Quasi Maximum Likelihood Theory

CHUNG-MING KUAN

Department of Finance & CRETA

June 17, 2010

Lecture Outline

1. Kullback-Leibler Information Criterion
2. Asymptotic Properties of the QMLE
   - Asymptotic Normality
   - Information Matrix Equality
3. Large Sample Tests: Nested Models
   - Wald Test
   - LM (Score) Test
   - Likelihood Ratio Test
4. Large Sample Tests: Non-Nested Models
   - Wald Encompassing Test
   - Score Encompassing Test
   - Pseudo-True Score Encompassing Test

Lecture Outline (cont'd)

5. Example I: Discrete Choice Models
   - Binary Choice Models
   - Multinomial Models
   - Ordered Multinomial Models
6. Example II: Limited Dependent Variable Models
   - Truncated Regression Models
   - Censored Regression Models
   - Sample Selection Models
7. Example III: Time Series Models

Quasi-Maximum Likelihood (QML) Theory

- Drawbacks of the least-squares method:
  - It leaves no room for modeling other conditional moments, such as conditional variance, of the dependent variable.
  - It fails to accommodate certain characteristics of the dependent variable, such as binary response and data truncation.
- The quasi-maximum likelihood (QML) method:
  - Specifying a likelihood function that admits specifications of different conditional moments and/or distribution characteristics.
  - Model misspecification is allowed, cf. conventional ML method.

Kullback-Leibler Information Criterion (KLIC)

- Given $\mathbb{P}(A) = p$, the message that A will surely occur would be more (less) valuable or more (less) surprising when p is small (large).
- The information content of the message above ought to be a decreasing function of p. An information function is

$$\iota(p) = \log(1/p),$$

which decreases from positive infinity ($p \approx 0$) to zero ($p = 1$).

- Clearly, $\iota(1-p) \neq \iota(p)$. The expected information is

$$I = p\,\iota(p) + (1-p)\,\iota(1-p) = p \log\left(\frac{1}{p}\right) + (1-p) \log\left(\frac{1}{1-p}\right),$$

which is also known as the entropy of the event A.

- The information that $\mathbb{P}(A)$ changes from p to q would be useful when p and q are very different. The information content is

$$\iota(p) - \iota(q) = \log(q/p),$$

which is positive (negative) when $q > p$ ($q < p$).

- Given n mutually exclusive events $A_1, \ldots, A_n$, each with an information value $\log(q_i/p_i)$, the expected information value is

$$I = \sum_{i=1}^{n} q_i \log\left(\frac{q_i}{p_i}\right).$$

- Kullback-Leibler Information Criterion (KLIC) of g relative to f is

$$\mathbb{I}(g:f) = \int_{\mathbb{R}} \log\left(\frac{g(\zeta)}{f(\zeta)}\right) g(\zeta)\, \mathrm{d}\zeta,$$

where g is the density function of z and f is another density function.

Theorem 9.1

$\mathbb{I}(g:f) \geq 0$; the equality holds if, and only if, $g = f$ almost everywhere.

Proof: As $\log(1+x) < x$ for all $x > -1$, we have

$$\log\left(\frac{g}{f}\right) = -\log\left(1 + \frac{f-g}{g}\right) > 1 - \frac{f}{g}.$$

It follows that

$$\int \log\left(\frac{g(\zeta)}{f(\zeta)}\right) g(\zeta)\, \mathrm{d}\zeta > \int \left(1 - \frac{f(\zeta)}{g(\zeta)}\right) g(\zeta)\, \mathrm{d}\zeta = 0.$$

Remark: The KLIC measures the "closeness" between f and g, but is not a metric because it is not reflexive in general, i.e., $\mathbb{I}(g:f) \neq \mathbb{I}(f:g)$, and does not obey the triangle inequality.

Let $\mathbf{z}_t = (y_t \ \mathbf{w}_t')'$ be $\nu \times 1$ and $\mathbf{z}^t = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_t\}$.

- Given a sample of T observations, specifying a complete probability model for $\mathbf{z}^T$ may be a formidable task in practice.
- It is practically more convenient to focus on the conditional density $g_t(y_t \mid \mathbf{x}_t)$, where $\mathbf{x}_t$ include some elements of $\mathbf{w}_t$ and $\mathbf{z}^{t-1}$. We thus specify a quasi-likelihood function $f_t(y_t \mid \mathbf{x}_t; \boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$.
- he KLIC of g_t relative to f_t is

$$\mathbb{I}(g_t : f_t; \boldsymbol{\theta}) = \int_{\mathbb{R}} \log \left(\frac{g_t(y_t \mid \mathbf{x}_t)}{f_t(y_t \mid \mathbf{x}_t; \boldsymbol{\theta})} \right) g_t(y_t \mid \mathbf{x}_t) \, \mathrm{d}\, y_t.$$

- For a sample of T obs, the average of T individual KLICs is:

$$\bar{\mathbb{I}}_T(\{g_t : f_t\}; \boldsymbol{\theta}) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}(g_t : f_t; \boldsymbol{\theta})$$

$$= \frac{1}{T} \sum_{t=1}^{T} \big(\mathbb{E}[\log g_t(y_t \mid \mathbf{x}_t)] - \mathbb{E}[\log f_t(y_t \mid \mathbf{x}_t; \boldsymbol{\theta})] \big).$$

- Minimizing $\bar{\mathbb{I}}_T(\{g_t : f_t\}; \boldsymbol{\theta})$ is equivalent to maximizing

$$\bar{L}_T(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\log f_t(y_t \mid \mathbf{x}_t; \boldsymbol{\theta})];$$

The maximizer of $\bar{L}_T(\boldsymbol{\theta})$, $\boldsymbol{\theta}^*$, is the minimizer of the average KLIC.

- If there exists a $\boldsymbol{\theta}_o \in \Theta$ such that $f_t(y_t \mid \mathbf{x}_t; \boldsymbol{\theta}_o) = g_t(y_t \mid \mathbf{x}_t)$ for all t, we say that $\{f_t\}$ is correctly specified for $\{y_t \mid \mathbf{x}_t\}$. In this case, $\mathbb{I}(g_t : f_t; \boldsymbol{\theta}_o) = 0$, so that $\bar{\mathbb{I}}_T(\{g_t : f_t\}; \boldsymbol{\theta})$ is minimized at $\boldsymbol{\theta}^* = \boldsymbol{\theta}_o$.
- Maximizing the sample counterpart of $\bar{L}_T(\boldsymbol{\theta})$:

$$L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta}) := \frac{1}{T} \sum_{t=1}^{T} \log f_t(y_t \mid \mathbf{x}_t; \boldsymbol{\theta}),$$

the average of the individual quasi-log-likelihood functions, the resulting solution, $\tilde{\boldsymbol{\theta}}_T$, is known as the quasi-maximum likelihood estimator (QMLE) of $\boldsymbol{\theta}$. When $\{f_t\}$ is specified correctly for $\{y_t \mid \mathbf{x}_t\}$, the QMLE is the standard MLE.

- Concentrating on certain conditional attribute of y_t, we have, for example, $y_t \mid \mathbf{x}_t \sim \mathcal{N}\left(\mu_t(\mathbf{x}_t;\boldsymbol{\beta}), \sigma^2\right)$, or more generally,

$$y_t \mid \mathbf{x}_t \sim \mathcal{N}\left(\mu_t(\mathbf{x}_t;\boldsymbol{\beta}), h(\mathbf{x}_t;\boldsymbol{\alpha})\right).$$

- Suppose $y_t \mid \mathbf{x}_t \sim \mathcal{N}\left(\mu_t(\mathbf{x}_t;\boldsymbol{\beta}), \sigma^2\right)$ and let $\boldsymbol{\theta} = (\boldsymbol{\beta}' \ \sigma^2)'$. The maximizer of $T^{-1}\sum_{t=1}^{T} \log f_t(y_t \mid \mathbf{x}_t;\boldsymbol{\theta})$ also solves

$$\min_{\boldsymbol{\beta}} \frac{1}{T}\sum_{t=1}^{T} \left[y_t - \mu_t(\mathbf{x}_t;\boldsymbol{\beta})\right]' \left[y_t - \mu_t(\mathbf{x}_t;\boldsymbol{\beta})\right].$$

  That is, the NLS estimator is a QMLE under the specification of conditional normality with conditional homoskedasticity.

- Even when $\{\mu_t\}$ is correctly specified for the conditional mean, there is no guarantee that the specification of σ^2 is correct.

Asymptotic Properties of the QMLE

- The quasi-log-likelihood function is, in general, a nonlinear function in $\boldsymbol{\theta}$. The QMLE $\tilde{\boldsymbol{\theta}}_T$ must be computed numerically using a nonlinear optimization algorithm.
- **[ID-3]** There exists a unique $\boldsymbol{\theta}^*$ that minimizes the KLIC.
- Consistency: Suppose that $L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta})$ tends to $\bar{L}_T(\boldsymbol{\theta})$ in probability, uniformly in $\boldsymbol{\theta} \in \Theta$, i.e., $L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta})$ obeys a WULLN. Then, it is natural to expect the QMLE $\tilde{\boldsymbol{\theta}}_T$ to converge in probability to $\boldsymbol{\theta}^*$, the minimizer of the average KLIC, $\bar{\mathbb{I}}_T(\{g_t : f_t\}; \boldsymbol{\theta})$.

Asymptotic Normality

When $\boldsymbol{\theta}^*$ is in the interior of Θ, the mean-value expansion of $\nabla L_T(y^T, \mathbf{x}^T; \tilde{\boldsymbol{\theta}}_T)$ about $\boldsymbol{\theta}^*$ is

$$\nabla L_T(y^T, \mathbf{x}^T; \tilde{\boldsymbol{\theta}}_T) = \nabla L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta}^*) + \nabla^2 L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta}_T^\dagger)(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*),$$

where the left-hand side is zero because $\tilde{\boldsymbol{\theta}}_T$ must satisfy the first order condition.

Let $\mathbf{H}_T(\boldsymbol{\theta}) = \mathbb{E}[\nabla^2 L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta})]$. When $\nabla^2 L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta})$ obeys a WULLN,

$$\nabla^2 L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta}_T^\dagger) - \mathbf{H}_T(\boldsymbol{\theta}^*) \xrightarrow{\mathbb{P}} \mathbf{0}.$$

When $\mathbf{H}_T(\boldsymbol{\theta}^*)$ is nonsingular,

$$\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = -\mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\sqrt{T}\,\nabla L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta}^*) + o_{\mathbb{P}}(1).$$

Let $\mathbf{B}_T(\boldsymbol{\theta}) = \mathrm{var}\big(\sqrt{T}\nabla L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta})\big)$ be the information matrix. When $\nabla \log f_t(y_t \mid \mathbf{x}_t; \boldsymbol{\theta})$ obeys a CLT,

$$\mathbf{B}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}\big(\nabla L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta}^*) - \mathbb{E}[\nabla L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta}^*)]\big) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{I}_k).$$

Knowing $\mathbb{E}[\nabla L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta})] = \nabla\, \mathbb{E}[L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta})]$ is zero at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, the KLIC minimizer, we have

$$\mathbf{B}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}\,\nabla L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta}^*) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{I}_k).$$

This shows that $\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*)$ is asymptotically equivalent to

$$-\mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\mathbf{B}_T(\boldsymbol{\theta}^*)^{1/2}\left[\mathbf{B}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}\,\nabla L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta}^*)\right],$$

which has an asymptotic normal distribution.

Theorem 9.2

Letting $\mathbf{C}_T(\boldsymbol{\theta}^*) = \mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\mathbf{B}_T(\boldsymbol{\theta}^*)\mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}$,

$$\mathbf{C}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{I}_k).$$

Information Matrix Equality

When the likelihood is specified correctly in a proper sense, the information matrix equality holds:

$$\mathbf{H}_T(\boldsymbol{\theta}_o) + \mathbf{B}_T(\boldsymbol{\theta}_o) = \mathbf{0}.$$

In this case, $\mathbf{C}_T(\boldsymbol{\theta}_o)$ simplifies to $-\mathbf{H}_T(\boldsymbol{\theta}_o)^{-1}$ or $\mathbf{B}_T(\boldsymbol{\theta}_o)^{-1}$.

For the specification of $\{y_t|\mathbf{x}_t\}$, define the score functions:

$$\mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}) = \nabla \log f_t(y_t|\mathbf{x}_t; \boldsymbol{\theta}) = f_t(y_t|\mathbf{x}_t; \boldsymbol{\theta})^{-1} \nabla f_t(y_t|\mathbf{x}_t; \boldsymbol{\theta}),$$

so that $\nabla f_t(y_t|\mathbf{x}_t; \boldsymbol{\theta}) = \mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}) f_t(y_t|\mathbf{x}_t; \boldsymbol{\theta})$.

By permitting interchange of differentiation and integration,

$$\int_{\mathbb{R}} \mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}) f_t(y_t|\mathbf{x}_t; \boldsymbol{\theta}) \, \mathrm{d}y_t = \nabla \int_{\mathbb{R}} f_t(y_t|\mathbf{x}_t; \boldsymbol{\theta}) \, \mathrm{d}y_t = \mathbf{0}.$$

If $\{f_t\}$ is correctly specified for $\{y_t|\mathbf{x}_t\}$, $\mathbb{E}[\mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}_o)|\mathbf{x}_t] = \mathbf{0}$, where the conditional expectation is taken with respect to $g_t(y_t|\mathbf{x}_t) = f_t(y_t|\mathbf{x}_t; \boldsymbol{\theta}_o)$. Thus, mean score is zero: $\mathbb{E}[\mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}_o)] = \mathbf{0}$.

As $\nabla f_t = \mathbf{s}_t f_t$, we have $\nabla^2 f_t = (\nabla \mathbf{s}_t) f_t + (\mathbf{s}_t \mathbf{s}_t') f_t$, so that

$$\begin{aligned}
&\int_{\mathbb{R}} \left[\nabla \mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}) + \mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}) \mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta})'\right] f_t(y_t|\mathbf{x}_t; \boldsymbol{\theta}) \, \mathrm{d}y_t \\
&\quad = \nabla^2 \int_{\mathbb{R}} f_t(y_t|\mathbf{x}_t; \boldsymbol{\theta}) \, \mathrm{d}y_t \\
&\quad = \mathbf{0}.
\end{aligned}$$

We have shown

$$\mathbb{E}[\nabla \mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}_o)|\mathbf{x}_t] + \mathbb{E}[\mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}_o)\mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}_o)'|\mathbf{x}_t] = \mathbf{0}.$$

It follows that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\nabla \mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}_o)] + \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}_o)\mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}_o)'] \\
&\quad = \mathbf{H}_T(\boldsymbol{\theta}_o) + \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}_o)\mathbf{s}_t(y_t, \mathbf{x}_t; \boldsymbol{\theta}_o)'] \\
&\quad = \mathbf{0}.
\end{aligned}$$

i.e., the expected Hessian matrix is negative of the average of individual information matrices, which need not be the information matrix $\mathbf{B}_T(\theta_o)$.

By definition, the information matrix is

$$\begin{aligned}
\mathbf{B}_T(\boldsymbol{\theta}_o) &= \frac{1}{T}\mathbb{E}\left[\left(\sum_{t=1}^{T}\mathbf{s}_t(y_t,\mathbf{x}_t;\boldsymbol{\theta}_o)\right)\left(\sum_{t=1}^{T}\mathbf{s}_t(y_t,\mathbf{x}_t;\boldsymbol{\theta}_o)'\right)\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathbf{s}_t(y_t,\mathbf{x}_t;\boldsymbol{\theta}_o)\mathbf{s}_t(y_t,\mathbf{x}_t;\boldsymbol{\theta}_o)'] \\
&\quad + \frac{1}{T}\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T}\mathbb{E}[\mathbf{s}_{t-\tau}(y_{t-\tau},\mathbf{x}_{t-\tau};\boldsymbol{\theta}_o)\mathbf{s}_t(y_t,\mathbf{x}_t;\boldsymbol{\theta}_o)'] \\
&\quad + \frac{1}{T}\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T}\mathbb{E}[\mathbf{s}_t(y_t,\mathbf{x}_t;\boldsymbol{\theta}_o)\mathbf{s}_{t+\tau}(y_{t+\tau},\mathbf{x}_{t+\tau};\boldsymbol{\theta}_o)'].
\end{aligned}$$

That is, the information matrix involves the variances and autocovariances of individual score functions.

A specification of $\{y_t|\mathbf{x}_t\}$ is said to have dynamic misspecification if it is not correctly specified for $\{y_t|\mathbf{w}_t, \mathbf{z}^{t-1}\}$; that is, there does not exist any $\boldsymbol{\theta}_o$ such that $f_t(y_t|\mathbf{x}_t;\boldsymbol{\theta}_o) = g_t(y_t|\mathbf{w}_t, \mathbf{z}^{t-1})$.

When $f_t(y_t|\mathbf{x}_t;\boldsymbol{\theta}_o) = g_t(y_t|\mathbf{w}_t, \mathbf{z}^{t-1})$,

$$\mathbb{E}[\mathbf{s}_t(y_t, \mathbf{x}_t;\boldsymbol{\theta}_o)|\mathbf{x}_t] = \mathbb{E}[\mathbf{s}_t(y_t, \mathbf{x}_t;\boldsymbol{\theta}_o)|\mathbf{w}_t, \mathbf{z}^{t-1}] = \mathbf{0},$$

and by the law of iterated expectations,

$$\begin{aligned}&\mathbb{E}\left[\mathbf{s}_t(y_t, \mathbf{x}_t;\boldsymbol{\theta}_o)\mathbf{s}_{t+\tau}(y_{t+\tau}, \mathbf{x}_{t+\tau};\boldsymbol{\theta}_o)'\right] \\ &\quad = \mathbb{E}\left[\mathbf{s}_t(y_t, \mathbf{x}_t;\boldsymbol{\theta}_o)\,\mathbb{E}[\mathbf{s}_{t+\tau}(y_{t+\tau}, \mathbf{x}_{t+\tau};\boldsymbol{\theta}_o)'|w_{t+\tau}, \mathbf{z}^{t+\tau-1}]\right] = \mathbf{0},\end{aligned}$$

for $\tau \geq 1$. In this case,

$$\mathbf{B}_T(\boldsymbol{\theta}_o) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{s}_t(y_t, \mathbf{x}_t;\boldsymbol{\theta}_o)\mathbf{s}_t(y_t, \mathbf{x}_t;\boldsymbol{\theta}_o)'\right].$$

Theorem 9.3

Suppose that there exists a $\boldsymbol{\theta}_o$ such that $f_t(y_t|\mathbf{x}_t;\boldsymbol{\theta}_o) = g_t(y_t|\mathbf{x}_t)$ and there is no dynamic misspecification. Then,

$$\mathbf{H}_T(\boldsymbol{\theta}_o) + \mathbf{B}_T(\boldsymbol{\theta}_o) = \mathbf{0},$$

where $\mathbf{H}_T(\boldsymbol{\theta}_o) = T^{-1}\sum_{t=1}^{T} \mathbb{E}[\nabla \mathbf{s}_t(y_t, \mathbf{x}_t;\boldsymbol{\theta}_o)]$ and

$$\mathbf{B}_T(\boldsymbol{\theta}_o) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathbf{s}_t(y_t, \mathbf{x}_t;\boldsymbol{\theta}_o)\mathbf{s}_t(y_t, \mathbf{x}_t;\boldsymbol{\theta}_o)'].$$

When Theorem 9.3 holds, $\mathbf{B}_T(\boldsymbol{\theta}_o)^{1/2}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_o) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$. That is, the QMLE is asymptotically efficient because it achieves the Cramér-Rao lower bound asymptotically.

Example: Assume: $y_t|\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t'\boldsymbol{\beta}, \sigma^2)$ for all t, so that

$$L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta}) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{T}\sum_{t=1}^{T}\frac{(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2}{2\sigma^2}.$$

Straightforward calculation yields

$$\nabla L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} \frac{\mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})}{\sigma^2} \\ -\frac{1}{2\sigma^2} + \frac{(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2}{2(\sigma^2)^2} \end{bmatrix},$$

$$\nabla^2 L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} -\frac{\mathbf{x}_t\mathbf{x}_t'}{\sigma^2} & -\frac{\mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})}{(\sigma^2)^2} \\ -\frac{(y_t - \mathbf{x}_t'\boldsymbol{\beta})\mathbf{x}_t'}{(\sigma^2)^2} & \frac{1}{2(\sigma^2)^2} - \frac{(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2}{(\sigma^2)^3} \end{bmatrix}.$$

solving $\nabla L_T(y^T, \mathbf{x}^T; \boldsymbol{\theta}) = \mathbf{0}$ we obtain the QMLEs for $\boldsymbol{\beta}$ and σ^2.

If the specification above is correct for $\{y_t|\mathbf{x}_t\}$, there exists $\boldsymbol{\theta}_o = (\boldsymbol{\beta}_o' \ \sigma_o^2)'$ such that $y_t|\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t'\boldsymbol{\beta}_o, \sigma_o^2)$. Taking expectation with respect to the true distribution,

$$\mathbb{E}[\mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})] = \mathbb{E}[\mathbf{x}_t(\mathbb{E}(y_t|\mathbf{x}_t) - \mathbf{x}_t'\boldsymbol{\beta})] = \mathbb{E}(\mathbf{x}_t\mathbf{x}_t')(\boldsymbol{\beta}_o - \boldsymbol{\beta}),$$

which is zero when evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_o$. Similarly,

$$\begin{aligned}\mathbb{E}[(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2] &= \mathbb{E}[(y_t - \mathbf{x}_t'\boldsymbol{\beta}_o + \mathbf{x}_t'\boldsymbol{\beta}_o - \mathbf{x}_t'\boldsymbol{\beta})^2] \\ &= \mathbb{E}[(y_t - \mathbf{x}_t'\boldsymbol{\beta}_o)^2] + \mathbb{E}[(\mathbf{x}_t'\boldsymbol{\beta}_o - \mathbf{x}_t'\boldsymbol{\beta})^2] \\ &= \sigma_o^2 + \mathbb{E}[(\mathbf{x}_t'\boldsymbol{\beta}_o - \mathbf{x}_t'\boldsymbol{\beta})^2],\end{aligned}$$

where the second term on the right-hand side is zero if it is evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_o$.

The results above show that

$$\begin{aligned}\mathbf{H}_T(\boldsymbol{\theta}) &= \mathbb{E}[\nabla^2 L_T(\boldsymbol{\theta})] \\ &= \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} -\dfrac{\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')}{\sigma^2} & -\dfrac{\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')(\boldsymbol{\beta}_o-\boldsymbol{\beta})}{(\sigma^2)^2} \\ -\dfrac{(\boldsymbol{\beta}_o-\boldsymbol{\beta})'\,\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')}{(\sigma^2)^2} & \dfrac{1}{2(\sigma^2)^2} - \dfrac{\sigma_o^2+\mathbb{E}[(\mathbf{x}_t'\boldsymbol{\beta}_o-\mathbf{x}_t'\boldsymbol{\beta})^2]}{(\sigma^2)^3} \end{bmatrix},\end{aligned}$$

and

$$\mathbf{H}_T(\boldsymbol{\theta}_o) = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} -\dfrac{\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')}{\sigma_o^2} & \mathbf{0} \\ \mathbf{0}' & -\dfrac{1}{2(\sigma_o^2)^2} \end{bmatrix}.$$

Without dynamic misspecification, the information matrix $\mathbf{B}_T(\boldsymbol{\theta})$ is

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\begin{array}{cc} \frac{(y_t-\mathbf{x}_t'\boldsymbol{\beta})^2\mathbf{x}_t\mathbf{x}_t'}{(\sigma^2)^2} & -\frac{\mathbf{x}_t(y_t-\mathbf{x}_t'\boldsymbol{\beta})}{2(\sigma^2)^2}+\frac{\mathbf{x}_t(y_t-\mathbf{x}_t'\boldsymbol{\beta})^3}{2(\sigma^2)^3} \\ -\frac{(y_t-\mathbf{x}_t'\boldsymbol{\beta})\mathbf{x}_t'}{2(\sigma^2)^2}+\frac{(y_t-\mathbf{x}_t'\boldsymbol{\beta})^3\mathbf{x}_t'}{2(\sigma^2)^3} & \frac{1}{4(\sigma^2)^2}-\frac{(y_t-\mathbf{x}_t'\boldsymbol{\beta})^2}{2(\sigma^2)^3}+\frac{(y_t-\mathbf{x}_t'\boldsymbol{\beta})^4}{4(\sigma^2)^4} \end{array}\right].$$

Given conditional normality, its conditional third and fourth central moments are zero and $3(\sigma_o^2)^2$, respectively. Then,

$$\mathbb{E}[(y_t-\mathbf{x}_t'\boldsymbol{\beta})^3] = 3\sigma_o^2\,\mathbb{E}[(\mathbf{x}_t'\boldsymbol{\beta}_o-\mathbf{x}_t'\boldsymbol{\beta})]+\mathbb{E}[(\mathbf{x}_t'\boldsymbol{\beta}_o-\mathbf{x}_t'\boldsymbol{\beta})^3],$$

which is zero when evaluated at $\boldsymbol{\beta}=\boldsymbol{\beta}_o$. Similarly,

$$\mathbb{E}[(y_t-\mathbf{x}_t'\boldsymbol{\beta})^4] = 3(\sigma_o^2)^2+6\sigma_o^2\,\mathbb{E}[(\mathbf{x}_t'\boldsymbol{\beta}_o-\mathbf{x}_t'\boldsymbol{\beta})^2]+\mathbb{E}[(\mathbf{x}_t'\boldsymbol{\beta}_o-\mathbf{x}_t'\boldsymbol{\beta})^4],$$

which is $3(\sigma_o^2)^2$ when evaluated at $\boldsymbol{\beta}=\boldsymbol{\beta}_o$.

It is now easily seen that the information matrix equality holds because

$$\mathbf{B}_T(\boldsymbol{\theta}_o) = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} \frac{\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')}{\sigma_o^2} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2(\sigma_o^2)^2} \end{bmatrix}.$$

A typical consistent estimator of $\mathbf{H}_T(\boldsymbol{\theta}_o)$ is

$$\widetilde{\mathbf{H}}_T = \begin{bmatrix} -\frac{\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'}{T\tilde{\sigma}_T^2} & \mathbf{0} \\ \mathbf{0}' & -\frac{1}{2(\tilde{\sigma}_T^2)^2} \end{bmatrix}.$$

When the information matrix equality holds, a consistent estimator of $\mathbf{C}_T(\boldsymbol{\theta}_o)$ is $-\widetilde{\mathbf{H}}_T^{-1}$.

Example: The specification is $y_t|\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t'\boldsymbol{\beta}, \sigma^2)$, but the true conditional behavior is $y_t|\mathbf{x}_t \sim \mathcal{N}\big(\mathbf{x}_t'\boldsymbol{\beta}_o, h(\mathbf{x}_t, \boldsymbol{\alpha}_o)\big)$. That is, our specification is correct only for the conditional mean. Due to misspecification, the KLIC minimizer is $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_o' \ (\sigma^*)^2)'$. Then,

$$\mathbf{H}_T(\boldsymbol{\theta}^*) = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} -\dfrac{\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')}{(\sigma^*)^2} & \mathbf{0} \\ \mathbf{0}' & \dfrac{1}{2(\sigma^*)^4} - \dfrac{\mathbb{E}[h(\mathbf{x}_t,\boldsymbol{\alpha}_o)]}{(\sigma^*)^6} \end{bmatrix},$$

and

$$\mathbf{B}_T(\boldsymbol{\theta}^*) = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} \dfrac{\mathbb{E}[h(\mathbf{x}_t,\boldsymbol{\alpha}_o)\mathbf{x}_t\mathbf{x}_t']}{(\sigma^*)^4} & \mathbf{0} \\ \mathbf{0}' & \dfrac{1}{4(\sigma^*)^4} - \dfrac{\mathbb{E}[h(\mathbf{x}_t,\boldsymbol{\alpha}_o)]}{2(\sigma^*)^6} + \dfrac{3\,\mathbb{E}[h(\mathbf{x}_t,\boldsymbol{\alpha}_o)^2]}{4(\sigma^*)^8} \end{bmatrix}.$$

The information matrix equality does not hold here, despite that the conditional mean function is specified correctly.

The upper-left block of $\widetilde{\mathbf{H}}_T$ is $-\sum_{t=1}^T \mathbf{x}_t\mathbf{x}_t'/(T\tilde{\sigma}_T^2)$, which remains a consistent estimator of the corresponding block in $\mathbf{H}_T(\boldsymbol{\theta}^*)$. The information matrix $\mathbf{B}_T(\boldsymbol{\theta}^*)$ can be consistently estimated by a block-diagonal matrix with the upper-left block:

$$\frac{\sum_{t=1}^T \hat{e}_t^2\mathbf{x}_t\mathbf{x}_t'}{T(\tilde{\sigma}_T^2)^2}.$$

The upper-left block of $\widetilde{\mathbf{C}}_T = \widetilde{\mathbf{H}}_T^{-1}\widetilde{\mathbf{B}}_T\widetilde{\mathbf{H}}_T^{-1}$ is thus

$$\left(\frac{1}{T}\sum_{t=1}^T \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^T \hat{e}_t^2\mathbf{x}_t\mathbf{x}_t'\right)\left(\frac{1}{T}\sum_{t=1}^T \mathbf{x}_t\mathbf{x}_t'\right)^{-1}.$$

This is precisely the the Eicker-White estimator.

Wald Test

Consider the null hypothesis $\mathbf{R}\boldsymbol{\theta}^* = \mathbf{r}$, where $\mathbf{R}$ is $q \times k$ matrix with full row rank. The Wald test checks if $\mathbf{R}\tilde{\boldsymbol{\theta}}_T$ is sufficiently "close" to $\mathbf{r}$.

By the asymptotic normality result:

$$\mathbf{C}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{I}_k),$$

we have under the null hypothesis that

$$[\mathbf{R}\mathbf{C}_T(\boldsymbol{\theta}^*)\mathbf{R}']^{-1/2}\sqrt{T}(\mathbf{R}\tilde{\boldsymbol{\theta}}_T - \mathbf{r}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{I}_q).$$

This result remains valid when $\mathbf{C}_T(\boldsymbol{\theta}^*)$ is replaced by a consistent estimator: $\widetilde{\mathbf{C}}_T = \widetilde{\mathbf{H}}_T^{-1}\widetilde{\mathbf{B}}_T\widetilde{\mathbf{H}}_T^{-1}$. The Wald test is

$$\mathcal{W}_T = T(\mathbf{R}\tilde{\boldsymbol{\theta}}_T - \mathbf{r})'(\mathbf{R}\widetilde{\mathbf{C}}_T\mathbf{R}')^{-1}(\mathbf{R}\tilde{\boldsymbol{\theta}}_T - \mathbf{r}) \xrightarrow{D} \chi^2(q).$$

Example: Specification: $y_t|\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t'\boldsymbol{\beta}, \sigma^2)$. Writing $\boldsymbol{\theta} = (\sigma^2 \ \boldsymbol{\beta}')'$ and $\boldsymbol{\beta} = (\mathbf{b}_1' \ \mathbf{b}_2')'$, where $\mathbf{b}_1$ is $(k-s)\times 1$, and $\mathbf{b}_2$ is $s \times 1$. We are interested in the hypothesis $\mathbf{b}_2^* = \mathbf{R}\boldsymbol{\theta}^* = \mathbf{0}$, where $\mathbf{R} = [\mathbf{0} \ \mathbf{R}_1]$ and $\mathbf{R}_1 = [\mathbf{0} \ \mathbf{I}_s]$ is $s \times k$.

With $\tilde{\boldsymbol{\beta}}_{2,T} = \mathbf{R}\tilde{\boldsymbol{\theta}}_T$, the Wald test is

$$\mathcal{W}_T = T\tilde{\boldsymbol{\beta}}_{2,T}'(\mathbf{R}\widetilde{\mathbf{C}}_T\mathbf{R}')^{-1}\tilde{\boldsymbol{\beta}}_{2,T}.$$

When the information matrix equality holds, $\tilde{\mathbf{C}}_T = -\widetilde{\mathbf{H}}_T^{-1}$ so that

$$\mathbf{R}\widetilde{\mathbf{C}}_T\mathbf{R}' = -\mathbf{R}\widetilde{\mathbf{H}}_T^{-1}\mathbf{R}' = \tilde{\sigma}_T^2\mathbf{R}_1(\mathbf{X}'\mathbf{X}/T)^{-1}\mathbf{R}_1'.$$

The Wald test becomes

$$\mathcal{W}_T = T\tilde{\boldsymbol{\beta}}_{2,T}'[\mathbf{R}_1(\mathbf{X}'\mathbf{X}/T)^{-1}\mathbf{R}_1']^{-1}\tilde{\boldsymbol{\beta}}_{2,T}/\tilde{\sigma}_T^2 \xrightarrow{D} \chi^2(s).$$

LM (Score) Test

Maximizing $L_T(\boldsymbol{\theta})$ subject to the constraint $\mathbf{R}\boldsymbol{\theta} = \mathbf{r}$ yields the Lagrangian:

$$L_T(\boldsymbol{\theta}) + \boldsymbol{\theta}'\mathbf{R}'\boldsymbol{\lambda},$$

where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers. The constrained QMLEs are $\ddot{\boldsymbol{\theta}}_T$ and $\ddot{\boldsymbol{\lambda}}_T$; we want to check if $\ddot{\boldsymbol{\lambda}}_T$ is sufficiently "close" to zero.

First note the saddle-point condition: $\nabla L_T(\ddot{\boldsymbol{\theta}}_T) + \mathbf{R}'\ddot{\boldsymbol{\lambda}}_T = \mathbf{0}$. The mean-value expansion of $\nabla L_T(\ddot{\boldsymbol{\theta}}_T)$ about $\boldsymbol{\theta}^*$ yields

$$\nabla L_T(\boldsymbol{\theta}^*) + \nabla^2 L_T(\boldsymbol{\theta}_T^\dagger)(\ddot{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) + \mathbf{R}'\ddot{\boldsymbol{\lambda}}_T = \mathbf{0},$$

where $\boldsymbol{\theta}_T^\dagger$ is the mean value between $\ddot{\boldsymbol{\theta}}_T$ and $\boldsymbol{\theta}^*$.

Recall from the discussion of the Wald test that

$$\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = -\mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\sqrt{T}\nabla L_T(\boldsymbol{\theta}^*) + o_{\mathbb{P}}(1),$$

we obtain

$$\begin{aligned}\mathbf{0} &= \mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\sqrt{T}\nabla L_T(\boldsymbol{\theta}^*) - \mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\nabla^2 L_T(\boldsymbol{\theta}_T^\dagger)\sqrt{T}(\ddot{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) \\ &\qquad - \mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\sqrt{T}\mathbf{R}'\ddot{\boldsymbol{\lambda}}_T \\ &= \sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) - \sqrt{T}(\ddot{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) - \mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\mathbf{R}'\sqrt{T}\ddot{\boldsymbol{\lambda}}_T + o_{\mathbb{P}}(1).\end{aligned}$$

Pre-multiplying both sides by $\mathbf{R}$ and noting that $\mathbf{R}(\ddot{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) = \mathbf{0}$, we have

$$\sqrt{T}\ddot{\boldsymbol{\lambda}}_T = [\mathbf{R}\mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\mathbf{R}']^{-1}\mathbf{R}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) + o_{\mathbb{P}}(1).$$

For $\mathbf{\Lambda}_T(\boldsymbol{\theta}^*) = [\mathbf{R}\mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\mathbf{R}']^{-1}\mathbf{R}\mathbf{C}_T(\boldsymbol{\theta}^*)\mathbf{R}'[\mathbf{R}\mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\mathbf{R}']^{-1}$, the normalized Lagrangian multiplier is

$$\begin{aligned}\mathbf{\Lambda}_T(\boldsymbol{\theta}^*)^{-1/2}\sqrt{T}\ddot{\boldsymbol{\lambda}}_T &= \mathbf{\Lambda}_T(\boldsymbol{\theta}^*)^{-1/2}[\mathbf{R}\mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\mathbf{R}']^{-1}\mathbf{R}\sqrt{T}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) \\ &\xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{I}_q).\end{aligned}$$

It follows that

$$\ddot{\mathbf{\Lambda}}_T^{-1/2}\sqrt{T}\ddot{\boldsymbol{\lambda}}_T \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{I}_q).$$

where $\ddot{\mathbf{\Lambda}}_T = (\mathbf{R}\ddot{\mathbf{H}}_T^{-1}\mathbf{R}')^{-1}\mathbf{R}\ddot{\mathbf{C}}_T\mathbf{R}'(\mathbf{R}\ddot{\mathbf{H}}_T^{-1}\mathbf{R}')^{-1}$, and $\ddot{\mathbf{H}}_T$ and $\ddot{\mathbf{C}}_T$ are consistent estimators based on the constrained QMLE $\ddot{\boldsymbol{\theta}}_T$. The LM test is

$$\mathcal{LM}_T = T\ddot{\boldsymbol{\lambda}}_T'\ddot{\mathbf{\Lambda}}_T^{-1}\ddot{\boldsymbol{\lambda}}_T \xrightarrow{D} \chi^2(q).$$

In the light of the saddle-point condition: $\nabla L_T(\ddot{\boldsymbol{\theta}}_T) + \mathbf{R}'\ddot{\boldsymbol{\lambda}}_T = \mathbf{0}$,

$$\begin{aligned}\mathcal{LM}_T &= T\ddot{\boldsymbol{\lambda}}_T'\mathbf{R}\ddot{\mathbf{H}}_T^{-1}\mathbf{R}'(\mathbf{R}\ddot{\mathbf{C}}_T\mathbf{R}')^{-1}\mathbf{R}\ddot{\mathbf{H}}_T^{-1}\mathbf{R}'\ddot{\boldsymbol{\lambda}}_T \\ &= T[\nabla L_T(\ddot{\boldsymbol{\theta}}_T)]'\ddot{\mathbf{H}}_T^{-1}\mathbf{R}'(\mathbf{R}\ddot{\mathbf{C}}_T\mathbf{R}')^{-1}\mathbf{R}\ddot{\mathbf{H}}_T^{-1}[\nabla L_T(\ddot{\boldsymbol{\theta}}_T)],\end{aligned}$$

which mainly depends on the score function ∇L_T evaluated at $\ddot{\boldsymbol{\theta}}_T$.

When the information matrix equality holds,

$$\begin{aligned}\mathcal{LM}_T &= -T\ddot{\boldsymbol{\lambda}}_T'\mathbf{R}\ddot{\mathbf{H}}_T^{-1}\mathbf{R}'\ddot{\boldsymbol{\lambda}}_T \\ &= -T[\nabla L_T(\ddot{\boldsymbol{\theta}}_T)]'\ddot{\mathbf{H}}_T^{-1}[\nabla L_T(\ddot{\boldsymbol{\theta}}_T)].\end{aligned}$$

The LM test is also known as the score test in the statistics literature.

Example: Specification: $y_t|\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t'\boldsymbol{\beta}, \sigma^2)$. Let $\boldsymbol{\theta} = (\sigma^2 \ \boldsymbol{\beta}')'$ and $\boldsymbol{\beta} = (\mathbf{b}_1' \ \mathbf{b}_2')'$, where $\mathbf{b}_1$ is $(k-s)\times 1$, and $\mathbf{b}_2$ is $s \times 1$. The null hypothesis is $\mathbf{b}_2^* = \mathbf{R}\boldsymbol{\theta}^* = \mathbf{0}$, where $\mathbf{R} = [\mathbf{0} \ \mathbf{R}_1]$ is $s \times (k+1)$ and $\mathbf{R}_1 = [\mathbf{0} \ \mathbf{I}_s]$ is $s \times k$.

From the saddle-point condition, $\nabla L_T(\ddot{\boldsymbol{\theta}}_T) = -\mathbf{R}'\ddot{\boldsymbol{\lambda}}_T$, and hence

$$\nabla L_T(\ddot{\boldsymbol{\theta}}_T) = \begin{bmatrix} \nabla_{\sigma^2} L_T(\ddot{\boldsymbol{\theta}}_T) \\ \nabla_{\mathbf{b}_1} L_T(\ddot{\boldsymbol{\theta}}_T) \\ \nabla_{\mathbf{b}_2} L_T(\ddot{\boldsymbol{\theta}}_T) \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{0} \\ -\ddot{\boldsymbol{\lambda}}_T \end{bmatrix}.$$

Thus, the LM test is mainly based on:

$$\nabla_{\mathbf{b}_2} L_T(\ddot{\boldsymbol{\theta}}_T) = \frac{1}{T\,\ddot{\sigma}_T^2} \sum_{t=1}^{T} \mathbf{x}_{2t}\ddot{\epsilon}_t = X_2'\ddot{\boldsymbol{\epsilon}}/(T\,\ddot{\sigma}_T^2),$$

where $\ddot{\sigma}_T^2 = \ddot{\boldsymbol{\epsilon}}'\ddot{\boldsymbol{\epsilon}}/T$, with $\ddot{\boldsymbol{\epsilon}}$ the vector of constrained residuals.

The LM test statistic now reads:

$$T \begin{bmatrix} 0 \\ \mathbf{0} \\ X_2'\ddot{\epsilon}/(T\,\ddot{\sigma}_T^2) \end{bmatrix}' \ddot{\mathbf{H}}_T^{-1}\mathbf{R}'(\mathbf{R}\ddot{\mathbf{C}}_T\mathbf{R}')^{-1}\mathbf{R}\ddot{\mathbf{H}}_T^{-1} \begin{bmatrix} 0 \\ \mathbf{0} \\ X_2'\ddot{\epsilon}/(T\,\ddot{\sigma}_T^2) \end{bmatrix},$$

which converges in distribution to $\chi^2(s)$ under the null.

When the information matrix equality holds,

$$\begin{aligned} \mathcal{LM}_T &= -T[\nabla L_T(\ddot{\boldsymbol{\theta}}_T)]'\ddot{\mathbf{H}}_T^{-1}[\nabla L_T(\ddot{\boldsymbol{\theta}}_T)] \\ &= T[\mathbf{0}'\ \ddot{\epsilon}'\mathbf{X}_2/T](\mathbf{X}'\mathbf{X}/T)^{-1}[\mathbf{0}'\ \ddot{\epsilon}'\mathbf{X}_2/T]'/\ddot{\sigma}_T^2 \\ &= T[\ddot{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ddot{\epsilon}/\ddot{\epsilon}'\ddot{\epsilon}] = TR^2, \end{aligned}$$

where R^2 is from the auxiliary regression of $\ddot{\epsilon}_t$ on $\mathbf{x}_{1t}$ and $\mathbf{x}_{2t}$.

Example (Breusch-Pagan Test):

Let $h\colon \mathbb{R} \to (0,\infty)$ be a differentiable function. Consider the specification:

$$y_t|\mathbf{x}_t, \boldsymbol{\zeta}_t \sim \mathcal{N}\big(\mathbf{x}_t'\boldsymbol{\beta}, h(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})\big),$$

where $\boldsymbol{\zeta}_t'\boldsymbol{\alpha} = \alpha_0 + \sum_{i=1}^{p} \zeta_{ti}\alpha_i$. Under this specification,

$$L_T(y^T, \mathbf{x}^T, \boldsymbol{\zeta}^T; \boldsymbol{\theta}) = -\frac{1}{2}\log(2\pi) - \frac{1}{2T}\sum_{t=1}^{T} \log\big(h(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})\big) - \frac{1}{T}\sum_{t=1}^{T} \frac{(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2}{2h(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})}.$$

The null hypothesis is $\alpha_1 = \cdots = \alpha_p = 0$ so that $h(\alpha_0) = \sigma_0^2$, i.e., conditional homoskedasticity.

For the LM test, the corresponding score vector is

$$\nabla_\alpha L_T(y^T, \mathbf{x}^T, \zeta^T; \boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{h_1(\zeta_t'\boldsymbol{\alpha})\zeta_t}{2h(\zeta_t'\boldsymbol{\alpha})}\left(\frac{(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2}{h(\zeta_t'\boldsymbol{\alpha})} - 1\right)\right],$$

where $h_1(\eta) = \mathrm{d}\, h(\eta)/\,\mathrm{d}\,\eta$. Under the null, $h_1(\zeta_t'\boldsymbol{\alpha}) = h_1(\alpha_0) =: c$.

The constrained specification is $y_t|\mathbf{x}_t, \zeta_t \sim \mathcal{N}(\mathbf{x}_t'\boldsymbol{\beta}, \sigma^2)$, and the constrained QMLEs are the OLS estimator $\ddot{\boldsymbol{\beta}}_T$ and $\ddot{\sigma}_T^2 = \sum_{t=1}^T \hat{e}_t^2/T$. The score vector evaluated at the constrained QMLEs is

$$\nabla_\alpha L_T(y_t, \mathbf{x}_t, \zeta_t; \ddot{\boldsymbol{\theta}}_T) = \frac{c}{T}\sum_{t=1}^{T}\left[\frac{\zeta_t}{2\ddot{\sigma}_T^2}\left(\frac{\hat{e}_t^2}{\ddot{\sigma}_T^2} - 1\right)\right].$$

It can be shown that the $(p+1)\times(p+1)$ block of the Hessian matrix corresponding to $\boldsymbol{\alpha}$ is

$$\frac{1}{T}\sum_{t=1}^{T}\left[\frac{-(y_t-\mathbf{x}_t'\boldsymbol{\beta})^2}{h^3(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})}+\frac{1}{2h^2(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})}\right][h_1(\boldsymbol{\zeta}'\boldsymbol{\alpha})]^2\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t'$$
$$+\left[\frac{(y_t-\mathbf{x}_t'\boldsymbol{\beta})^2}{2h^2(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})}-\frac{1}{2h(\boldsymbol{\zeta}_t'\boldsymbol{\alpha})}\right]h_2(\boldsymbol{\zeta}'\boldsymbol{\alpha})\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t',$$

where $h_2(\eta)=\mathrm{d}\,h_1(\eta)/\,\mathrm{d}\eta$. Evaluating the expectation of this block at $\boldsymbol{\theta}_o=(\boldsymbol{\beta}_o'\ \alpha_0\ \mathbf{0}')'$ and noting that $\sigma_o^2=h(\alpha_0)$, we have

$$-\left(\frac{c^2}{2[(\sigma_o^2)^2}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t')\right).$$

This block of the expected Hessian matrix can be consistently estimated by

$$-\left(\frac{c^2}{2(\ddot{\sigma}_T^2)^2}\right)\left(\frac{1}{T}\sum_{t=1}^{T}(\zeta_t\zeta_t')\right).$$

Setting $d_t = \hat{e}_t^2/\ddot{\sigma}_T^2 - 1$, the LM statistic under the information matrix equality is

$$\mathcal{LM}_T = \frac{1}{2}\left(\sum_{t=1}^{T} d_t\zeta_t'\right)\left(\sum_{t=1}^{T}\zeta_t\zeta_t'\right)^{-1}\left(\sum_{t=1}^{T}\zeta_t d_t\right) \xrightarrow{D} \chi^2(p),$$

where the numerator is the (centered) regression sum of squares (RSS) of regressing d_t on ζ_t.

Remarks:

- In the Breusch-Pagan test, the function h does not show up in the statistic. As such, this test is capable of testing general conditional heteroskedasticity without specifying a functional form of h.
- When $\boldsymbol{\zeta}_t$ contains the squares and all cross-product terms of the non-constant elements of $\mathbf{x}_t$: x_{it}^2 and $x_{it}x_{jt}$ (let there be n of such terms), the resulting Breusch-Pagan test is also the test of heteroskedasticity of unknown form due to White (1980) and has the limiting distribution $\chi^2(n)$.
- The Breusch-Pagan test is valid when the information matrix equality holds. Thus, the Breusch-Pagan test is not robust to dynamic misspecification, e.g., when the errors are serially correlated.

Koenker (1981): As $T^{-1}\sum_{t=1}^{T}\hat{e}_t^4 \xrightarrow{\mathbb{P}} 3(\sigma_o^2)^2$ under the null,

$$\frac{1}{T}\sum_{t=1}^{T} d_t^2 = \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{e}_t^4}{(\ddot{\sigma}_T^2)^2} - \frac{2}{T}\sum_{t=1}^{T}\frac{\hat{e}_t^2}{\ddot{\sigma}_T^2} + 1 \xrightarrow{\mathbb{P}} 2.$$

Thus, the test below is asymptotically equivalent to the original Breusch-Pagan test:

$$\mathcal{LM}_T = T\left(\sum_{t=1}^{T} d_t\zeta_t'\right)\left(\sum_{t=1}^{T}\zeta_t\zeta_t'\right)^{-1}\left(\sum_{t=1}^{T}\zeta_t d_t\right)\Big/\sum_{t=1}^{T} d_t^2,$$

which can be computed as TR^2, where R^2 is obtained from regressing d_t on ζ_t. As $\sum_{i=1}^{T} d_i = 0$, the centered and non-centered R^2 are equivalent. Thus, the Breusch-Pagan test can be computed as TR^2 from the regression of $\hat{e}_t^2$ on ζ_t. (Why?)

Example (Breusch-Godfrey Test):

Given the specification $y_t|\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t'\boldsymbol{\beta}, \sigma^2)$, we are interested in testing if $y_t - \mathbf{x}_t'\boldsymbol{\beta}$ are serially uncorrelated. For the AR(1) error:

$$y_t - \mathbf{x}_t'\boldsymbol{\beta} = \rho(y_{t-1} - \mathbf{x}_{t-1}'\boldsymbol{\beta}) + u_t,$$

with $|\rho| < 1$ and $\{u_t\}$ a white noise. The null hypothesis is $\rho^* = 0$.

Consider a general specification that admits serial correlations:

$$y_t|y_{t-1}, \mathbf{x}_t, \mathbf{x}_{t-1} \sim \mathcal{N}(\mathbf{x}_t'\boldsymbol{\beta} + \rho(y_{t-1} - \mathbf{x}_{t-1}'\boldsymbol{\beta}), \sigma_u^2).$$

Under the null, $y_t|\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t'\boldsymbol{\beta}, \sigma^2)$, and the constrained QMLE of $\boldsymbol{\beta}$ is the OLS estimator $\hat{\boldsymbol{\beta}}_T$. Testing the null hypothesis amounts to testing whether $y_{t-1} - \mathbf{x}_{t-1}'\boldsymbol{\beta}$ should be included in the mean specification.

When the information matrix equality holds, the LM test can be computed as TR^2, where R^2 is from the regression of $\hat{e}_t = y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}_T$ on $\mathbf{x}_t$ and $y_{t-1} - \mathbf{x}_{t-1}'\boldsymbol{\beta}$. Replacing $\boldsymbol{\beta}$ with its constrained estimator $\hat{\boldsymbol{\beta}}_T$, we can also obtain R^2 from the regression of $\hat{e}_t$ on $\mathbf{x}_t$ and $\hat{e}_{t-1}$. This test has the limiting $\chi^2(1)$ distribution under the null.

The Breusch-Godfrey test can be extended straightforwardly to check if the errors follow an AR(p) process. By regressing $\hat{e}_t$ on $\mathbf{x}_t$ and $\hat{e}_{t-1}, \ldots, \hat{e}_{t-p}$, the resulting TR^2 is the LM test when the information matrix equality holds and has a limiting $\chi^2(p)$ distribution.

Remark: If there is neglected conditional heteroskedasticity, the information matrix equality would fail, and the Breusch-Godfrey test no longer has a limiting χ^2 distribution.

If the specification is $y_t - \mathbf{x}_t'\boldsymbol{\beta} = u_t + \alpha u_{t-1}$, i.e., the errors follow an MA(1) process, we can write

$$y_t|\mathbf{x}_t, u_{t-1} \sim \mathcal{N}(\mathbf{x}_t'\boldsymbol{\beta} + \alpha u_{t-1}, \sigma_u^2).$$

The null hypothesis is $\alpha^* = 0$. Again, the constrained specification is the standard linear regression model $y_t = \mathbf{x}_t'\boldsymbol{\beta}$, and the constrained QMLE of $\boldsymbol{\beta}$ is still the OLS estimator $\hat{\boldsymbol{\beta}}_T$.

The LM test of $\alpha^* = 0$ can be computed as TR^2 with R^2 obtained from the regression of $\hat{u}_t = y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}_T$ on $\mathbf{x}_t$ and $\hat{u}_{t-1}$. This is identical to the LM test for AR(1) errors. Similarly, the Breusch-Godfrey test for MA(p) errors is also the same as that for AR(p) errors.

Likelihood Ratio Test

The likelihood ratio (LR) test compares the log-likelihoods of the constrained and unconstrained specifications:

$$\mathcal{LR}_T = -2T[L_T(\ddot{\theta}_T) - L_T(\tilde{\theta}_T)].$$

Recall from the discussion of the LM test:

$$\sqrt{T}(\ddot{\theta}_T - \theta^*) = \sqrt{T}(\tilde{\theta}_T - \theta^*) - \mathbf{H}_T(\theta^*)^{-1}\mathbf{R}'\sqrt{T}\ddot{\lambda}_T + o_{\mathbb{P}}(1),$$

and $\sqrt{T}\ddot{\lambda}_T = [\mathbf{R}\mathbf{H}_T(\theta^*)^{-1}\mathbf{R}']^{-1}\mathbf{R}\sqrt{T}(\tilde{\theta}_T - \theta^*) + o_{\mathbb{P}}(1)$, we have

$$\sqrt{T}(\tilde{\theta}_T - \ddot{\theta}_T) = \mathbf{H}_T(\theta^*)^{-1}\mathbf{R}'[\mathbf{R}\mathbf{H}_T(\theta^*)^{-1}\mathbf{R}']^{-1}\mathbf{R}\sqrt{T}(\tilde{\theta}_T - \theta^*) + o_{\mathbb{P}}(1).$$

By Taylor expansion of $L_T(\ddot{\boldsymbol{\theta}}_T)$ about $\tilde{\boldsymbol{\theta}}_T$, we have

$$\begin{aligned}
\mathcal{LR}_T &= -2T\left[L_T(\ddot{\boldsymbol{\theta}}_T) - L_T(\tilde{\boldsymbol{\theta}}_T)\right] \\
&= -2T\nabla L_T(\tilde{\boldsymbol{\theta}}_T)(\ddot{\boldsymbol{\theta}}_T - \tilde{\boldsymbol{\theta}}_T) \\
&\qquad - T(\ddot{\boldsymbol{\theta}}_T - \tilde{\boldsymbol{\theta}}_T)'\mathbf{H}_T(\tilde{\boldsymbol{\theta}}_T)(\ddot{\boldsymbol{\theta}}_T - \tilde{\boldsymbol{\theta}}_T) + o_{\mathbb{P}}(1) \\
&= -T(\ddot{\boldsymbol{\theta}}_T - \tilde{\boldsymbol{\theta}}_T)'\mathbf{H}_T(\boldsymbol{\theta}^*)(\ddot{\boldsymbol{\theta}}_T - \tilde{\boldsymbol{\theta}}_T) + o_{\mathbb{P}}(1),
\end{aligned}$$

because $\nabla L_T(\tilde{\boldsymbol{\theta}}_T) = \mathbf{0}$. It follows that

$$\mathcal{LR}_T = -T(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*)'\mathbf{R}'[\mathbf{R}\mathbf{H}_T(\boldsymbol{\theta}^*)^{-1}\mathbf{R}']^{-1}\mathbf{R}(\tilde{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*) + o_{\mathbb{P}}(1).$$

The first term on the RHS would have an asymptotic $\chi^2(q)$ distribution provided that the information matrix equality holds. (Why?)

Remarks:

- The 3 classical large sample tests check different aspects of the likelihood function. The Wald test checks if $\tilde{\boldsymbol{\theta}}_T$ is close to $\boldsymbol{\theta}^*$; the LM test checks if $\nabla LT(\ddot{\boldsymbol{\theta}}_T)$ is close to zero; the LR test checks if the constrained and unconstrained log-likelihood values are close to each other.
- The Wald and LM tests are based on unconstrained and constrained estimation results, respectively, but the LR test requires both.
- The Wald and LM test can be made robust to misspecification by employing a suitable consistent estimator of the asymptotic covariance matrix. Yet, the LR test is valid only when the information matrix equality holds.

Testing Non-Nested Models

Consider testing non-nested specifications:

$$H_0: \; y_t|\mathbf{x}_t, \boldsymbol{\xi}_t \sim f(y_t|\mathbf{x}_t; \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p,$$

$$H_1: \; y_t|\mathbf{x}_t, \boldsymbol{\xi}_t \sim \varphi(y_t|\boldsymbol{\xi}_t; \boldsymbol{\psi}), \quad \boldsymbol{\psi} \in \Psi \subseteq \mathbb{R}^q,$$

where $\mathbf{x}_t$ and $\boldsymbol{\xi}_t$ are two sets of variables. These specification are non-nested because they can not be derived from each other by imposing restrictions on the parameters.

The encompassing principle of Mizon (1984) and Mizon and Richard (1986) asserts that, if the model under the null is true, it should encompass the model under the alternative, such that a statistic of the alternative model should be close to its pseudo-true value, the probability limit evaluated under the null model.

Wald Encompassing Test

An encompassing test for non-nested hypotheses is based on the difference between a chosen statistic and the sample counterpart of its pseudo-true value. When the chosen statistic is the QMLE of the alternative model, the resulting test is the Wald encompassing test (WET).

Consider now the non-nested specifications of the conditional mean function:

$$H_0: \; y_t|\mathbf{x}_t, \boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{x}_t'\boldsymbol{\beta}, \sigma^2), \quad \boldsymbol{\beta} \in \mathcal{B} \subseteq \mathbb{R}^k,$$

$$H_1: \; y_t|\mathbf{x}_t, \boldsymbol{\xi}_t \sim \mathcal{N}(\boldsymbol{\xi}_t'\boldsymbol{\delta}, \sigma^2), \quad \boldsymbol{\delta} \in \mathcal{D} \subseteq \mathbb{R}^r,$$

where $\mathbf{x}_t$ and $\boldsymbol{\xi}_t$ do not have elements in common. Let $\hat{\boldsymbol{\beta}}_T$ and $\hat{\boldsymbol{\delta}}_T$ denote the QMLEs of the parameters in the null and alternative models.

Under the null: $y_t|\mathbf{x}_t, \boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{x}_t'\boldsymbol{\beta}_o, \sigma_o^2)$, $\mathbb{E}(\boldsymbol{\xi}_t y_t) = \mathbb{E}(\boldsymbol{\xi}_t \mathbf{x}_t')\boldsymbol{\beta}_o$, and hence

$$\hat{\boldsymbol{\delta}}_T = \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\xi}_t \boldsymbol{\xi}_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\xi}_t y_t\right) \xrightarrow{\mathbb{P}} \mathbf{M}_{\xi\xi}^{-1}\mathbf{M}_{\xi x}\boldsymbol{\beta}_o,$$

where

$$\mathbf{M}_{\xi\xi} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(\boldsymbol{\xi}_t \boldsymbol{\xi}_t'), \qquad \mathbf{M}_{\xi x} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(\boldsymbol{\xi}_t \mathbf{x}_t').$$

Clearly, the pseudo-true parameter $\boldsymbol{\delta}(\boldsymbol{\beta}_o) = \mathbf{M}_{\xi\xi}^{-1}\mathbf{M}_{\xi x}\boldsymbol{\beta}_o$ would not be the probability limit of $\hat{\boldsymbol{\delta}}_T$ if $\mathbf{x}_t'\boldsymbol{\beta}$ is an incorrect specification of the conditional mean. Thus, whether $\hat{\boldsymbol{\delta}}_T$ and the sample counterpart of $\boldsymbol{\delta}(\boldsymbol{\beta}_o)$ is sufficiently close to zero constitutes an evidence for or against the null hypothesis.

The WET is then based on:

$$\hat{\delta}_T - \hat{\delta}(\hat{\beta}_T) = \hat{\delta}_T - \left(\frac{1}{T}\sum_{t=1}^{T}\xi_t\xi_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\xi_t\mathbf{x}_t'\right)\hat{\beta}_T$$

$$= \left(\frac{1}{T}\sum_{t=1}^{T}\xi_t\xi_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\xi_t(y_t - \mathbf{x}_t'\hat{\beta}_T)\right),$$

which in effect checks if $\epsilon_t = y_t - \mathbf{x}_t'\beta$ and ξ_t are correlated. Letting $\hat{e}_t = y_t - \mathbf{x}_t'\hat{\beta}_T$, we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_t\hat{e}_t = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_t\epsilon_t - \left(\frac{1}{T}\sum_{t=1}^{T}\xi_t\mathbf{x}_t'\right)\sqrt{T}(\hat{\beta}_T - \beta_o),$$

which is

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_t\epsilon_t - \left(\frac{1}{T}\sum_{t=1}^{T}\xi_t\mathbf{x}_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{x}_t\epsilon_t\right).$$

Setting $\hat{\boldsymbol{\xi}}_t = \mathbf{M}_{\xi x}\mathbf{M}_{xx}^{-1}\mathbf{x}_t$, we can write

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{\xi}_t\hat{e}_t = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_t)\epsilon_t + o_{\mathbb{P}}(1).$$

Under the null hypothesis,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{\xi}_t\hat{e}_t \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_o),$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_o &= \sigma_o^2\left(\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_t)(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_t)'\right]\right)\\ &= \sigma_o^2\left(\mathbf{M}_{\xi\xi} - \mathbf{M}_{\xi x}\mathbf{M}_{xx}^{-1}\mathbf{M}_{x\xi}\right).\end{aligned}$$

Consequently,

$$T^{1/2}[\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)] \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \mathbf{M}_{\xi\xi}^{-1}\boldsymbol{\Sigma}_o\mathbf{M}_{\xi\xi}^{-1}\right),$$

and hence

$$T\left[\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)\right]'\mathbf{M}_{\xi\xi}\boldsymbol{\Sigma}_o^{-1}\mathbf{M}_{\xi\xi}\left[\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)\right] \xrightarrow{D} \chi^2(r).$$

A consistent estimator for $\boldsymbol{\Sigma}_o$ is

$$\widehat{\boldsymbol{\Sigma}}_T = \hat{\sigma}_T^2 \left[\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\boldsymbol{\xi}_t'\right) - \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\mathbf{x}_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\boldsymbol{\xi}_t'\right)\right].$$

The WET statistic reads

$$\begin{aligned}\mathcal{WE}_T = T\left[\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)\right]' & \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\boldsymbol{\xi}_t'\right)\widehat{\boldsymbol{\Sigma}}_T^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\boldsymbol{\xi}_t'\right)\left[\hat{\boldsymbol{\delta}}_T - \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)\right] \\ & \xrightarrow{D} \chi^2(r).\end{aligned}$$

When $\mathbf{x}_t$ and $\boldsymbol{\xi}_t$ have s ($s < r$) elements in common, $\sum_{t=1}^{T}\boldsymbol{\xi}_t\hat{e}_t$ have s elements that are identically zero. Thus, $\text{rank}(\boldsymbol{\Sigma}_o) = r^* \leq r - s$, and the WET should be computed with $\widehat{\boldsymbol{\Sigma}}_T^{-1}$ replaced by $\widehat{\boldsymbol{\Sigma}}_T^{-}$, the generalized inverse $\widehat{\boldsymbol{\Sigma}}_T$.

Score Encompassing Test

Under H_1: $y_t|\mathbf{x}_t, \boldsymbol{\xi}_t \sim \mathcal{N}(\boldsymbol{\xi}_t'\boldsymbol{\delta}, \sigma^2)$, the score function evaluated at the pseudo-true parameter $\boldsymbol{\delta}(\boldsymbol{\beta}_o)$ is (apart from a constant σ^{-2})

$$\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t[y_t - \boldsymbol{\xi}_t'\boldsymbol{\delta}(\boldsymbol{\beta}_o)] = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\left[y_t - \boldsymbol{\xi}_t'\left(\mathbf{M}_{\xi\xi}^{-1}\mathbf{M}_{\xi x}\boldsymbol{\beta}_o\right)\right].$$

When the pseudo-true parameter is replaced by its estimator $\hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\beta}}_T)$, the score function becomes

$$\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\left[y_t - \boldsymbol{\xi}_t'\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\boldsymbol{\xi}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\mathbf{x}_t'\right)\hat{\boldsymbol{\beta}}_T\right]$$

$$= \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\hat{e}_t.$$

The score encompassing test (SET) is based on

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \boldsymbol{\xi}_t \hat{e}_t \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_o),$$

and the SET statistic is

$$\frac{1}{T} \left(\sum_{t=1}^{T} \hat{e}_t \boldsymbol{\xi}_t' \right) \widehat{\boldsymbol{\Sigma}}_T^{-1} \left(\sum_{t=1}^{T} \boldsymbol{\xi}_t \hat{e}_t \right) \xrightarrow{D} \chi^2(r).$$

- The WET and SET are based on the same ingredient and both require evaluating the pseudo-true parameter.
- The WET and SET are difficult to implement when the pseudo-true parameter is not readily derived; this may happen when, e.g., the QMLE does not have an analytic form. (Examples?)

Pseudo-True Score Encompassing Test

Chen and Kuan (2002) propose the pseudo-true score encompassing (PSE) test which is based on the pseudo-true value of the score function under the alternative. Although it may be difficult to evaluate the pseudo-true value of a QMLE when it does not have a closed form, it would be easier to evaluate the pseudo-true score because the analytic from of the score function is usually available.

The null and alternative hypotheses are:

$$H_0: \; y_t|\mathbf{x}_t, \boldsymbol{\xi}_t \sim f(\mathbf{x}_t; \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p,$$

$$H_1: \; y_t|\mathbf{x}_t, \boldsymbol{\xi}_t \sim \varphi(\boldsymbol{\xi}_t, \boldsymbol{\psi}), \quad \boldsymbol{\psi} \in \Psi \subseteq \mathbb{R}^q.$$

Let $s_{f,t}(\boldsymbol{\theta}) = \nabla \log f(y_t|\mathbf{x}_t;\boldsymbol{\theta})$, $s_{\varphi,t}(\boldsymbol{\psi}) = \nabla \log \varphi(y_t|\boldsymbol{\xi}_t;\boldsymbol{\psi})$, and

$$\nabla L_{f,T}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} s_{f,t}(\boldsymbol{\theta}), \qquad \nabla L_{\varphi,T}(\boldsymbol{\psi}) = \frac{1}{T}\sum_{t=1}^{T} s_{\varphi,t}(\boldsymbol{\psi}).$$

The pseudo-true score function of $\nabla L_{\varphi,T}(\boldsymbol{\psi})$ is

$$J_\varphi(\boldsymbol{\theta},\boldsymbol{\psi}) := \lim_{T\to\infty} \mathbb{E}_{f(\boldsymbol{\theta})}\left[\nabla L_{\varphi,T}(\boldsymbol{\psi})\right].$$

As the pseudo-true parameter $\boldsymbol{\psi}(\boldsymbol{\theta}_o)$ is the KLIC minimizer when the null hypothesis is specified correctly, we have

$$J_\varphi(\boldsymbol{\theta}_o, \boldsymbol{\psi}(\boldsymbol{\theta}_o)) = \mathbf{0}.$$

Thus, whether $J_\varphi(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\psi}}_T)$ is close to zero constitutes an evidence for or against the null hypothesis.

Following Wooldridge (1990), we can incorporate the null model into the the score function and write

$$\nabla L_{\varphi,T}(\boldsymbol{\theta},\boldsymbol{\psi}) = \frac{1}{T}\sum_{t=1}^{T} d_{1,t}(\boldsymbol{\theta},\boldsymbol{\psi}) + \frac{1}{T}\sum_{t=1}^{T} d_{2,t}(\boldsymbol{\theta},\boldsymbol{\psi}) c_t(\boldsymbol{\theta}),$$

where $\mathbb{E}_{f(\boldsymbol{\theta})}[c_t(\boldsymbol{\theta})|\mathbf{x}_t,\boldsymbol{\xi}_t] = \mathbf{0}$. As such,

$$J_{\varphi}(\boldsymbol{\theta},\boldsymbol{\psi}) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{f(\boldsymbol{\theta})}\left[d_{1,t}(\boldsymbol{\theta},\boldsymbol{\psi})\right],$$

and its sample counterpart is

$$\widehat{J}_{\varphi}(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T) = \frac{1}{T}\sum_{t=1}^{T} d_{1,t}(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T) = -\frac{1}{T}\sum_{t=1}^{T} d_{2,t}(\hat{\boldsymbol{\theta}}_T,\hat{\boldsymbol{\psi}}_T) c_t(\hat{\boldsymbol{\theta}}_T),$$

where the second equality follows because $\nabla L_{\varphi,T}(\hat{\boldsymbol{\psi}}_T) = \mathbf{0}$.

For non-nested, linear specifications for the conditional mean, we can incorporate the null model ($y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t$) into the score and obtain

$$\nabla_{\boldsymbol{\delta}} L_{\varphi,T}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\xi}_t(\mathbf{x}_t'\boldsymbol{\beta} - \boldsymbol{\xi}_t'\boldsymbol{\delta}) + \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\xi}_t\varepsilon_t$$

with $d_{1,t}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \boldsymbol{\xi}_t(\mathbf{x}_t'\boldsymbol{\beta} - \boldsymbol{\xi}_t'\boldsymbol{\delta})$, $d_{2,t}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \boldsymbol{\xi}_t$, and $c_t(\boldsymbol{\theta}) = \varepsilon_t$. Then,

$$\widehat{J}_{\varphi}(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\psi}}_T) = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\xi}_t(\mathbf{x}_t'\hat{\boldsymbol{\beta}}_T - \boldsymbol{\xi}_t'\hat{\boldsymbol{\delta}}_T) = -\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\xi}_t\hat{e}_t,$$

as in the SET discussed earlier.

For nonlinear specifications, it can be seen that

$$\widehat{J}_\varphi(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\psi}}_T) = \frac{1}{T}\sum_{t=1}^{T} \nabla_{\boldsymbol{\delta}}\mu(\boldsymbol{\xi}_t, \hat{\boldsymbol{\delta}}_T)\left[m(\mathbf{x}_t, \hat{\boldsymbol{\beta}}_T) - \mu(\boldsymbol{\xi}_t, \hat{\boldsymbol{\delta}}_T)\right]$$

$$= -\frac{1}{T}\sum_{t=1}^{T} \nabla_{\boldsymbol{\delta}}\mu(\boldsymbol{\xi}_t, \hat{\boldsymbol{\delta}}_T)\hat{e}_t,$$

with $\hat{e}_t = y_t - m(\mathbf{x}_t, \hat{\boldsymbol{\beta}}_T)$ the nonlinear OLS residuals. Yet, it is not easy to compute the SET here because evaluating the pseudo-true value of $\hat{\boldsymbol{\delta}}_T$ is a formidable task.

The linear expansion of $T^{1/2}\widehat{J}_\varphi(\hat{\boldsymbol{\theta}}_T, \hat{\psi}_T)$ about $(\boldsymbol{\theta}_o, \psi(\boldsymbol{\theta}_o))$ is

$$\sqrt{T}\widehat{J}_\varphi(\hat{\boldsymbol{\theta}}_T, \hat{\psi}_T) \approx -\frac{1}{\sqrt{T}}\sum_{t=1}^{T} d_{2,t}(\boldsymbol{\theta}_o, \psi(\boldsymbol{\theta}_o))c_t(\boldsymbol{\theta}_o) - \mathbf{A}_o\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_o),$$

where $\mathbf{A}_o = \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} \mathbb{E}_{f(\boldsymbol{\theta}_o)}\left[d_{2,t}(\boldsymbol{\theta}_o, \psi(\boldsymbol{\theta}_o))\nabla_{\boldsymbol{\theta}} c_t(\boldsymbol{\theta}_o)\right]$. Note that the other terms in the expansion that involve c_t would vanish in the limit because they have zero mean. Recall also that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_o) = -H_T(\boldsymbol{\theta}_o)^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \mathbf{s}_{f,t}(\boldsymbol{\theta}_o) + o_{\mathbb{P}}(1),$$

where $\mathbb{E}_{f(\boldsymbol{\theta}_o)}[s_{f,t}(\boldsymbol{\theta}_o)|\mathbf{x}_t, \boldsymbol{\xi}_t] = \mathbf{0}$.

Collecting terms we have

$$\sqrt{T}\widehat{J}_\varphi(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\psi}}_T) = -\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{b}_t(\boldsymbol{\theta}_o, \boldsymbol{\psi}(\boldsymbol{\theta}_o)) + o_{\mathbb{P}}(1),$$

where $\mathbf{b}_t(\boldsymbol{\theta}_o, \boldsymbol{\psi}(\boldsymbol{\theta}_o)) = d_{2,t}(\boldsymbol{\theta}_o, \boldsymbol{\psi}(\boldsymbol{\theta}_o))c_t(\boldsymbol{\theta}_o) - \mathbf{A}_o\mathbf{H}_T(\boldsymbol{\theta}_o)^{-1}\mathbf{s}_{f,t}(\boldsymbol{\theta}_o)$ and

$$\mathbb{E}_{f(\boldsymbol{\theta}_o)}[\mathbf{b}_t(\boldsymbol{\theta}_o, \boldsymbol{\psi}(\boldsymbol{\theta}_o))|\mathbf{x}_t, \boldsymbol{\xi}_t] = \mathbf{0}.$$

By invoking a suitable CLT, $T^{1/2}\widehat{J}_\varphi(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\psi}}_T)$ has a limiting normal distribution with the asymptotic covariance matrix:

$$\boldsymbol{\Sigma}_o = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{f(\boldsymbol{\theta}_o)}\left[\mathbf{b}_t(\boldsymbol{\theta}_o, \boldsymbol{\psi}(\boldsymbol{\theta}_o))\mathbf{b}_t(\boldsymbol{\theta}_o, \boldsymbol{\psi}(\boldsymbol{\theta}_o))'\right],$$

which can be consistently estimated by

$$\widehat{\boldsymbol{\Sigma}}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{b}_t(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\psi}}_T)\mathbf{b}_t(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\psi}}_T)'.$$

It follows that the PSE test is

$$\mathcal{PSE}_T = T\widehat{J}_\varphi(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\psi}}_T)'\widehat{\boldsymbol{\Sigma}}_T^- \widehat{J}_\varphi(\hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\psi}}_T) \xrightarrow{D} \chi^2(k),$$

where k is the rank of $\widehat{\boldsymbol{\Sigma}}_T$ and $\widehat{\boldsymbol{\Sigma}}_T^-$ is the generalized inverse of $\widehat{\boldsymbol{\Sigma}}_T$. The PSE test can be understood as an extension of the conditional mean encompassing test of Wooldridge (1990).

Example: Consider non-nested specifications of conditional variance:

$$H_0: \; y_t|\mathbf{x}_t, \boldsymbol{\xi}_t \sim \mathcal{N}\big(0, h(\mathbf{x}_t, \boldsymbol{\alpha})\big),$$

$$H_1: \; y_t|\mathbf{x}_t, \boldsymbol{\xi}_t \sim \mathcal{N}\big(0, \kappa(\boldsymbol{\xi}_t, \boldsymbol{\gamma})\big).$$

When h_t and κ_t are evaluated at the respective QMLEs $\tilde{\boldsymbol{\alpha}}_T$ and $\tilde{\boldsymbol{\gamma}}_T$, we write $\hat{h}_t$ and $\hat{\kappa}_t$. It can be verified that

$$\mathbf{s}_{h,t}(\boldsymbol{\alpha}) = \frac{\nabla_{\boldsymbol{\alpha}} h_t}{2h_t^2}(y_t^2 - h_t),$$

$$\mathbf{s}_{\kappa,t}(\boldsymbol{\gamma}) = \frac{\nabla_{\boldsymbol{\gamma}} \kappa_t}{2\kappa_t^2}(y_t^2 - \kappa_t) = \underbrace{\frac{\nabla_{\boldsymbol{\gamma}} \kappa_t}{2\kappa_t^2}(h_t - \kappa_t)}_{d_{1,t}} + \underbrace{\frac{\nabla_{\boldsymbol{\gamma}} \kappa_t}{2\kappa_t^2}}_{d_{2,t}} \underbrace{(y_t^2 - h_t)}_{c_t},$$

where $\mathbb{E}_{f(\boldsymbol{\theta})}(y_t^2 - h_t|\mathbf{x}_t, \boldsymbol{\xi}_t) = 0$.

The sample counterpart of the pseudo-true score function is thus

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\nabla_{\gamma}\hat{\kappa}_t}{2\hat{\kappa}_t^2}(y_t^2-\hat{h}_t).$$

Thus, the PSE test amounts to checking whether $\nabla_{\gamma}\kappa_t/(2\kappa_t^2)$ are correlated with the "generalized" errors $(y_t^2-h_t)$.

Binary Choice Models

Consider the binary dependent variable:

$$y_t = \begin{cases} 1, & \text{with probability } \mathbb{P}(y_t = 1|\mathbf{x}_t), \\ 0, & \text{with probability } 1 - \mathbb{P}(y_t = 1|\mathbf{x}_t). \end{cases}$$

The density function of y_t given $\mathbf{x}_t$ is:

$$g(y_t|\mathbf{x}_t) = \mathbb{P}(y_t = 1|\mathbf{x}_t)^{y_t}[1 - \mathbb{P}(y_t = 1|\mathbf{x}_t)]^{1-y_t}.$$

Approximating $\mathbb{P}(y_t = 1|\mathbf{x}_t)$ by $F(\mathbf{x}_t; \boldsymbol{\theta})$, the quasi-likelihood function is

$$f(y_t|\mathbf{x}_t; \boldsymbol{\theta}) = F(\mathbf{x}_t; \boldsymbol{\theta})^{y_t}[1 - F(\mathbf{x}_t; \boldsymbol{\theta})]^{1-y_t}.$$

The QMLE $\tilde{\boldsymbol{\theta}}_T$ is obtained by maximizing

$$L_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\big[y_t \log F(\mathbf{x}_t; \boldsymbol{\theta}) + (1 - y_t)\log(1 - F(\mathbf{x}_t; \boldsymbol{\theta}))\big].$$

- Probit model:

$$F(\mathbf{x}_t;\boldsymbol{\theta}) = \Phi(\mathbf{x}_t'\boldsymbol{\theta}) = \int_{-\infty}^{\mathbf{x}_t'\boldsymbol{\theta}} \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\,\mathrm{d}v,$$

  where Φ denotes the standard normal distribution function.

- Logit model:

$$F(\mathbf{x}_t;\boldsymbol{\theta}) = G(\mathbf{x}_t'\boldsymbol{\theta}) = \frac{1}{1+\exp(-\mathbf{x}_t'\boldsymbol{\theta})} = \frac{\exp(\mathbf{x}_t'\boldsymbol{\theta})}{1+\exp(\mathbf{x}_t'\boldsymbol{\theta})},$$

  where G is the logistic distribution function with mean zero and variance $\pi^2/3$. Note that the logistic distribution is more peaked around its mean and has slightly thicker tails than the standard normal distribution.

Global Concavity

The log-likelihood function L_T is globally concave provided that $\nabla^2_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta})$ is negative definite for $\boldsymbol{\theta} \in \Theta$. By a 2nd-order Taylor expansion,

$$\begin{aligned} L_T(\boldsymbol{\theta}) &= L_T(\tilde{\boldsymbol{\theta}}_T) + \nabla_{\boldsymbol{\theta}} L_T(\tilde{\boldsymbol{\theta}}_T)(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_T) \\ &\qquad + (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_T)' \nabla^2_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}^\dagger)(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_T) \\ &= L_T(\tilde{\boldsymbol{\theta}}_T) + (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_T)' \nabla^2_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}^\dagger)(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_T), \end{aligned}$$

where $\boldsymbol{\theta}^\dagger$ is between $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}_T$. Global concavity implies that $L_T(\boldsymbol{\theta})$ must be less than $L_T(\tilde{\boldsymbol{\theta}}_T)$ for any $\boldsymbol{\theta} \in \Theta$. Thus, $\tilde{\boldsymbol{\theta}}_T$ must be a global maximizer.

For the logit model, as $G'(u) = G(u)[1 - G(u)]$, we have

$$\begin{aligned}\nabla_\theta L_T(\theta) &= \frac{1}{T}\sum_{t=1}^{T}\left[y_t\frac{G'(\mathbf{x}_t'\theta)}{G(\mathbf{x}_t'\theta)} - (1-y_t)\frac{G'(\mathbf{x}_t'\theta)}{1-G(\mathbf{x}_t'\theta)}\right]\mathbf{x}_t \\ &= \frac{1}{T}\sum_{t=1}^{T}\left\{y_t[1-G(\mathbf{x}_t'\theta)] - (1-y_t)G(\mathbf{x}_t'\theta)\right\}\mathbf{x}_t \\ &= \frac{1}{T}\sum_{t=1}^{T}[y_t - G(\mathbf{x}_t'\theta)]\mathbf{x}_t,\end{aligned}$$

from which we can solve for the QMLE $\tilde{\theta}_T$. Note that

$$\nabla_\theta^2 L_T(\theta) = -\frac{1}{T}\sum_{t=1}^{T} G(\mathbf{x}_t'\theta)[1-G(\mathbf{x}_t'\theta)]\mathbf{x}_t\mathbf{x}_t',$$

which is negative definite, so that L_T is globally concave in θ.

For the probit model,

$$\nabla_\theta L_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left[y_t\frac{\phi(\mathbf{x}_t'\theta)}{\Phi(\mathbf{x}_t'\theta)} - (1-y_t)\frac{\phi(\mathbf{x}_t'\theta)}{1-\Phi(\mathbf{x}_t'\theta)}\right]\mathbf{x}_t$$

$$= \frac{1}{T}\sum_{t=1}^{T}\frac{y_t - \Phi(\mathbf{x}_t'\theta)}{\Phi(\mathbf{x}_t'\theta)[1-\Phi(\mathbf{x}_t'\theta)]}\phi(\mathbf{x}_t'\theta)\mathbf{x}_t,$$

where ϕ is the standard normal density function. It can be verified that

$$\nabla_\theta^2 L_T(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\left[y_t\frac{\phi(\mathbf{x}_t'\theta) + \mathbf{x}_t'\theta\Phi(\mathbf{x}_t'\theta)}{\Phi^2(\mathbf{x}_t'\theta)}\right.$$

$$\left. + (1-y_t)\,\frac{\phi(\mathbf{x}_t'\theta) - \mathbf{x}_t'\theta[1-\Phi(\mathbf{x}_t'\theta)]}{[1-\Phi(\mathbf{x}_t'\theta)]^2}\right]\phi(\mathbf{x}_t'\theta)\mathbf{x}_t\mathbf{x}_t',$$

which is also negative definite; see e.g., Amemiya (1985, pp. 273–274).

As $\mathbb{E}(y_t \mid \mathbf{x}_t) = \mathbb{P}(y_t = 1 \mid \mathbf{x}_t)$, we can write

$$y_t = F(\mathbf{x}_t; \boldsymbol{\theta}) + e_t.$$

Note that when F is correctly specified for the conditional mean, y_t is conditionally heteroskedastic with

$$\text{var}(y_t|\mathbf{x}_t) = \mathbb{P}(y_t = 1|\mathbf{x}_t)[1 - \mathbb{P}(y_t = 1|\mathbf{x}_t)].$$

Thus, the probit and logit models are also different nonlinear mean specifications with conditional heteroskedasticity.

- The NLS estimator of $\boldsymbol{\theta}$ is inefficient because the NLS objective function ignores conditional heteroskedasticity.
- A weighted NLS estimator that takes into account the conditional variance is still inefficient. (Why?)

As $\mathbb{E}(y_t \mid \mathbf{x}_t) = \mathbb{P}(y_t = 1 \mid \mathbf{x}_t)$, we can write

$$y_t = F(\mathbf{x}_t; \boldsymbol{\theta}) + e_t.$$

Note that when F is correctly specified for the conditional mean, y_t is conditionally heteroskedastic with

$$\text{var}(y_t|\mathbf{x}_t) = \mathbb{P}(y_t = 1|\mathbf{x}_t)[1 - \mathbb{P}(y_t = 1|\mathbf{x}_t)].$$

Thus, the probit and logit models are also different nonlinear mean specifications with conditional heteroskedasticity.

- The NLS estimator of $\boldsymbol{\theta}$ is inefficient because the NLS objective function ignores conditional heteroskedasticity.
- A weighted NLS estimator that takes into account the conditional variance is still inefficient. (Why?)

- Marginal response: For the probit model,

$$\frac{\partial \Phi(\mathbf{x}_t'\boldsymbol{\theta})}{\partial x_{tj}} = \phi(\mathbf{x}_t'\boldsymbol{\theta})\theta_j;$$

for the logit model,

$$\frac{\partial G(\mathbf{x}_t'\boldsymbol{\theta})}{\partial x_{tj}} = G(\mathbf{x}_t'\boldsymbol{\theta})[1 - G(\mathbf{x}_t'\boldsymbol{\theta})]\theta_j.$$

These marginal effects all depend on $\mathbf{x}_t$.

- It is typical to evaluate the marginal response based on a particular value of $\mathbf{x}_t$, such as $\mathbf{x}_t = \mathbf{0}$ or $\mathbf{x}_t = \bar{\mathbf{x}}$, the sample average of $\mathbf{x}_t$.
- When $\mathbf{x}_t = \mathbf{0}$, $\phi(0) \approx 0.4$ and $G(0)[1 - G(0)] = 0.25$. This suggests that the QMLE for the logit model is approximately 1.6 times the QMLE for the probit model when $\mathbf{x}_t$ are close to zero.

- Letting $p = \mathbb{P}(y = 1 \mid \mathbf{x})$, $p/(1-p)$ is the odds ratio: the probability of $y = 1$ relative to the probability of $y = 0$.
- For the logit model, the log-odds-ratio is

$$\ln\left(\frac{p_t}{1-p_t}\right) = \ln(\exp(\mathbf{x}_t'\boldsymbol{\theta})) = \mathbf{x}_t'\boldsymbol{\theta}.$$

so that

$$\frac{\partial \ln(p_t/(1-p_t))}{\partial x_{tj}} = \theta_j.$$

As the effect of a regressor on the log-odds-ratio, each coefficient in the logit model is also understood as a semi-elasticity.

Measure of Goodness of Fit

- Digression: Given the objective function Q_T, let $Q_T(c)$ denote the value of Q_T when the model contains only a constant term, Q_T^* the largest possible value of Q_T (if exists), and $Q_T(\hat{\boldsymbol{\theta}}_T)$ the value of Q_T for the fitted model. Then, R^2 of relative gain is:

$$R^2_{\text{RG}} = \frac{Q_T(\hat{\boldsymbol{\theta}}_T) - Q_T(c)}{Q_T^* - Q_T(c)} = 1 - \frac{Q_T^* - Q_T(\hat{\boldsymbol{\theta}}_T)}{Q_T^* - Q_T(c)}.$$

- For binary choice models, $Q_T = L_T$ and the largest possible $L_T^* = 0$ when $\mathbb{P}(y_t = 1|\mathbf{x}_t) = 1$. The measure of McFadden (1974) is

$$R^2_{\text{RG}} = 1 - \frac{L_T(\hat{\boldsymbol{\theta}}_T)}{L_T(\bar{y})} = 1 - \frac{\sum_{t=1}^T \left[y_t \ln \hat{p}_t + (1-y_t)\ln(1-\hat{p}_t)\right]}{T\left[\bar{y}\ln\bar{y} + (1-\bar{y})\ln(1-\bar{y})\right]},$$

where $\hat{p}_t = G(\mathbf{x}_t'\hat{\boldsymbol{\theta}}_T)$ or $\hat{p}_t = \Phi(\mathbf{x}_t'\hat{\boldsymbol{\theta}}_T)$.

Latent Index Model

Assume that y_t is determined by the latent index variable y_t^*:

$$y_t = \begin{cases} 1, & y_t^* > 0, \\ 0, & y_t^* \leq 0, \end{cases}$$

where $y_t^* = \mathbf{x}_t'\boldsymbol{\beta} + e_t$. Thus,

$$\mathbb{P}(y_t = 1|\mathbf{x}_t) = \mathbb{P}(y_t^* > 0|\mathbf{x}_t) = \mathbb{P}(e_t > -\mathbf{x}_t\boldsymbol{\beta}|\mathbf{x}_t),$$

which is also $\mathbb{P}(e_t < \mathbf{x}_t\boldsymbol{\beta}|\mathbf{x}_t)$ provided that e_t is symmetric about zero. The probit specification Φ or the logit specification G can be understood as specifications of the conditional distribution of e_t.

Multinomial Models

Consider $J+1$ mutually exclusive choices that do not have a natural ordering, e.g., employment status and commuting mode. The dependent variable y_t takes on $J+1$ values such that

$$y_t = \begin{cases} 0, & \text{with probability } \mathbb{P}(y_t = 0|\mathbf{x}_t), \\ 1, & \text{with probability } \mathbb{P}(y_t = 1|\mathbf{x}_t), \\ \vdots & \\ J, & \text{with probability } \mathbb{P}(y_t = J|\mathbf{x}_t). \end{cases}$$

Define the new binary variable $d_{t,j}$ for $j = 0, 1, \ldots, J$ as

$$d_{t,j} = \begin{cases} 1, & \text{if } y_t = j, \\ 0, & \text{otherwise,} \end{cases}$$

note that $\sum_{j=0}^{J} d_{t,j} = 1$.

The density function of $d_{t,0}, \ldots, d_{t,J}$ given $\mathbf{x}_t$ is then

$$g(d_{t,0}, \ldots, d_{t,J} | \mathbf{x}_t) = \prod_{j=0}^{J} \mathbb{P}(y_t = j | \mathbf{x}_t)^{d_{t,j}}.$$

Approximating $\mathbb{P}(y_t = j | \mathbf{x}_t)$ by $F_j(\mathbf{x}_t; \boldsymbol{\theta})$ we obtain the quasi-log-likelihood function:

$$L_T(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=0}^{J} d_{t,j} \ln F_j(\mathbf{x}_t; \boldsymbol{\theta}).$$

The first order condition is

$$\nabla_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=0}^{J} d_{t,j} \frac{1}{F_j(\mathbf{x}_t; \boldsymbol{\theta})} \left[\nabla_{\boldsymbol{\theta}} F_j(\mathbf{x}_t; \boldsymbol{\theta}) \right] = \mathbf{0},$$

from which we can solve for the QMLE $\tilde{\boldsymbol{\theta}}_T$.

Multinomial Logit Model

Common specifications of the conditional probabilities are:

$$F_j(\mathbf{x}_t;\boldsymbol{\theta}) = G_{t,j} = \frac{\exp(\mathbf{x}_t'\boldsymbol{\theta}_j)}{\sum_{k=0}^{J}\exp(\mathbf{x}_t'\boldsymbol{\theta}_k)}, \quad j = 0,\ldots,J,$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_0' \ \boldsymbol{\theta}_1' \ \ldots \ \boldsymbol{\theta}_j')'$, and $\mathbf{x}_t$ does not depend on choices. Note, however, that the parameters are not identified because, for example,

$$\begin{aligned} &\frac{\exp[\mathbf{x}_t'(\boldsymbol{\theta}_0+\boldsymbol{\gamma})]}{\exp[\mathbf{x}_t'(\boldsymbol{\theta}_0+\boldsymbol{\gamma})]+\sum_{k=1}^{J}\exp(\mathbf{x}_t'\boldsymbol{\theta}_k)} \\ = &\frac{\exp(\mathbf{x}_t'\boldsymbol{\theta}_0)}{\exp(\mathbf{x}_t'\boldsymbol{\theta}_0)+\sum_{k=1}^{J}\exp[\mathbf{x}_t'(\boldsymbol{\theta}_k-\boldsymbol{\gamma})]}. \end{aligned}$$

For normalization, we set $\boldsymbol{\theta}_0 = \mathbf{0}$, so that

$$F_0(\mathbf{x}_t; \boldsymbol{\theta}) = G_{t,0} = \frac{1}{1 + \sum_{k=1}^{J} \exp(\mathbf{x}_t' \boldsymbol{\theta}_k)},$$

$$F_j(\mathbf{x}_t; \boldsymbol{\theta}) = G_{t,j} = \frac{\exp(\mathbf{x}_t' \boldsymbol{\theta}_j)}{1 + \sum_{k=1}^{J} \exp(\mathbf{x}_t' \boldsymbol{\theta}_k)}, \quad j = 1, \ldots, J,$$

with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1' \ \boldsymbol{\theta}_2' \ \ldots \ \boldsymbol{\theta}_J')'$. This leads to the multinomial logit model, and the quasi-log-likelihood function is

$$L_T(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{J} d_{t,j} \mathbf{x}_t' \boldsymbol{\theta}_j - \frac{1}{T} \sum_{t=1}^{T} \log \left(1 + \sum_{k=1}^{J} \exp(\mathbf{x}_t' \boldsymbol{\theta}_k) \right).$$

It is easy to see that $\nabla_{\boldsymbol{\theta}_j} G_{t,k} = -G_{t,k} G_{t,j} \mathbf{x}_t$ for $k \neq j$, and

$$\nabla_{\boldsymbol{\theta}_j} G_{t,j} = G_{t,j}\big[1 - G_{t,j}\big]\mathbf{x}_t.$$

It follows that

$$\nabla_{\boldsymbol{\theta}_j} L_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}(d_{t,j} - G_{t,j})\mathbf{x}_t, \quad j = 1, \ldots, J.$$

The Hessian matrix contains:

$$\nabla_{\boldsymbol{\theta}_j \boldsymbol{\theta}_i'} L_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}(G_{t,j} G_{t,i})\mathbf{x}_t \mathbf{x}_t', \quad i \neq j,\ i, j = 1, \ldots, J,$$

$$\nabla_{\boldsymbol{\theta}_j \boldsymbol{\theta}_j'} L_T(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T} G_{t,j}(1 - G_{t,j})\mathbf{x}_t \mathbf{x}_t', \quad j = 1, \ldots, J.$$

The marginal response of $G_{t,j}$ to the change of $\mathbf{x}_t$ are

$$\nabla_{\mathbf{x}_t} G_{t,0} = -G_{t,0} \sum_{i=1}^{J} G_{t,i} \boldsymbol{\theta}_i,$$

$$\nabla_{\mathbf{x}_t} G_{t,j} = G_{t,j} \left(\boldsymbol{\theta}_j - \sum_{i=1}^{J} G_{t,i} \boldsymbol{\theta}_i \right), \quad j = 1, \ldots, J.$$

Thus, all coefficient vectors $\boldsymbol{\theta}_i$, $i = 1, \ldots, J$, enter $\nabla_{\mathbf{x}_t} G_{t,j}$. The log-odds ratios (relative to the base choice $j = 0$) are:

$$\ln(G_{t,j}/G_{t,0}) = \mathbf{x}_t'(\boldsymbol{\theta}_j - \boldsymbol{\theta}_0) = \mathbf{x}_t'\boldsymbol{\theta}_j, \quad j = 1, \ldots, J,$$

as $\boldsymbol{\theta}_0 = 0$ in $G_{t,0}$ by construction. Thus, each coefficient $\theta_{j,k}$ is also understood as the effect of x_{tk} on the log-odds-ratio $\ln(G_{t,j}/G_{t,0})$.

Conditional Logit Model

In the conditional logit model, there are multiple choices with choice-dependent or alternative-varying regressors. For example, in choosing among several commuting modes, the regressors may include the in-vehicle time and waiting time that vary with the vehicle.

Let $\mathbf{x}_t = (\mathbf{x}'_{t,0}\ \mathbf{x}'_{t,1}\ \ldots\ \mathbf{x}'_{t,J})'$. The specifications for the conditional probabilities are

$$G_{t,j} = \frac{\exp(\mathbf{x}'_{t,j}\boldsymbol{\theta})}{\sum_{k=0}^{J}\exp(\mathbf{x}'_{t,k}\boldsymbol{\theta})}, \qquad j = 0, 1, \ldots, J.$$

In contrast with the multinomial model, $\boldsymbol{\theta}$ is common for all j, so that there is no identification problem.

The quasi-log-likelihood function is

$$L_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=0}^{J} d_{t,j}\mathbf{x}_{t,j}'\boldsymbol{\theta} - \frac{1}{T}\sum_{t=1}^{T}\log\left(\sum_{k=0}^{J}\exp(\mathbf{x}_{t,k}'\boldsymbol{\theta})\right).$$

The gradient and Hessian matrix are

$$\nabla_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=0}^{J}(d_{t,j} - G_{t,j})\mathbf{x}_{t,j},$$

$$\begin{aligned}\nabla^2_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}) &= -\frac{1}{T}\sum_{t=1}^{T}\left[\sum_{j=0}^{J} G_{t,j}\mathbf{x}_{t,j}\mathbf{x}_{t,j}' - \left(\sum_{j=0}^{J} G_{t,j}\mathbf{x}_{t,j}\right)\left(\sum_{j=0}^{J} G_{t,j}\mathbf{x}_{t,j}'\right)\right] \\ &= -\frac{1}{T}\sum_{t=1}^{T}\sum_{j=0}^{J} G_{t,j}(\mathbf{x}_{t,j} - \bar{\mathbf{x}}_t)(\mathbf{x}_{t,j} - \bar{\mathbf{x}}_t)',\end{aligned}$$

where $\bar{\mathbf{x}}_t = \sum_{j=0}^{J} G_{t,j}\mathbf{x}_{t,j}$ is the weighted average of $\mathbf{x}_{t,j}$.

It can be shown that the Hessian matrix is negative definite so that the quasi-log-likelihood function is globally concave.

Each choice probability $G_{t,j}$ is affected not only by $\mathbf{x}_{t,j}$ but also the regressors for other choices, $\mathbf{x}_{t,i}$, because

$$\nabla_{\mathbf{x}_{t,i}} G_{t,j} = -G_{t,j} G_{t,i} \boldsymbol{\theta}, \quad i \neq j, \ i = 0, \ldots, J,$$

$$\nabla_{\mathbf{x}_{t,j}} G_{t,j} = G_{t,j}(1 - G_{t,j})\boldsymbol{\theta}, \quad j = 0, \ldots, J.$$

Note that for positive $\boldsymbol{\theta}$, an increase in $\mathbf{x}_{t,j}$ increases the probability of the j^{th} choice but decreases the probabilities of other choices.

Random Utility Interpretation

McFadden (1974): Consider the random utility of the choice j:

$$U_{t,j} = V_{t,j} + \varepsilon_{t,j}, \quad j = 0, 1, \ldots, J.$$

The alternative i would be chosen if

$$\begin{aligned}\mathbb{P}(y_t = i|\mathbf{x}_t) &= \mathbb{P}(U_{t,i} > U_{t,j}, \text{ for all } j \neq i|\mathbf{x}_t) \\ &= \mathbb{P}(\varepsilon_{t,j} - \varepsilon_{t,i} \leq V_{t,j} - V_{t,i}, \text{ for all } j \neq i|\mathbf{x}_t).\end{aligned}$$

Letting $\tilde{\varepsilon}_{t,ji} = \varepsilon_{t,j} - \varepsilon_{t,i}$ and $\tilde{V}_{t,ji} = V_{t,j} - V_{t,i}$. Then we have (for $j = 0$),

$$\begin{aligned}\mathbb{P}(y_t = 1|\mathbf{x}_t) &= \mathbb{P}(\tilde{\varepsilon}_{t,j0} \leq -\tilde{V}_{t,j0}, \quad j = 1, 2, \ldots, J|\mathbf{x}_t) \\ &= \int_{\infty}^{-\tilde{V}_{t,10}} \cdots \int_{\infty}^{-\tilde{V}_{t,J0}} f(\tilde{\varepsilon}_{t,10}, \ldots \tilde{\varepsilon}_{t,J0}) \, \mathrm{d}\tilde{\varepsilon}_{t,10} \cdots \mathrm{d}\tilde{\varepsilon}_{t,J0}.\end{aligned}$$

This setup is consistent with the theory of decision making. Different models are obtained by imposing different assumptions on the joint distribution of $\varepsilon_{t,j}$.

Suppose that $\varepsilon_{t,j}$ are independent random variables across j with the type I extreme value distribution: $\exp[-\exp(-\varepsilon_{t,j})]$, and the density:

$$f(\varepsilon_{t,j}) = \exp(-\varepsilon_{t,j}) \exp[-\exp(-\varepsilon_{t,j})], \quad j = 0, 1, \ldots, J.$$

It can be shown that

$$\mathbb{P}(y_t = j | \mathbf{x}_t) = \frac{\exp(V_{t,j})}{\sum_{i=0}^{J} \exp(V_{t,i})}.$$

We obtain the conditional logit model when $V_{t,j} = \mathbf{x}_{t,j}'\boldsymbol{\theta}$ and multinomial logit model when $V_{t,j} = \mathbf{x}_t'\boldsymbol{\theta}_j$.

Remarks:

1. In the conditional logit model, the choices must be quite different such that they are independent of each other. This is known as independence of irrelevant alternatives (IIA) which is a restrictive condition in practice.
2. The multinomial probit model is obtained by assuming that $\varepsilon_{t,j}$ are jointly normally distributed. This model permits correlations among the choices, yet it requires numerical or simulation method to evaluate the J-fold integral for choice probabilities.

Nested Logit Models

McFadden (1978): Generalized extreme value (GEV) distribution with

$$F(\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_J) = \exp\left[-G\left(e^{-\varepsilon_0}, e^{-\varepsilon_1}, \ldots, e^{-\varepsilon_J}\right)\right],$$

where G is non-negative and homogeneous of degree one and satisfies other conditions. With this distribution assumption, the choice probabilities of the random utility model are

$$e^{V_j} \frac{G_j\left(e^{-V_0}, e^{-V_1}, \ldots, e^{-V_J}\right)}{G\left(e^{-V_0}, e^{-V_1}, \ldots, e^{-V_J}\right)}, \quad j = 0, 1, \ldots, J,$$

where G_j denotes the derivative of G with respect to the j^{th} argument. In particular, when $G\left(e^{-V_0}, e^{-V_1,}, \ldots, e^{-V_J}\right) = \sum_{k=0}^{J} \exp(-V_k)$, we obtain the multinomial logit model.

Consider the case that there are J groups of choices, in which the j^{th} group has $1, \ldots, K_j$ choices. The random utility is:

$$U_{t,jk} = V_{t,jk} + \varepsilon_{t,jk}, \quad j = 1, \ldots, J, \ k = 1, \ldots, K_j.$$

For example, one must first choose between taking public transportation and driving and then select a vehicle in the chosen group. The nested logit model postulates the GEV distribution for $\varepsilon_{t,jk}$:

$$\exp\left[-G\left(e^{-\varepsilon_{11}}, \ldots, e^{-\varepsilon_{1K_1}}, \ldots, e^{-\varepsilon_{J1}}, \ldots, e^{-\varepsilon_{JK_J}}\right)\right]$$
$$= \exp\left[-\sum_{j=1}^{J}\left(\sum_{k=1}^{K_j}\left(e^{-\varepsilon_{jk}}\right)^{1/r_j}\right)^{r_j}\right],$$

where $r_j = [1 - \text{corr}(\varepsilon_{jk}, \varepsilon_{jm})]^{1/2}$. When the choices in the j^{th} group are uncorrelated, we have $r_j = 1$ and hence the multinomial logit model.

Define the binary variable $d_{t,jk}$ as

$$d_{t,jk} = \begin{cases} 1, & \text{if } y_t = jk, \\ 0, & \text{otherwise,} \end{cases} \qquad j = 1, \ldots, J, \ k = 1, \ldots, K_j.$$

A particular choice jk is chosen with the probability

$$p_{t,jk} = \mathbb{P}(d_{t,jk} = 1 | \mathbf{x}_t) = \mathbb{P}(U_{t,jk} > U_{t,mn} \ \forall m, n | \mathbf{x}_t).$$

Let $\mathbf{x}_t$ be the collection of $\mathbf{z}_{t,j}$ and $\mathbf{w}_{t,jk}$ and

$$V_{t,jk} = \mathbf{z}_{t,j}' \boldsymbol{\alpha} + \mathbf{w}_{t,jk}' \gamma_j, \quad j = 1, \ldots, J, \ k = 1, \ldots, K_j.$$

Thus, we have group-specific regressors and regressors that that depend on both j and k.

Let $I_{t,j} = \ln\left(\sum_{k=1}^{K_j} \exp(\mathbf{w}'_{t,jk}\gamma_j/r_j)\right)$ denote the inclusive value, where r_j are known as scale parameters. We can write

$$p_{t,jk} = \underbrace{\frac{\exp(\mathbf{z}'_{t,j}\alpha + r_j I_{t,j})}{\sum_{m=1}^{J} \exp(\mathbf{z}'_{t,m}\alpha + r_m I_{t,m})}}_{p_{t,j}} \times \underbrace{\frac{\exp(\mathbf{w}'_{t,jk}\gamma_j/r_j)}{\sum_{n=1}^{K_j} \exp(\mathbf{w}'_{t,jn}\gamma_j/r_j)}}_{p_{t,k|j}};$$

details can be found in Cameron and Trivedi (2005, pp. 526–527). Note that for a given j, $p_{t,k|j}$ is a conditional logit model with the parameter γ_j/r_j.

The approximating density is

$$f = \prod_{j=1}^{J} \prod_{k=1}^{K_j} \left(p_{t,j} \times p_{t,k|j} \right)^{d_{t,jk}} = \prod_{j=1}^{J} \left(p_{t,j}^{d_{t,jk}} \prod_{k=1}^{K_j} p_{t,k|j}^{d_{t,jk}} \right),$$

and the quasi-log-likelihood function is

$$L_T = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{J} d_{t,jk} \ln(p_{t,j}) + \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{k=1}^{K_j} d_{t,jk} \ln(p_{t,k|j}).$$

Maximizing this likelihood function yields the full information maximum likelihood (FIML) estimator. There is also a sequential (limited information maximum likelihood) estimator; we omit the details.

Ordered Multinomial Models

Suppose that the categories of data have an ordering such that

$$y_t = \begin{cases} 0, & y_t^* \le c_1 \\ 1, & c_1 < y_t^* \le c_2 \\ \vdots & \\ J, & c_J < y_t^*. \end{cases}$$

where y_t^* are latent variables. For example, language proficiency status and education level have a natural ordering. Such models may be estimated using multinomial logit models; yet taking into account this structure yields simpler models.

Setting $y_t^* = \mathbf{x}_t'\boldsymbol{\theta} + e_t$, the conditional probabilities are:

$$\begin{aligned}
\mathbb{P}(y_t = j|\mathbf{x}_t) &= \mathbb{P}(c_j < y_t^* < c_{j+1} \mid \mathbf{x}_t) \\
&= \mathbb{P}(c_j - \mathbf{x}_t'\boldsymbol{\theta} < e_t < c_{j+1} - \mathbf{x}_t'\boldsymbol{\theta} \mid \mathbf{x}_t) \\
&= F(c_{j+1} - \mathbf{x}_t'\boldsymbol{\theta}) - F(c_j - \mathbf{x}_t'\boldsymbol{\theta}), \qquad j = 0, 1, \ldots, J,
\end{aligned}$$

where $c_0 = -\infty$, $c_{J+1} = \infty$, and F is the distribution of e_t. When F is the logistic (standard normal) distribution function, it is the ordered logit (probit) model. The quasi-log-likelihood function is

$$L_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=0}^{J} \ln\big(F(c_{j+1} - \mathbf{x}_t'\boldsymbol{\theta}) - F(c_j - \mathbf{x}_t'\boldsymbol{\theta})\big).$$

Pratt (1981) showed that the Hessian matrix is negative definite so that the quasi-log-likelihood function is globally concave.

The marginal response of the choice probabilities to the change of regressors are

$$\nabla_{\mathbf{x}_t} \mathbb{P}(y_t = j|\mathbf{x}_t) = \left[f(c_j - \mathbf{x}_t'\boldsymbol{\theta}) - f(c_{j+1} - \mathbf{x}_t'\boldsymbol{\theta})\right]\boldsymbol{\theta}.$$

For the ordered probit model, these responses are

$$\begin{cases} -\phi(c_1 - \mathbf{x}_t'\boldsymbol{\theta})\boldsymbol{\theta}, & j = 0 \\ \left[\phi(c_j - \mathbf{x}_t'\boldsymbol{\theta}) - \phi(c_{j+1} - \mathbf{x}_t'\boldsymbol{\theta})\right]\boldsymbol{\theta}, & j = 1, \ldots, J-1, \\ \phi(c_J - \mathbf{x}_t'\boldsymbol{\theta})\boldsymbol{\theta}, & j = J. \end{cases}$$

Truncated Regression Models

When a variable can only be observed within a limited range, it is known as a limited dependent variable. The data are truncated if they are completely lost outside a given range; for example, income data may be truncated when they are below certain level.

When y_t is truncated from below at c, i.e., $y_t > c$, the conditional density of truncated y_t is

$$\tilde{g}(y_t|y_t > c, \mathbf{x}_t) = g(y_t|\mathbf{x}_t)/\,\mathbb{P}(y_t > c|\mathbf{x}_t).$$

Similarly, the conditional density of y_t truncated from above at c is

$$\tilde{g}(y_t|y_t < c, \mathbf{x}_t) = g(y_t|\mathbf{x}_t)/\,\mathbb{P}(y_t < c|\mathbf{x}_t).$$

Our illustration is based on y_t truncated from below.

Approximating $g(y_t|\mathbf{x}_t)$ by

$$f(y_t|\mathbf{x}_t;\boldsymbol{\beta},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_t-\mathbf{x}_t'\boldsymbol{\beta})^2}{2\sigma^2}\right) = \frac{1}{\sigma}\phi\Big(\frac{y_t-\mathbf{x}_t'\boldsymbol{\beta}}{\sigma}\Big),$$

and setting $u_t = (y_t - \mathbf{x}_t'\boldsymbol{\beta})/\sigma$, we have

$$\mathbb{P}(y_t > c|\mathbf{x}_t) = \frac{1}{\sigma}\int_c^\infty \phi\Big(\frac{y_t-\mathbf{x}_t'\boldsymbol{\beta}}{\sigma}\Big)\,\mathrm{d}\,y_t = \int_{(c-\mathbf{x}_t'\boldsymbol{\beta})/\sigma}^\infty \phi(u_t)\,\mathrm{d}\,u_t$$

$$= 1 - \Phi\Big(\frac{c-\mathbf{x}_t'\boldsymbol{\beta}}{\sigma}\Big).$$

The truncated density function $\tilde{g}(y_t|y_t > c, \mathbf{x}_t)$ is then approximated by

$$f(y_t|y_t > c, \mathbf{x}_t;\boldsymbol{\beta},\sigma^2) = \frac{\phi[(y_t-\mathbf{x}_t'\boldsymbol{\beta})/\sigma]}{\sigma\left[1-\Phi\big((c-\mathbf{x}_t'\boldsymbol{\beta})/\sigma\big)\right]}.$$

The quasi-log-likelihood function is

$$-\frac{1}{2}[\log(2\pi)+\log(\sigma^2)]-\frac{1}{2T\sigma^2}\sum_{t=1}^{T}(y_t-\mathbf{x}_t'\boldsymbol{\beta})^2 - \frac{1}{T}\sum_{t=1}^{T}\log\left[1-\Phi\Big(\frac{c-\mathbf{x}_t'\boldsymbol{\beta}}{\sigma}\Big)\right].$$

Reparameterizing by $\boldsymbol{\alpha}=\boldsymbol{\beta}/\sigma$ and $\gamma=\sigma^{-1}$, we have

$$L_T(\boldsymbol{\theta})=-\frac{\log(2\pi)}{2}+\log(\gamma)-\frac{1}{2T}\sum_{t=1}^{T}(\gamma y_t-\mathbf{x}_t'\boldsymbol{\alpha})^2 - \frac{1}{T}\sum_{t=1}^{T}\log\big[1-\Phi(\gamma c-\mathbf{x}_t'\alpha)\big],$$

where $\boldsymbol{\theta}=(\boldsymbol{\alpha}'\ \gamma)'$.

The first-order condition is

$$\nabla_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T}\left[(\gamma y_t - \mathbf{x}_t'\boldsymbol{\alpha}) - \frac{\phi(\gamma c - \mathbf{x}_t'\boldsymbol{\alpha})}{1-\Phi(\gamma c - \mathbf{x}_t'\boldsymbol{\alpha})}\right]\mathbf{x}_t \\ \frac{1}{\gamma} - \frac{1}{T}\sum_{t=1}^{T}\left[(\gamma y_t - \mathbf{x}_t'\boldsymbol{\alpha})y_t - \frac{\phi(\gamma c - \mathbf{x}_t'\boldsymbol{\alpha})}{1-\Phi(\gamma c - \mathbf{x}_t'\boldsymbol{\alpha})}c\right] \end{bmatrix} = \mathbf{0},$$

from which we can solve for the QMLE of $\boldsymbol{\theta}$. It can also be verified that $L_T(\boldsymbol{\theta})$ is globally concave in $\boldsymbol{\theta}$.

When $f(y_t|y_t > c, \mathbf{x}_t; \boldsymbol{\theta}_o)$ is correctly specified, the conditional mean of truncated y_t is

$$\begin{aligned} \mathbb{E}(y_t|y_t > c, \mathbf{x}_t) &= \int_c^\infty y_t f(y_t|y_t > c, \mathbf{x}_t; \boldsymbol{\theta}_o)\,\mathrm{d}y_t \\ &= \int_c^\infty y_t \left(\frac{\phi[(y_t - \mathbf{x}_t'\boldsymbol{\beta}_o)/\sigma_o]}{\sigma_o\left[1 - \Phi\big((c - \mathbf{x}_t'\boldsymbol{\beta}_o)/\sigma_o\big)\right]}\right)\,\mathrm{d}y_t. \end{aligned}$$

Letting $u_{t,o} = (y_t - \mathbf{x}_t'\boldsymbol{\beta}_o)/\sigma_o$ and $c_{t,o} = (c - \mathbf{x}_t'\boldsymbol{\beta}_o)/\sigma_o$, we have

$$\begin{aligned}
\mathbb{E}(y_t | y_t > c, \mathbf{x}_t) &= \int_{c_{t,o}}^{\infty} (\sigma_o u_{t,o} + \mathbf{x}_t'\boldsymbol{\beta}_o) \frac{\phi(u_{t,o})}{[1 - \Phi(c_{t,o})]} \, \mathrm{d}\, u_{t,o} \\
&= \frac{\sigma_o}{1 - \Phi(c_{t,o})} \int_{c_{t,o}}^{\infty} u_{t,o} \phi(u_{t,o}) \, \mathrm{d}\, u_{t,o} + \mathbf{x}_t'\boldsymbol{\beta}_o \\
&= \frac{\sigma_o}{1 - \Phi(c_{t,o})} \int_{c_{t,o}}^{\infty} -\phi'(u_{t,o}) \, \mathrm{d}\, u_{t,o} + \mathbf{x}_t'\boldsymbol{\beta}_o \\
&= \sigma_o \frac{\phi(c_{t,o})}{1 - \Phi(c_{t,o})} + \mathbf{x}_t'\boldsymbol{\beta}_o.
\end{aligned}$$

That is, even when y_t has a linear conditional mean function, its truncated mean function is necessarily nonlinear. The OLS estimator of regressing y_t on $\mathbf{x}_t$ is inconsistent for $\boldsymbol{\beta}_o$. Although NLS estimation may be employed, QML estimation is typically preferred in practice.

Define the hazard function λ as $\lambda(u) = \phi(u)/[1 - \Phi(u)]$, which is also known as the inverse Mill's ratio. Then,

$$\frac{\mathrm{d}\,\lambda(u)}{\mathrm{d}\,u} = -\frac{\phi(u)u}{1-\Phi(u)} + \left(\frac{\phi(u)}{1-\Phi(u)}\right)^2 = -\lambda(u)u + \lambda(u)^2.$$

The truncated mean is $\mathbb{E}(y_t|y_t > c, \mathbf{x}_t) = \mathbf{x}_t'\boldsymbol{\beta}_o + \sigma_o\lambda(c_{t,o})$, and it can be shown that the truncated variance is

$$\mathrm{var}(y_t|y_t > c, \mathbf{x}_t) = \sigma_o^2\left[1 + \lambda(c_{t,o})c_{t,o} - \lambda^2(c_{t,o})\right],$$

instead of σ_o^2. Note that the marginal response of $\mathbb{E}(y_t|y_t > c, \mathbf{x}_t)$ to a change of $\mathbf{x}_t$ is proportional to conditional variance:

$$\boldsymbol{\beta}_o\left[1 + \lambda(c_{t,o})c_{t,o} - \lambda^2(c_{t,o})\right] = \frac{\boldsymbol{\beta}_o}{\sigma_o^2}\,\mathrm{var}(y_t|y_t > c, \mathbf{x}_t).$$

Consider now the case that the dependent variable y_t is truncated from above at c, i.e., $y_t < c$. The truncated density function $\tilde{g}(y_t|y_t < c, \mathbf{x}_t)$ can be approximated by

$$f(y_t|y_t < c, \mathbf{x}_t; \boldsymbol{\beta}, \sigma^2) = \frac{\phi[(y_t - \mathbf{x}_t'\boldsymbol{\beta})/\sigma]}{\sigma\,\Phi\big((c - \mathbf{x}_t'\boldsymbol{\beta})/\sigma\big)}.$$

The QMLE can be obtained by maximizing the resulting quasi-log-likelihood function L_T. The truncated conditional mean in this case is

$$\mathbb{E}(y_t|y_t < c, \mathbf{x}_t) = \mathbf{x}_t'\boldsymbol{\beta}_o - \sigma_o \frac{\phi\big((c - \mathbf{x}_t'\boldsymbol{\beta}_o)/\sigma_o\big)}{\Phi\big((c - \mathbf{x}_t'\boldsymbol{\beta}_o)/\sigma_o\big)},$$

with the inverse Mill's ratio $\lambda(u) = -\phi(u)/\Phi(u)$.

Censored Regression Models

In many applications, a variable may be censored, rather than truncated. For example, the price of a product is censored at the cheapest available price. Consider

$$y_t = \begin{cases} y_t^*, & y_t^* > 0, \\ 0, & y_t^* \le 0, \end{cases}$$

where $y_t^* = \mathbf{x}_t'\boldsymbol{\beta} + e_t$ is an index variable. It does not matter whether the threshold value of y_t^* is zero or a non-zero constant c.

Let g and g^* denote the densities of y_t and y_t^* conditional on $\mathbf{x}_t$. When $y_t^* > 0$, $g(y_t|\mathbf{x}_t) = g^*(y_t^*|\mathbf{x}_t)$, and when $y_t^* \le 0$, censoring yields

$$\mathbb{P}(y_t = 0|\mathbf{x}_t) = \int_{-\infty}^{0} g^*(y_t^*|\mathbf{x}_t)\,\mathrm{d}y_t^*.$$

The density g is a hybrid of g^* and $\mathbb{P}(y_t = 0|\mathbf{x}_t)$. Define

$$d_t = \begin{cases} 1, & \text{if } y_t > 0, \\ 0, & \text{if } y_t = 0, \end{cases}$$

Then,

$$g(y_t|\mathbf{x}_t) = g^*(y_t^*|\mathbf{x}_t)^{d_t}[\mathbb{P}(y_t = 0|\mathbf{x}_t)]^{1-d_t}.$$

In the standard Tobit model (Tobin, 1958), $g^*(y_t^*|\mathbf{x}_t)$ is approximated by

$$f(y_t^*|\mathbf{x}_t; \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t^* - \mathbf{x}_t'\boldsymbol{\beta})^2}{2\sigma^2}\right) = \frac{1}{\sigma}\phi\Big(\frac{y_t^* - \mathbf{x}_t'\boldsymbol{\beta}}{\sigma}\Big),$$

and $\mathbb{P}\{y_t = 0|\mathbf{x}_t\}$ is approximated by

$$\frac{1}{\sigma}\int_{-\infty}^{0} \phi((y_t^* - \mathbf{x}_t'\boldsymbol{\beta})/\sigma)\, \mathrm{d}\, y_t^* = \int_{-\infty}^{-\mathbf{x}_t'\boldsymbol{\beta}/\sigma} \phi(v_t)\, \mathrm{d}\, v_t = 1 - \Phi\Big(\frac{\mathbf{x}_t'\boldsymbol{\beta}}{\sigma}\Big).$$

The approximating desnity is

$$f(y_t|\mathbf{x}_t;\boldsymbol{\beta},\sigma^2) = \left(\frac{1}{\sigma}\phi\left(\frac{y_t^* - \mathbf{x}_t'\boldsymbol{\beta}}{\sigma}\right)\right)^{d_t}\left(1-\Phi\left(\frac{\mathbf{x}_t'\boldsymbol{\beta}}{\sigma}\right)\right)^{1-d_t}.$$

The quasi-log-likelihood function is thus

$$-\frac{T_1}{2T}[\log(2\pi)+\log(\sigma^2)] + \frac{1}{T}\sum_{\{t:y_t=0\}}\log(1-\Phi(\mathbf{x}_t'\boldsymbol{\beta}/\sigma)) - \frac{1}{2T}\sum_{\{t:y_t>0\}}[(y_t-\mathbf{x}_t'\boldsymbol{\beta})/\sigma]^2,$$

where T_1 is the number of t such that $y_t > 0$.

Letting $\boldsymbol{\alpha} = \boldsymbol{\beta}/\sigma$ and $\gamma = \sigma^{-1}$, the QMLE of $\boldsymbol{\theta} = (\boldsymbol{\alpha}' \ \gamma)'$ is obtained by maximizing

$$L_T(\boldsymbol{\theta}) = \frac{1}{T} \sum_{\{t:y_t=0\}} \log\big(1 - \Phi(\mathbf{x}_t'\boldsymbol{\alpha})\big) + \frac{T_1}{T} \log \gamma$$
$$- \frac{1}{2T} \sum_{\{t:y_t>0\}} (\gamma y_t - \mathbf{x}_t'\boldsymbol{\alpha})^2,$$

which is globally concave in $\boldsymbol{\alpha}$ and γ. The first order condition is

$$\nabla_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}) = \frac{1}{T} \begin{bmatrix} -\sum_{\{t:y_t=0\}} \frac{\phi(\mathbf{x}_t'\boldsymbol{\alpha})}{1-\Phi(\mathbf{x}_t'\boldsymbol{\alpha})} \mathbf{x}_t + \sum_{\{t:y_t>0\}} (\gamma y_t - \mathbf{x}_t'\boldsymbol{\alpha}) \mathbf{x}_t \\ \frac{T_1}{\gamma} - \sum_{\{t:y_t>0\}} (\gamma y_t - \mathbf{x}_t'\boldsymbol{\alpha}) y_t. \end{bmatrix}$$
$$= \mathbf{0},$$

from which the QMLE can be computed.

The conditional mean of censored y_t is

$$\begin{aligned}\mathbb{E}(y_t|\mathbf{x}_t) &= \mathbb{E}(y_t|y_t > 0, \mathbf{x}_t)\ \mathbb{P}(y_t > 0|\mathbf{x}_t) \\ &\qquad + \mathbb{E}(y_t|y_t = 0, \mathbf{x}_t)\ \mathbb{P}(y_t = 0|\mathbf{x}_t) \\ &= \mathbb{E}(y_t^*|y_t^* > 0, \mathbf{x}_t)\ \mathbb{P}(y_t^* > 0|\mathbf{x}_t).\end{aligned}$$

When $f(y_t^*|\mathbf{x}_t; \boldsymbol{\beta}, \sigma^2)$ is correctly specified for $g^*(y_t^*|\mathbf{x}_t)$,

$$\mathbb{P}(y_t^* > 0|\mathbf{x}_t) = \mathbb{P}\left(\frac{y_t^* - \mathbf{x}_t'\boldsymbol{\beta}_o}{\sigma_o} > \frac{-\mathbf{x}_t'\boldsymbol{\beta}_o}{\sigma_o}\middle| \mathbf{x}_t\right) = \Phi(\mathbf{x}_t'\boldsymbol{\beta}_o/\sigma_o).$$

This leads to the following conditional density:

$$\tilde{g}(y_t^*|y_t^* > 0, \mathbf{x}_t) = \frac{g^*(y_t^*|\mathbf{x}_t)}{\mathbb{P}(y_t^* > 0|\mathbf{x}_t)} = \frac{\phi[(y_t^* - \mathbf{x}_t'\boldsymbol{\beta}_o)/\sigma_o]}{\sigma_o\,\Phi(\mathbf{x}_t'\boldsymbol{\beta}_o/\sigma_o)}.$$

Thus, given $y_t^* > 0$,

$$\begin{aligned}\mathbb{E}(y_t^* | y_t^* > 0, \mathbf{x}_t) &= \int_0^\infty y_t^* \tilde{g}(y_t^* | y_t^* > 0, \mathbf{x}_t)\, \mathrm{d}\, y_t^* \\ &= \mathbf{x}_t'\boldsymbol{\beta}_o + \sigma_o \frac{\phi(\mathbf{x}_t'\boldsymbol{\beta}_o/\sigma_o)}{\Phi(\mathbf{x}_t'\boldsymbol{\beta}_o/\sigma_o)};\end{aligned}$$

The conditional mean of censored y_t is thus

$$\begin{aligned}\mathbb{E}(y_t | \mathbf{x}_t) &= \mathbb{E}(y_t^* | y_t^* > 0, \mathbf{x}_t)\ \mathbb{P}(y_t^* > 0 | \mathbf{x}_t) \\ &= \mathbf{x}_t'\boldsymbol{\beta}_o \Phi(\mathbf{x}_t'\boldsymbol{\beta}_o/\sigma_o) + \sigma_o \phi(\mathbf{x}_t'\boldsymbol{\beta}_o/\sigma_o).\end{aligned}$$

This shows that $\mathbf{x}_t'\boldsymbol{\beta}$ can not be the correct specification of the conditional mean of censored y_t. Regressing y_t on $\mathbf{x}_t$ thus results in an inconsistent estimator for $\boldsymbol{\beta}_o$.

Consider the case that y_t is censored from above:

$$y_t = \begin{cases} y_t^*, & y_t^* < 0, \\ 0, & y_t^* \geq 0. \end{cases}$$

When $f(y_t^*|\mathbf{x}_t; \boldsymbol{\beta}, \sigma^2)$ is correctly specified for $g^*(y_t^*|\mathbf{x}_t)$, $\mathbb{P}(y_t^* < 0|\mathbf{x}_t) = 1 - \Phi(\mathbf{x}_t'\boldsymbol{\beta}_o/\sigma_o)$, and

$$\tilde{g}(y_t^*|y_t^* < 0, \mathbf{x}_t) = \frac{g^*(y_t^*|\mathbf{x}_t)}{\mathbb{P}(y_t^* < 0|\mathbf{x}_t)} = \frac{\phi[(y_t^* - \mathbf{x}_t'\boldsymbol{\beta}_o)/\sigma_o]}{\sigma_o\,[1 - \Phi(\mathbf{x}_t'\boldsymbol{\beta}_o/\sigma_o)]}.$$

Given $y_t^* < 0$,

$$\mathbb{E}(y_t^*|y_t^* < 0, \mathbf{x}_t) = \mathbf{x}_t'\boldsymbol{\beta}_o - \sigma_o \frac{\phi(\mathbf{x}_t'\boldsymbol{\beta}_o/\sigma_o)}{1 - \Phi(\mathbf{x}_t'\boldsymbol{\beta}_o/\sigma_o)}.$$

The conditional mean of y_t censored from above is

$$\mathbb{E}(y_t|\mathbf{x}_t) = \mathbf{x}_t'\boldsymbol{\beta}_o[1 - \Phi(\mathbf{x}_t'\boldsymbol{\beta}_o/\sigma_o)] - \sigma_o\phi(\mathbf{x}_t'\boldsymbol{\beta}_o/\sigma_o).$$

Remark: The results in Tobit model rely heavily on the distributional assumptions. If the postulated normality is incorrect or even if homoskedasticity does not hold, the QMLE also loses consistency.

Sample Selection Models

In the study of how wages and individual characteristics affect working hours, the data of working hours are only observed for those who select to work. This is the problem of sample selection or incidental truncation.

Consider two variables y_1 (indicator of working) and y_2 (working hour):

$$y_{1,t} = \begin{cases} 1, & y_{1,t}^* > 0, \\ 0, & y_{1,t}^* \leq 0. \end{cases}$$

$$y_{2,t} = y_{2,t}^*, \quad \text{if } y_{1,t} = 1,$$

where $y_{1,t}^* = \mathbf{x}_{1,t}'\boldsymbol{\beta} + e_{1,t}$ and $y_{2,t}^* = \mathbf{x}_{2,t}'\boldsymbol{\gamma} + e_{2,t}$ are unobserved index variables. The selection problem arises because the former affects the latter, and $y_{2,t}$ are incidentally truncated when $y_{1,t} = 0$.

In general, the likelihood function of y_t contains 2 parts:

$$\left[P(y_{1,t}^* \leq 0|\mathbf{x}_t)\right]^{1-y_{1,t}} \left[f(y_{2,t}|y_{1,t}^* > 0, \mathbf{x}_t)\, \mathbb{P}(y_{1,t}^* > 0|\mathbf{x}_t)\right]^{y_{1,t}};$$

see Amemiya (1985, pp. 385–387) for details.

Type 2 Tobit model: Under conditional normality,

$$\begin{pmatrix} y_{1,t}^* \\ y_{2,t}^* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{x}_{1,t}'\boldsymbol{\beta}_o \\ \mathbf{x}_{2,t}'\boldsymbol{\gamma}_o \end{pmatrix}, \begin{bmatrix} \sigma_{1,o}^2 & \sigma_{12,o} \\ \sigma_{12,o} & \sigma_{2,o}^2 \end{bmatrix}\right).$$

We know

$$\mathbb{E}(y_{2,t}^*|y_{1,t}^*) = \mathbf{x}_{2,t}'\boldsymbol{\gamma}_o + \sigma_{12,o}(y_{1,t}^* - \mathbf{x}_{1,t}'\boldsymbol{\beta}_o)/\sigma_{1,o}^2,$$

and $\operatorname{var}(y_{2,t}^*|y_{1,t}^*) = \sigma_{2,o}^2 - \sigma_{12,o}^2/\sigma_{1,o}^2$.

Recall from truncated regression that

$$\mathbb{E}(y_{1,t}^*|y_{1,t}^* > c) = \mathbf{x}_{1,t}'\boldsymbol{\beta}_o + \sigma_{1,o}\phi(c_{t,o})/[1-\Phi(c_{t,o})],$$

where $c_{t,o} = (c - \mathbf{x}_{1,t}'\boldsymbol{\beta}_o)/\sigma_{1,o}$. When the truncation parameter $c = 0$, $\phi(c_{t,o}) = \phi(\mathbf{x}_{1,t}'\boldsymbol{\beta}_o/\sigma_{1,o})$ and $1-\Phi(c_{t,o}) = \Phi(\mathbf{x}_{1,t}'\boldsymbol{\beta}_o/\sigma_{1,o})$. It follows that

$$\mathbb{E}(y_{1,t}^*|y_{1,t}^* > 0) = \mathbf{x}_{1,t}'\boldsymbol{\beta}_o + \sigma_{1,o}\lambda\left(\frac{\mathbf{x}_{1,t}'\boldsymbol{\beta}_o}{\sigma_{1,o}}\right),$$

with $\lambda(u) = \phi(u)/\Phi(u)$. Consequently,

$$\begin{aligned}\mathbb{E}(y_{2,t}|y_{1,t}^* > 0, \mathbf{x}_t) &= \mathbf{x}_{2,t}'\boldsymbol{\gamma}_o + \frac{\sigma_{12,o}}{\sigma_{1,o}^2}\left[\mathbb{E}(y_{1,t}^*|y_{1,t}^* > 0, \mathbf{x}_t) - \mathbf{x}_{1,t}'\boldsymbol{\beta}_o\right] \\ &= \mathbf{x}_{2,t}'\boldsymbol{\gamma}_o + \frac{\sigma_{12,o}}{\sigma_{1,o}}\lambda\left(\frac{\mathbf{x}_{1,t}'\boldsymbol{\beta}_o}{\sigma_{1,o}}\right).\end{aligned}$$

Again from the truncation regression result, we have

$$\text{var}(y_{1,t}^*|y_{1,t}^* > 0, \mathbf{x}_t) = \sigma_{1,o}^2 \left[1 - \lambda\left(\frac{\mathbf{x}_{1,t}'\boldsymbol{\beta}_o}{\sigma_{1,o}}\right)\frac{\mathbf{x}_{1,t}'\boldsymbol{\beta}_o}{\sigma_{1,o}} - \lambda\left(\frac{\mathbf{x}_{1,t}'\boldsymbol{\beta}_o}{\sigma_{1,o}}\right)^2\right].$$

Writing

$$y_{2,t}^* = \mathbf{x}_{2,t}'\boldsymbol{\gamma}_o + \sigma_{12,o}(y_{1,t}^* - \mathbf{x}_{1,t}'\boldsymbol{\beta}_o)/\sigma_{1,o}^2 + v_t,$$

where $v_t|y_{1,t}^* \sim \mathcal{N}(0, \sigma_{2,o}^2 - \sigma_{12,o}^2/\sigma_{1,o}^2)$. Then $\text{var}(y_{2,t}|y_{1,t}^* > 0, \mathbf{x}_t)$ is

$$\begin{aligned}
&\frac{\sigma_{12,o}^2}{\sigma_{1,o}^4}\text{var}(y_{1,t}^*|y_{1,t}^* > 0, \mathbf{x}_t) + \text{var}(v_t|y_{1,t}^* > 0, \mathbf{x}_t) \\
&= \frac{\sigma_{12,o}^2}{\sigma_{1,o}^2}\left[1 - \lambda\left(\frac{\mathbf{x}_{1,t}'\boldsymbol{\beta}_o}{\sigma_{1,o}}\right)\frac{\mathbf{x}_{1,t}'\boldsymbol{\beta}_o}{\sigma_{1,o}} - \lambda\left(\frac{\mathbf{x}_{1,t}'\boldsymbol{\beta}_o}{\sigma_{1,o}}\right)^2\right] + \left(\sigma_{2,o}^2 - \frac{\sigma_{12,o}^2}{\sigma_{1,o}^2}\right) \\
&= \sigma_{2,o}^2 - \frac{\sigma_{12,o}^2}{\sigma_{1,o}^2}\left[\lambda\left(\frac{\mathbf{x}_{1,t}'\boldsymbol{\beta}_o}{\sigma_{1,o}}\right)\frac{\mathbf{x}_{1,t}'\boldsymbol{\beta}_o}{\sigma_{1,o}} + \lambda\left(\frac{\mathbf{x}_{1,t}'\boldsymbol{\beta}_o}{\sigma_{1,o}}\right)^2\right].
\end{aligned}$$

Heckman's Two-Step Estimator

We thus have a complex nonlinear specification:

$$y_{2,t} = \mathbf{x}_{2,t}'\boldsymbol{\gamma} + \frac{\sigma_{12}}{\sigma_1}\lambda\left(\frac{\mathbf{x}_{1,t}'\boldsymbol{\beta}}{\sigma_1}\right) + e_t,$$

with conditional heteroskedasticity. Clearly, OLS regression of $y_{2,t}$ on $\mathbf{x}_t$ is inconsistent, unless $\sigma_{12,o} = 0$. The sample-selection bias may be very severe in finite samples.

Heckman's two-step procedure yields consistent estimate:

1. Compute the QMLE of $\boldsymbol{\alpha} = \boldsymbol{\beta}/\sigma_1$ using the probit model of $y_{1,t}$ and denote the estimator as $\tilde{\boldsymbol{\alpha}}_T$.
2. Regress $y_{2,t}$ on $\mathbf{x}_{2,t}$ and $\tilde{\lambda}_t$, where $\tilde{\lambda}_t = \lambda(\mathbf{x}_{1,t}'\tilde{\boldsymbol{\alpha}}_T)$.

The variance estimate can be computed as

$$\hat{\sigma}_2^2 = \frac{1}{T}\sum_{t=1}^{T}\left[\tilde{e}_t^2 + \frac{\hat{\sigma}_{12}^2}{\hat{\sigma}_1^2}\tilde{\lambda}_t\left(\mathbf{x}_{1,t}'\tilde{\boldsymbol{\alpha}}_T + \tilde{\lambda}_t\right)\right],$$

where $\tilde{e}_t$ are the residuals of the regression in the 2nd step.

Remarks:

- We can test whether sample selection is relevant by checking the coefficient of $\tilde{\lambda}_t$ in the 2nd step is zero.
- Both the OLS standard errors and heteroskedasticity-consistent standard errors in the regression of the 2nd step are incorrect, due to the presence of the parameter estimate $\tilde{\boldsymbol{\alpha}}_T$ in $\tilde{\lambda}_t$. There are some complicated ways to handle this problem (check your software); bootstrap is an alternative.

Time Series Models