Maximum Likelihood Methods

- Some of the models used in econometrics specify the complete probability distribution of the outcomes of interest rather than just a regression function.
- Sometimes this is because of special features of the outcomes under study - for example because they are discrete or censored, or because there is serial dependence of a complex form.
- When the complete probability distribution of outcomes given covariates is specified we can develop an expression for the probability of observation of the responses we see as a function of the unknown parameters embedded in the specification.
- We can then ask what values of these parameters maximise this probability for the data we have. The resulting statistics, functions of the observed data, are called *maximum likelihood estimators*. They possess important optimality properties and have the advantage that they can be produced in a rule directed fashion.

Estimating a Probability

- Suppose $Y_1, \ldots Y_n$ are binary independently and identically distributed random variables with $P[Y_i = 1] = p$, $P[Y_i = 0] = 1-p$ for all i.
- We might use such a model for data recording the occurrence or otherwise of an event for n individuals, for example being in work or not, buying a good or service or not, etc.
- Let $y_1, \ldots, y_n$ indicate the data values obtained and note that in this model

$$\begin{aligned} P[Y_1 = y_1 \cap \cdots \cap Y_n = y_n] &= \prod_{i=1}^{n} p^{y_i}(1-p)^{(1-y_i)} \\ &= p^{\sum_{i=1}^{n} y_i}(1-p)^{\sum_{i=1}^{n}(1-y_i)} \\ &= L(p; y). \end{aligned}$$

  With any set of data $L(p; y)$ can be calculated for any value of p between 0 and 1. The result is the probability of observing the data to hand for each chosen value of p.
- One strategy for estimating p is to use that value that maximises this probability. The resulting estimator is called the *maximum likelihood estimator* (MLE) and the maximand, $L(p; y)$, is called the *likelihood function*.

Log Likelihood Function

- The maximum of the *log likelihood function*, $l(p; y) = \log L(p, y)$, is at the same value of p as is the maximum of the likelihood function (because the log function is monotonic).
- It is often easier to maximise the log likelihood function (LLF). For the problem considered here the LLF is

$$l(p; y) = \left(\sum_{i=1}^{n} y_i \right) \log p + \sum_{i=1}^{n} (1 - y_i) \log(1 - p).$$

Let

$$\hat{p} = \arg\max_p L(p; y) = \arg\max_p l(p; y).$$

On differentiating we have the following.

$$\begin{aligned} l_p(p; y) &= \frac{1}{p} \sum_{i=1}^{n} y_i - \frac{1}{1-p} \sum_{i=1}^{n} (1 - y_i) \\ l_{pp}(p; y) &= -\frac{1}{p^2} \sum_{i=1}^{n} y_i - \frac{1}{(1-p)^2} \sum_{i=1}^{n} (1 - y_i). \end{aligned}$$

Note that $l_{pp}(p; y)$ is always negative for admissable p so the optimisation problem has a unique solution corresponding to a maximum. The solution to $l_p(\hat{p}; y) = 0$ is

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

just the mean of the observed values of the binary indicators, equivalently the proportion of 1's observed in the data.

Likelihood Functions and Estimation in General

- Let Y_i, $i = 1, \ldots, n$ be continuously distributed random variables with joint probability density function $f(y_1, \ldots, y_n, \theta)$.
- The probability that Y falls in infinitesimal intervals of width $dy_1, \ldots dy_n$ centred on values $y_1, \ldots, y_n$ is

  $$A = f(y_1, \ldots, y_n, \theta) dy_1 dy_2 \ldots dy_n$$

  Here only the joint density function depends upon θ and the value of θ that maximises $f(y_1, \ldots, y_n, \theta)$ also maximises A.
- In this case the likelihood function is defined to be the joint *density* function of the Y_i's.
- When the Y_i's are discrete random variables the likelihood function is the joint probability mass function of the Y_i's, and in cases in which there are discrete and continuous elements the likelihood function is a combination of probability density elements and probability mass elements.
- In all cases the likelihood function is a function of the observed data values that is equal to, or proportional to, the probability of observing these particular values, where the constant of proportionality does not depend upon the parameters which are to be estimated.

Likelihood Functions and Estimation in General

- When Y_i, $i = 1, \ldots, n$ are *independently* distributed the joint density (mass) function is the *product* of the marginal density (mass) functions of each Y_i, the likelihood function is

$$L(y; \theta) = \prod_{i=1}^{n} f_i(y_i; \theta),$$

and the log likelihood function is the *sum*:

$$l(y; \theta) = \sum_{i=1}^{n} \log f_i(y_i; \theta).$$

There is a subscript i on f to allow for the possibility that each Y_i has a distinct probability distribution.

- This situation arises when modelling conditional distributions of Y given some covariates x. In particular, $f_i(y_i; \theta) = f_i(y_i|x_i; \theta)$, and often $f_i(y_i|x_i; \theta) = f(y_i|x_i; \theta)$.

- In time series and panel data problems there is often dependence among the Y_i's. For any list of random variables $Y = \{Y_1, \ldots, Y_n\}$ define the $i-1$ element list $Y_{i-} = \{Y_1, \ldots, Y_{i-1}\}$. The joint density (mass) function of Y can be written as

$$f(y) = \prod_{i=2}^{n} f_{y_i|y_{i-}}(y_i|y_{i-}) f_{y_1}(y_1),$$

Invariance

- Note that (parameter free) monotonic transformations of the Y_i's (for example, a change of units of measurement, or use of logs rather than the original y data) usually leads to a change in the value of the maximised likelihood function when we work with continuous distributions.
- If we transform from y to z where $y = h(z)$ and the joint density function of y is $f_y(y;\theta)$ then the joint density function of z is
$$f_z(z;\theta) = \left|\frac{\partial h(z)}{\partial z}\right| f_y(h(z);\theta).$$
- For any given set of values, y^*, the value of θ that maximises the likelihood function $f_y(y^*,\theta)$ also maximises the likelihood function $f_z(z^*;\theta)$ where $y^* = h(z^*)$, so the maximum likelihood estimator is invariant with respect to such changes in the way the data are presented.
- However the maximised likelihood functions will differ by a factor equal to $\left|\frac{\partial h(z)}{\partial z}\right|_{z=z^*}$.
- The reason for this is that we omit the infinitesimals $dy_1, \ldots dy_n$ from the likelihood function for continuous variates and these change when we move from y to z because they are denominated in the units in which y or z are measured.

Maximum Likelihood: Properties

- Maximum likelihood estimators possess another important *invariance property*. Suppose two researchers choose different ways in which to parameterise the same model. One uses θ, and the other uses $\lambda = h(\theta)$ where this function is one-to-one. Then faced with the same data and producing estimators $\hat{\theta}$ and $\hat{\lambda}$, it will always be the case that $\hat{\lambda} = h(\hat{\theta})$.
- There are a number of important consequences of this:
  - For instance, if we are interested in the ratio of two parameters, the MLE of the ratio will be the ratio of the ML estimators.
  - Sometimes a re-parameterisation can improve the numerical properties of the likelihood function. Newton's method and its variants may in practice work better if parameters are rescaled.

Maximum Likelihood: Improving Numerical Properties

- An example of this often arises when, in index models, elements of x involve squares, cubes, etc., of some covariate, say x_1. Then maximisation of the likelihood function may be easier if instead of x_1^2, x_1^3, etc., you use $x_1^2/10$, $x_1^3/100$, etc., with consequent rescaling of the coefficients on these covariates. You can always recover the MLEs you would have obtained without the rescaling by rescaling the estimates.
- There are some cases in which a re-parameterisation can produce a globally concave likelihood function where in the original parameterisation there was not global concavity.
- An example of this arises in the "Tobit" model.
  - This is a model in which each Y_i is $N(x_i'\beta, \sigma^2)$ with negative realisations replaced by zeros. The model is sometimes used to model expenditures and hours worked, which are necessarily non-negative.
  - In this model the likelihood as parameterised here is not globally concave, but re-parameterising to $\lambda = \beta/\sigma$, and $\gamma = 1/\sigma$, produces a globally concave likelihood function.
  - The invariance property tells us that having maximised the "easy" likelihood function and obtained estimates $\hat{\lambda}$ and $\hat{\gamma}$, we can recover the maximum likelihood estimates we might have had difficulty finding in the original parameterisation by calculating $\hat{\beta} = \hat{\lambda}/\hat{\gamma}$ and $\hat{\sigma} = 1/\hat{\gamma}$.

Properties Of Maximum Likelihood Estimators

- First we just sketch the main results:
  - Let $l(\theta; Y)$ be the log likelihood function now regarded as a random variable, a function of a set of (possibly vector) random variables $Y = \{Y_1, \ldots, Y_n\}$.
  - Let $l_\theta(\theta; Y)$ be the gradient of this function, itself a vector of random variables (scalar if θ is scalar) and let $l_{\theta\theta}(\theta; Y)$ be the matrix of second derivatives of this function (also a scalar if θ is a scalar).
  - Let
$$\hat{\theta} = \arg\max_{\theta} \quad l(\theta; Y).$$
In order to make inferences about θ using $\hat{\theta}$ we need to determine the distribution of $\hat{\theta}$. We consider developing a large sample approximation. The limiting distribution for a quite wide class of maximum likelihood problems is as follows:
$$n^{1/2}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_0)$$
where
$$V_0 = -\operatorname{plim}_{n\to\infty}(n^{-1} l_{\theta\theta}(\theta_0; Y))^{-1}$$
and θ_0 is the unknown parameter value. To get an approximate distribution that can be used in practice we use $(n^{-1} l_{\theta\theta}(\hat{\theta}; Y))^{-1}$ or some other consistent estimator of V_0 in place of V_0.

Properties Of Maximum Likelihood Estimators

- We apply the method for dealing with M-estimators.
- Suppose $\hat{\theta}$ is uniquely determined as the solution to the first order condition

$$l_\theta(\hat{\theta}; Y) = 0$$

and that $\hat{\theta}$ is a consistent estimator of the unknown value of the parameter, θ_0. Weak conditions required for consistency are quite complicated and will not be given here.

- Taking a Taylor series expansion around $\theta = \theta_0$ and then evaluating this at $\theta = \hat{\theta}$ gives

$$0 \simeq l_\theta(\theta_0; Y) + l_{\theta\theta'}(\theta_0; Y)(\hat{\theta} - \theta_0)$$

and rearranging and scaling by powers of the sample size n

$$n^{1/2}(\hat{\theta} - \theta_0) \simeq -\left(n^{-1} l_{\theta\theta'}(\theta; Y)\right)^{-1} n^{-1/2} l_\theta(\theta; Y).$$

As in our general treatment of M-estimators if we can show that

$$n^{-1} l_{\theta\theta'}(\theta_0; Y) \xrightarrow{p} A(\theta_0)$$

and

$$n^{-1/2} l_\theta(\theta_0; Y) \xrightarrow{d} N(0, B(\theta_0))$$

then

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, A(\theta_0)^{-1} B(\theta_0) A(\theta_0)^{-1\prime}).$$

Maximum Likelihood: Limiting Distribution

- What is the limiting distribution of $n^{-1/2} l_\theta(\theta_0; Y)$?
- First note that in problems for which the Y_i's are independently distributed, $n^{-1/2} l_\theta(\theta_0; Y)$ is a scaled mean of random variables and we may be able to find conditions under which a central limit theorem applies, indicating a limiting *normal* distribution.
- We must now find the mean and variance of this distribution.

  Since $L(\theta; Y)$ is a joint probability density function (we just consider the continuous distribution case here),

  $$\int L(\theta; y) dy = 1$$

  where multiple integration is over the support of Y. If this support *does not depend upon* θ, then

  $$\frac{\partial}{\partial \theta} \int L(\theta; y) dy = \int L_\theta(\theta; y) dy = 0.$$

  But, because $l(\theta; y) = \log L(\theta; y)$, and $l_\theta(\theta; y) = L_\theta(\theta; y)/L(\theta; y)$, we have

  $$\int L_\theta(\theta; y) dy = \int l_\theta(\theta; y) L(\theta; y) dy = E\left[l_\theta(\theta; Y)\right]$$

  and so $E\left[l_\theta(\theta; Y)\right] = 0$.
- This holds for any value of θ, in particular for θ_0 above. If the variance of $l_\theta(\theta_0; Y)$ converges to zero as n becomes large then $l_\theta(\theta_0; Y)$ will converge in probability to zero and the mean of the limiting distribution of $n^{-1/2} l_\theta(\theta_0; Y)$ will be zero.

Maximum Likelihood: Limiting Distribution

- We turn now to the variance of the limiting distribution. We have just shown that

$$\int l_\theta(\theta; y)L(\theta; y)dy = 0.$$

Differentiating again

$$\begin{aligned}
\frac{\partial}{\partial \theta'} \int l_\theta(\theta; y)L(\theta; y)dy &= \int \left(l_{\theta\theta'}(\theta; y)L(\theta; y) + l_\theta(\theta; y)L_{\theta'}(\theta; y)\right) dy \\
&= \int \left(l_{\theta\theta'}(\theta; y) + l_\theta(\theta; y)l_\theta(\theta; y)'\right) L(\theta; y)dy \\
&= E\left[l_{\theta\theta'}(\theta; Y) + l_\theta(\theta; Y)l_\theta(\theta; Y)'\right] \\
&= 0.
\end{aligned}$$

Separating the two terms in the penultimate line,

$$E\left[l_\theta(\theta; Y)l_\theta(\theta; Y)'\right] = -E\left[l_{\theta\theta'}(\theta; Y)\right] \tag{4}$$

and note that, since $E\left[l_\theta(\theta; Y)\right] = 0$,

$$Var[l_\theta(\theta; Y)] = E\left[l_\theta(\theta; Y)l_\theta(\theta; Y)'\right]$$

and so

$$\begin{aligned}
Var[l_\theta(\theta; Y)] &= -E\left[l_{\theta\theta'}(\theta; Y)\right] \\
\Rightarrow Var[n^{-1/2}l_\theta(\theta; Y)] &= -E\left[n^{-1}l_{\theta\theta'}(\theta; Y)\right]
\end{aligned}$$

giving

$$B(\theta_0) = -\operatorname*{plim}_{n\to\infty} n^{-1}l_{\theta\theta'}(\theta_0; Y).$$

The matrix

$$I(\theta) = -E\left[l_{\theta\theta}(\theta; Y)\right]$$

plays a central role in likelihood theory - it is called the *Information Matrix*.

Finally, because $B(\theta_0) = -A(\theta_0)$

$$A(\theta)^{-1}B(\theta)A(\theta)^{-1\prime} = -\left(\operatorname*{plim}_{n\to\infty} n^{-1}l_{\theta\theta'}(\theta; Y)\right)^{-1}.$$

Of course a number of conditions are required to hold for the results above to hold. These include the boundedness of third order derivatives of the log likelihood function, independence or at most weak dependence of the Y_i's, existence of moments of derivatives of the log likelihood, or at least of probability limits of suitably scaled versions of them, and lack of dependence of the support of the Y_i's on θ.

The result in equation (4) above leads, under suitable conditions concerning convergence, to

$$\operatorname*{plim}_{n\to\infty}\left(n^{-1}l_\theta(\theta;Y)l_\theta(\theta;Y)'\right) = -\operatorname*{plim}_{n\to\infty}\left(n^{-1}l_{\theta\theta'}(\theta;Y)\right).$$

This gives an alternative way of "estimating" V_0, namely

$$\hat{V}_0^o = \left\{n^{-1}l_\theta(\hat{\theta};Y)l_\theta(\hat{\theta};Y)'\right\}^{-1}$$

which compared with

$$\tilde{V}_0^o = \left\{-n^{-1}l_{\theta\theta'}(\hat{\theta};Y)\right\}^{-1}$$

has the advantage that only first derivatives of the log likelihood function need to be calculated. Sometimes $\hat{V}_0^o$ is referred to as the "outer product of gradient" (OPG) estimator. Both these estimators use the "observed" values of functions of derivatives of the LLF and. It may be possible to derive explicit expressions for the expected values of these functions. Then one can estimate V_0 by

$$\begin{aligned}\hat{V}_0^e &= \left\{E[n^{-1}l_\theta(\theta;Y)l_\theta(\theta;Y)']|_{\theta=\hat{\theta}}\right\}^{-1}\\ &= \left\{-E[n^{-1}l_{\theta\theta'}(\theta;Y)]|_{\theta=\hat{\theta}}\right\}^{-1}.\end{aligned}$$

These two sorts of estimators are sometimes referred to as "observed information" ($\hat{V}_0^o$, $\tilde{V}_0^o$) and "expected information" ($\hat{V}_0^e$) estimators.

Maximum likelihood estimators possess optimality property, namely that, among the class of consistent and asymptotically normally distributed estimators, the variance matrix of their limiting distribution is the smallest that can be achieved in the sense that other estimators in the class have limiting distributions with variance matrices exceeding the MLE's by a positive semidefinite matrix.

Estimating a Conditional Probability

- Suppose $Y_1, \ldots Y_n$ are binary independently and identically distributed random variables with

$$\begin{aligned} P[Y_i &= 1|X = x_i] = p(x_i, \theta) \\ P[Y_i &= 0|X = x_i] = 1 - p(x, \theta). \end{aligned}$$

  This is an obvious extension of the model in the previous section.

- The likelihood function for this problem is

$$\begin{aligned} P[Y_1 = y_1 \cap \cdots \cap Y_1 = y_1|x] &= \prod_{i=1}^{n} p(x_i, \theta)^{y_i}(1 - p(x_i, \theta))^{(1-y_i)} \\ &= L(\theta; y). \end{aligned}$$

  where y denotes the complete set of values of y_i and dependence on x is suppressed in the notation. The log likelihood function is

$$l(\theta; y) = \sum_{i=1}^{n} y_i \log p(x_i, \theta) + \sum_{i=1}^{n} (1 - y_i) \log(1 - p(x_i, \theta))$$

  and the maximum likelihood estimator of θ is

$$\hat{\theta} = \arg\max_{\theta} \quad l(\theta; y).$$

  So far this is an obvious generalisation of the simple problem met in the last section.

Estimating a Conditional Probability

- To implement the model we choose a form for the function $p(x, \theta)$, which must of course lie between zero and one.
  - One common choice is

$$p(x, \theta) = \frac{\exp(x'\theta)}{1 + \exp(x'\theta)}$$

    which produces what is commonly called a *logit model*.
  - Another common choice is

$$\begin{aligned} p(x, \theta) &= \Phi(x'\theta) = \int_{-\infty}^{x'\theta} \phi(w)dw \\ \phi(w) &= (2\pi)^{-1/2} \exp(-w^2/2) \end{aligned}$$

    in which Φ is the standard normal distribution function. This produces what is known as a *probit model*.

- Both models are widely used. Note that in both cases a single index model is specified, the probability functions are monotonic increasing, probabilities arbitrarily close to zero or one are obtained when $x'\theta$ is sufficiently large or small, and there is a symmetry in both of the models in the sense that $p(-x, \theta) = 1 - p(x, \theta)$. Any or all of these properties might be inappropriate in a particular application but there is rarely discussion of this in the applied econometrics literature.

More on Logit and Probit

- Both models can also be written as a linear model involving a latent variable.
- We define a **latent variable** Y_i^*, which is unobserved, but determined by the following model:

$$Y_i^* = X_i'\theta + \varepsilon_i$$

We observe the variable Y_i which is linked to Y_i^* as:

$$\begin{cases} Y_i = 0 & \text{if } Y_i^* < 0 \\ Y_i = 1 & \text{if } Y_i^* \geq 0 \end{cases}$$

- The probability of observing $Y_i = 1$ is:

$$\begin{aligned} p_i = P(Y_i = 1) &= P(Y_i^* \geq 0) \\ &= P(X_i'\theta + \varepsilon_i \geq 0) \\ &= P(\varepsilon_i \geq -X_i'\theta) \\ &= 1 - F_\varepsilon(-X_i'\theta) \end{aligned}$$

where F_ε is the cumulative distribution function of the random variable ε.

Odds-Ratio

- Define the ratio $p_i/(1-p_i)$ as the **odds-ratio**. This is the ratio of the probability of outcome 1 over the probability of outcome 0. If this ratio is equal to 1, then both outcomes have equal probability ($p_i = 0.5$). If this ratio is equal to 2, say, then outcome 1 is twice as likely than outcome 0 ($p_i = 2/3$).
- In the logit model, the log odds-ratio is linear in the parameters:

$$\ln \frac{p_i}{1-p_i} = X_i'\theta$$

- In the logit model, θ is the marginal effect of X on the log odds-ratio. A unit increase in X leads to an increase of θ % in the odds-ratio.

Marginal Effects

- Logit model:

$$\begin{aligned}\frac{\partial p_i}{\partial X} &= \frac{\theta \exp(X_i'\theta)(1+\exp(X_i'\theta)) - \theta \exp(X_i'\theta)^2}{(1+\exp(X_i'\theta))^2} \\ &= \frac{\theta \exp(X_i'\theta)}{(1+\exp(X_i'\theta))^2} \\ &= \theta p_i(1-p_i)\end{aligned}$$

  A one unit increase in X leads to an increase of $\theta p_i(1-p_i)$.

- Probit model:

$$\frac{\partial p_i}{\partial X_i} = \theta \phi(X_i'\theta)$$

  A one unit increase in X leads to an increase of $\theta \phi(X_i'\theta)$.

ML in Single Index Models

- We can cover both cases by considering general single index models, so for the moment rewrite $p(x, \theta)$ as $g(w)$ where $w = x'\theta$.

- The first derivative of the log likelihood function is:

$$\begin{aligned} l_\theta(\theta; y) &= \sum_{i=1}^{n} \frac{g_w(x_i'\theta)x_i}{g(x_i'\theta)} y_i - \frac{g_w(x_i'\theta)x_i}{1 - g(x_i'\theta)}(1 - y_i) \\ &= \sum_{i=1}^{n} \left(y_i - g(x_i'\theta)\right) \frac{g_w(x_i'\theta)}{g(x_i'\theta)\left(1 - g(x_i'\theta)\right)} x_i \end{aligned}$$

  Here $g_w(w)$ is the derivative of $g(w)$ with respect to w.

- The expression for the second derivative is rather messy. Here we just note that its expected value given x is quite simple, namely

$$E[l_{\theta\theta}(\theta; y)|x] = -\sum_{i=1}^{n} \frac{g_w(x_i'\theta)^2}{g(x_i'\theta)\left(1 - g(x_i'\theta)\right)} x_i x_i',$$

  the negative of which is the Information Matrix for general single index binary data models.

Asymptotic Properties of the Logit Model

- For the logit model there is major simplification

$$\begin{aligned} g(w) &= \frac{\exp(w)}{1+\exp(w)} \\ g_w(w) &= \frac{\exp(w)}{(1+\exp(w))^2} \\ \Rightarrow \frac{g_w(w)}{g(w)\left(1-g(w)\right)} &= 1. \end{aligned}$$

Therefore in the logit model the MLE satisfies

$$\sum_{i=1}^{n}\left(y_i - \frac{\exp(x_i'\hat{\theta})}{1+\exp(x_i'\hat{\theta})}\right)x_i = 0,$$

the Information Matrix is

$$I(\theta) = \sum_{i=1}^{n}\frac{\exp(x_i'\theta)}{(1+\exp(x_i'\theta))^2}x_i x_i',$$

the MLE has the limiting distribution

$$\begin{aligned} n^{1/2}(\hat{\theta}_n - \theta) &\xrightarrow{d} N(0, V_0) \\ V_0 &= \left(\operatorname*{plim}_{n\to\infty} n^{-1}\sum_{i=1}^{n}\frac{\exp(x_i'\theta)}{(1+\exp(x_i'\theta))^2}x_i x_i'\right)^{-1}, \end{aligned}$$

and we can conduct approximate inference using the following approximation

$$n^{1/2}(\hat{\theta}_n - \theta) \simeq N(0, V_0)$$

using the estimator

$$\hat{V}_0 = \left(n^{-1}\sum_{i=1}^{n}\frac{\exp(x_i'\hat{\theta})}{\left(1+\exp(x_i'\hat{\theta})\right)^2}x_i x_i'\right)^{-1}$$

when producing approximate hypothesis tests and confidence intervals.

Asymptotic Properties of the Probit Model

- In the probit model

$$\begin{aligned} g(w) &= \Phi(w) \\ g_w(w) &= \phi(w) \\ \Rightarrow \frac{g_w(w)}{g(w)\left(1-g(w)\right)} &= \frac{\phi(w)}{\Phi(w)(1-\Phi(w))}. \end{aligned}$$

Therefore in the probit model the MLE satisfies

$$\sum_{i=1}^{n}\left(y_i - \Phi(x_i'\hat{\theta})\right)\frac{\phi(x_i'\hat{\theta})}{\Phi(x_i'\hat{\theta})(1-\Phi(x_i'\hat{\theta}))}x_i = 0,$$

the Information Matrix is

$$I(\theta) = \sum_{i=1}^{n}\frac{\phi(x_i'\theta)^2}{\Phi(x_i'\theta)(1-\Phi(x_i'\theta))}x_i x_i',$$

the MLE has the limiting distribution

$$\begin{aligned} n^{1/2}(\hat{\theta}_n - \theta) &\xrightarrow{d} N(0, V_0) \\ V_0 &= \left(\operatorname*{plim}_{n\to\infty} n^{-1}\sum_{i=1}^{n}\frac{\phi(x_i'\theta)^2}{\Phi(x_i'\theta)(1-\Phi(x_i'\theta))}x_i x_i'\right)^{-1}, \end{aligned}$$

and we can conduct approximate inference using the following approximation

$$n^{1/2}(\hat{\theta}_n - \theta) \simeq N(0, V_0)$$

using the estimator

$$\hat{V}_0 = \left(n^{-1}\sum_{i=1}^{n}\frac{\phi(x_i'\hat{\theta})^2}{\Phi(x_i'\hat{\theta})(1-\Phi(x_i'\hat{\theta}))}x_i x_i'\right)^{-1}$$

when producing approximate tests and confidence intervals.

Example: Logit and Probit

- We have data from households in Kuala Lumpur (Malaysia) describing household characteristics and their concern about the environment. The question is

  "Are you concerned about the environment? Yes / No".

  We also observe their age, sex (coded as 1 men, 0 women), income and quality of the neighborhood measured as air quality. The latter is coded with a dummy variable *smell*, equal to 1 if there is a bad smell in the neighborhood. The model is:

  $$Concern_i = \beta_0 + \beta_1 age_i + \beta_2 sex_i + \beta_3 \log income_i + \beta_4 smell_i + u_i$$

- We estimate this model with three specifications, LPM, logit and probit:

Probability of being concerned by Environment

Variable	LPM		Logit		Probit	
	Est.	t-stat	Est.	t-stat	Est.	t-stat
age	.0074536	3.9	.0321385	3.77	.0198273	3.84
sex	.0149649	0.3	.06458	0.31	.0395197	0.31
log income	.1120876	3.7	.480128	3.63	.2994516	3.69
smell	.1302265	2.5	.5564473	2.48	.3492112	2.52
constant	-.683376	-2.6	-5.072543	-4.37	-3.157095	-4.46
			Some Marginal Effects			
Age	.0074536		.0077372		.0082191	
log income	.1120876		.110528		.1185926	
smell	.1302265		.1338664		.1429596	

Multinomial Logit

- The logit model was dealing with two qualitative outcomes. This can be generalized to multiple outcomes:
  - choice of transportation: car, bus, train...
  - choice of dwelling: house, apartment, social housing.
- The multinomial logit: Denote the outcomes as $j = 1, \ldots, J$ and p_j the probability of outcome j.

$$p_j = \frac{\exp(X'\theta^j)}{\sum_{k=1}^{J} \exp(X'\theta^k)}$$

where θ^j is a vector of parameter associated with outcome j.

Identification

- If we multiply all the coefficients by a factor λ this does not change the probabilities p_j, as the factor cancel out. This means that there is under identification. We have to normalize the coefficients of one outcome, say, J to zero. All the results are interpreted as deviations from the baseline choice.
- We write the probability of choosing outcome $j = 1, \ldots, J-1$ as:

$$p_j = \frac{\exp(X'\theta^j}{1 + \sum_{k=1}^{J-1} \exp(X'\theta^k)}$$

- We can express the logs odds-ratio as:

$$\ln \frac{p_j}{p_J} = X'\theta^j$$

- The odds-ratio of choice j versus J is only expressed as a function of the parameters of choice j, but not of those other choices: Independence of Irrelevant Alternatives (IIA).

Independence of Irrelevant Alternatives

An anecdote which illustrates a violation of this property has been attributed to Sidney Morgenbesser:

After finishing dinner, Sidney Morgenbesser decides to order dessert. The waitress tells him he has two choices: apple pie and blueberry pie. Sidney orders the apple pie.

After a few minutes the waitress returns and says that they also have cherry pie at which point Morgenbesser says "In that case I'll have the blueberry pie."

Independence of Irrelevant Alternatives

- Consider travelling choices, by car or with a red bus. Assume for simplicity that the choice probabilities are equal:

$$P(car) = P(\text{red bus}) = 0.5 \quad \Longrightarrow \quad \frac{P(car)}{P(\text{red bus})} = 1$$

- Suppose we introduce a blue bus, (almost) identical to the red bus. The probability that individuals will choose the blue bus is therefore the same as for the red bus and the odd ratio is:

$$P(\text{blue bus}) = P(\text{red bus}) \quad \Longrightarrow \quad \frac{P(\text{blue bus})}{P(\text{red bus})} = 1$$

- However, the IIA implies that odds ratios are the same whether of not another alternative exists. The only probabilities for which the three odds ratios are equal to one are:

$$P(car) = P(\text{blue bus}) = P(\text{red bus}) = 1/3$$

  However, the prediction we ought to obtain is:

$$P(\text{red bus}) = P(\text{blue bus}) = 1/4 \quad P(car) = 0.5$$

Marginal Effects: Multinomial Logit

- θ^j can be interpreted as the marginal effect of X on the log odds-ratio of choice j to the baseline choice.
- The marginal effect of X on the probability of choosing outcome j can be expressed as:

$$\frac{\partial p_j}{\partial X} = p_j[\theta^j - \sum_{k=1}^{J} p_k \theta^k]$$

  Hence, the marginal effect on choice j involves not only the coefficients relative to j but also the coefficients relative to the other choices.

- Note that we can have $\theta^j < 0$ and $\partial p_j / \partial X > 0$ or vice versa.

  Due to the non linearity of the model, the sign of the coefficients does not indicate the direction nor the magnitude of the effect of a variable on the probability of choosing a given outcome. One has to compute the marginal effects.

Example

- We analyze here the choice of dwelling: house, apartment or low cost flat, the latter being the baseline choice. We include as explanatory variables the age, sex and log income of the head of household:

Variable	Estimate	Std. Err.	Marginal Effect
	Choice of House		
age	.0118092	.0103547	-0.002
sex	-.3057774	.2493981	-0.007
log income	1.382504	.1794587	0.18
constant	-10.17516	1.498192	
	Choice of Apartment		
age	.0682479	.0151806	0.005
sex	-.89881	.399947	-0.05
log income	1.618621	.2857743	0.05
constant	-15.90391	2.483205	

Ordered Models

- In the multinomial logit, the choices were not ordered. For instance, we cannot rank cars, busses or train in a meaningful way. In some instances, we have a natural ordering of the outcomes even if we cannot express them as a continuous variable:
  - Yes / Somehow / No.
  - Low / Medium / High
- We can analyze these answers with ordered models.

Ordered Probit

- We code the answers by arbitrary assigning values:

$$Y_i = 0 \text{ if No}, \quad Y_i = 1 \text{ if Somehow}, \quad Y_i = 2 \text{ if Yes}$$

- We define a latent variable Y_i^* which is linked to the explanatory variables:

$$Y_i^* = X_i'\theta + \varepsilon_i$$

$$\begin{array}{ll} Y_i = 0 & \text{if } Y_i^* < 0 \\ Y_i = 1 & \text{if } Y_i^* \in [0, \mu[\\ Y_i = 2 & \text{if } Y_i^* \geq \mu \end{array}$$

  μ is a threshold and an auxiliary parameter which is estimated along with θ.

- We assume that ε_i is distributed normally.

- The probability of each outcome is derived from the normal cdf:

$$\begin{aligned} P(Y_i = 0) &= \Phi(-X_i'\theta) \\ P(Y_i = 1) &= \Phi(\mu - X_i'\theta) - \Phi(-X_i'\theta) \\ P(Y_i = 2) &= 1 - \Phi(\mu - X_i'\theta) \end{aligned}$$

Ordered Probit

- Marginal Effects:

$$\frac{\partial P(Y_i = 0)}{\partial X_i} = -\theta\phi(-X_i'\theta)$$
$$\frac{\partial P(Y_i = 1)}{\partial X_i} = \theta\left(\phi(X_i'\theta) - \phi(\mu - X_i'\theta)\right)$$
$$\frac{\partial P(Y_i = 2)}{\partial X_i} = \theta\phi(\mu - X_i'\theta)$$

- Note that if $\theta > 0$, $\partial P(Y_i = 0)/\partial X_i < 0$ and $\partial P(Y_i = 2)/\partial X_i > 0$:
  - If X_i has a positive effect on the latent variable, then by increasing X_i, fewer individuals will stay in category 0.
  - Similarly, more individuals will be in category 2.
  - In the intermediate category, the fraction of individual will either increase or decrease, depending on the relative size of the inflow from category 0 and the outflow to category 2.

Ordered Probit: Example

- We want to investigate the determinants of health.
- Individuals are asked to report their health status in three categories: poor, fair or good.
- We estimate an ordered probit and calculate the marginal effects at the mean of the sample.

Variable	Coeff	sd. err.	Marginal Effects			Sample
			Poor	Fair	Good	Mean
Age 18-30	-1.09**	.031	-.051**	-.196**	.248**	.25
Age 30-50	-.523**	.031	-.031**	-.109**	.141**	.32
Age 50-70	-.217**	.026	-.013**	-.046**	.060**	.24
Male	-.130**	.018	-.008**	-.028**	.037**	.48
Income low third	.428**	.027	.038**	.098**	-.136**	.33
Income medium third	.264**	.022	.020**	.059**	-.080**	.33
Education low	.40**	.028	.031**	.091**	-.122**	.43
Education Medium	.257**	.026	.018**	.057**	-.076**	.37
Year of interview	-.028	.018	-.001	-.006	.008	1.9
Household size	-.098**	.008	-.006**	-.021**	.028**	2.5
Alcohol consumed	.043**	.041	.002**	.009**	-.012**	.04
Current smoker	.160**	.018	.011**	.035**	-.046**	.49
cut1	.3992**	.058				
cut2	1.477**	.059				

Age group	Proportion		
	Poor Health	Fair Health	Good Health
Age 18-30	.01	.08	.90
Age 30-50	.03	.13	.83
Age 50-70	.07	.28	.64
Age 70 +	.15	.37	.46

Ordered Probit: Example

- Marginal Effects differ by individual characteristics.
- Below, we compare the marginal effects from an ordered probit and a multinomial logit.

Variable	Marginal Effects for Good Health			
	Ordered Probit at mean	X	Ordered Probit at X	Multinomial Logit at X
Age 18-30	.248**	1	.375**	.403**
Age 30-50	.141**	0	.093**	.077**
Age 50-70	.060**	0	.046**	.035**
Male	.037**	1	.033**	.031**
Income low third	-.136**	1	-.080**	-.066**
Income medium third	-.080**	0	-.071**	-.067**
Education low	-.122**	1	-.077**	-.067**
Education Medium	-.076**	0	-.069**	-.064**
Year of interview	.008	1	.006	.003
Household size	.028**	2	.023**	.020**
Alcohol consumed	-.012**	0	-.010**	-.011**
Current smoker	-.046**	0	-.041**	-.038**

Models for Count Data

- The methods developed above are useful when we want to model the occurrence or otherwise of an event. Sometimes we want to model the number of times an event occurs. In general it might be any nonnegative integer. Count data are being used increasingly in econometrics.
- An interesting application is to the modelling of the returns to R&D investment in which data on numbers of patents filed in a series of years by a sample of companies is studied and related to data on R&D investments.
- Binomial and Poisson probability models provide common starting points in the development of count data models.
- If $Z_1, \ldots, Z_m$ are identically and independently distributed binary random variables with $P[Z_i = 1] = p$, $P[Z_i = 0] = 1 - p$, then the sum of the Z_i's has a Binomial distribution,

$$Y = \sum_{i=1}^{m} Z_i \sim Bi(m, p)$$

and

$$P[Y = j] = \frac{m!}{j!(m-j)!} p^j (1-p)^{m-j}, \qquad j \in \{0, 1, 2, \ldots, m\}$$

Models for Count Data

- As m becomes large, $m^{1/2}(m^{-1}Y - p)$ becomes approximately normally distributed, $N(0, p(1-p))$, and as m becomes large while $mp = \lambda$ remains constant, Y comes to have a Poisson distribution,

  $$Y \sim Po(\lambda)$$

  and

  $$P[Y = j] = \frac{\lambda^j}{j!}\exp(-\lambda), \qquad j \in \{0, 1, 2, \dots\}.$$

- In each case letting p or λ be functions of covariates creates a model for the conditional distribution of a count of events given covariate values.
- The Poisson model is much more widely used, in part because there is no need to specify or estimate the parameter m.
- In the application to R&D investment one might imagine that a firm seeds a large number of research projects in a period of time, each of which has only a small probability of producing a patent. This is consonant with the Poisson probability model but note that one might be concerned about the underlying assumption of independence across projects built into the Poisson model.

Models for Count Data

- The estimation of the model proceeds by maximum likelihood. The Poisson model is used as an example. Suppose that we specify a single index model:

$$P[Y_i = y_i|x_i] = \frac{\lambda(x_i'\theta)^{y_i}}{y_i!}\exp(-\lambda(x_i'\theta)), \qquad j \in \{0, 1, 2, \dots\}.$$

- The log likelihood function is

$$l(\theta, y) = \sum_{i=1}^{n} y_i \log \lambda(x_i'\theta) - \lambda(x_i'\theta) - \log y_i!$$

  with first derivative

$$\begin{aligned} l_\theta(\theta, y) &= \sum_{i=1}^{n} \left(y_i \frac{\lambda_w(x_i'\theta)}{\lambda(x_i'\theta)} - \lambda_w(x_i'\theta) \right) x_i \\ &= \sum_{i=1}^{n} \left(y_i - \lambda(x_i'\theta) \right) \frac{\lambda_w(x_i'\theta)}{\lambda(x_i'\theta)} x_i \end{aligned}$$

  where $\lambda_w(w)$ is the derivative of $\lambda(w)$ with respect to w.

- The MLE satisfies

$$\sum_{i=1}^{n} \left(y_i - \lambda(x_i'\hat{\theta}) \right) \frac{\lambda_w(x_i'\hat{\theta})}{\lambda(x_i'\hat{\theta})} x_i = 0.$$

Models for Count Data

- The second derivative matrix is

$$l_{\theta\theta}(\theta, y) = \sum_{i=1}^{n} (y_i - \lambda(x_i'\theta)) \left(\frac{\lambda_{ww}(x_i'\theta)}{\lambda(x_i'\theta)} - \left(\frac{\lambda_w(x_i'\theta)}{\lambda(x_i'\theta)} \right)^2 \right) x_i x_i' - \sum_{i=1}^{n} \frac{\lambda_w(x_i'\theta)^2}{\lambda(x_i'\theta)} x_i x_i'$$

where, note, the first term has expected value zero. Therefore the Information Matrix for this conditional Poisson model is

$$I(\theta) = \sum_{i=1}^{n} \frac{\lambda_w(x_i'\theta)^2}{\lambda(x_i'\theta)} x_i x_i'.$$

The limiting distribution of the MLE is (under suitable conditions)

$$\begin{aligned} n^{1/2}(\hat{\theta} - \theta_0) &\xrightarrow{d} N(0, V_0) \\ V_0 &= \left(\operatorname*{plim}_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\lambda_w(x_i'\theta)^2}{\lambda(x_i'\theta)} x_i x_i' \right)^{-1} \end{aligned}$$

and we can make approximate inference about θ_0 using

$$(\hat{\theta} - \theta_0) \simeq N\left(0, n^{-1} V_0\right)$$

with V_0 estimated by

$$\hat{V}_0 = \left(n^{-1} \sum_{i=1}^{n} \frac{\lambda_w(x_i'\hat{\theta})^2}{\lambda(x_i'\hat{\theta})} x_i x_i' \right)^{-1}.$$

- In applied work a common choice is $\lambda(w) = \exp(w)$ for which

$$\frac{\lambda_w(w)}{\lambda(w)} = 1 \qquad \frac{\lambda_w(w)^2}{\lambda(w)} = \exp(w).$$