Ei Thandar Myint
ARE 212

5. Logit

This problem is meant to help draw connections between GMM estimators and maximum likelihood estimators, with a particular focus on the 'logit' model.

The development of a maximum likelihood estimator typically begins with an assumption that some random variable has a (conditional) distribution which is known up a k-vector of parameters β. Consider the case in which we observe N independent realizations of a Bernoulli random variable Y, with $\Pr(Y = 1|X) = \sigma(\beta^\top X)$, and $\Pr(Y = 0|X) = 1 - \sigma(\beta^\top X)$.

(1) Show that under this model $\mathbb{E}(Y_i - \sigma(X\beta)|X) = 0$. Assume that σ is a known function, and use this fact to develop a GMM estimator of β. Is your estimator just- or over-identified?

$Y_i \sim \text{Bernoulli}(\sigma)$.

$$E[Y_i|X_i] = 1 \cdot \Pr(Y_i = 1|X_i) + 0 \cdot \Pr(Y_i = 0|X_i) = \sigma(\beta^T X_i).$$

$$E[\sigma(\beta^T X_i)|X_i] = \sigma(\beta^T X_i)$$

Thus, $E[Y_i - \sigma(X\beta)|X] = \sigma(X\beta) - \sigma(X\beta) = 0$ $\square$

Suppose σ is a known function.

The moment condition is: $E[Y_i - \sigma(\beta^T X_i)|X_i) = 0$.

By taking the law of iterated expectations,

$$E[E[Y_i - \sigma(\beta^T X_i)|X_i]] = E[Y_i - \sigma(\beta^T X_i)] = 0.$$

This implies: $E[X_i[Y_i - \sigma(\beta^T X_i)] = 0$.

By taking the sample analog of the moment condition,

$$g_N(\beta) = \frac{1}{N} \sum_{i=1}^{N} X_i (Y_i - \sigma(\beta^T X_i))$$

To estimate β, GMM chooses the parameter that minimizes the squared weighted sum of these moment conditions:

$$\hat{\beta} = \arg\min_{\beta} \left(g_N(\beta)^T W g_N(\beta) \right)$$

where W is a positive definite weighting matrix.

[W is often chosen to be the inverse of the covariance matrix of the moment conditions, which provides the most efficient estimator in the class of GMM estimators].

β has k parameters. We need to estimate these k parameters.

Each component of X_i provides a moment condition derived from

$X_i(y_i - \sigma(\beta^T X_i))$.

If X_i is k-dimensional, which is equal to the number of parameters in β, then the system is just-identified.

If there are more moment conditions than parameters, it is over-identified.

(2) Show that the likelihood can be written as

$$L(\beta|y, X) = \prod_{i=1}^{N} \sigma(\beta^\top X_i)^{y_i} \left(1 - \sigma(\beta^\top X_i)\right)^{1-y_i}.$$

$$Pr(Y_i = y_i \mid X_i) = \sigma(\beta^T X_i)^{y_i} \left(1 - \sigma(\beta^T X_i)\right)^{1-y_i}$$

If $y_i = 1$, $Pr(Y_i = 1 \mid X_i) = \sigma(\beta^T X_i)$

If $y_i = 0$, $Pr(Y_i = 0 \mid X_i) = 1 - \sigma(\beta^T X_i)$.

The likelihood function is:

$$\mathcal{L}(\beta \mid y, X) = \prod_{i=1}^{N} Pr(Y_i = y_i \mid X_i).$$

$$\Rightarrow \mathcal{L}(\beta \mid y, X) = \prod_{i=1}^{N} \sigma(\beta^T X_i)^{y_i} \left(1 - \sigma(\beta^T X_i)\right)^{1-y_i}$$

(3) To obtain the maximum likelihood estimator (MLE) one can chose b to maximize $\log L(b|y,X)$. When the likelihood is well-behaved, the MLE estimator satisfies the first order conditions (also called the "scores") from this maximization problem, in which case this is called a "type I" MLE. Let $\sigma(z) = \frac{1}{1+e^{-z}}$ (this is sometimes called the logistic function, or the sigmoid function), and obtain the scores $S_N(b)$ for this estimation problem. Show that $\mathbb{E}S_N(\beta) = 0$. Demonstrate that these moment conditions can serve as the basis for a GMM estimator of β, and compare this estimator to the GMM estimator you developed above. Which is more efficient, and why?

Let $\sigma(z) = \dfrac{1}{1+e^{-z}}$.

Note that $1 - \sigma(z) = 1 - \dfrac{1}{1+e^{-z}} = \dfrac{1+e^{-z}-1}{1+e^{-z}} = \dfrac{e^{-z}}{1+e^{-z}}$

Thus,

$$\sigma'(z) = \frac{-1}{(1+e^{-z})^2}(-e^{-z}) = \frac{e^{-z}}{(1+e^{-z})^2} = \sigma(z)(1-\sigma(z)).$$

$$\mathcal{L}(\beta|y,X) = \prod_{i=1}^{N} \sigma(\beta^T X_i)^{y_i} (1-\sigma(\beta^T X_i))^{1-y_i}$$

$$\log \mathcal{L}(\beta|y,X) = \sum_{i=1}^{N} \Big[y_i \underbrace{\log(\sigma(\beta^T X_i))}_{(1)} + (1-y_i) \underbrace{\log(1-\sigma(\beta^T X_i))}_{(2)} \Big]$$

The score function is obtained by $\dfrac{\partial \log \mathcal{L}}{\partial \beta} = S_N(\beta)$

$$\frac{\partial}{\partial \beta} \log \mathcal{L}(\beta|y,X) = \sum_{i=1}^{N} [y_i - \sigma(\beta^T X_i)] X_i = S_N(\beta).$$

[work shown below].

Differentiate (1):

$$\frac{\partial}{\partial \beta} \log(\sigma(\beta^T X_i)) = \frac{1}{\sigma(\beta^T X_i)} \sigma'(\beta^T X_i) \cdot X_i$$

$$= \frac{1}{\cancel{\sigma(\beta^T X_i)}} \cancel{\sigma(\beta^T X_i)} (1 - \sigma(\beta^T X_i)) \cdot X_i$$

$$= (1 - \sigma(\beta^T X_i)) X_i$$

Similarly, differentiate (2):

$$\frac{\partial}{\partial \beta} \log(1 - \sigma(\beta^T X_i)) = -\sigma(\beta^T X_i) X_i$$

Thus,

$$\frac{\partial}{\partial \beta} \log \mathcal{L}(\beta | y, X) = \sum_{i=1}^{N} [y_i (1 - \sigma(\beta^T X_i)) X_i + (1 - y_i)(-\sigma(\beta^T X_i)) X_i]$$

$$= \sum_{i=1}^{N} [y_i X_i - \cancel{y_i X_i \sigma(\beta^T X_i)} - X_i \sigma(\beta^T X_i) + \cancel{y_i X_i \sigma(\beta^T X_i)}]$$

$$= \sum_{i=1}^{N} [y_i - \sigma(\beta^T X_i)] X_i$$

Proof that $E[S_N(\beta)] = 0$:

$$E[S_N(\beta)] = E\left[\sum_{i=1}^{N} [y_i - \sigma(\beta^T X_i)] X_i\right]$$

$$= \sum_{i=1}^{N} [E[y_i - \sigma(\beta^T X_i)] X_i] \underset{\text{by LIE from part 1.}}{=} \sum_{i=1}^{N} 0 = 0.$$

The scores $S_N(\beta)$ can be directly used as moment conditions for GMM estimator:

$$g_N(\beta) = \frac{1}{N} S_N(\beta) = \frac{1}{N} \sum_{i=1}^{N} [y_i - \sigma(\beta^T X_i)] X_i.$$

$$\Rightarrow \hat{\beta}_{GMM} = \arg\min_{\beta} (g_N(\beta)^T W g_N(\beta))$$

where W is a weighting matrix.

(GMM1) (GMM3)

- Comparing GMM estimator from part (1) vs this GMM estimator:

  - The functional forms of both GMM estimators are the same. However, the sources of moment conditions differ.
  - In GMM1, moment conditions are derived assuming that the model residuals should average to 0, which is the direct application of the logistic regression model's fit to the data.
  - In GMM 3, moment conditions come from the score function of the MLE, which also implies that the model's residuals should average to 0.
  - The choice of W can affect the efficiency of both estimators. If W is chosen as the inverse of the covariance matrix of moment conditions, both estimators can be asymptotically equivalent and efficient.
  - otherwise, GMM3 might be more efficient as this comes from MLE and aligns with the Fisher information achieving the Cramer-Rao lower bound.