2.8}$$

so we see that assumptions about $\text{Var}(y|\mathbf{x})$ affect the asymptotic variance of $\widehat{\beta}$ from Theorem 2.1 through $\mathbf{K}$. In this section we consider a number of possible assumptions for $\text{Var}(y|\mathbf{x})$, leading to different asymptotic variance matrix calculations for the Poisson regression estimator $\widehat{\boldsymbol{\beta}}$. The first, and strongest, of these assumptions is the **Poisson variance assumption**, which holds automatically if the Poisson regression model from (2.4) is correctly specified.

[10] Strictly speaking we have not shown that $\boldsymbol{\beta}_o$ is the *unique* maximizer of the expected log likelihood. There can be *multiple solutions* if the regressors $\mathbf{x}$ are perfectly co-linear, as in OLS regression.

[11] Although $\text{Var}[\mathbf{s}_i(\boldsymbol{\beta}_o)] = \mathbb{E}[\mathbf{s}_i(\boldsymbol{\beta}_o)\mathbf{s}_i(\boldsymbol{\beta}_o)'] - \mathbb{E}[\mathbf{s}_i(\boldsymbol{\beta}_o)]\mathbb{E}[\mathbf{s}_i(\boldsymbol{\beta}_o)]'$, the second term drops out since $\mathbb{E}[\mathbf{s}_i(\boldsymbol{\beta}_o)] = \mathbf{0}$ by the first-order condition for MLE.

Assumption 2.2 (Poisson Variance Assumption)**.** $Var(y|\mathbf{x}) = \mathbb{E}(y|\mathbf{x})$

Substituting Assumption 2.2 into (2.8), we obtain $\mathbf{K} = \mathbb{E}\left[\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_o\right)\mathbf{x}_i\mathbf{x}_i'\right]$ under Assumption 2.1. Hence, the Poisson variance assumption implies $\mathbf{K} = \mathbf{J}$, so that $\mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1} = \mathbf{J}^{-1}$. In this case,

$$\widehat{\mathbf{J}}^{-1} = \left[\frac{1}{N}\sum_{i=1}^{N}\exp\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)\mathbf{x}_i\mathbf{x}_i'\right]^{-1} \tag{2.9}$$

provides a consistent estimator of the asymptotic variance matrix of $\widehat{\boldsymbol{\beta}}$. Notice that Assumption 2.2 *could* still hold even if distribution of $y_i|\mathbf{x}_i$ is not Poisson. All that is required here is that the conditional mean equals the conditional variance. In practice however, we would be unlikely to entertain the Poisson variance assumption unless we considered it plausible that the data truly did arise from a Poisson distribution. A weaker assumption on $\text{Var}(y|\mathbf{x})$ is the **Quasi-Poisson** assumption.

Assumption 2.3 (Quasi-Poisson Assumption)**.** $Var(y|\mathbf{x}) = \sigma^2\mathbb{E}(y|\mathbf{x})$

Assumption 2.3 allows the conditional variance to differ from the conditional mean by a scalar factor σ^2. When $\sigma^2 > 1$ we have $\text{Var}(y|\mathbf{x}) > \mathbb{E}(y|\mathbf{x})$, a situation called **overdispersion**. This is extremely common in practice: the conditional variance of real world count data is typically larger than the conditional mean. When $\sigma^2 < 1$ we have $\text{Var}(y|\mathbf{x}) < \mathbb{E}(y|\mathbf{x})$, a situation called **underdispersion**. And in the special case of $\sigma^2 = 1$, Assumption 2.3 reduces the to the Poisson variance assumption. Under Assumption 2.3, $\mathbf{K} = \sigma^2\mathbb{E}\left[\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_o\right)\mathbf{x}_i\mathbf{x}_i'\right] = \sigma^2\mathbf{J}$ and hence $\mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1} = \sigma^2\mathbf{J}^{-1}$. Equation 2.9 provides a consistent estimator of $\mathbf{J}^{-1}$ regardless of the assumption we make about $\text{Var}(y|\mathbf{x})$. Thus, to construct a consistent estimator of the asymptotic variance matrix of $\widehat{\boldsymbol{\beta}}$ under the quasi-Poisson assumption, we need only find a consistent estimator of σ^2. Re-arranging Assumption 2.3 and substituting the definition of $\text{Var}(y|\mathbf{x})$ in terms of $\mathbb{E}(y|\mathbf{x})$, we see that

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

But since $1/\mathbb{E}(y|\mathbf{x})$ is a function of $\mathbf{x}$, we can pull it inside of the leftmost conditional expectation, yielding

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

under Assumption 2.1. Now, since σ^2 is a constant, it equals $\mathbb{E}[\sigma^2]$. Thus, by iterated expectations,

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

this motivates the following simple estimator of σ^2

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

which can be shown to be consistent as $N \to \infty$. Combined with $\widehat{\mathbf{J}}^{-1}$ from (2.9), this gives a consistent estimator of the asymptotic variance matrix of $\widehat{\boldsymbol{\beta}}$ under Assumption 2.3.

The weakest possible assumption we can make for $\text{Var}(y|\mathbf{x})$ is no assumption at all! This gives rise to the **robust** asymptotic variance matrix estimator $\widehat{\mathbf{J}}^{-1}\widehat{\mathbf{K}}\widehat{\mathbf{J}}^{-1}$ where

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

and $\widehat{\mathbf{J}}^{-1}$ is as defined in (2.9). Using the robust asymptotic variance matrix estimator in Poisson regression is analogous to using heteroskedasticity-robust standard errors in a linear regression model. As in the linear regression case, however, the robust asymptotic variance matrix estimator can be very noisy in small samples.

2.7 Why Poisson Regression Rather Than NLLS?

Throughout this chapter we have assumed that $\mathbb{E}(y|\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta}_o)$, Assumption 2.1. It turns out that, if the conditional mean is correctly specified, then *both* Poisson regression and nonlinear least squares (NLLS) are consistent and asymptotically normal. They differ, however, in their respective asymptotic variance matrices. So why prefer Poisson regression over NLLS? We don't have time to go into this in detail, but here's the basic idea. Count data are typically heteroskedastic. If $\text{Var}(y|\mathbf{x})$ varies with $\mathbf{x}$, then NLLS will tend to be relatively inefficient, i.e. it will have a relatively high asymptotic variance. If the Poisson model is correct, Poisson regression has the lowest variance among all estimators that leave the distribution of $\mathbf{x}$ unspecified. If the Poisson model is incorrect but $\text{Var}(y|\mathbf{x}) = \sigma^2\mathbb{E}(y|\mathbf{x})$ then it can be shown that Poisson regression is *still* more efficient than NLLS.

Chapter 3

Binary Outcome Models

Many outcomes of interest in economics are binary: either zero or one. Suppose, for example, that $y = 1$ if a person is employed and zero otherwise. Perhaps we observe a vector of regressors $\mathbf{x} = \{\text{age, sex, education, experience, ...}\}$ and would like to determine the predictive relationship between education and employment, holding all other regressors constant.[1] In this chapter we'll consider three models for this situation: the **linear probability model**, **logistic regression**, and **probit regression**.

3.1 Properties of Binary Outcome Models

We begin with a lemma providing some basic properties of binary outcome models.

Lemma 3.1. *Suppose that y is binary and let $p(\mathbf{x}) \equiv \mathbb{P}(y = 1|\mathbf{x})$. Then,*

(i) $\mathbb{E}(y|\mathbf{x}) = p(\mathbf{x})$, *and*

(ii) $Var(y|\mathbf{x}) = p(\mathbf{x})\left[1 - p(\mathbf{x})\right]$.

Proof of Lemma 3.1. For part (i),

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

For part (ii),

$$\mathbb{E}(y^2|\mathbf{x}) = \left\{0^2 \times [1 - p(\mathbf{x})] + 1^2 \times p(\mathbf{x})\right\} = p(\mathbf{x})$$

[1] Notice my careful use of the term *predictive relationship*. In this chapter, and indeed in this set of notes, we will not take any stand on whether a particular relationship is causal or merely predictive. Causal inference is extremely interesting and important, but we unfortunately don't have time to cover it here. For notes on this topic from my second-year MPhil lectures, see treatment-effects.com. For a discussion of "treatment effects" versus "structural" approaches to limited dependent variables models such as the ones in this chapter, see Angrist (2001) and the associated rejoinders.

and thus, we obtain

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

□

Lemma 3.1 reveals a crucial fact about binary outcome models: the conditional mean *completely determines* the conditional variance. This is quite different from the situation we encountered when studying count data in chapter 2. While the Poisson model does imply that the mean and variance are equal, real-world counts are typically over-dispersed: the variance exceeds the mean. As long as we correctly specify the conditional mean function for our count dataset, Poisson regression remains consistent *regardless* of the true conditional variance. Indeed, we explored a number of possible specifications for $\text{Var}(y|\mathbf{x})$, any of which could comfortably co-exist with a given specification for $\mathbb{E}(y|\mathbf{x})$, and explained how to obtain correct standard errors in each case. When y is binary, however, it is *impossible* to correctly specify $\mathbb{E}(y|\mathbf{x})$ while mis-specifying $\text{Var}(y|\mathbf{x})$. Because the conditional mean and conditional variance are linked, if we get one wrong, we will necessarily get the other wrong. Another consequence of Lemma 3.1 is that binary outcome models *necessarily* exhibit heteroskedasticity. Whenever $\mathbb{E}(y|\mathbf{x})$ depends on $\mathbf{x}$, so does $\text{Var}(y|\mathbf{x})$.

3.2 The Linear Probability Model (LPM)

All of the models for binary outcomes that we study in this chapter amount to assuming a particular function form for $p(\mathbf{x})$ in terms of a vector of unknown coefficients $\boldsymbol{\beta}$. The linear probability model assumes that $p(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$. By Lemma 3.1, this implies that

$$\mathbb{E}(y|\mathbf{x}) = p(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}, \quad \text{Var}(y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}\left(1 - \mathbf{x}'\boldsymbol{\beta}\right).$$

The most important thing to realize about the LPM is that it's just a fancy name for *linear regression with a binary outcome*. To see why, suppose that we define $u \equiv y - \mathbf{x}'\boldsymbol{\beta}$. Under the LPM, $\mathbb{E}(y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ and hence

$$\mathbb{E}(u|\mathbf{x}) = \mathbb{E}(y - \mathbf{x}'\boldsymbol{\beta}|\mathbf{x}) = \mathbb{E}(y|\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta} = \mathbf{x}'\boldsymbol{\beta} - \mathbf{x}'\boldsymbol{\beta} = 0.$$

This means that we can express the LPM as

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

which we recognize as a linear regression model. Since $\mathbb{E}(u|\mathbf{x}) = 0$ OLS estimation of $y = \mathbf{x}'\boldsymbol{\beta} + u$ is unbiased and consistent provided that the LPM assumption $p(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ is

correct, of which more anon. The second most important thing to realize about the LPM is that the error term u is *necessarily* heteroskedastic. Since $\mathbb{E}(u|\mathbf{x}) = 0$, by Lemma 3.1 and the definition of conditional variance, we obtain

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

Since u is heteroskedastic, inference for the LPM should use robust standard errors.

A key question remains: is the LPM actually a *reasonable* model for binary outcomes? If $p(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ then changing x_j by Δ always changes $p(\mathbf{x})$ by $\beta_j\Delta$ When x_j is a regressor without an upper or lower bound, this means that our predicted probabilities $p(\mathbf{x})$ could easily turn out to be greater than one or less than zero! So while the LPM is often a reasonable approximation, it cannot be *literally* true as a model for $p(\mathbf{x})$ except in special cases.[2] You will explore one of these special cases on your problem set.

3.3 Index Models: Logit & Probit

As we have seen, a key of the LPM is that it can yield predicted probabilities that lie outside of $[0, 1]$. To avoid this problem, we need to constrain our chosen functional form for $p(\mathbf{x})$ to be between zero and one. **Index models** achieve this as follows.

Assumption 3.1 (Index Model)**.** *Suppose that* $p(\mathbf{x}) = G(\mathbf{x}'\boldsymbol{\beta})$*, where* $\mathbf{x}$ *includes a constant, and* G *satisfies*

(i) $0 \leq G(\cdot) \leq 1$*,*

(ii) G *is differentiable and strictly increasing,*

(iii) $\lim_{z\to-\infty} G(z) = 0$*, and* $\lim_{z\to\infty} G(z) = 1$*.*

In Assumption 3.1, we call $\mathbf{x}'\boldsymbol{\beta}$ the **linear index** and G the **index function**. Notice that the four requirements for the index function G are *identical* to the conditions required for a function to be the cumulative distribution function (CDF) of a continuous random variable. This gives us an easy way to construct an index model: simply choose a continuous random variable and use its CDF. We'll discuss some common choices for G in a moment, but first it is worth asking *why* the conditions of Assumption 3.1 make sense. Part (i) should be clear. Since $G(\mathbf{x}'\boldsymbol{\beta})$ is supposed to be a probability, the function G should only take values between zero and one. But what about the remaining conditions?

[2] For a discussion of the LPM as an approximation when studying causal effects, see Angrist (2001).

To understand the value of part (ii), consider *partial effects* of $\mathbf{x}$, i.e. the derivatives of $p(\mathbf{x})$ with respect to $\mathbf{x}$. By (ii), G is differentiable. Let g denote it's derivative. Then,

$$\frac{\partial}{\partial x_j} p(\mathbf{x}) = g(\mathbf{x}'\boldsymbol{\beta})\beta_j, \quad g(z) \equiv \frac{d}{dz} G(z). \tag{3.1}$$

Notice that the partial effect of x_j depends on the value of $\mathbf{x}$ at which we evaluate g. Since G is assumed to be strictly increasing, g is strictly positive: $g(\mathbf{x}'\boldsymbol{\beta}) > 0$ for any value of $\mathbf{x}$ or $\boldsymbol{\beta}$. This implies that the *sign* of the partial effect is completely determined by β_j, a very convenient property. Without Assumption 3.1 (ii), this would not hold.

Whereas Assumption 3.1 (i) ensures that the predictions from an index model must be valid probabilities, (iii) allows these probabilities to be arbitrarily close to zero or one. A model without this feature would be insufficiently flexible, as it could only yield predicted probabilities in a limited range. Although it was not given a number, we snuck one further condition into Assumption 3.1, namely the requirement that $\mathbf{x}$ include a constant. To see why this is important, consider an index model with a single regressor x and no constant, $p(x) = G(\beta x)$. Regardless of the value of β, this model implies that $\mathbb{P}(y = 1|x = 0) = G(0)$. But this is a very strange restriction. Just as it practically always makes sense to include an intercept in a linear regression model, it makes sense to do so in an index model. Adding one to this example would yield $p(x) = G(\beta_0 + \beta_1 x)$ so that $\mathbb{P}(y = 1|x = 0) = \beta_0$ becomes a parameter that we estimate from the data rather than arbitrarily setting equal to $G(0)$.

While it is possible to construct an index model from any continuous CDF G, two choices are common in practice. The **logistic regression** model, or logit for short, takes

$$G(z) = \Lambda(z) \equiv \frac{\exp(z)}{1 + \exp(z)}$$

where Λ denotes the CDF of a "standard logistic" random variable. In contrast, the **probit regression** model, or probit for short, takes

$$G(z) = \Phi(z) \equiv \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-t^2/2\right)\, dt$$

where Φ is the CDF of a standard normal distribution.[3] Figure 3.1 compares the standard logistic and standard normal distributions. Both the logistic density λ and the normal density φ are bell-shaped and symmetric about zero, but compared to normal, the logistic has a greater spread. This is because the standard logistic random variable has a variance of $\pi^2/3 \approx 3.3$ compared to 1 for the standard normal.

[3] A lesser-known alternative to probit called **robit regression** takes G to be the CDF of a Student-t distribution. See Liu (2004) for details.

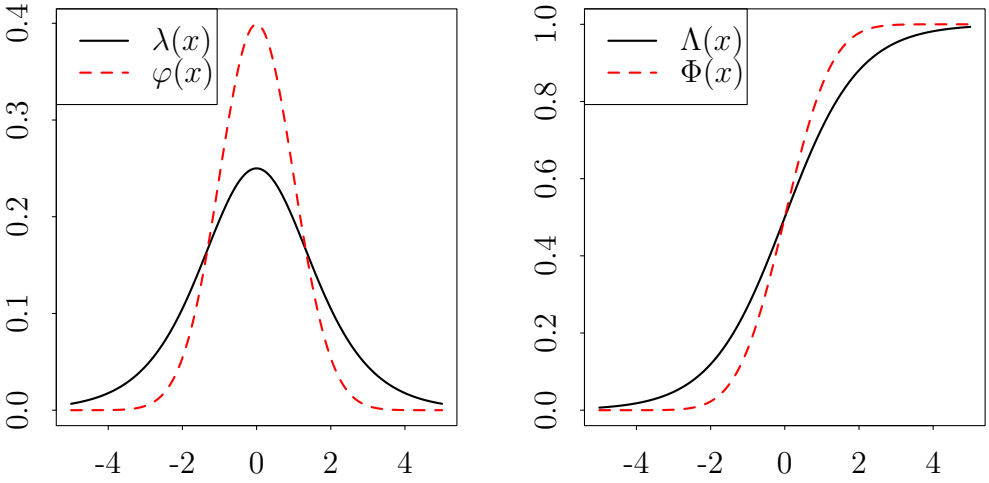

Figure 3.1: At left: λ is the density of a standard logistic and φ of a standard normal RV. At right: Λ is the CDF of a standard logistic and Φ of a standard normal RV.

3.4 Partial Effects

We have now seen three different models for binary outcomes: the LPM, logit and probit. Each of them depends on the same covariate vector $\mathbf{x}$ through the linear index $\mathbf{x}'\boldsymbol{\beta}$, but the *interpretation* of $\boldsymbol{\beta}$ in terms of partial effects differs across the three specifications. Since the linear probability model is simply linear regression, the partial effects for this model are given by

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

In other words, increasing x_j by one unit increases our prediction of the probability that $y = 1$ by β_j. For logit and probit, the partial effects are more complicated. For any index model (see Assumption 3.1) we have

$$\frac{\partial}{\partial x_j} p(\mathbf{x}) = \frac{\partial}{\partial x_j} G(\mathbf{x}'\boldsymbol{\beta}) = g(\mathbf{x}'\boldsymbol{\beta})\beta_j, \quad g(z) \equiv \frac{d}{dz} G(z) \tag{3.2}$$

so the magnitude of a partial effect depends on the value of $\mathbf{x}$ at which it is evaluated. As explained above, the sign of a partial effect coincides with that of β_j because g is strictly greater than zero. Now, since

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

specializing (3.2) to logit gives

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

And since

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

specializing to probit yields

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

Given that partial effects for logit and probit vary with $\mathbf{x}$, how can we summarize them? There are several possibilities. The first is to consider the **maximum partial effect**. Since the logistic density λ is unimodel and symmetric around zero (see Figure 3.1), the maximum partial effects for these models occur when $\mathbf{x}'\boldsymbol{\beta} = 0$. Thus, the maximum partial effect for logit is

$$\lambda(0)\beta_j = \frac{\beta_j \exp(0)}{[1+\exp(0)]^2} = \frac{\beta_j}{4}.$$

This is sometimes called the **the divide-by-four rule**: if β_j is the coefficient on x_j in a logistic regression model, then the partial effect of x_j *cannot exceed* $\beta_j/4$ or equivalently $0.25 \times \beta_j$. Because the normal density φ is also unimodel and symmetric around zero, we can apply the same reasoning: the maximum partial effect for probit is

$$\varphi(0)\beta_j = \frac{\beta_j \exp(0)}{\sqrt{2\pi}} = \frac{\beta_j}{\sqrt{2\pi}} \approx 0.4 \times \beta_j.$$

Though not quite as catchy as the divide-by-four rule the preceding expression still gives us a useful result: if β_j is the coefficient on x_j in a probit regresison model, then the partial effect of x_j *cannot exceed* $0.4 \times \beta_j$.

We can also summarize an index model by considering **relative effects**, the ratio of the partial effects of x_j and x_h. By (3.2), we see that these do *not* depend on $\mathbf{x}$:

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

So, for example, if β_3 is twice as large as β_4, this means that the partial effect of x_3 is twice as large as that of x_4 regardless of the value of of $\mathbf{x}$ at which we evaluate them.

A third and final way to summarize the results of an index model is by calculating **average partial effects** (APEs). The idea here is quite intuitive: if the partial effects vary with $\mathbf{x}$, then a reasonable way to summarize them is by *averaging* over the distribution of $\mathbf{x}$ in the population, i.e.

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

The APE tells us the average value of the APE for the people in our population. This

is *not* the same thing, however, as the APE evaluated at the average value of $\mathbf{x}$ in the population. Stated mathematically,

$$\mathbb{E}[g(\mathbf{x}'\boldsymbol{\beta})]\beta_j \neq g(\mathbb{E}[\mathbf{x}]'\boldsymbol{\beta})\beta_j$$

because $\mathbb{E}[f(Z)]$ does not in general equal $f(\mathbb{E}[Z])$, except in the special case where f is a linear function.

To estimate maximum partial effects and relative partial effects, we simply substitute estimates of the relevant parameters, $\widehat{\beta}_j$ and $\widehat{\beta}_h$. To estimate the average partial effect, we substitute estimates of $\boldsymbol{\beta}$ and replace the population expectation with a sample average:

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

3.5 Conditional MLE for Index Models

So how can we estimate the parameters of an index model? As we did for Poisson regression, we'll again rely on conditional maximum likelihood estimation. Suppose we observe a random sample $(y_1, \mathbf{x}_1), \ldots, (y_N, \mathbf{x}_N)$ where y_i is binary. Then the conditional likelihood of a single observation is given by

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

which we can write more compactly as

$$f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = G(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}\left[1 - G(\mathbf{x}_i'\boldsymbol{\beta})\right]^{1-y_i}.$$

It follows that the conditional log-likelihood of a single observation is

$$\ell_i(\boldsymbol{\beta}) \equiv \log f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = y_i \log\left[G(\mathbf{x}_i'\boldsymbol{\beta})\right] + (1 - y_i)\log\left[1 - G(\mathbf{x}_i'\boldsymbol{\beta})\right] \tag{3.3}$$

and, because we observe iid data, the conditional maximum likelihood estimator is

$$\widehat{\boldsymbol{\beta}} \equiv \underset{\boldsymbol{\beta}\in\boldsymbol{\Theta}}{\arg\max}\frac{1}{N}\sum_{i=1}^{N}\ell_i(\boldsymbol{\beta}).$$

So what exactly does $\widehat{\boldsymbol{\beta}}$ estimate? Applying the theory for mis-specified conditional maximum likelihood estimation that we discussed in chapters 1–2, define

$$\boldsymbol{\beta}_o \equiv \underset{\boldsymbol{\beta}\in\boldsymbol{\Theta}}{\arg\max}\,\mathbb{E}\left[\ell(\boldsymbol{\beta})\right].$$

If our index model is correctly specified, then $\mathbb{E}(y|\mathbf{x}) = p(\mathbf{x}) = G(\mathbf{x}'\boldsymbol{\beta}_o)$. If not, then $\boldsymbol{\beta}$ is still intepretable as the parameter value that minimizes the KL divergence from the unknown true conditional distribution $p_0(\mathbf{x})$ to our parametric model $G(\mathbf{x}'\boldsymbol{\beta})$, as discussed in chapters 1–2. Recalling the asymptotic results from these same chapters,

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

where $\mathbf{J}$ and $\mathbf{K}$ are defined in terms of the *Hessian* $\mathbf{H}_i$ and *score* $\mathbf{s}_i$ according to

$$\mathbf{J} = -\mathbb{E}\left[\mathbf{H}_i(\boldsymbol{\beta}_o)\right] \qquad \mathbf{H}_i(\boldsymbol{\beta}) \equiv \frac{\partial \mathbf{s}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}$$
$$\mathbf{K} = \mathbb{E}\left[\mathbf{s}_i(\boldsymbol{\beta}_o)\mathbf{s}_i(\boldsymbol{\beta}_o)'\right] \qquad \mathbf{s}_i(\boldsymbol{\beta}) \equiv \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

The *robust* variance matrix $\mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1}$ is quite complicated for index models, so we'll let the computer calculate it for us. Under the assumption that our index model is correctly specified, i.e. $p_o(\mathbf{x}) = G(\mathbf{x}'\boldsymbol{\beta}_o)$, the asymptotic variance matrix simplifies to $\mathbf{J}^{-1}$ by the information matrix equality.[4] In this case, we have

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

It turns out to be fairly unpleasant to calculate $\mathbf{J}$ directly, so we use a different approach. By iterated expectations, we can write

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

For correctly specified index models $\mathbb{E}[\mathbf{H}_i(\boldsymbol{\beta}_o)|\mathbf{x}_i]$ turns out to be much easier to work with than its unconditional counterpart. While we only consider the application of this idea to index models, this conditional approach to calculating $\mathbf{J}$ is more broadly applicable to correctly specified maximum likelihood models.

Theorem 3.1. *If $G(\mathbf{x}'\boldsymbol{\beta}_o)$ satisfies Assumption 3.1 and $p_o(\mathbf{x}) = G(\mathbf{x}'\boldsymbol{\beta}_o)$, then*

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

Proof of Theorem 3.1. In the first step, we calculate the score vector. By (3.3),

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

[4] See chapter 1 for a proof.

and hence, writing each term with a common denominator, we obtain

$$\mathbf{s}_i = \frac{g(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i}{G(\mathbf{x}_i'\boldsymbol{\beta})\left[1 - G(\mathbf{x}_i'\boldsymbol{\beta})\right]}\left\{\left[1 - G(\mathbf{x}_i'\boldsymbol{\beta})\right] y_i - G(\mathbf{x}_i'\boldsymbol{\beta})(1 - y_i)\right\} = \frac{g(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i\left[y_i - G(\mathbf{x}_i'\boldsymbol{\beta})\right]}{G(\mathbf{x}_i'\boldsymbol{\beta})\left[1 - G(\mathbf{x}_i'\boldsymbol{\beta})\right]}.$$

In the second step, we begin calculating the Hessian by applying the product rule to our expression for $\mathbf{s}_i$ as follows,

$$\begin{aligned}\mathbf{H}_i(\boldsymbol{\beta}) \equiv \frac{\partial \mathbf{s}_i}{\partial \boldsymbol{\beta}'} &= \frac{\partial}{\partial \boldsymbol{\beta}'}\left\{\left[y_i - G(\mathbf{x}_i'\boldsymbol{\beta})\right]\left[\frac{g(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i}{G(\mathbf{x}_i'\boldsymbol{\beta})\left\{1 - G(\mathbf{x}_i'\boldsymbol{\beta})\right\}}\right]\right\} \\ &= \frac{-g(\mathbf{x}_i'\boldsymbol{\beta})^2\mathbf{x}_i\mathbf{x}_i'}{G(\mathbf{x}_i'\boldsymbol{\beta})\left\{1 - G(\mathbf{x}_i'\boldsymbol{\beta})\right\}} + \left[y_i - G(\mathbf{x}_i'\boldsymbol{\beta})\right]\mathbf{M}(\mathbf{x}_i, \boldsymbol{\beta})\end{aligned}$$

defining the shorthand

$$\mathbf{M}(\mathbf{x}_i, \boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}'}\left\{\frac{g(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i}{G(\mathbf{x}_i'\boldsymbol{\beta})\left[1 - G(\mathbf{x}_i'\boldsymbol{\beta})\right]}\right\}.$$

Now, a *direct* approach to calculating $\mathbf{J}$ would require us to use the quotient and product rules to evaluate the derivative defined by $\mathbf{M}(\mathbf{x}_i, \boldsymbol{\beta})$, substitute this into the expression for $\mathbf{H}_i(\boldsymbol{\beta})$ and then take expectations. This is very messy, and there's a good chance that we'd make a mistake. Fortunately, there's a simpler approach.

By definition, $-\mathbf{J}$ is the expectation of $\mathbf{H}_i(\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta}_o$. In the third step, we calculate the *conditional* expectation of $\mathbf{H}_i(\boldsymbol{\beta}_o)$ given $\mathbf{x}$, yielding

$$\begin{aligned}\mathbb{E}\left[\mathbf{H}_i(\boldsymbol{\beta})|\mathbf{x}_i\right] &= \frac{-g(\mathbf{x}_i'\boldsymbol{\beta})^2\mathbf{x}_i\mathbf{x}_i'}{G(\mathbf{x}_i'\boldsymbol{\beta})\left\{1 - G(\mathbf{x}_i'\boldsymbol{\beta})\right\}} + \mathbb{E}\left[y_i - G(\mathbf{x}_i'\boldsymbol{\beta})|\mathbf{x}_i\right]\mathbf{M}(\mathbf{x}_i, \boldsymbol{\beta}) \\ &= \frac{-g(\mathbf{x}_i'\boldsymbol{\beta})^2\mathbf{x}_i\mathbf{x}_i'}{G(\mathbf{x}_i'\boldsymbol{\beta})\left\{1 - G(\mathbf{x}_i'\boldsymbol{\beta})\right\}}\end{aligned}$$

since $\mathbf{M}(\mathbf{x}_i, \boldsymbol{\beta})$ is a function of $\mathbf{x}_i$ and $\mathbb{E}[y_i - G(\mathbf{x}_i'\boldsymbol{\beta}_o)|\mathbf{x}_i] = 0$ under the assumption that our index model is correctly specified. Finally, by iterated expectations,

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

From Theorem 3.1, we obtain the following asymptotic distribution for the conditional maximum likelihood estimator of a correctly-specified index model

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

and it can be shown that the following expression provides a consistent estimator of the

asymptotic variance matrix:

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

3.6 Pseudo R-squared

There is no single agreed-upon measure of in-sample fit for binary outcome models. A measure that is often reported in applied work in economics is the so-called **pseudo R-squared**.[5] Although this measure has little else in common with the more familiar R-squared from linear regression, the pseudo R-squared is unitless, and takes on a value between zero and one where larger values indicate a better in-sample fit.

Pseudo R-squared is constructed from the maximized sample log-likelihood of *two* models: the "full" model, whose fit we will evaluate, and a "null" model that contains only an intercept. Let $\ell(\widehat{\boldsymbol{\beta}})$ denote the sample log-likelihood of the full model, evaluated at its MLE, $\widehat{\boldsymbol{\beta}}$, and $\ell(\bar{y})$ denote the sample log-likelihood of the null model evaluated at its MLE, $\bar{y}$.[6] Then, the pseudo R-squared is defined by

$$\widetilde{R}^2 \equiv 1 - \frac{\ell(\widehat{\beta})}{\ell(\bar{y})} = \frac{\ell(\bar{y}) - \ell(\widehat{\boldsymbol{\beta}})}{\ell(\bar{y})}.$$

On your problem set, you will show that this measure is always between zero and one.

Intuitively, $\widetilde{R}^2$ compares the fit of the full model to that of the null model, where fit is measured by the respective log-likelihoods. If the two fit equally well, then the ratio of log-likelihoods is one, so $\widetilde{R}^2 = 0$. The better the fit of the full model compared to the null model, the closer $\widetilde{R}^2$ will be to one. Figure 3.2 provides some intuition for this measure in a simple example. Note that this is related to but *not the same* as a likelihood ratio test of the full model against the null model. An LR test statistic is constructed from $\ell(\bar{y}) - \ell(\widehat{\boldsymbol{\beta}})$, whereas $\widetilde{R}^2$ *divides* this quantity by $\ell(\bar{y})$.

[5] See Windmeijer (1995) for discussion of various alternative measures.

[6] Consider an index model with only an intercept β_0. Regardless of the choice of index function G, the MLE for this parameter equals the sample mean of y. Try verifying this as a practice problem.

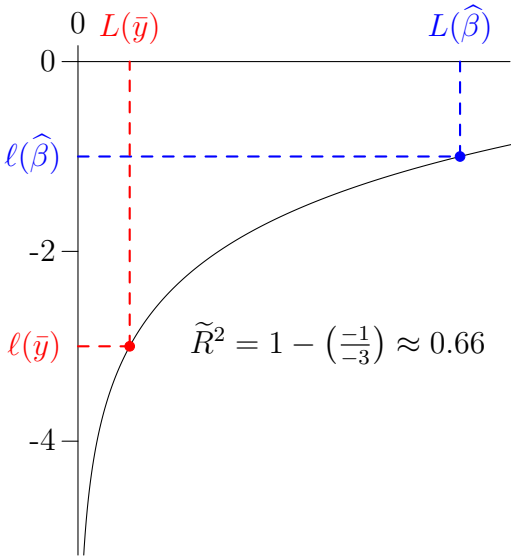

Figure 3.2: A hypothetical example of calculating pseudo R-squared $\widetilde{R}^2$. The log-likelihood $\ell(\widehat{\beta})$ of the unrestricted model is -1 while that of the model with only an intercept, $\ell(\bar{y})$, is -3. This gives a pseudo R-squared of approximately 0.66.

Chapter 4

Random Utility Models

In chapter 3 we examined binary outcome models from a purely *statistical* perspective, namely as a model for $\mathbb{P}(y = 1|\mathbf{x})$ when y is a binary random variable. It turns out that there is another way of thinking about these models, one that relates them to economic models of rational choice. This is the so-called **random utility** or **discrete choice** approach. Random utility models provide not only a novel interpretation of probit and logit, but a framework for constructing richer models for discrete outcomes. The discussion and notation below mainly follow chapters 1–3 of Train (2009). See also Chapter 15 of Cameron and Trivedi (2005).

4.1 Overview of Random Utility Models

We'll begin with some basic terminology. In a discrete choice model a **decision-maker**—a household, person, or firm—choices from a collection of **alternatives**—products or actions—to maximize her utility. The set of all alternatives that are available to the decision-maker is called the **choice set**. Because I'm growing weary of typing "decision-maker" over and over, I'll give mine a name: Alice. We only consider models that satisfy the following conditions.

Assumption 4.1 (Choice Sets)**.**

(i) Choices are mutually exclusive.

(ii) The choice set is exhaustive.

(iii) The number of alternatives is finite.

Part (i) of Assumption 4.1 says that Alice chooses only *one* of the alternatives. Part (ii) says that the choice set contains every alternative or, to put it another way, that the Alice always chooses *something* from the choice set. In fact, these two conditions are not restrictive. If necessary, we can always redefine the choice set so that it satisfies them

automatically. For example, suppose Alice is at a restaurant that serves beer and pizza. Unless this is a very strange resaturant, the choice between beer and pizza is not mutually exclusive: nothing is to stop Alice from ordering both. In this case the appropriate choice set is not {Beer, Pizza} but rather

{Beer only, Pizza only, Beer & Pizza}.

Unless beer and pizza are the only things on the menu, however, this choice set is not exhaustive. But this too is easy to correct: simply add a "default" option, for example

{Beer only, Pizza only, Beer and Pizza, Something Else}.

Now we have a choice set that satisfies all the requirements of Assumption 4.1. To make further progress, we need to define some notation. As mentioned above, this discussion follows Train (2009). Some of the notation is a bit different from what we've used in earlier chapters, so stay alert! There are N decision-makers, indexed by $n = 1, 2, \ldots, N$. Each decision-maker chooses between J alternatives, indexed by $j = 1, \ldots, J$. Let U_{nj} denote the utility that decision-maker n obtains if she chooses alternative j. We assume that choices are *rational*: Alice chooses the alternative that maximizes her utility.

Assumption 4.2 (Rational Choice)**.** *Decision-maker n chooses alternative i if and only if $U_{ni} > U_{nj}$ for any $j \neq i$.*

Utility, of course, is unobserved. So if we want to model choice, we need to specify what we *do* observe. The basic idea behind random utility models is to back out preferences from observed *choices* and *attributes*, given the assumption of rational choice.

Assumption 4.3 (Observables)**.** *The researcher observes:*

(i) the attributes x_{nj} of each alternative,

(ii) the attributes s_n of each decision-maker, and

(iii) the choice that each decision-maker makes.

Attributes of an alternative could be the price of beer, the kind of pizza, and so on, while attributes of a decision-maker could be age, sex, education etc. We assume that the 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}. We call V_{nj} the **representative utility**. To drive home the distinction, I will sometimes call U_{nj} *true* utility. In effect, V_{nj} is the part of U_{nj} that is "explained" by x_{nj} and s_n.

To obtain an econometric model to which we can apply probabilistic reasoning, we need a source of randomness. Rather than treating a given person's choices as random,

we instead view randomness as a feature of our *sampling procedure*, arising from differences between people that are not captured in our model of representative utility. Such differences are often called **unobserved heterogeneity**. The idea is as follows. Define the **error term** $\varepsilon_{nj} \equiv U_{nj} - V_{nj}$ to be the difference between true utility U_{nj} and representative utility V_{nj} for decision-maker n and alternative j. Because there are J alternatives, there are J error terms for each decision-maker, namely

$$\boldsymbol{\varepsilon}' \equiv [\ \varepsilon_{n1} \quad \dots \quad \varepsilon_{nJ}\ ].$$

The vector of errors $\boldsymbol{\varepsilon}$ represents unobserved factors that affect choices but are not captured by the representative utilities. We will treat these errors as *random* in the following sense. Alice and Bob, along with all of their fellow decision-makers $n = 1, \dots, N$ are a random sample from some population. In this population, some decision-makers with the same attributes make different choices. This means that they must have different error terms. We can represent these differences in the population using a *probability density* function $f(\boldsymbol{\varepsilon}_n)$. From this perspective, drawing a decision-maker at random from the population is equivalent to making a random draw of $\boldsymbol{\varepsilon}$ according to f. Alice's choice isn't random; the fact that we the fact that we sampled her from the population is.

Using this formulation, we can calculate **choice probabilities**: the probability that decision-maker n chooses alternative j. In particular,

$$P_{ni} \equiv \mathbb{P}(U_{ni} > U_{nj} \quad \forall j \neq i) = \int_{\mathbb{R}^J} \mathbb{1}\left\{\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \quad \forall j \neq i\right\} f(\boldsymbol{\varepsilon}_n)\, d\boldsymbol{\varepsilon}_n. \quad (4.1)$$

We now have all the ingredients needed to give an overview of random utility models:

1. Write down a parametric model for $V_{nj}(x_{nj}, s_n)$ with unknown parameters $\boldsymbol{\theta}$.
2. Choose a distribution f for the errors (unobserved heterogeneity) $\boldsymbol{\varepsilon}_n$.
3. Calculate the choice probabilities as a function of parameters $\boldsymbol{\theta}$.
4. Use observed choices and attributes to find the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$.

As we will show below, logit and probit can both be viewed as special cases of the general random utility approach when the choice set contains only two elements and we choose particular error distributions. More broadly, random utility models provide a framework for estimating much richer discrete choice models.

Notice that the expression for the choice probabilities from (4.1) implicitly treats the **representative utilities as fixed constants** rather than random variables. In reality, of course, they are not fixed. Instead, what we have in mind is a *conditional model*: V_{nj} depends on observed attributes x_{nj} and s_n and an unknown parameter vector $\boldsymbol{\theta}$ and all probabilities are implicitly conditioned on (x_{nj}, s_n). Both to match Train (2009) and to

avoid cluttering the notation, I often suppress explicit conditioning on (x_{nj}, s_n) in the discussion that follows. If you find this more confusing than helpful, feel free to add it back in when you write out your notes!

With the exception of a short break for beer and pizza, our discussion so far has been very abstract. To make things more concrete, let's close this section with a very simple example. Suppose there are exactly two ways to get to work: by *car* and by *bus*. We observe two attributes: the cost in *time* T and *money* M of each mode of transport. Now, suppose that we specify the following model for representative utilities, with unknown parameters (β, γ)

$$\begin{aligned} V_{\text{car}} &= \beta T_{\text{car}} + \gamma M_{\text{car}} & U_{\text{car}} &= V_{\text{car}} + \varepsilon_{\text{car}} \\ V_{\text{bus}} &= \beta T_{\text{bus}} + \gamma M_{\text{bus}} & U_{\text{bus}} &= V_{\text{bus}} + \varepsilon_{\text{bus}}. \end{aligned}$$

In this admittedly very simple example, the choice probabilities can be written as

$$\begin{aligned} P_{\text{car}} &= \mathbb{P}(\varepsilon_{\text{bus}} - \varepsilon_{\text{car}} < V_{\text{car}} - V_{\text{bus}}) \\ P_{\text{bus}} &= \mathbb{P}(\varepsilon_{\text{car}} - \varepsilon_{\text{bus}} < V_{\text{bus}} - V_{\text{car}}) = 1 - P_{\text{car}}. \end{aligned}$$

This model allows *observed heterogeneity* to enter through T and M, depending on what we know about each decision-maker. For example, perhaps Bob is 70 and gets a discount on public transport so his M_{bus} is low while Alice lives far from the bus stop, so her T_{bus} is high. In contrast, unobserved heterogeneity enters through the error terms. Perhaps James hates to drive $(\varepsilon_{\text{car}} - \varepsilon_{\text{bus}} < 0)$ but Steve loves driving $(\varepsilon_{\text{car}} - \varepsilon_{\text{bus}} > 0)$.

4.2 The Likelihood for Random Utility Models

As described in the previous section, a random utility model combines a specification for $V_{nj}(x_{nj}, s_n)$ in terms of an unknown parameter vector $\boldsymbol{\theta}$ with an assumed density f for the vector of errors $\boldsymbol{\varepsilon}_n$. We now explain how to write down the likelihood for such a model, allowing us to estimate and carry out inference for the parameter vector $\boldsymbol{\theta}$. First a bit of notation: let $y_n \in \{1, \ldots, J\}$ denote decision-makers n's choice, and $\mathbf{z}_n$ denote the vector of *all attributes* for decision-maker n, potentially including both attributes that are fixed across alternatives, s_n, and attributes that are not, x_{nj}.

Given the observed covariates $\mathbf{z}_n$, by (4.1) the choice probabilities P_{ni} can viewed as a *function* of $\boldsymbol{\theta}$, namely $\mathbb{P}(y_n = i|\mathbf{z}_n; \boldsymbol{\theta})$. In some cases we can work out a closed form expression for this function. In the conditional logit model, described in more detail in

section 4.5, $V_{nj} = \mathbf{x}_{nj}'\boldsymbol{\beta}$ and the choice probabilities are

$$P_{ni} = \mathbb{P}(y_n = i|\mathbf{z}_n; \boldsymbol{\theta}) = \frac{\exp(\mathbf{x}_{ni}'\boldsymbol{\beta})}{\sum_{j=1}^{J} \exp(\mathbf{x}_{nj}'\boldsymbol{\beta})}.$$

Unfortunately, models like the conditional logit are quite rare. For most densities f that we might choose for the error terms $\boldsymbol{\varepsilon}_n$, e.g. a multivariate normal density, the integral from (4.1) has no closed form. In such cases, we cannot calculate the choice probabilities P_{ni} directly and must instead rely on numerical approximations or simulation-based methods. But even when we cannot write down a simple formula for the choice probabilities, they are *still* a function of $\boldsymbol{\theta}$ given observed attributes.

The hard part about random utility models is calculating the choice probabilities. Given P_{ni}, however, the likelihood is straightforward. Conditional on the attributes $\mathbf{z}_n$, the observed choice y_n is a random variable with support set $\{1, 2, \ldots, J\}$ and

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

where $\mathbb{1}\{y_n = j\}$ equals 1 if $y_n = j$ and zero otherwise. Thus, given a random sample of N observations $(y_n, \mathbf{z}_n)$, the log-likelihood function is given by

$$\ell_N(\boldsymbol{\theta}) = \sum_{n=1}^{N} \sum_{j=1}^{J} \mathbb{1}\{y_n = j\} \log P_{nj}, \tag{4.2}$$

which, as explained above, depends on $\boldsymbol{\theta}$ through the choice probabilities P_{nj}. Note that (4.2) holds for *any random utility model*. The differences between different models come from different choice probabilities P_{nj}, which in turn come from different specifications for V_{nj} and the error density f.

4.3 Identification of Choice Models

Our goal when writing down a random utility model is to **identify** and estimate the parameters from our specification for the representative utilities V_{nj}. In the car versus bus example from above we supposed that $V_{nj} = \beta T_{nj} + \gamma M_{nj}$ where T_{nj} and M_{nj} were the *time* and *monetary* costs of a given mode of transportation j for a particular decision-maker n. Here, the parameters of interest are (β, γ). We say that a parameter is *identified* if it could be uniquely determined by observing the whole population of data from which our sample was drawn.[1] Fundamentally, identification is about what we can and cannot

[1] As Lewbel (2019) points out "econometric identification really means just one thing … yet well over two dozen different terms for identification now appear in the econometrics literature." If you're confused about the way this term is used in any papers or references you've come across, I strongly recommend

learn from data. In the context of random utility models, the relevant question is: given that we cannot observe utilities, what can we learn from choices and attributes? There are two crucial insights for answering this question, both of which should be at least somewhat familiar from your coursework in microeconomics:

1. Only *differences* in utility matter.

2. The *scale* of utility is irrelevant.

We will explore the consequences of each of these in turn.

4.3.1 Only differences in utility matter.

We assumed that Alice chooses the alternative with the highest utility, namely the i such that $U_{ni} > U_{nj}$ for all $j \neq i$. Subtracting U_{nj} from both sides of the inequality, this is the *same thing* as choosing the i such that $U_{ni} - U_{nj} > 0$ for all $j \neq i$. In other words, all that matters is how much better or worse a given alternative is than the others. In terms of choice probabilities,

$$\mathbb{P}(U_{ni} > U_{nj} \quad \forall j \neq i) = \mathbb{P}\left(U_{ni} - U_{nj} > 0 \quad \forall j \neq i\right).$$

Because only differences in utility matter, only differences of *errors* matter for calculating choice probabilities. Let $\widetilde{\varepsilon}_{njk} \equiv \varepsilon_{nj} - \varepsilon_{nk}$ be the difference of errors ε_{nj} and ε_{nk}. Taking this idea further, let $\widetilde{\boldsymbol{\varepsilon}}_{ni}$ be the vector of all unique differences taking ε_{ni} as the "base case." For example, in a setting with three alternatives and error terms $\boldsymbol{\varepsilon}_n' = (\varepsilon_{n1}, \varepsilon_{n2}, \varepsilon_{n3})$, we could calculate two difference relative to the first error term, namely $\widetilde{\boldsymbol{\varepsilon}}_{n1}' = (\varepsilon_{n2} - \varepsilon_{n1}, \varepsilon_{n3} - \varepsilon_{n1})$. If there are J alternatives, then there are J errors and $(J-1)$ unique differences. Defining g to be the joint density of $\widetilde{\boldsymbol{\varepsilon}}_{ni}$, we can calculate the choice probabilities as

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

Notice that this is a $(J-1)$-dimensional integral rather than a J-dimensional integral because, again, the number of differences is one fewer than the number of error terms.

So what does all of this have to do with identification? The simple insight that only differences of utility matter for choices has two important consequences. First, we cannot identify a different intercept for each alternative. Second, only differences of effects for decision-maker attributes are identified. We'll consider each of these in the context of our car versus bus example from above.

taking a look at the aforementioned review article.

The representative utility $V_{nj} = \beta T_{nj} + \gamma M_{nj}$ from our car versus bus example did not include intercepts. Suppose we add them in, yielding

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

What has changed by adding α_{car} and α_{bus}? As you may recall from your lectures on linear regression, adding an intercept to a model is a way of "absorbing" a nonzero mean for the error term. Suppose $U_{nj} = \mathbf{x}'_{nj}\boldsymbol{\beta} + \varepsilon^*_{nj}$ where $\mathbf{x}_{nj}$ *excludes* a constant. If we define $\alpha_j \equiv \mathbb{E}[\varepsilon^*_{nj}]$ and $\varepsilon_{nj} \equiv \varepsilon^*_{nj} - \alpha_j$, then $U_{nj} = \alpha_j + \mathbf{x}'_{nj}\boldsymbol{\beta} + \varepsilon_{nj}$ where $\mathbb{E}[\varepsilon_{nj}] = 0$ by construction. So compared to our original car versus bus example from above, this new version has $\mathbb{E}[\varepsilon_{\text{car}}] = \mathbb{E}[\varepsilon_{\text{bus}}] = 0$. Now, when deciding whether to travel by bus or car, Alice only needs to consider the difference of utilities $U_{\text{bus}} - U_{\text{car}}$. Subtracting the preceding equations,

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

Notice that the two intercepts α_{bus} and α_{car} only enter this expression as a *difference*. This means that, as far as Alice's decision is concerned, all that matters is the value of $\alpha_{\text{bus}} - \alpha_{\text{car}}$, not the values of α_{bus} and α_{car} separately. To make this more concrete: Alice doesn't care whether $\alpha_{\text{bus}} = 3$ and $\alpha_{\text{car}} = 2$ or whether $\alpha_{\text{bus}} = 101$ and $\alpha_{\text{car}} = 100$. In each case $\alpha_{\text{bus}} - \alpha_{\text{car}} = 1$. Because the specific values of each intercept can never affect Alice's choices, and we only observe choices rather than utilities, there is no way that we can separately identify α_{bus} and α_{car}. We can only learn the *difference* of these, since it is the differences that can actually affect Alice's choices.[2] More generally, in a model with J alternatives **we can only identify the** $(J-1)$ **differences of intercepts**. Another way of thinking about this is by designating one of the alternatives as the "base case" as we did when constructing the differences of errors above. We can only learn the differences of intercepts relative to this base case.

Our car versus bus example from above contained only attributes that varied depending on whether Alice traveled by bus or car: T_{nj} and M_{nj}. In the language of Assumption 4.3, these are attributes x_{nj} of the alternatives.[3] What happens if we include an attribute s_n of the decision-maker, in other words an observed characteristic

[2]Here's a simpler example with the same structure. Suppose $X_1, \ldots, X_n \sim N(\mu_1 - \mu_2, \sigma^2)$. No matter how large n may be, we can never learn μ_1 and μ_2 separately in this example. In contrast, it's easy to learn their difference: $\bar{X}_n$ is an unbiased and consistent estimator of $\mu_1 - \mu_2$.

[3]The terminology in Assumption 4.3 can be a little confusing. If Bob is a senior citizen and Alice is not, then his M_{bus} will be lower than hers. In spite of this, we do not call M_{nj} a decision-maker attribute. The distinction is between attributes that vary across alternatives and *might also* vary across decision-makers, x_{nj}, and attributes that vary across decision-makers but *not* across alternatives, s_n. The clue is in the subscripts: s_n does not have a j index.

that *does not vary* with mode of transportation? Annual income is a good example: this varies across decision-makers—Alice may earn more than Bob—but does not vary across alternatives for the same decision-maker—Alice's income doesn't depend on which mode of transport she chooses. Defining Y to be income, suppose that

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

Are θ_{car} and θ_{bus} both identified? In other words: can we identify the effects of income Y separately for Bus and Car? Again, only differences in utility matter. Subtracting,

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

As in the example with α_{bus} and α_{car} from above, the income effects θ_{car} and θ_{bus} only matter for Alice's choices through the difference $(\theta_{\text{bus}} - \theta_{\text{car}})$. For the same reason, only this *difference* is identified: not the individual parameters themselves. In general, **we can only identify differences of effects for decision-maker attributes**. We cannot identify the effect of income on the utility of taking the bus, but we can identify how much *larger* this effect is than the effect of income on the utility of driving.

4.3.2 The scale of utility is irrelevant.

Under Assumption 4.2, Alice chooses alternative i if and only if $U_{ni} > U_{nj}$ for all $j \neq i$. For any constant $\lambda > 0$, however, this is equivalent to $\lambda U_{ni} > \lambda U_{nj}$. This shows that rescaling the utility of each alternative by a positive constant has no effect on Alice's choices. To put it another way **the scale of utility is irrelevant**.

This seemingly trivial observation has important consequences for the identification of random utility models. Suppose we specify a model of the form $U_{nj} = \mathbf{x}'_{nj}\boldsymbol{\beta} + \varepsilon_{nj}$ where $\text{Var}(\varepsilon_{nj}) = \sigma^2$. Because the scale of utility is irrelevant, we are free to multiply both sides of the model by any positive constant. Choosing to multiply by $1/\sigma$ gives the re-scaled model $U^*_{nj} = \mathbf{x}'_{nj}(\boldsymbol{\beta}/\sigma) + \varepsilon^*_{nj}$ where $U^*_{nj} \equiv U_{nj}/\sigma$ and $\varepsilon^*_{nj} \equiv \varepsilon_{nj}/\sigma$ so that $\text{Var}(\varepsilon^*_{nj}) = 1$. Because the scale of utility has no effect on Alice's choices, there is no way for us to tell these two models apart given the data we observe. This means that we cannot identify the scale of the parameter vector $\boldsymbol{\beta}$ separately from the variance of the error term ε_{nj}. Another way of saying this is that $\text{Var}(\varepsilon_{nj})$ **determines the scale** of $\boldsymbol{\beta}$.

As we will show in the next section, any of the index models from chapter 3 can be viewed as a random utility model. From this perspective, the rescaling required to compare logit and probit coefficients is merely an example of the general phenomenon that the error variance determines the scale of the model coefficients. This is because the standard logistic and standard normal distributions have different variances.

4.4 Index Models as Special Cases: Logit & Probit

Consider a setting with two alternatives, e.g. Alice chooses between taking the bus, alternative 1, or some other way of getting to work, alternative 2. Let $y_n = 1$ if Alice chooses the first alternative, bus, and zero otherwise. Now suppose that our model of representative utility is $V_{nj} = \mathbf{s}_n'\boldsymbol{\gamma}_j$. In other words, suppose that the representative utility depends on decision-maker attributes only. As we discussed in the previous section, only differences in errors matter for choices. Finally, suppose that $(\varepsilon_{n2} - \varepsilon_{n1})$ has CDF G and is independent of $\mathbf{s}_n$. Subtracting,

$$\begin{aligned} U_{n1} - U_{n2} &= (\mathbf{s}_n'\boldsymbol{\gamma}_1 - \mathbf{s}_n'\boldsymbol{\gamma}_2) + (\varepsilon_{n1} - \varepsilon_{n2}) = \mathbf{s}_n'(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2) + (\varepsilon_{n1} - \varepsilon_{n2}) \\ &= \mathbf{s}_n'\boldsymbol{\gamma} + (\varepsilon_{n1} - \varepsilon_{n2}) \end{aligned}$$

where we define $\boldsymbol{\gamma} \equiv \boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2$. Since Alice chooses alternative 1 if and only if $U_{n1} - U_{n2}$, the choice probability for this alternative is given by

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

When G is the standard logistic CDF Λ, this is *precisely* the logit model from chapter 3; when G is the standard normal CDF, it is the probit model.

At the beginning of this section, we assumed that the representative utilities only involved decision-maker attributes. When there are only two alternatives, however, we can *convert* attributes of the alternatives into decision-maker attributes in a fairly flexible way. Suppose that $\mathbf{x}_{1n}$ is a vector of attributes for the first alternative, and $\mathbf{x}_{2n}$ for the second. Then the difference of attributes $\mathbf{s}_n \equiv \mathbf{x}_{1n} - \mathbf{x}_{2n}$ does not vary across alternatives. For example, suppose that Alice must choose between taking the bus and driving. If T_{bus} and T_{car} are her door-to-door commuting time for each mode of transport, then we can allow V_{nj} to depend on $T_{\text{bus}} - T_{\text{car}}$ while still including only decision-maker attributes.

It's interesting to know that probit and logit can be viewed as random utility models, but the real value of the methods outlined above is in allowing us to specify richer and more interesting models that go *beyond* binary choice. In the next section, we explore some common models that extend logit to settings with more than two alternatives.

4.5 The Logit Family of Choice Models

As discussed in section 4.2, for most densities f that one might think to specify for the error terms $\boldsymbol{\varepsilon}_n$, the integral from (4.1) has no closed form. The so-called "logit family" of choice models constitute an important exception to this general rule, which explains their popularity in the days before inexpensive high-performance computing. Although

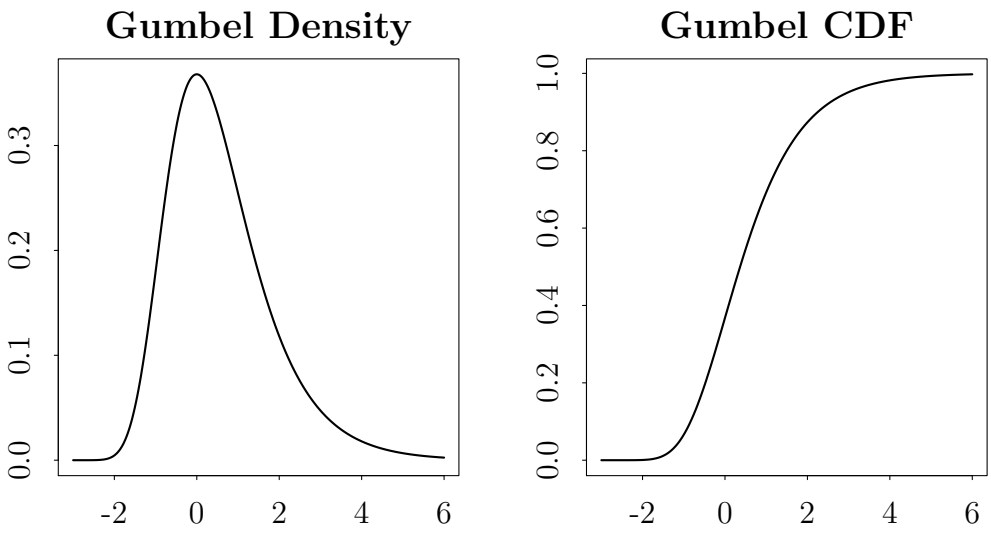

Figure 4.1: The density and CDF of Gumbel, aka Type I Extreme Value, random variable. Left panel: $f(z) = \exp(-z)\exp\{-\exp(-z)\}$. Right panel: $F(z) = \exp\{-\exp(-z)\}$.

computing is much cheaper and faster today, models from the logit family remain popular today. Indeed, logistic regression, introduced in chapter 3, is one of the many members of this venerable family of models. In this section we'll look at logit models in general and explore three special cases in detail: multinomial logit, conditional logit, and mixed logit. In the following section we'll examine an important limitation of logit models: the so-called *independence of irrelevant alternatives*.

The logit family of random utility models is constructed by assuming that the errors ε_{nj} are independent draws from a distribution with CDF $F(z) = \exp\{-\exp(-z)\}$. Under this assumption, there is a simple and intuitive closed-form solution for P_{ni}, as detailed in the following theorem. For a proof, see section 4.7.

Theorem 4.1. *Suppose that* $\varepsilon_{n1}, \ldots \varepsilon_{nJ} \sim$ *iid* F *where* $F(z) = \exp\{-\exp(-z)\}$. *Then,*

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

The distribution F from Theorem 4.1 is known as the **Gumbel** or **Type I Extreme Value** distribution. Figure 4.1 plots its CDF and density function. As a corollary of Theorem 4.1, the difference of two independent Gumbel random variables is a standard logistic random variable. In the case with two errors $(\varepsilon_{n1}, \varepsilon_{n2})$, we obtain

$$\mathbb{P}(\varepsilon_{n2} - \varepsilon_{n1} < V_{n1} - V_{n2}) = \frac{\exp(V_{n1})}{\exp(V_{n1}) + \exp(V_{n2})} = \frac{\exp(V_{n1} - V_{n2})}{1 + \exp(V_{n1} - V_{n2})}$$

after dividing the numerator and denominator by $\exp(V_{n2})$. The right-hand side of this expression equals $\Lambda(V_{n1} - V_{n2})$, where $\Lambda(z) \equiv \exp(z)/[1+\exp(z)]$ is the CDF of a standard logistic random variable, as defined in chapter 3.

By specifying the representative utilities V_{nj} in different ways, Theorem 4.1 allows us to construct a number of popular discrete choice models for settings with more than two alternatives. Here we will focus on two of them: the **multinomial logit** model, which takes $V_{nj} = \mathbf{s}_n'\boldsymbol{\gamma}_j$, and the **conditional logit** model, which takes $V_{nj} = \mathbf{x}_{nj}'\boldsymbol{\beta}$. To help us understand the differences between these two models, we will consider the following simple example.[4] Suppose that Alice wants to go fishing this weekend. She can choose to fish at the beach, from a pier, on a private boat, or on a charter boat. Call these four, mutually exclusive possibilities 1, 2, 3, and 4.

4.5.1 Multinomial Logit

The multinomial logit model arises when all attributes are fixed across alternatives, so that $V_{nj} = \mathbf{s}_n'\boldsymbol{\gamma}_j$. As discussed in subsection 4.3.1, only the *differences*, $(\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_i)$, of coefficients for decision-maker attributes are identified. Typically the coefficient associated with the first alternative, $\boldsymbol{\gamma}_1$ is normalized to zero. In the fishing example, a multinomial logit model could allow Alice's choice of where to go fishing to depend on her income, her age, and her level of fishing experience, because these do not vary depending on where she chooses to fish.

Partial effects for the multinomial logit model are tedious to derive and difficult interpret, because the attributes $\mathbf{s}_n$ are common to all alternatives. A simpler approach is to consider partial effects for **relative risk**, the effects of $\mathbf{s}_n$ on *ratios* of choice probabilities, taking one alternative as the "base case." Suppose we designate the first alternative as our base category. Normalizing $\boldsymbol{\gamma}_1 = \mathbf{0}$, we have $\exp(\mathbf{s}_n'\boldsymbol{\gamma}_1) = \exp(0) = 1$. Hence,

$$\frac{P_{ni}}{P_{n1}} = \frac{\exp\left(\mathbf{s}_n\boldsymbol{\gamma}_i\right)}{\sum_{j=1}^{J}\exp\left(\mathbf{s}_n\boldsymbol{\gamma}_j\right)} \times \frac{\sum_{j=1}^{J}\exp\left(\mathbf{s}_n\boldsymbol{\gamma}_j\right)}{\exp\left(\mathbf{s}_n\boldsymbol{\gamma}_1\right)} = \frac{\exp(\mathbf{s}_n\boldsymbol{\gamma}_i)}{\exp(\mathbf{s}_n\boldsymbol{\gamma}_1)} = \exp(\mathbf{s}_n\boldsymbol{\gamma}_i)$$

Taking logs, it follows that $\log\left(P_{ni}/P_{n1}\right) = \log\left[\exp(\mathbf{s}_n\boldsymbol{\gamma}_i)\right] = \mathbf{s}_n'\boldsymbol{\gamma}_i$. Thus, $\gamma_i^{(k)}$ is the marginal effect of $s_n^{(k)}$ on the *relative probability* that $y = i$ compared to $y = 1$, measured on the log scale. In the fishing example, suppose that the k^{th} attribute is income. Then taking fishing at the beach as the base category, $\gamma_2^{(k)}$ is the effect, measured on the log scale, of a one unit increase in income on the probability that Alice will fish from a pier *relative* to the probability that she will fish from the beach.

4.5.2 Conditional Logit

The conditional logit model includes only attributes that vary across alternatives, so that $V_{nj} = \mathbf{x}_{nj}'\boldsymbol{\beta}$ Notice that the coefficient vector $\boldsymbol{\beta}$ in the conditional logit model is fixed

[4] A third model called mixed logit *combines* the multinomial and conditional logit specifications, taking $V_{nj} = \mathbf{s}_n'\boldsymbol{\gamma}_j + \mathbf{x}_{nj}'\boldsymbol{\beta}$. For more on the mixed logit model, see Cameron and Trivedi (2005, Chapter 15) and Train (2009, Chapter 6).

across alternatives while the attributes $\mathbf{x}_{nj}$ vary. This is precisely the reverse of the multinomial logit model, in which coefficients vary across alternatives but attributes to not. In the fishing example, $\mathbf{x}_{nj}$ could include the price and abundance of fish for each alternative mode of fishing. Perhaps Alice can fish at the beach for free but is unlikely to catch anything there, while she's much more likely to catch something on a charter boat but has to pay \$50 to be allowed on the boat.

Because all of the attributes in a conditional logit model are specific to a particular alternative, it's much easier to work out partial effects in this model compared to the multinomial logit model discussed above. On your second problem set, you'll show that the **own-attribute** effects are given by

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

while the **cross-attribute** effects are

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

for $j \neq i$. Notice from these expressions that if increasing $\mathbf{x}_{nj}^{(k)}$ makes $y = j$ more likely, it must make $y = i$ less likely. This is because $P_{nj}(1 - P_{nj})$ and $P_{nj}P_{ni}$ are both positive, so the sign of the own-effects correspond to those of the coefficients $\boldsymbol{\beta}$ while those of the cross-effects have the opposite sign.

4.6 the Independence of Irrelevant Alternatives

While the closed-form expressions for the logit choice probabilities from Theorem 4.1 are extremely convenient, this convenience comes at a cost. Consider two alternatives: i and k. Under a logit model,

$$P_{ni} = \frac{\exp(V_{ni})}{\sum_{j=1}^{J}\exp(V_{nj})}, \quad P_{nk} = \frac{\exp(V_{nk})}{\sum_{j=1}^{J}\exp(V_{nj})}$$

by Theorem 4.1 and thus, taking the ratio of the two choice probabilities,

$$\frac{P_{ni}}{P_{nk}} = \exp(V_{ni} - V_{nk}). \tag{4.3}$$

Equation 4.3 says that the probability of choosing alternative i relative to that of choosing alternative k depends only the representative utilities of i and k. Because it implies that the representative utility $V_{n\ell}$ of any *other* alternative ℓ has no bearing on Alice's choice between i and k, this condition is called the **independence of irrelevant alternatives**, or **IIA** for short. IIA arises in logit models because the errors $\varepsilon_{n1}, \ldots, \varepsilon_{nJ}$ are assumed

to be independent.[5] In many real world examples, however, it is more plausible to allow dependence between the errors across alternatives. In words: "some alternatives are more similar than others" and models that ignore this can make nonsensical predictions.[6]

Consider an example in which each voter n must choose a presidential candidate $j \in \{\text{Trump}, \text{Sanders}, \text{Warren}\}$.[7] Suppose that our model of representative utility is

$$V_{nj} = (\text{Demographics}_n)'\boldsymbol{\gamma}_j + (\text{Ideology}_{nj})'\boldsymbol{\beta}$$

where Ideology_{nj} is a vector whose elements measure the ideological similarity between voter n and candidate j on a number of issues, and Demographics_n is a vector of demographic information, e.g. age, sex, household income, etc. Now, suppose we 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. Since these voters all have the same regressors, V_{nj} doesn't vary over n within the group, so we can drop the n subscript and simply write $\{V_{\text{Trump}}, V_{\text{Sanders}}, V_{\text{Warren}}\}$.

First consider a two-way race in which this group of voters must choose between Sanders and Trump. Say that 2/3 of them vote for Sanders. Then we have

$$P_{\text{Sanders}}/P_{\text{Trump}} = 2.$$

Now let's add Warren back into the mix, and consider a choice between all three candidates. Under IIA, $P_{\text{Sanders}}/P_{\text{Trump}}$ is *unaffected* by our addition of Warren to the choice set: the relative choice probabilities of any two alternatives do not depend on the representative utility of any other alternative that may be present. Because Sanders and Warren are ideologically similar, the representative utility of choosing Warren is approximately equal to that of choosing Sanders. Thus, under (4.3),

$$P_{\text{Warren}}/P_{\text{Sanders}} = \exp\left(V_{\text{Warren}} - V_{\text{Sanders}}\right) \approx 1.$$

Because the choice probabilities must sum to one across the full choice set, we have the following conditions on the choice probabilities in our three-way race under IIA:

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

[5] For details, see the proof of Theorem 4.1 in section 4.7.

[6] The classic example in which IIA fails is a choice between driving, taking a red bus, and taking a blue bus. Personally, I find the "red bus, blue bus example" too artificial to be illuminating, but if you would like to know more about it, see Train (2009, pg. 46).

[7] While it's unlikely that we'll ever see a three-way presidential race between Donald Trump, Bernie Sanders, and Elizabeth Warren, it's at least *possible*: Sanders, could run as an independent.

Substituting the first two conditions into third gives

$$P_{\text{Trump}} + 2P_{\text{Trump}} + 2P_{\text{Trump}} = 1 \implies P_{\text{Trump}} = 1/5$$

and because the choice probabilities for Warren and Sanders are approximately equal, it follows that $P_{\text{Warren}} = P_{\text{Sanders}} = 2/5$.

To summarize: in our two-way race, 2/3 of our voters preferred Sanders and 1/3 preferred Trump. Under IIA, this implied that the same voters should break 2/5 for Sanders, 2/5 for Warren, and 1/5 for Trump in a three-way race, given that Sanders and Warren are ideologically similar. But given what we know about the world these candidates, this makes *absolutely no sense*. In reality we would expect that adding Warren to the race would split the anti-Trump vote, leading to vote shares of around 1/3 Trump, 1/3 Sanders, and 1/3 Warren. To put it another way: IIA implies that a considerable share of people who prefer Trump to Sanders in a two-way race will *switch* to Warren in a three-way race, a wildly implausible prediction given what we know about these three politicians. The problem with this model is that the logit specification assumes $\varepsilon_{\text{Warren}}, \varepsilon_{\text{Sanders}}$, and $\varepsilon_{\text{Trump}}$ are independent. In reality they're not: voters who have a high value for $\varepsilon_{\text{Trump}}$ likely have very low values for $\varepsilon_{\text{Sanders}}$ and $\varepsilon_{\text{Warren}}$. Similarly, voters who have a high value for $\varepsilon_{\text{Sanders}}$ likely also have a high value for $\varepsilon_{\text{Warren}}$

To solve this problem, we need to go beyond models from the logit family. A popular model that does not impose IIA is the **probit model**, a generalization of probit regression—see chapter 3 and section 4.4—to a setting with three or more alternatives. Under this model, we assume that $\boldsymbol{\varepsilon}_n$ follows a J-dimensional multivariate normal distribution with a mean vector of zero and covariance matrix Ω. Crucially Ω need not be diagonal, allowing for correlation between the errors that we ruled out in Theorem 4.1. As usual, there is no free lunch. The added flexibility of the probit model introduces two complications. First, the choice probabilities are no longer available in closed form. Instead they must be approximated either by simulation or $(J-1)$-fold numerical integration. Second, the parameters that allow for correlation between the error terms must be *estimated* along with the parameter vector $\boldsymbol{\theta}$. For more on the probit model, see Train (2009, Chapter 5).

4.7 Appendix: Deriving Logit Choice Probabilities

The proof of Theorem 4.1 isn't very technical, but it's somewhat lengthy, which is why I've relegated it to this appendix. While I won't require you to be able to reproduce it on an exam, you may find it helpful to work through this proof. Doing so will deepen your understanding of random utility models, and give you a chance to practice applying some results from your lectures on probability from earlier in the academic year.

Proof of Theorem 4.1. As explained in section 4.1 above, I implicitly condition on (x_{nj}, s_n) throughout, so that the representative utilities can be treated as constants. To simplify the notation, I also drop the n subscripts: ε_{nj} becomes ε_j while V_{nj} becomes V_j.

Suppose that $\varepsilon_1, \ldots, \varepsilon_J \sim$ iid with CDF $F(z) = \exp\{-\exp(-z)\}$. Our goal is to find an explicit formula for the choice probabilities

$$P_i \equiv \mathbb{P}(\varepsilon_j - \varepsilon_i < V_i - V_j \quad \forall j \neq i) = \mathbb{P}(\varepsilon_j < \varepsilon_i + V_i - V_j \quad \forall j \neq i)$$

where the second equality follows by moving ε_i to the right-hand side of the inequality inside of the probability statement. Rather than calculating this directly, we'll take an indirect approach. First we *condition* on ε_i, and calculate

$$P_i(\varepsilon_i) \equiv \mathbb{P}(\varepsilon_j < \varepsilon_i + V_i - V_j \quad \forall j \neq i \,|\, \varepsilon_i).$$

Notice that this probability is a function of ε_i, the random variable upon which we condition. Given ε_i, we can treat $(\varepsilon_i + V_i - V_j)$ as a constant. For any $j \neq i$, call this constant c_j. In words, $P_i(\varepsilon_i)$ is the *joint probability* that $\varepsilon_j < c_j$ for all j other than i, conditional on ε_i. Since the errors are independent, it follows that

$$\begin{aligned}
P_i(\varepsilon_i) &= \mathbb{P}(\varepsilon_1 < c_1,\, \varepsilon_2 < c_2,\, \ldots,\, \varepsilon_{i-1} < c_{i-1},\, \varepsilon_{i+1} < c_{i+1},\, \ldots, \varepsilon_J < c_J \,|\, \varepsilon_i) \\
&= \mathbb{P}(\varepsilon_1 < c_1,\, \varepsilon_2 < c_2,\, \ldots,\, \varepsilon_{i-1} < c_{i-1},\, \varepsilon_{i+1} < c_{i+1},\, \ldots,\, \varepsilon_J < c_J) \\
&= \mathbb{P}(\varepsilon_1 < c_1)\mathbb{P}(\varepsilon_2 < c_2) \cdots \mathbb{P}(\varepsilon_{i-1} < c_{i-1})\mathbb{P}(\varepsilon_{i+1} < c_{i+1}) \cdots \mathbb{P}(\varepsilon_J < c_J) \\
&= \prod_{j \neq i} \mathbb{P}(\varepsilon_j < c_j).
\end{aligned}$$

And since F is the CDF of ε_j while $c_j \equiv \varepsilon_i + V_i - V_j$, we have shown that

$$P_i(\varepsilon_i) = \prod_{j \neq i} F(\varepsilon_i + V_i - V_j). \tag{4.4}$$

But $P_i(\varepsilon_i)$ isn't what we were after. To calculate the desired *unconditional* choice probability P_i, we apply the law of total probability. Specifically, integrating the conditional choice probability $P(\varepsilon_i)$ over the marginal distribution of ε_i, we obtain

$$P_i = \int_{-\infty}^{\infty} P_i(\varepsilon_i) f(\varepsilon_i)\, d\varepsilon_i = \int_{-\infty}^{+\infty} \left[\prod_{j \neq i} F(z + V_i - V_j) \right] f(z)\, dz \tag{4.5}$$

where f denotes the probability density function of ε_i.

The rest of the proof is just algebra: we simplify (4.5) and show that it equals the

expression from the statement of the theorem. Differentiating F,

$$f(z) = F'(z) = \exp(-z)\exp\{-\exp(-z)\}.$$

Substituting this into the function inside the integral on the right-hand side of (4.5) gives

$$\left[\prod_{j\neq i} F(z + V_i - V_j)\right] f(z) = \left[\prod_{j\neq i} \exp\{-\exp(V_j - V_i - z)\}\right] \exp(-z)\exp\{-\exp(-z)\}.$$

But since $\exp\{-\exp(-z)\} = \exp\{-\exp(V_i - V_i - z\}$, we can pull this factor inside the square brackets in the preceding equality, yielding

$$\left[\prod_{j\neq i} F(z + V_i - V_j)\right] f(z) = \left[\prod_{j=1}^{J} \exp\{-\exp(V_j - V_i - z)\}\right] \exp(-z)$$

where the product is now taken over all j rather than $j \neq i$. Using the properties of exponents, we can re-write this product as

$$\begin{aligned}\prod_{j=1}^{J} \exp\{-\exp(V_j - V_i - z)\} &= \exp\left\{\sum_{j=1}^{J} -\exp\left(V_j - V_i - z\right)\right\} \\ &= \exp\left\{-\exp(-z)\sum_{j=1}^{J}\exp\left(V_j - V_i\right)\right\}.\end{aligned}$$

Putting the pieces together, we have shown that

$$P_i = \int_{-\infty}^{\infty} \exp\left\{-\exp(-z)\sum_{j=1}^{J}\exp\left(V_j - V_i\right)\right\}\exp(-z)\,dz. \tag{4.6}$$

Comparing (4.6) to (4.5), it may not be immediately apparent that we've made our integration problem any simpler. Notice, however, that the sum $\sum_{j=1}^{J}\exp(V_j - V_i)$ does not involve z. As far as this integral is concerned, it's simply a constant. Call it K for short. Using this shorthand, we obtain

$$P_i = \int_{-\infty}^{\infty} \exp\{-K\exp(-z)\}\exp(-z)\,dz.$$

Now things are starting to look more promising. To evaluate this integral we'll use the change of variables $t = \exp(-z)$. As z diverges to $+\infty$, t converges to zero. And as z diverges to $-\infty$, t diverges to $+\infty$. Accordingly, after the change of variables our lower limit of integration will become ∞ while our upper limit will become 0. Since

$dt = -\exp(-z)\,dz$, it follows that

$$\begin{aligned} P_i &= \int_{-\infty}^{\infty} \exp\left\{-K\exp(-z)\right\}\exp(-z)\,dz = \int_{\infty}^{0} -\exp(-Kt)\,dt \\ &= \int_{0}^{\infty} \exp(-Kt)\,dt = \left.\frac{\exp(-Kt)}{-K}\right|_{0}^{\infty} \end{aligned}$$

where the third equality uses the fact that reversing the limits of integration is equivalent to multiplying the integral by -1. Since $K \equiv \sum_{j=1}^{J} \exp(V_j - V_i)$ is positive, we see that

$$P_i = \left.\frac{\exp(-Kt)}{-K}\right|_{0}^{\infty} = 0 - \left[\frac{\exp(0)}{-K}\right] = \frac{1}{K}.$$

Finally, substituting the definition of K and using the properties of exponents,

$$\begin{aligned} P_i = \frac{1}{K} &= \frac{1}{\sum_{j=1}^{J}\exp(V_j - V_i)} = \frac{1}{\sum_{j=1}^{J}\exp(V_j)\exp(-V_i)} \\ &= \frac{1}{\exp(-V_i)\sum_{j=1}^{J}\exp(V_j)} = \frac{\exp(V_i)}{\sum_{j=1}^{J}\exp(V_j)}. \end{aligned}$$

□

Chapter 5

Sample Selection Models

So far we have always assumed that $(y_1, \mathbf{x}_1), \ldots, (y_N, \mathbf{x}_N)$ are a random sample from the population of interest. What if they aren't? A sample that is *not* representative of the population we hope to study is called a **selected sample**, and methods to handle this problem are typically called **sample selection models**.

5.1 Examples of Sample Selection

Graduate Admissions Admissions committees read applications to try to screen out candidates who are unlikely to succeed in graduate school. But perhaps their judgement is flawed. Maybe they give too much weight to letters of reference from famous economists and not enough to undergraduate grades; or perhaps they're biased in favor of applicants who studied at their own *alma mater*. Suppose we wanted to improve admissions decisions at Oxford. Let y be a person's overall mark in the 1st year of the Economics MPhil, and $\mathbf{x}$ be a vector representing the information in her application: undergraduate grades, the strength of her letters of reference, etc. Ideally we'd like to know the predictive relationship between $\mathbf{x}$ and y: $\mathbb{E}(y|\mathbf{x})$, but there's a problem. While we observe $\mathbf{x}$ for everyone who applies, we only observe y for candidates who enrolled at Oxford. To improve our admissions decisions, we need to learn the relationship between $\mathbf{x}$ and y in the population of *all applicants* but our data for y constitute a selected sample. Alice was admitted to Oxford and decided to attend; Bob was not admitted; Chiara was admitted but went to MIT instead. This means that we only observe y for Alice. Based on what we know about them, Alice, Bob, and Chiara are likely very different. Regressing y on $\mathbf{x}$ using data for people like Alice may tell us little or nothing about whether we should admit people like Bob or Chiara.

Wage Offers Another classic example of sample selection comes from Gronau (1974). Suppose we are interested in learning how wage offers w^o vary with a person's characteristics $\mathbf{x}$, e.g. age, experience, and education. Ideally we'd like to learn $\mathbb{E}(w^o|\mathbf{x})$ for the

whole population, but we can only observe wage offers for people who were both offered a job and *accepted* it, i.e. those who are currently employed. For reasons similar to those described in the admissions example, it's very likely that $\mathbb{E}(w^o|\mathbf{x})$ most likely does not equal $\mathbb{E}(w^o|\mathbf{x}, \text{Employed})$.

5.2 The Heckman Selection Model

The classic econometric model for sample selection is the so-called **Heckman Selection Model**. This model allows both selection on observables and unobservables. In other words, it allows people who attend Oxford to differ from those who do not attend both in ways that we can observe from their application files and in ways that we cannot. To accomplish this impressive feat, we rely on fairly strong parametric assumptions. These assumptions can be weakened to a certain extent, but not as much as we might like. Selection on unobservables is fundamentally a very hard nut to crack! Our first assumption gives the basic structure of the selection model we work with below. The notation $\mathbb{1}\{A\}$ represents the **indicator** of the event A: the function that equals one if A occurs and zero otherwise.

Assumption 5.1 (Sample Selection Model)**.** *Suppose that*

$$y_1 = \mathbf{x}_1'\boldsymbol{\beta}_1 + u_1 \tag{5.1}$$

$$y_2 = \mathbb{1}\{\mathbf{x}'\boldsymbol{\delta}_2 + v_2 > 0\} \tag{5.2}$$

where y_2 and $\mathbf{x}' \equiv (\mathbf{x}_1', \mathbf{x}_2')$ are always observed but y_1 is only observed when $y_2 = 1$.

We call (5.1) the **outcome equation** and (5.2) the **participation equation**. The outcome y_1 is only observed for people who participate: $y_2 = 1$. Our goal is to learn the parameter $\boldsymbol{\beta}_1$ that relates $\mathbf{x}_1$ to y_1 despite only observing y_1 for a selected sample. In the admissions example, exam results are only available for students who complete the first year. The vector $\mathbf{x}$ contains all the information that we can observe about candidates when they apply to the MPhil. The first component, $\mathbf{x}_1$, contains the covariates that enter both the participation and outcome equations, while the second component, $\mathbf{x}_2$ contains the covariates that only enter the participation equation. It's possible that $\mathbf{x}_2$ is empty: in this case all of the covariates enter both equations. When $\mathbf{x}_2$ is *not empty*, we say that there is an **exclusion restriction**: there is at least one covariate that matters for participation but not for outcomes, effectively an instrumental variable. As we will discuss further below, exclusion restrictions aren't strictly required to apply the Heckman selection model, but if available they are extremely helpful.

Assumption 5.1 assumes that both the participation and outcome equations are linear in parameters and depend on observed covariates $\mathbf{x}$ but says mothing about the

unobservable error terms (u_1, v_2). Selection on observables enters the model through $\mathbf{x}_1$: application files of people who attend Oxford are different from those of people who do not attend in ways that matter for first-year MPhil grades. In other words, the observables $\mathbf{x}_1$ enters both the selection and outcome equations. To allow selection on *unobservables*, we need to allow an unobserved variable to enter both equations. This is equivalent to allowing (u_1, v_2) to be statistically *dependent*, in that knowing something about v_2, which affects participation, tells us something about u_1, which affects outcomes. The remaining assumptions of the Heckman Selection Model concern the unobserved errors.

Assumption 5.2 (Exogeneity)**.** (u_1, v_2) *are mean zero and jointly independent of* $\mathbf{x}$*.*

The first part of Assumption 5.2 is not restrictive. It merely requires that $\mathbb{E}(u_1)$ and $\mathbb{E}(v_2)$ are both zero. This is satisfied automatically if we include a constant in both the participation and outcome equations, i.e. if the first element of $\mathbf{x}_1$ is 1. In contrast, the second part of Assumption 5.2 is somewhat stronger than the usual exogeneity assumption that we make in a regression model. Notice that the participation equation (5.2) is a binary outcome model, just like the ones we examined in chapter 3 above. The Heckman Selection Model assumes that it is a probit model.

Assumption 5.3 (Probit Participation Model)**.** $v_2 \sim Normal(0, 1)$

The final assumption concerns the dependence between the unobservable in the outcome equation, u_1, and the unobservable in the participation equation, v_2.

Assumption 5.4. $\mathbb{E}(u_1|v_2) = \gamma_1 v_2$ *where* γ_1 *is an unknown constant.*

Assumption 5.4 maintains that the conditional mean function of u_1 given v_2 is *linear*. As will be further clarified in our derivations below, the unknown slope parameter γ_1 controls the nature and extent of sample selection bias. If γ_1 is positive then, among candidates with equally strong applications, those who choose to attend Oxford tend to perform *better* on exams than those who do not choose to attend. The larger the value of γ_1, the stronger the effect. Assumption 5.4 can be relaxed to allow for more general forms of dependence between u_1 and v_2. In fact you will consider just such an extension in your problem set! Depending on which references you consult, you may see Assumptions 5.3 and 5.4 replaced by the assumption that (u_1, v_2) are *jointly* normally distributed. This joint normality assumption implies the assumptions we use here, but is strictly stronger: we do *not* require u_1 to be normally distributed.

5.3 Two Key Lemmas

Under Assumptions 5.1–5.4, $\boldsymbol{\beta}_1$ is identified from a pair of straightforward regressions using data for the selected sample only, as we will show below. Rather than jumping

straight to the answer, we'll take things in steps, beginning with a lemma that reveals the anatomy of the selection problem.

Lemma 5.1. *Under Assumptions 5.1, 5.2, and 5.4,*

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

We'll show why Lemma 5.1 follows from our assumptions shortly. But before doing so, we'll answer a more important question: what does it mean? Both $\boldsymbol{\beta}_1$ and γ_1 are constants while $\mathbf{x}' = (\mathbf{x}_1', \mathbf{x}_2')$ is a vector of observables. Now, $\mathbb{E}(v_2|\mathbf{x}, y_2 = 1)$ is a *function* of $\mathbf{x}$, call it $h(\mathbf{x})$, that maps the regressor vector $\mathbf{x}$ to a scalar. If $h(\cdot)$ were known, we could simply define $w \equiv h(\mathbf{x})$ and treat this as an *additional regressor*, yielding

$$\mathbb{E}(y_1|\mathbf{x}, y_2 = 1) = \mathbb{E}\left(y_1|\mathbf{x}_1, w, y_2 = 1\right) = \mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma_1 w \tag{5.3}$$

since $\mathbb{E}(y_1|\mathbf{x}, y_2)$ only depends on $\mathbf{x}$ through $\mathbf{x}_1$ and $w \equiv h(\mathbf{x})$, as shown in Lemma 5.1. Equation 5.3 has two important consequences. First, it shows that regressing y_1 on $(\mathbf{x}_1, w)$ for the subset of individuals with $y_2 = 1$ would suffice to identify both $\boldsymbol{\beta}_1$ and γ_1. This consequence is so significant that it bears repeating: regressing y_1 on $\mathbf{x}_1$ and w using data for the *selected population* allows us to identify the parameter $\boldsymbol{\beta}_1$ that governs the relationship between y and $\mathbf{x}_1$ in the *whole population*. In other words, we can learn the effect of education on wage offers in the population as a whole despite observing wages for people who are employed only. To do this, we need only find a way to calculate the "additional regressor" $w \equiv h(\mathbf{x}) \equiv \mathbb{E}(v_2|\mathbf{x}, y_2 = 1)$. Second, (5.3) shows that any bias arising from sample selection comes from dependence between the errors in the participation and outcome equations. If $\gamma_1 = 0$, so that $\mathbb{E}(u_1|v_2) = \mathbb{E}(u_1) = 0$, there is no sample selection bias. In this case $\mathbb{E}(y_1|\mathbf{x}, y_2 = 1) = \mathbf{x}_1'\boldsymbol{\beta}$ so regressing y_1 on $\mathbf{x}_1$ for the subset of individuals with $y_2 = 1$ identifies $\boldsymbol{\beta}_1$. Now that we understand the significance of Lemma 5.1, we're ready to delve into the proof.

Proof of Lemma 5.1. This proof proceeds in three steps. In step one we show that u_1 and $\mathbf{x}$ are conditionally independent given v_2. In words: conditioning on $(v_2, \mathbf{x})$ gives the same information about u_1 as conditioning on v_2 only. In step two, we use the result of step one to show that $\mathbb{E}(y_1|\mathbf{x}, v_2) = \mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma_1 v_2$. In step three we combine the results of steps one and two to complete the proof.

For step one we use Assumption 5.2, which states that (u_1, v_2) are jointly independent of $\mathbf{x}$. Let $f_{1,2|\mathbf{x}}$ be the conditional density of (u_1, v_2) given $\mathbf{x}$ and $f_{1,2}$ be the unconditional density of (u_1, v_2). By independence, $f_{1,2|\mathbf{x}} = f_{1,2}$. Similarly, let $f_{1|\mathbf{x}}, f_{2|\mathbf{x}}$ be the densities of $u_1|\mathbf{x}$ and $v_2|\mathbf{x}$ and f_1, f_2 be the corresponding unconditional densities. Since joint independence implies marginal independence, $f_{1|\mathbf{x}} = f_1$ and $f_{2|\mathbf{x}} = f_2$. Now let $f_{1|2,\mathbf{x}}$ be the conditional density of u_1 given $(v_2, \mathbf{x})$. By the definition of a conditional density and

the independence results we have just discussed,

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

Since $f_{1|2,\mathbf{x}}$ does not depend on $\mathbf{x}$, u_1 is conditionally independent of $\mathbf{x}$ given v_2.

For step two, we begin by substituting the outcome equation, (5.1), for y_1 to yield

$$\mathbb{E}(y_1|\mathbf{x},v_2) = \mathbb{E}(\mathbf{x}_1'\boldsymbol{\beta}_1 + u_1|\mathbf{x},v_2) = \mathbf{x}_1'\boldsymbol{\beta}_1 + \mathbb{E}(u_1|\mathbf{x},v_2)$$

using the fact that $\mathbf{x}_1$ is a subset of $\mathbf{x}$ and hence is "known" conditional on $\mathbf{x}$. Now, by step one $\mathbb{E}(u_1|\mathbf{x},v_2) = \mathbb{E}(u_1|v_2)$ and by Assumption 5.4 $\mathbb{E}(u_1|v_2) = \gamma_1 v_2$. It follows that $\mathbb{E}(y_1|\mathbf{x},v_2) = \mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma_1 v_2$, completing step 2.

Now we have all the ingredients needed to prove the lemma. By iterated expectations,

$$\mathbb{E}(y_1|\mathbf{x},y_2) = \mathbb{E}_{v_2|(\mathbf{x},y_2)}\left[\mathbb{E}\left(y_1|\mathbf{x},y_2,v_2\right)\right]. \tag{5.4}$$

Now, by the participation equation, (5.2), y_2 is a deterministic function of $(\mathbf{x},v_2)$. It follows that conditioning on y_2 after we have already conditioned on $(\mathbf{x},v_2)$ is *redundant*. Hence, using the result of step two, we can write the inner expectation from (5.4) as

$$\mathbb{E}(y_1|\mathbf{x},y_2,v_2) = \mathbb{E}(y_1|\mathbf{x},v_2) = \mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma_1 v_2. \tag{5.5}$$

Finally, substituting (5.5) into (5.4), we obtain

$$\mathbb{E}(y_1|\mathbf{x},y_2) = \mathbb{E}\left[\mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma_1 v_2|\mathbf{x},y_2\right] = \mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma_1\mathbb{E}(v_2|\mathbf{x},y_2)$$

since $\mathbf{x}_1$ is a subset of $\mathbf{x}$ and γ_1 is a constant. The preceding equality holds for any value of y_2. Setting $y_2 = 1$ gives the desired result. □

So far we have only used Assumptions 5.1–5.2 and 5.4. As explained in the discussion above, all that remains is for us to determine the function $h(\mathbf{x}) \equiv \mathbb{E}(v_2|\mathbf{x}|y_2 = 1)$. This is where Assumption 5.3 finally makes its appearance. Under the assumption that v_2 is a standard normal random variable, this function takes a simple form: it equals the ratio of the standard normal density and CDF, each evaluated at $\mathbf{x}'\boldsymbol{\delta}_2$.

Lemma 5.2. *Under Assumptions 5.1–5.3,*

$$\mathbb{E}(v_2|\mathbf{x},y_2 = 1) = \frac{\varphi(\mathbf{x}'\boldsymbol{\delta}_2)}{\Phi(\mathbf{x}'\boldsymbol{\delta}_2)}$$

where $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal probability density function and CDF.

The function $\lambda(c) \equiv \varphi(c)/\Phi(c)$ that features in Lemma 5.2 is called the **inverse Mills**

Ratio.[1] The lemma shows that $\mathbb{E}(v_2|\mathbf{x}, y_2 = 1) = \lambda(\mathbf{x}'\boldsymbol{\delta}_2)$. Thus, if we can identify the parameter $\boldsymbol{\delta}$ from the participation equation, we will have completely pinned down $\mathbb{E}(v_2|\mathbf{x}, y_2 = 1)$, solving the sample selection problem. The proof of Lemma 5.2 relies on a fact concerning truncated normal distributions. In particular, if $z \sim N(0,1)$ then $\mathbb{E}(z|z > c) = \varphi(c)/[1 - \Phi(c)]$ for any constant c. A proof of this appears in section 5.5.

Proof of Lemma 5.2. Let $\boldsymbol{x}$ be a realization of the random vector $\mathbf{x}$. Conditional on $\{\mathbf{x} = \boldsymbol{x}\}$, by (5.2) the event $\{y_2 = 1\}$ is equivalent to $\{v_2 > -\boldsymbol{x}'\boldsymbol{\delta}_2\}$ and thus

$$\mathbb{P}(v_2 \leq t|\mathbf{x} = \boldsymbol{x}, y_2 = 1) = \mathbb{P}(v_2 \leq t|\mathbf{x} = \boldsymbol{x}, v_2 > -\boldsymbol{x}'\boldsymbol{\delta}_2).$$

Applying the definition of conditional probability to the right hand side gives

$$\mathbb{P}(v_2 \leq t|\mathbf{x} = \boldsymbol{x}, y_2 = 1) = \frac{\mathbb{P}\left(\{v_2 \leq t\} \cap \{v_2 > -\boldsymbol{x}'\boldsymbol{\delta}_2\} \,|\mathbf{x} = \boldsymbol{x}\right)}{\mathbb{P}(v_2 > -\boldsymbol{x}'\boldsymbol{\delta}_2|\mathbf{x} = \boldsymbol{x})}.$$

Like t, the product $-\boldsymbol{x}'\boldsymbol{\delta}_2$ is simply a constant. Call it c for short. Using this shorthand, the numerator of the preceding equality is $\mathbb{P}(\{v_2 \leq t\} \cap \{v_2 > c\} \,|\mathbf{x} = \boldsymbol{x})$ while the denominator is $\mathbb{P}(v_2 > c|\mathbf{x} = \boldsymbol{x})$. Each of these is simply the probability of v_2 falling in a particular interval given that $\mathbf{x} = \boldsymbol{x}$. By Assumption 5.2, however, v_2 and $\mathbf{x}$ are *independent*. It follows that these conditional probabilities given $\mathbf{x}$ equal the corresponding unconditional probabilities, and thus

$$\mathbb{P}(v_2 \leq t|\mathbf{x} = \boldsymbol{x}, y_2 = 1) = \frac{\mathbb{P}\left(\{v_2 \leq t\} \cap \{v_2 > c\}\right)}{\mathbb{P}(v_2 > c)} = \mathbb{P}(v_2 \leq t|v_2 > c)$$

again using the definition of conditional probability. Thus, we have shown that the distribution of v_2 given $\{\mathbf{x} = \boldsymbol{x}, y_2 = 1\}$ coincides with the distribution of v_2 given $\{v_2 > c\}$, again using the shorthand $c \equiv -\boldsymbol{x}'\boldsymbol{\delta}_2$. Thus, to complete the proof we simply need to calculate $\mathbb{E}(v_2|v_2 > c)$. By Assumption 5.3, $v_2 \sim N(0,1)$ so we require the conditional expectation of a standard normal random variable given that it has exceeded a specified threshold c. Applying Lemma 5.2 from the appendix to this chapter,

$$\mathbb{E}(v_2|\mathbf{x} = \boldsymbol{x}, y_2 = 1) = \mathbb{E}(v_2|v_2 > c) = \frac{\varphi(c)}{1 - \Phi(c)}.$$

Because the standard normal density function φ is symmetric about zero, $\varphi(c) = \varphi(-c)$. For the same reason, $1 - \Phi(c) = \Phi(-c)$. Therefore,

$$\mathbb{E}(v_2|\mathbf{x} = \boldsymbol{x}, y_2 = 1) = \frac{\varphi(c)}{1 - \Phi(c)} = \frac{\varphi(-c)}{\Phi(-c)} = \frac{\varphi(\boldsymbol{x}'\boldsymbol{\delta}_2)}{\Phi(\boldsymbol{x}'\boldsymbol{\delta}_2)}. \qquad \square$$

[1] I use the notation $\lambda(\cdot)$ for the inverse Mills Ratio because it is standard. Note that this is *not the same thing* as the function $\lambda(\cdot)$ from chapter 3!

5.4 The Heckman Two-step Estimator

The preceding section contained a large number of details, so before proceeding let's take stock of what we've learned. Lemma 5.1 established that

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

while Lemma 5.2 showed that

$$\mathbb{E}(v_2|\mathbf{x}, y_2 = 1) = \lambda(\mathbf{x}'\boldsymbol{\delta}_2), \quad \lambda(c) \equiv \frac{\varphi(c)}{\Phi(c)}.$$

Substituting the second expression into the first, we see that

$$\mathbb{E}(y_1|\mathbf{x}, y_2 = 1) = \mathbf{x}_1'\boldsymbol{\beta}_1 + \gamma_1\lambda(\mathbf{x}'\boldsymbol{\delta}_2), \quad \lambda(c) \equiv \frac{\varphi(c)}{\Phi(c)} \tag{5.6}$$

Given a random sample of N observations $(y_{2i}, \mathbf{x}_i)$ in which we only observe y_{i1} if $y_{2i} = 1$, Equation 5.6 immediately suggests a simple estimation procedure, commonly called the **Heckman two-step** estimator, or **heckit**:

Step 1 Estimate $\lambda(\mathbf{x}'\boldsymbol{\delta}_2)$.

- Run the probit regression $\mathbb{P}(y_{2i} = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_i'\boldsymbol{\delta}_2)$ using the full sample of observations for y_{2i} and $\mathbf{x}_i$.
- Set $\widehat{\lambda}_i \equiv \lambda(\mathbf{x}_i'\widehat{\boldsymbol{\delta}}_2)$ where $\widehat{\boldsymbol{\delta}}_2$ is the probit estimate for $\boldsymbol{\delta}_2$.

Step 2 Estimate (β_1, γ_1).

- Run an OLS regression of y_{i1} on $(\mathbf{x}_{1i}, \widehat{\lambda}_i)$ using the selected sample of observations for which y_{i1} is observed.
- Report the second-step OLS estimates $(\widehat{\boldsymbol{\beta}}_1, \widehat{\gamma}_1)$.

While we will not go into the details here, it can be shown that the Heckman two-step estimator is consistent and asymptotically normal, in particular

$$\begin{bmatrix} \widehat{\boldsymbol{\delta}}_2 \\ \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\gamma}_1 \end{bmatrix} \to_p \begin{bmatrix} \boldsymbol{\delta}_2 \\ \boldsymbol{\beta}_1 \\ \gamma_1 \end{bmatrix} \quad \text{and} \quad \sqrt{N}\begin{bmatrix} \widehat{\boldsymbol{\delta}}_2 - \boldsymbol{\delta}_2 \\ \widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \widehat{\gamma}_1 - \gamma_1 \end{bmatrix} \to_d \text{Normal}(\mathbf{0}, \boldsymbol{\Omega}) \quad \text{as } N \to \infty.$$

Calculating the asymptotic covariance matrix $\boldsymbol{\Omega}$ is somewhat involved. Because the second-step OLS regression does not take into account the fact that $\widehat{\lambda}_i$ is a *generated regressor*, i.e. that it is estimated from the first-step probit, the usual regression standard errors are incorrect. To ensure that you obtain the correct standard errors, it's preferable to use a packaged heckit routine rather than "rolling your own."

So what is the big picture here? How exactly does heckit solve the selection bias problem? If we regress y_{1i} on $\mathbf{x}_{1i}$ for the selected sample, there is an omitted variable. Under the Heckit assumptions, we showed that this omitted variable is *precisely* $\lambda(\mathbf{x}_i'\boldsymbol{\delta}_2)$. Hence, a regression of y_{1i} on $\mathbf{x}_{1i}$ and $\lambda(\mathbf{x}_i'\boldsymbol{\delta}_2)$ for the selected sample is correctly specified, and recovers our parameters of interest. If you have been paying close attention, however, you may have noticed something strange: the regression equation in (5.6) includes $\mathbf{x}_1$ in *two places*. It appears first as $\mathbf{x}_1'\boldsymbol{\beta}_1$ and again as a constituent of $\mathbf{x}' \equiv (\mathbf{x}_1', \mathbf{x}_2')$ in $\lambda(\mathbf{x}'\boldsymbol{\delta}_2)$. Indeed, as mentioned above, we can allow $\mathbf{x}_2$ to be completely empty, in which case we say that there is no *exclusion restriction*. Consider the simplest possible case, where $\mathbf{x}_1 = (1, x)$ and $\mathbf{x}_2$ is empty so that

$$\mathbb{E}(y_1|\mathbf{x}, y_2 = 1) = \beta_0 + \beta_1 x + \gamma_1 w, \quad w \equiv \lambda(\delta_0 + \delta_1 x).$$

If $\lambda(\cdot)$ were a linear function, then the regressors x and w would be perfectly linearly dependent and γ_1 and β_1 would not be separately identified. Because λ is in fact a *nonlinear function*, w and x will not be perfectly linearly dependent. Thus, in the case without exclusion restrictions, identification in the Heckman selection model comes solely from the nonlinearity of λ. Depending on the values at which λ is evaluated, however, λ can be *close* to linear, leading to extremely imprecise estimates. If we have an exclusion restriction, e.g. a variable x_2 that enters the participation equation but not the outcome equation, then

$$\mathbb{E}(y_1|\mathbf{x}, y_2 = 1) = \beta_0 + \beta_1 x + \gamma_1 w, \quad w \equiv \lambda(\delta_0 + \delta_1 x_1 + \delta_2 x_2).$$

In this case, identification no longer comes solely from the nonlinearity of λ. Even if λ were a linear function, this regression would still be identified as long as x_2 and x_1 are not perfectly correlated: x_2 induces variation in w that is unrelated to x_1. As a rule, the Heckman two-step estimator tends to perform better in settings where an exclusion restriction is available. This of course raises the important question: where does an exclusion restriction come from? Unfortunately there is no general answer to this question. In effect, an exclusion restriction is like an instrumental variable. The question of whether a particular regressor is excluded from the outcome equation can only be evaluated in the context of a particular applied example.

5.5 Appendix: The Mean of a Truncated Normal

Lemma 5.3. *Suppose that $z \sim N(0,1)$ and let φ and Φ be the standard normal density and CDF, respectively. Then for any constant c,*

$$\mathbb{E}(z|z > c) = \frac{\varphi(c)}{1 - \Phi(c)}.$$

Proof of Lemma 5.3. We first calculate the conditional CDF F of z given that $z > c$. By the definition of conditional probability,

$$\mathbb{P}(z \leq t|z > c) = \frac{\mathbb{P}\left(\{z \leq t\} \cap \{z > c\}\right)}{\mathbb{P}(z > c)} = \begin{cases} 0, & t \leq c \\ \mathbb{P}(c < z \leq t)/\mathbb{P}(z > c), & t > c. \end{cases}$$

Now, since z is standard normal,

$$\begin{aligned} \mathbb{P}(z > c) &= 1 - \mathbb{P}(z \leq c) = 1 - \Phi(c) \\ \mathbb{P}(c < z \leq t) &= \mathbb{P}(z \leq t) - \mathbb{P}(z \leq c) = \Phi(t) - \Phi(c) \end{aligned}$$

and hence

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

To find the corresponding conditional density of z given $z > c$, we differentiate the conditional CDF, yielding

$$f(t) \equiv \frac{d}{dt}F(t) = \mathbb{1}\{c < t\} \frac{\varphi(t)}{1 - \Phi(c)}.$$

The desired conditional expectation is $\mathbb{E}(z|z > c) \equiv \int_{-\infty}^{\infty} t f(t)\, dt$. Integrating,

$$\begin{aligned} \int_{-\infty}^{\infty} t f(t)\, dt &= \int_c^{\infty} \frac{t\varphi(t)}{1 - \Phi(c)}\, dt = \left[\frac{1}{1 - \Phi(c)}\right] \left(\frac{1}{\sqrt{2\pi}}\right) \int_c^{\infty} t \exp\left\{-\frac{t^2}{2}\right\} dt \\ &= \left[\frac{1}{1 - \Phi(c)}\right] \left(\frac{1}{\sqrt{2\pi}}\right) \left[-\exp\left\{-\frac{t^2}{2}\right\}\right]_c^{\infty} \\ &= \left[\frac{1}{1 - \Phi(c)}\right] \left(\frac{\exp\{-c^2/2\}}{\sqrt{2\pi}}\right) = \frac{\varphi(c)}{1 - \Phi(c)}. \end{aligned}$$

□

Appendix A

Errata

Errors in equations, proofs etc. from earlier versions of these notes are listed below along with the date and time when they were corrected. Minor non-mathematical typos from previous versions, e.g. duplicated words words or spelling misteaks, are not listed here. The errors are printed in red and the corrections in blue.

Bibliography

Angrist, J.D., 2001. Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. Journal of business & economic statistics 19, 2–28.

Cameron, A.C., Trivedi, P.K., 2005. Microeconometrics: methods and applications. Cambridge university press.

Lewbel, A., 2019. The identification zoo: Meanings of identification in econometrics. Journal of Economic Literature 57, 835–903.

Liu, C., 2004. Robit regression: a simple robust alternative to logistic and probit regression. Applied Bayesian Modeling and Casual Inference from Incomplete-Data Perspectives , 227–238.

Train, K.E., 2009. Discrete choice methods with simulation. Cambridge university press.

Windmeijer, F.A., 1995. Goodness-of-fit measures in binary choice models. Econometric reviews 14, 101–116.