DegreeLive

Bachelor of Commerce (B2001) | Level 3 • Part B Major | Monash University

Principles of econometrics

Course Code: ETC3400 • Module 4 Notes

Module 4: Maximum Likelihood Estimation, GMM & Limited Dependent Variable Models

4.1 The Likelihood Principle, Score Vector, Fisher Information & Cramér-Rao Bound

While OLS and 2SLS require only first- and second-moment assumptions, fully parametric econometric inference specifies the joint probability density (or mass) function of the data conditional on exogenous covariates. Let $\{(y_i, x_i)\}_{i=1}^{n}$ be an independent random sample where the conditional density of y_i given x_i is $f(y_i \mid x_i; \theta)$, indexed by a $k \times 1$ unknown parameter vector $\theta \in \Theta \subset \mathrm{R}^k$ with true value θ_0.

Under the **Likelihood Principle**, all statistical evidence about θ contained in the realised sample is captured by the **likelihood function** $L(\theta; y, X) = \prod_{i=1}^{n} f(y_i \mid x_i; \theta)$ viewed as a function of θ. Because products are numerically unstable and algebraically inconvenient, estimation maximises the **log-likelihood function**:

$$\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta) = \sum_{i=1}^{n} \ln f(y_i \mid x_i; \theta)$$

Where $\ell_i(\theta) = \ln f(y_i \mid x_i; \theta)$ is the log-likelihood contribution of observation i.

Score Vector and Regularity Conditions

Assuming the support of y_i does not depend on θ and $\ell_i(\theta)$ is twice continuously differentiable in an open neighbourhood of θ_0, define the $k \times 1$ **score vector** $s(\theta)$ and $k \times k$ **Hessian matrix** $H(\theta)$ as:

$$s(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^{n} s_i(\theta) = \sum_{i=1}^{n} \frac{\partial \ell_i(\theta)}{\partial \theta}, \qquad H(\theta) = \frac{\partial^2 \ell(\theta)}{\partial \theta \, \partial \theta'} = \sum_{i=1}^{n} \frac{\partial^2 \ell_i(\theta)}{\partial \theta \, \partial \theta'}$$

Where $s_i(\theta)$ is the individual score contribution. Differentiating the identity $\int f(y \mid x; \theta)\, dy = 1$ with respect to θ and interchanging differentiation and integration (valid under standard Cramér-Rao regularity conditions) yields two fundamental identities evaluated at the true parameter θ_0:

1. **Zero Expected Score:**

$$E\left[s_i(\theta_0) \mid x_i\right] = \int \frac{\partial \ln f(y \mid x_i; \theta_0)}{\partial \theta} f(y \mid x_i; \theta_0)\, dy = \frac{\partial}{\partial \theta} \int f(y \mid x_i; \theta)\, dy \Big|_{\theta=\theta_0} = 0$$

DegreeLive

2. **Information Matrix Equality:** Differentiating $\int s_i(\theta) f(y \mid x_i; \theta)\, dy = 0$ a second time with respect to θ' gives the equivalence between the negative expected Hessian and the outer product of the score:

$$\mathrm{I}_1(\theta_0) = -E\left[\frac{\partial^2 \ell_i(\theta_0)}{\partial \theta \partial \theta'}\right] = E\left[s_i(\theta_0) s_i(\theta_0)'\right]$$

Where $\mathrm{I}_1(\theta_0)$ is the single-observation **Fisher Information Matrix**, and the total sample Fisher Information is $\mathrm{I}_n(\theta_0) = n\,\mathrm{I}_1(\theta_0)$ under i.i.d. sampling.

Cramér-Rao Lower Bound (CRLB) and Minimum Variance Unbiased Estimation

Theorem (Cramér-Rao Lower Bound): Under regularity conditions, let $\widetilde{\theta}$ be any unbiased estimator of θ_0 ($E(\widetilde{\theta}) = \theta_0$ for all $\theta_0 \in \Theta$). Then the covariance matrix of $\widetilde{\theta}$ satisfies the matrix inequality:

$$\mathrm{Var}(\widetilde{\theta}) - \mathrm{I}_n(\theta_0)^{-1} \geq 0$$

Where $\mathrm{I}_n(\theta_0)^{-1}$ is the Cramér-Rao Lower Bound. For a scalar smooth transformation $\tau(\theta)$ with gradient $d = \partial \tau(\theta)/\partial \theta$, any unbiased estimator $\widetilde{\tau}$ satisfies $\mathrm{Var}(\widetilde{\tau}) \geq d' \mathrm{I}_n(\theta_0)^{-1} d$. If an unbiased estimator attains the CRLB identically, it is the **Minimum Variance Unbiased Estimator (MVUE)**. By the Lehmann-Scheffé theorem, even when no estimator attains the CRLB in finite samples, any unbiased estimator that is a function of a complete sufficient statistic is uniquely the MVUE.

4.2 Asymptotic Distribution Theory of MLE, Numerical Optimisation & Simulation

The **Maximum Likelihood Estimator (MLE)** $\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \ell(\theta)$ solves the k score equations $s(\hat{\theta}_{MLE}) = 0$ whenever an interior maximum exists. Even when $\hat{\theta}_{MLE}$ is biased in finite samples, it enjoys optimal large-sample properties under standard identification and compactness conditions.

Consistency, Asymptotic Normality, Efficiency, and Invariance

- **Consistency via Kullback-Leibler Divergence:** By the WLLN, $\frac{1}{n}\ell(\theta) \xrightarrow{p} E[\ln f(y_i \mid x_i; \theta)]$. By Jensen's inequality applied to the strictly concave logarithm, for any $\theta \neq \theta_0$ with distinct density, $E[\ln f(y_i \mid x_i; \theta)] - E[\ln f(y_i \mid x_i; \theta_0)] = -D_{KL}(f_{\theta_0} \| f_\theta) < 0$. Hence the population objective is uniquely maximised at θ_0, giving $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$.
- **Asymptotic Normality and Cramér-Rao Efficiency:** Expanding $s(\hat{\theta}_{MLE}) = 0$ by the Mean Value Theorem around θ_0 yields:

$$0 = \frac{1}{n}s(\theta_0) + \left(\frac{1}{n}H(\bar{\theta})\right) \quad n(\hat{\theta}_{MLE} - \theta_0)$$

Where $\bar{\theta}$ lies on the line segment between $\hat{\theta}_{MLE}$ and θ_0. Since $\frac{1}{n}s(\theta_0) \xrightarrow{d} N(0, \mathrm{I}_1(\theta_0))$ by the CLT and $-\frac{1}{n}H(\bar{\theta}) \xrightarrow{p} \mathrm{I}_1(\theta_0)$ by the WLLN and consistency, Slutsky's Theorem yields:

$$n(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0,\ \mathrm{I}_1(\theta_0)^{-1})$$

Because $\mathrm{I}_1(\theta_0)^{-1}$ equals the asymptotic Cramér-Rao Lower Bound, MLE is **asymptotically efficient** among all regular consistent estimators.

- **Invariance Property:** For any transformation $\gamma = g(\theta)$ (even if many-to-one), the MLE of γ is simply $\hat{\gamma}_{MLE} = g(\hat{\theta}_{MLE})$, with asymptotic variance $G\,\mathrm{I}_n(\hat{\theta})^{-1}G'$ via the Delta Method.
- **Quasi-MLE (QMLE) and Sandwich Covariance:** If the density is misspecified (e.g., applying a Gaussian or Poisson log-likelihood when true data are overdispersed) but the conditional mean remains correctly specified, the Information Matrix Equality fails ($-E[H_i] \neq E[s_i s_i']$). Valid asymptotic inference then uses the Huber-White sandwich covariance matrix $A_0^{-1}B_0A_0^{-1}$, where $A_0 = -E[H_i(\theta_0)]$ and $B_0 = E[s_i(\theta_0)s_i(\theta_0)']$.

Numerical Optimisation Algorithms for MLE

Outside exponential-family linear models, the score equations $s(\theta) = 0$ are nonlinear and lack closed-form solutions. Econometric software locates $\hat{\theta}_{MLE}$ iteratively using gradient-based hill-climbing updates $\theta^{(m+1)} = \theta^{(m)} + \alpha_m P_m^{-1} s(\theta^{(m)})$, where $\alpha_m > 0$ is a line-search step size and P_m is a positive definite curvature matrix:

Algorithm	Curvature Matrix P_m	Key Advantage & Convergence Property
Newton-Raphson (NR)	$-H(\theta^{(m)}) = -\frac{\partial^2 \ell}{\partial\theta\partial\theta'}$	Quadratic convergence near optimum; requires analytic or numeric Hessian
Fisher Scoring	$\mathrm{I}_n(\theta^{(m)}) = -E[H(\theta^{(m)})]$	Guaranteed positive definite throughout iterations; identical to NR for canonical GLMs
BHHH (Berndt-Hall-Hall-Hausman)	$\sum_{i=1}^{n} s_i(\theta^{(m)})s_i(\theta^{(m)})'$	Uses only first derivatives (outer product of gradient); always positive semi-definite
BFGS (Quasi-Newton)	Secant approximation to $-H(\theta^{(m)})$	Superlinear convergence without computing second derivatives; industry default

Degreelive

Monte Carlo Simulation for Evaluating Likelihood Estimators

Because asymptotic normality only describes behaviour as $n \to \infty$, Monte Carlo simulation provides essential finite-sample validation. Across R independent replications ($r = 1, \ldots, R$, typically $R \geq 2{,}000$), artificial samples of size n are drawn from the exact data-generating process at fixed θ_0, numerical optimisation computes $\hat{\theta}^{(r)}$ and $\widehat{\text{se}}(\hat{\theta}^{(r)})$, and the analyst evaluates finite-sample bias $\bar{\hat{\theta}} - \theta_0$, root mean squared error (RMSE), and empirical 95% confidence interval coverage. Simulation also verifies whether the Likelihood Ratio, Wald, and Score tests maintain nominal size $\alpha = 0.05$ in small samples.

Worked Example 1: Newton-Raphson MLE and CRLB for Exponential Trade Durations

Suppose $n = 10$ inter-trade waiting times $y_i > 0$ (in hours) on an ASX equity follow an Exponential distribution $f(y_i; \theta) = \theta e^{-\theta y_i}$ ($\theta > 0$), with observed sample total $\sum_{i=1}^{10} y_i = 25.0$ ($\bar{y} = 2.5$). The log-likelihood, score, Hessian, and Fisher information are:

$$\ell(\theta) = 10 \ln \theta - 25.0\theta, \qquad s(\theta) = \frac{10}{\theta} - 25.0, \qquad H(\theta) = -\frac{10}{\theta^2}, \qquad \mathrm{I}_{10}(\theta) = \frac{10}{\theta^2}$$

Setting $s(\hat{\theta}) = 0$ yields the exact MLE $\hat{\theta}_{MLE} = 10/25.0 = 0.40$ with asymptotic standard error $\text{se}(\hat{\theta}_{MLE}) = \sqrt{\hat{\theta}^2/10} = 0.40/\sqrt{10} \approx 0.1265$. For the mean duration $\mu = \tau(\theta) = 1/\theta$, $\tau'(\theta) = -1/\theta^2$, so the CRLB is $(-1/\theta^2)^2/(10/\theta^2) = 1/(10\theta^2) = \mu^2/10$; since $\text{Var}(\bar{y}) = \mu^2/10$, the sample mean $\bar{y} = 2.5$ attains the CRLB and is the MVUE.

Illustrating Newton-Raphson from starting value $\theta^{(0)} = 0.25$ via $\theta^{(m+1)} = \theta^{(m)} - H(\theta^{(m)})^{-1} s(\theta^{(m)})$:

- **Iteration 1:** $s(0.25) = 40 - 25 = 15$, $H(0.25) = -160$, so $\theta^{(1)} = 0.25 - (15/-160) = 0.34375$.
- **Iteration 2:** $s(0.34375) = 29.0909 - 25 = 4.0909$, $H(0.34375) = -84.6354$, so $\theta^{(2)} = 0.34375 + 0.04834 = 0.39209$.
- **Iteration 3:** $s(0.39209) = 25.5043 - 25 = 0.5043$, $H(0.39209) = -65.0472$, so $\theta^{(3)} = 0.39209 + 0.00775 = 0.39984$, converging to 0.40000 on iteration 4.

4.3 Generalised Method of Moments (GMM)

When economic theory supplies population moment restrictions without specifying the complete probability density of the data—such as Euler equations in Australian household consumption models or overidentified instrumental variable regressions with heteroskedastic errors—**Generalised Method of Moments (GMM)** provides a unified semiparametric estimation framework.

Orthogonality Conditions and the Quadratic Objective Function

Suppose a $p \times 1$ parameter vector θ_0 satisfies an $\ell \times 1$ vector of population orthogonality (moment) conditions ($ell \geq p$):

Degreelive

$$E\left[g(w_i, \theta_0)\right] = 0$$

Where w_i denotes all observed data for unit i and $g(w_i, \theta)$ is a known ℓ-dimensional moment function (for linear IV, $g(w_i, \beta) = z_i(y_i - x_i'\beta)$). Define the $\ell \times 1$ sample moment average $\bar{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i, \theta)$.

When $\ell = p$ (just-identified), setting $\bar{g}_n(\hat{\theta}) = 0$ solves p equations in p unknowns uniquely. When $\ell > p$ (overidentified), no θ generally sets all ℓ sample moments simultaneously to zero. Instead, GMM chooses $\hat{\theta}_{GMM}$ to make $\bar{g}_n(\theta)$ as close to zero as possible by minimising a weighted quadratic norm:

$$Q_n(\theta) = \bar{g}_n(\theta)' W_n \bar{g}_n(\theta)$$

Where W_n is a symmetric positive definite $\ell \times \ell$ weighting matrix satisfying $W_n \xrightarrow{p} W$ (positive definite).

Asymptotic Distribution and Optimal Two-Step Efficient GMM

Let $G_0 = E\left[\frac{\partial g(w_i, \theta_0)}{\partial \theta'}\right]$ be the $\ell \times p$ expected Jacobian matrix (with full column rank p for local identification) and $S_0 = E[g(w_i, \theta_0) g(w_i, \theta_0)']$ be the $\ell \times \ell$ covariance matrix of the moment conditions. For any weighting matrix W, the GMM estimator is consistent and asymptotically normal:

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \xrightarrow{d} N\left(0,\ (G_0' W G_0)^{-1} G_0' W S_0 W G_0 (G_0' W G_0)^{-1}\right)$$

Hansen (1982) proved that the asymptotic covariance matrix is minimised in the matrix Loewner order by setting $W^* = S_0^{-1}$, weighting more precise (lower-variance) moment conditions more heavily and down-weighting noisy or redundant moments. With optimal weighting $W^* = S_0^{-1}$, the asymptotic distribution simplifies to:

$$\sqrt{n}(\hat{\theta}_{EGMM} - \theta_0) \xrightarrow{d} N\left(0,\ (G_0' S_0^{-1} G_0)^{-1}\right)$$

Where $\hat{\theta}_{EGMM}$ is the **Efficient GMM** estimator, implemented in two steps: (1) obtain an initial consistent estimate $\tilde{\theta}$ using a suboptimal W_n (such as I_ℓ or $(Z'Z/n)^{-1}$, which yields 2SLS in linear IV) and form $\hat{S} = \frac{1}{n}\sum_{i=1}^{n} g(w_i, \tilde{\theta}) g(w_i, \tilde{\theta})'$ (or Newey-West HAC $\hat{S}$ for time series); (2) re-estimate $\hat{\theta}_{EGMM}$ using $W_n = \hat{S}^{-1}$. In linear IV, this produces the closed-form efficient GMM estimator $\hat{\beta}_{EGMM} = (X'Z\hat{S}^{-1}Z'X)^{-1}X'Z\hat{S}^{-1}Z'y$, which outperforms 2SLS asymptotically whenever disturbances are heteroskedastic. Finally, evaluating the minimised efficient objective function provides **Hansen's J-test** of the $\ell - p$ overidentifying restrictions:

DegreeLive

$$J = n\,\bar{g}_n(\hat{\theta}_{EGMM})'\hat{S}^{-1}\bar{g}_n(\hat{\theta}_{EGMM}) \quad \overset{d}{} \quad \chi^2_{\ell-p}$$

4.4 Binary Choice Models: LPM, Logit & Probit

Many microeconometric outcomes are binary indicators $y_i \in \{0,1\}$, such as Australian labour force participation, ASX corporate dividend initiation, or household mortgage default. Modelling the response probability $P(y_i = 1 \mid x_i) = G(x_i'\beta)$ via OLS—the **Linear Probability Model (LPM)** where $G(z) = z$—suffers from three structural defects: predicted probabilities can fall outside $[0,1]$, marginal effects $\partial P/\partial x_{ij} = \beta_j$ are implausibly constant across all covariate values, and errors are inherently heteroskedastic with $\text{Var}(y_i \mid x_i) = x_i'\beta(1 - x_i'\beta)$.

Latent Variable Formulation, Logit, and Probit MLE

Binary choice models map the linear index $x_i'\beta$ into $(0,1)$ using a strictly increasing symmetric cumulative distribution function $G(\cdot)$. Equivalently, they arise from an unobserved continuous **latent utility** (or net propensity) equation:

$$y_i^* = x_i'\beta + \varepsilon_i, \qquad y_i = \mathbf{1}(y_i^* > 0) = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \le 0 \end{cases}$$

Where symmetry of the zero-mean distribution of ε_i gives $\Pr(y_i = 1 \mid x_i) = \Pr(\varepsilon_i > -x_i'\beta \mid x_i) = G(x_i'\beta)$. Two canonical choices of $G(\cdot)$ dominate applied work:

- **Probit Model:** Assumes $\varepsilon_i \mid x_i \sim N(0,1)$, so $G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(t)\,dt$, where $\phi(t) = (2\pi)^{-1/2}e^{-t^2/2}$ is the standard normal PDF. Normalising $\text{Var}(\varepsilon_i) = 1$ is required for scale identification because multiplying $(y_i^*, \beta, \varepsilon_i)$ by any positive constant leaves $y_i = \mathbf{1}(y_i^* > 0)$ unchanged.
- **Logit Model:** Assumes $\varepsilon_i \mid x_i$ follows a standard logistic distribution with variance $\pi^2/3$, yielding the closed-form logistic CDF and derivative:

  $$\Lambda(z) = \frac{\exp(z)}{1 + \exp(z)} = \frac{1}{1 + \exp(-z)}, \qquad \Lambda'(z) = \Lambda(z)\,[1 - \Lambda(z)]$$

  Because $\text{sd}(\varepsilon_{\text{logit}}) = \pi/\sqrt{3} \approx 1.814$ versus $\text{sd}(\varepsilon_{\text{probit}}) = 1$, Logit slope estimates are typically about 1.6 to 1.8 times larger than Probit slopes (Amemiya's approximate conversion: $\hat{\beta}_{\text{logit}} \approx 1.6\hat{\beta}_{\text{probit}}$), even though both models yield nearly identical predicted probabilities and marginal effects.

Since each observation $y_i \mid x_i$ is Bernoulli with success probability $G(x_i'\beta)$, the joint binary choice log-likelihood is:

DegreeLive

$$\ell(\beta) = \sum_{i=1}^{n} \{y_i \ln G(x_i'\beta) + (1 - y_i) \ln [1 - G(x_i'\beta)]\}$$

Both Probit and Logit log-likelihoods are globally strictly concave in β (provided X has full column rank and outcomes are not perfectly separated by a hyperplane), guaranteeing a unique global MLE easily found via Newton-Raphson. Model fit is commonly summarised by **McFadden's Pseudo-R^2**, defined as $R^2_{McF} = 1 - \ell(\hat{\beta})/\ell_0$, where ℓ_0 is the intercept-only log-likelihood.

Marginal Effects (MEM vs. AME) and Logit Odds Ratios

Because $G(\cdot)$ is nonlinear, $\hat{\beta}_j$ alone gives only the direction and statistical significance of covariate x_j, not the probability change. For a continuous variable x_j, the partial effect at covariate vector x_i is:

$$\frac{\partial \Pr(y_i = 1 \mid x_i)}{\partial x_{ij}} = g(x_i'\beta)\beta_j$$

Where $g(z) = G'(z)$ is $\phi(z)$ for Probit and $\Lambda(z)[1 - \Lambda(z)]$ for Logit. For a binary dummy variable $d_i \in \{0, 1\}$, the discrete marginal effect is $G(\beta_0 + \beta_1 d_i|_{d=1} + \ldots) - G(\beta_0 + \beta_1 d_i|_{d=0} + \ldots)$. Two summary reporting conventions exist:

- **Marginal Effect at the Mean (MEM):** Evaluates $g(\bar{x}'\hat{\beta})\hat{\beta}_j$ at the sample average covariate vector $\bar{x}$.
- **Average Marginal Effect (AME — Preferred):** Averages individual partial effects across all n observations, $\text{AME}_j = \frac{1}{n}\sum_{i=1}^{n} g(x_i'\hat{\beta})\hat{\beta}_j$, avoiding evaluation at impossible composite individuals (such as $\bar{d} = 0.48$ for gender).
- **Logit Odds Ratios:** In Logit, $\ln\left(\frac{\Pr(y_i=1|x_i)}{1-\Pr(y_i=1|x_i)}\right) = x_i'\beta$, so $\exp(\hat{\beta}_j)$ represents the multiplicative factor by which the odds of $y_i = 1$ change for a one-unit increase in x_{ij}.

Worked Example 2: Probit Marginal Effects Across Covariate Values

Suppose an Australian labour market Probit regression explaining employment participation ($y_i = 1$) gives $\hat{\beta}_0 = -1.20$ and $\hat{\beta}_1 = 0.30$ on post-secondary schooling years x_i. At $x_0 = 4$ years, the index is $z_0 = -1.20 + 0.30(4) = 0.00$, giving $\Pr(y_i = 1 \mid x_i = 4) = \Phi(0) = 0.5000$ and $\phi(0) = (2\pi)^{-1/2} \approx 0.3989$. The marginal effect at $x_0 = 4$ is:

$$\frac{\partial \Pr(y_i = 1 \mid x_i = 4)}{\partial x_i} = \phi(0)\hat{\beta}_1 = 0.3989 \times 0.30 = 0.1197$$

Degreelive

At $x_1 = 8$ years, the index rises to $z_1 = -1.20 + 0.30(8) = 1.20$, where $\Phi(1.20) \approx 0.8849$ and $\phi(1.20) = 0.3989e^{-0.72} \approx 0.1942$, reducing the marginal effect to $0.1942 \times 0.30 = 0.0583$ (5.83 percentage points per year) as participation approaches saturation.

4.5 Censored Regression: The Tobit (Type I) Model

In household expenditure, charitable giving, and hours-worked surveys, a non-negative continuous outcome frequently clusters at a corner value of zero for a substantial fraction of the sample while remaining smoothly distributed across positive values. Estimating y_i on x_i by OLS yields inconsistent estimates of the structural slopes because $E(y_i \mid x_i)$ is nonlinear.

The **Standard (Type I) Tobit model** (Tobin, 1958) specifies a Gaussian latent linear index censored at zero:

$$y_i^* = x_i'\beta + u_i, \qquad u_i \mid x_i \sim N(0, \sigma^2), \qquad y_i = \max(0, y_i^*)$$

Where y_i equals 0 whenever $y_i^* \leq 0$ and equals y_i^* whenever $y_i^* > 0$. Notice carefully the distinction between *data censoring* (true y_i^* is simply unrecorded below zero, e.g., top-coded ABS income brackets) and *corner solution outcomes* (zero represents an optimal economic choice, such as zero household alcohol spending or zero charitable donations).

Tobit Log-Likelihood and Conditional Expectations

For zero observations ($y_i = 0$), the probability contribution is $\Pr(y_i^* \leq 0 \mid x_i) = \Phi(-x_i'\beta/\sigma) = 1 - \Phi(x_i'\beta/\sigma)$. For positive observations ($y_i > 0$), the density contribution is $\frac{1}{\sigma}\phi\left((y_i - x_i'\beta)/\sigma\right)$. Combining both regimes yields the Tobit log-likelihood:

$$\ell(\beta, \sigma) = \sum_{y_i = 0} \ln\left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right] + \sum_{y_i > 0} \left\{-\ln\sigma + \ln\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right\}$$

Crucially, unlike linear OLS, Tobit MLE relies heavily on both normality and homoskedasticity of u_i; violations of either assumption render Tobit estimates inconsistent.

Inverse Mills Ratio and McDonald-Moffitt Marginal Effect Decomposition

Define $z_i = x_i'\beta/\sigma$ and the **Inverse Mills Ratio (IMR)** $\lambda(z) = \phi(z)/\Phi(z)$. Using properties of truncated normal distributions, the conditional and unconditional expectations of the observed outcome y_i are:

$$E(y_i \mid x_i, y_i > 0) = x_i'\beta + \sigma\lambda\left(\frac{x_i'\beta}{\sigma}\right), \qquad E(y_i \mid x_i) = \Phi\left(\frac{x_i'\beta}{\sigma}\right)x_i'\beta + \sigma\phi\left(\frac{x_i'\beta}{\sigma}\right)$$

DegreeLive

Differentiating with respect to a continuous regressor x_{ij} reveals three distinct marginal effects (and the **McDonald-Moffitt decomposition**):

1. **Unconditional (Total) Marginal Effect:** $\frac{\partial E(y_i|x_i)}{\partial x_{ij}} = \Phi(z_i)\beta_j$. Notice that β_j is the slope of latent y_i^*, whereas scaling by the participation probability $\Phi(z_i)$ yields the observable average marginal effect across the full population.
2. **Conditional (Intensive-Margin) Marginal Effect ($y_i > 0$):** $\frac{\partial E(y_i|x_i, y_i>0)}{\partial x_{ij}} = \beta_j \{1 - \lambda(z_i)\,[z_i + \lambda(z_i)]\}$, where the bracketed factor strictly lies in $(0, 1)$.
3. **Participation (Extensive-Margin) Effect:** $\frac{\partial \Pr(y_i>0|x_i)}{\partial x_{ij}} = \frac{\phi(z_i)}{\sigma}\beta_j$. By the product rule, the total marginal effect decomposes cleanly into intensive and extensive margins:

$$\frac{\partial E(y_i \mid x_i)}{\partial x_{ij}} = \Pr(y_i > 0 \mid x_i)\frac{\partial E(y_i \mid x_i, y_i > 0)}{\partial x_{ij}} + E(y_i \mid x_i, y_i > 0)\frac{\partial \Pr(y_i > 0 \mid x_i)}{\partial x_{ij}}$$

Where the first product term captures changes in expected expenditure among active participants and the second captures changes in the probability of positive participation.

Worked Example 3: Tobit Marginal Effect Decomposition

Suppose a Tobit model of annual household charitable donations y_i (in hundreds of AUD) yields $\hat{\sigma} = 4.0$ and $\hat{\beta}_{\text{inc}} = 0.80$ on household income. Evaluate at mean income where $z = x'\hat{\beta}/\hat{\sigma} = 0.50$, giving $\Phi(0.50) = 0.6915$, $\phi(0.50) = 0.3521$, and $\lambda(0.50) = 0.3521/0.6915 = 0.5092$.

- **Unconditional Marginal Effect:** $\Phi(0.50)\hat{\beta}_{\text{inc}} = 0.6915 \times 0.80 = 0.5532$ (A\$55.32 per unit income increase across all households).
- **Conditional Effect ($y_i > 0$):** $0.80[1 - 0.5092(0.50 + 0.5092)] = 0.80(1 - 0.5139) = 0.3889$ (A\$38.89 among active donors).
- **Participation Probability Effect:** $(0.3521/4.0) \times 0.80 = 0.0704$ (a 7.04 percentage point increase in donor probability).
- **McDonald-Moffitt Identity Check:** Conditional mean $E(y \mid y > 0) = 4.0(0.50 + 0.5092) = 4.0368$, so $0.6915(0.3889) + 4.0368(0.0704) = 0.2689 + 0.2842 = 0.5531$, matching the total effect 0.5532.

4.6 Sample Selection Models: Heckman Selection (Type II Tobit)

Whereas the Type I Tobit applies a single latent equation to both the threshold decision and the continuous amount, many economic problems involve two separate mechanisms: for instance, observed hourly wages y_i in the HILDA survey are observed only for individuals who choose to work ($s_i = 1$), and participation depends on reservation wages and household childcare constraints that differ from market wage determinants.

Incidental Truncation and Sample Selection Bias

Heckman's (1979) sample selection model specifies an **outcome equation** and a binary **selection equation** with jointly normal disturbances (u_i, v_i):

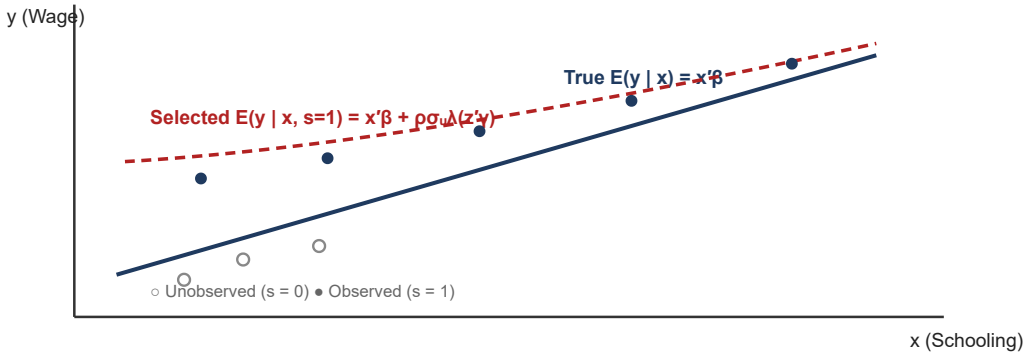
$$y_i = x_i'\beta + u_i \quad (\text{observed iff } s_i = 1), \qquad s_i = \mathbf{1}(z_i'\gamma + v_i > 0)$$

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \Big| z_i \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \rho\sigma_u \\ \rho\sigma_u & 1 \end{pmatrix}\right)$$

Where $\rho = \text{Corr}(u_i, v_i)$ measures correlation between unobservables in selection and outcome (e.g., unobserved motivation raising both employment probability and wages), and x_i is typically a strict subset of z_i. Taking the conditional expectation of y_i over the selected sample $s_i = 1$ yields:

$$E(y_i \mid z_i, s_i = 1) = x_i'\beta + E(u_i \mid v_i > -z_i'\gamma) = x_i'\beta + \rho\sigma_u\lambda(z_i'\gamma)$$

Where $\lambda(z_i'\gamma) = \phi(z_i'\gamma)/\Phi(z_i'\gamma)$ is the Inverse Mills Ratio. Whenever $\rho \neq 0$, running OLS on the selected subsample ($s_i = 1$) omits the nonlinear term $\rho\sigma_u\lambda(z_i'\gamma)$, producing **sample selection bias** as a special case of omitted variable bias.



Heckman Two-Step Estimator ("Heckit") vs. Full Information MLE

Heckman's two-step procedure corrects selection bias transparently:

1. **Step 1 (Probit Selection Equation):** Estimate $\hat{\gamma}$ by Probit MLE on the full sample of both $s_i = 1$ and $s_i = 0$ observations, and construct the estimated Inverse Mills Ratio $\hat{\lambda}_i = \phi(z_i'\hat{\gamma})/\Phi(z_i'\hat{\gamma})$ for each selected observation ($s_i = 1$).
2. **Step 2 (Augmented OLS on Selected Subsample):** Regress y_i on x_i and $\hat{\lambda}_i$ across observations with $s_i = 1$:

$$y_i = x_i'\beta + \theta_\lambda \hat{\lambda}_i + \text{error}_i, \qquad \theta_\lambda = \rho\sigma_u$$

Consistent estimation of β is achieved, and testing $H_0 : \theta_\lambda = 0$ via a standard t-test directly tests for the presence of sample selection bias (valid without adjustment under H_0, though general standard errors require Heckman's generated-regressor correction for $\hat{\gamma}$ estimation uncertainty and truncated-normal heteroskedasticity).

Although β is technically weakly identified via the nonlinearity of $\lambda(z_i'\gamma)$ even if $z_i = x_i$, $\lambda(z)$ is nearly linear over its central range, causing severe multicollinearity with x_i. Credible empirical identification therefore requires at least one **exclusion restriction**: an instrument in z_i that strongly shifts selection s_i (e.g., young children in the household or non-labour spouse income affecting labour force participation) without directly entering the structural wage equation x_i. Under joint bivariate normality, **Full Information Maximum Likelihood (FIMLE)** estimates $(\beta, \gamma, \sigma_u, \rho)$ jointly in one step and achieves full Cramér-Rao asymptotic efficiency, whereas the Heckit two-step estimator provides greater computational stability and weaker distributional sensitivity.

4.7 Count Data Models and Likelihood-Based Model Selection

When the dependent variable takes non-negative integer count values $y_i \in \{0, 1, 2, \dots\}$—such as the number of annual hospital visits in Australian health economics or patent filings by ASX firms—linear regression predicts negative counts and ignores heteroskedastic count variance. The benchmark specification is the **Poisson Regression Model**, where conditional on covariates x_i, y_i follows a Poisson distribution with conditional mean $\mu_i = \exp(x_i'\beta) > 0$:

$$\Pr(Y_i = y_i \mid x_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \qquad E(y_i \mid x_i) = \text{Var}(y_i \mid x_i) = \exp(x_i'\beta)$$

Where equality of conditional mean and variance is known as **equidispersion**. Maximising the Poisson log-likelihood $\ell(\beta) = \sum_{i=1}^{n}[y_i x_i'\beta - \exp(x_i'\beta) - \ln(y_i!)]$ produces the first-order score equation $\sum_{i=1}^{n}[y_i - \exp(x_i'\hat{\beta})]x_i = 0$. Crucially, because the Poisson distribution belongs to the linear exponential family, Poisson Quasi-MLE remains consistent for β whenever the conditional mean $E(y_i \mid x_i) = \exp(x_i'\beta)$ is correctly specified, even if data exhibit **overdispersion** ($\text{Var}(y_i \mid x_i) > E(y_i \mid x_i)$), provided Huber-White sandwich standard errors are used. When full parametric efficiency under overdispersion is desired, the **Negative Binomial (NegBin II)** model introduces unobserved Gamma heterogeneity v_i such that $\text{Var}(y_i \mid x_i) = \mu_i + \alpha\mu_i^2$, testable against Poisson via a boundary Likelihood Ratio test of $H_0 : \alpha = 0$.

Information Criteria for Non-Nested Likelihood Models

When comparing competing likelihood specifications that are not nested (such as Logit vs. Probit or Poisson vs. Hurdle models) on identical data, Likelihood Ratio tests do not apply. Analysts instead compare penalised maximised log-likelihoods using the **Akaike Information Criterion (AIC)** and **Bayesian Information Criterion (BIC)**:

$$\text{AIC} = -2\ell(\hat{\theta}) + 2k, \qquad \text{BIC} = -2\ell(\hat{\theta}) + k\ln(n)$$

Where k is the number of estimated free parameters and n is the sample size; lower values indicate superior penalised fit. For $n \geq 8$, $\ln(n) > 2$, so BIC penalises model complexity more strictly than AIC and is model-selection consistent (selecting the true parsimonious model with probability approaching 1 as $n \to \infty$).

Limited Dependent Model	Outcome Support	Primary Estimation Method	Key Structural Identification & Caveat
Logit / Probit	Binary $\{0, 1\}$	Binary Bernoulli MLE	Error scale normalised; report AMEs or odds ratios, not raw $\hat{\beta}$
Tobit (Type I)	Corner at 0, continuous > 0	Censored Normal MLE	Inconsistent under non-normality or heteroskedasticity
Heckman (Type II Tobit)	Continuous observed iff $s_i = 1$	Two-step Heckit or bivariate FIMLE	Requires valid exclusion restriction in selection equation z_i
Poisson / NegBin II	Integer counts $\{0, 1, 2, \dots\}$	Exponential-family QMLE / MLE	Poisson QMLE consistent for mean; use robust SEs or NegBin for overdispersion

Degreelive