1 Extremum estimator

An extremum estimators are a wide class of estimators for parametric models that are calculated through maximization (or minimization) of a certain objective function, which depends on the data. The M-estimator is a subcategory or a type of extremum estimators. The types of extremum estimators differs based on the objective functions. For instance

1. Classical Minimum Distance (CMD) extremum estimator is $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} Q_n(\mathbf{w};\theta)$ and $Q_n(\mathbf{w};\theta) = -n\mathbf{g}_n^{'}(\mathbf{w};\theta)\mathbf{W}_n\mathbf{g}_n(\mathbf{w};\theta)$ where $\mathbf{w}$ is your entire data and θ is the parameter vector. GMM is a special case of the CMD estimator.

2. Maximum likelihood (ML) (M-estimator) is $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} Q_n(\mathbf{w};\theta)$ and $Q_n(\mathbf{w};\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(\mathbf{w}_i;\theta)$ and $\log f(\mathbf{w}_i;\theta)$ is your log-likelihood function.

3. Nonlinear Least Squares (ML) (M-estimator) is $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} Q_n(\mathbf{w};\theta)$ and $Q_n(\mathbf{w};\theta) = \frac{1}{n}\sum_{i=1}^{n} -\hat{\epsilon}_i^2$ and $\hat{\epsilon}_i = y_i - \phi(\mathbf{x}_i;\theta)$, where $\phi(\mathbf{x}_i;\theta)$ is your nonlinear function.

1.1 Maximum Likelihood Estimation of the OLS

We have our linear regression model across n observations

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \epsilon \sim N(0, \sigma^2\mathbf{I}_n),$$

which can be written as

$$L(\beta, \sigma^2|\mathbf{y}) = f(\mathbf{y}|\beta, \sigma^2) \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}_n),$$

and the pdf of this density is

$$L(\beta, \sigma^2|\mathbf{y}) = (2\pi)^{-\frac{n}{2}} \det(\sigma^2\mathbf{I}_n)^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)'(\sigma^2\mathbf{I}_n)^{-1}(\mathbf{y} - \mathbf{X}\beta)],$$

$$ln(L(\beta, \sigma^2|\mathbf{y})) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta),$$

$$ln(L(\beta, \sigma^2|\mathbf{y})) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta),$$

Let's assume $\theta = (\beta, \sigma^2)$ is the parameter vector. Thus the the maximum likelihood estimation is

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}} ln(L(\theta|\mathbf{y})),$$

Then, all we need to do is take the F.O.C of $ln(L(\beta, \sigma^2|\mathbf{y}))$ respect to β and σ^2, and set them both to equal to 0.

$$\frac{\partial ln(L(\beta, \sigma^2|\mathbf{y}))}{\partial \beta} = -\frac{1}{2\sigma^2}(2\mathbf{X}'\mathbf{X}\beta - 2\mathbf{X}'\mathbf{y}) = 0,$$

$$\frac{\partial ln(L(\beta, \sigma^2|\mathbf{y}))}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = 0,$$

When we are taking the derivative respect to β, we need to apply this rule $\psi(\mathbf{x}) = \mathbf{x}\mathbf{A}\mathbf{x}'$ and the $\frac{\partial \psi(\mathbf{x})}{\mathbf{x}} = \mathbf{x}'(\mathbf{A} + \mathbf{A}')$ and if $\mathbf{A}$ is a symmetric matrix then $\frac{\partial \psi(\mathbf{x})}{\mathbf{x}} = 2\mathbf{x}'\mathbf{A}$.
We can solve the above system of equations since we have two unknowns and two equations,

$$\hat{\beta}_{MLE} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta}_{MLE})'(\mathbf{y} - \mathbf{X}\hat{\beta}_{MLE}).$$

2 Wald, LM and LR test

For the ML estimator, there are three types of tests:

1. Wald test: All you need is the unrestricted ML estimator $\hat{\theta}_{ML}$ and the asymptotic variance of the ML estimator $AVAR(\hat{\theta}_{ML})$. A simple example could be, let's assume we only one ML estimator $\hat{\theta}_0$ and we want test whether $\hat{\theta}_0$ is equal to a hypothesised value of θ_0. Then, $H_0 : \hat{\theta}_0 = \theta_0$, $H_A : \hat{\theta}_0 \neq \theta_0$ and Wald statistics is $W = \frac{(\hat{\theta}_0 - \theta_0)^2}{AVAR(\hat{\theta}_{ML})} \underset{d}{\to} \chi^2_1$.

2. Lagrange Multiplier (LM) test: Imagine you want to test find a ML estimator based on some certain constraint, for example $\tilde{\theta}_{ML} = \underset{\theta}{\text{argmin}} Q_n(\theta)$ s.t $\mathbf{a}(\theta) = \theta_0$. Basically the LM test is to test whether these contraints are binding. Then, $H_0 : \mathbf{a}(\theta) = \theta_0$, $H_A : \mathbf{a}(\theta) \neq \theta_0$ and LM statistics is $\frac{U(\theta_0)^2}{I(\theta_0)}$, where $U(\theta) = \frac{\partial \log L(\theta)}{\partial \theta}$ is gradient of the log-likelihood and $I(\theta_0) = -E[\frac{\partial^2 \log L(\theta)}{\partial \theta^2}|\theta]$ is the fisher information.

3. Likelihood ratio (LR) test uses both the unrestricted $\hat{\theta}_{ML}$ and restricted ML estimators $\tilde{\theta}_{ML} = \underset{\theta}{\text{argmin}} Q_n(\theta)$ s.t $\mathbf{a}(\theta) = 0$. Then, $H_0 : \mathbf{a}(\theta) = \theta_0$, $H_A : \mathbf{a}(\theta) \neq \theta_0$ and the LR test statistics is given by $2\{\log(\hat{\theta}_{ML}) -$

$\log(\tilde{\theta}_{ML})\} \underset{d}{\rightarrow} \chi^2_r$ where r is the no. of restrictions in the constraint and $\log(\hat{\theta}_{ML}) - \log(\tilde{\theta}_{ML})$ is the difference between the log-likelihoods between the unrestricted ML and restricted ML estimators.

3 Cramer-Rao Inequality

Cramér–Rao bound (CRB) expresses a lower bound on the variance of unbiased estimators of a deterministic (fixed, though unknown) parameter, stating that the variance of any such estimator is at least as high as the inverse of the Fisher information.

$$var(\hat{\theta}(z)) \geq \mathbf{I}(\theta_0)^{-1},$$

where $\hat{\theta}(z)$ is an unbiased estimator of θ with a finite variance-coariance matrix and $\mathbf{I}(\theta_0)$ is the fisher information matrix. The Information Matrix Equality assumes the fisher information matrix equals the negative of the expected value of the Hessian (matrix of second partial derivatives) of the log likelihood $\mathbf{I}(\theta_0) = -E[\frac{\partial^2 \log L(\theta)}{\partial \theta^2}]$. An unbiased estimator which achieves this lower bound is said to be asymptotically efficient since it has the smallest asymptotic variance in the class of consistent and asymptotically normal estimators.

4 Probit Model

The likelihood function for the probit model with a random effect is

$$L(\beta, \sigma^2|\mathbf{y}) = \prod_{i=1}^{n} [\Phi(\mathbf{x}_i^{'}\beta + \alpha_i)]^{y_i} [1 - \Phi(\mathbf{x}_i^{'}\beta + \alpha_i)]^{1-y_i},$$

where Φ is the CDF of a standard normal distribution. Then the log-likelihood

$$ln(L(\beta, \sigma^2|\mathbf{y})) = y_i \sum_{i}^{n_1} ln([\Phi(\mathbf{x}_i^{'}\beta + \alpha_i)]) + (1 - y_i) \sum_{i}^{n_2} ln([1 - \Phi(\mathbf{x}_i^{'}\beta + \alpha_i)]),$$

where n_1 is the total number of observations when $y_i = 1$ and n_2 is the total number of observations when $y_i = 0$, and $n = n_1 + n_2$. If we define the parameter vector $\theta = (\beta, \sigma^2)$, then the ML estimation is given by

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}} ln(L(\theta|\mathbf{y})),$$

where we need to take both the partial derivatives $\frac{\partial ln(L(\beta,\sigma^2|\mathbf{y}))}{\partial \beta}$ and $\frac{\partial ln(L(\beta,\sigma^2|\mathbf{y}))}{\partial \sigma^2}$, and set them equal to zero, and solve for the parameters. The above log-likelihood is non-linear and non-analytic, therefore we cannot simply solve it with a pen and paper. We have to use numerical optimisation techniques, such as Newton-

Rapson method, to obtain the ML estimators for the parameters of the probit model. This is also true for the logit model.

5 Sample selection model

Consider a sample selection model

$$y_i = \mathbf{x}_i^{'}\beta + u_{1,i},$$

$$\mathbf{z}_i^{'}\gamma + u_{z,i}, > 0$$

We can define an indicator function $d_i = 1(\mathbf{z}_i^{'}\gamma + u_{z,i}, > 0)$ for the sample selection bias which implies when $\mathbf{z}_i^{'}\gamma + u_{z,i}, > 0$ then $d_i = 1$ and vice versa. Let's consider a simple example, let's interpret y_i as how much we work and then d_i can be interpret as whether we work (or are employed) ($d_i = 1$) or not ($d_i = 0$). Intuitively, this model is implying that we can only determine how much a person works if they work or are employed. In the lecture notes, we assume that the errors are

$$\begin{bmatrix} u_{1,i} \\ u_{2,i} \end{bmatrix} \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & 1 \end{bmatrix}),$$

The likelihood function the sample selection model can be given as

$$L = \prod_{i=1}^{n} [Pr(\mathbf{z}_i^{'}\gamma + u_{z,i}, < 0)]^{1-d_i} [f(y_i|d_i = 1)Pr(\mathbf{z}_i^{'}\gamma + u_{z,i}, > 0)]^{d_i},$$

Note here $Pr(\mathbf{z}_i^{'}\gamma + u_{z,i}, < 0)$ is probability that person does not work which is equal to $Pr(\mathbf{z}_i^{'}\gamma + u_{z,i}, < 0) = \Phi(-\mathbf{z}_i^{'}\gamma) = 1 - \Phi(\mathbf{z}_i^{'}\gamma)$, where Φ is denoted as the CDF of the standard normal distribution. Thus, probability that person works will be $Pr(\mathbf{z}_i^{'}\gamma + u_{z,i}, > 0) = \Phi(\frac{\mathbf{z}_i^{'}\gamma + \rho(y_i - \mathbf{x}_i^{'}\beta)/\sigma}{\sqrt{1-\rho^2}})$. $f(y_i|d_i = 1)$ can be interpreted as the likelihood function for a particular person given that person works or are employed, and is equal to $f(y_i|d_i = 1) = \frac{1}{\sigma}\phi(\frac{y_i - \mathbf{x}_i^{'}\beta}{\sigma})$ where ϕ is denoted as the pdf of a standard normal distribution. Note $f(y_i|d_i = 1) \sim N(\mathbf{x}_i^{'}\beta, \sigma^2)$ and $\frac{1}{\sigma}\phi(\frac{y_i - \mathbf{x}_i^{'}\beta}{\sigma})$ means the same thing but is written in standard normal distribution form. However, now the questions state $\rho = 0$, which implies

$$\begin{bmatrix} u_{1,i} \\ u_{2,i} \end{bmatrix} \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 \\ 0 & 1 \end{bmatrix}),$$

there is no correlation in the errors between the two equations. Therefore, the two equations are now independent from each other. The first equation is just a linear model and likelihood function can be written as

$$y_i = \mathbf{x}_i^{'}\beta + u_{1,i}, u_{1,i} \sim N(0, \sigma^2),$$

$$f(y_i|\beta, \sigma^2) \sim N(\mathbf{x}_i^{'}\beta, \sigma^2),$$

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} f(y_i|\beta, \sigma^2), \tag{1}$$

and second equation becomes a probit model similar to the first question, where the likelihood is defined as

$$L(\gamma) = \prod_{i=1}^{n} [Pr(\mathbf{z}_i^{'}\gamma + u_{z,i}, < 0)]^{1-d_i} [Pr(\mathbf{z}_i^{'}\gamma + u_{z,i}, > 0)]^{d_i},$$

$$L(\gamma) = \prod_{i=1}^{n} [1 - \Phi(\mathbf{z}_i^{'}\gamma)]^{1-d_i} [\Phi(\mathbf{z}_i^{'}\gamma))]^{d_i}, \tag{2}$$

Then, all we need to do is undertake separate MLE for (1) and (2). Note the ML estimator for a linear model is the same as the OLS estimator $\hat{\beta}_{MLE} = \hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

6 Logit model

Let's consider a simple example where we have only two outcomes, success or fail. Let's denote p as the probability of success and $1-p$ as the probability of failure. The logit model is basically used to estimate this probability p. We can calculate some sort of probability odds ratio $\frac{p}{1-p}$ and if $p = 0.8$, this results $\frac{p}{1-p} = 4$, which implies that you will likely achieve success 4 times out of the 5 outcomes.

Thus, let's consider a one regressor logit model and the function form is

$$F(Y) = p = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)},$$

$$1 - p = 1 - \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)} = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1)},$$

Then, the odd ratio is

$$\frac{p}{1-p} = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)} \div \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1)}$$

$$\frac{p}{1-p} = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)} \times \frac{1 + \exp(\beta_0 + \beta_1 x_1)}{1} = \exp(\beta_0 + \beta_1 x_1),$$

Then if we take natural logs on both sides we get

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1,$$

where the linear model equals to the log odd ratio. Thus, the estimated coefficients β_0 and β_1 affects the log odd ratio. Therefore, in the example of the low birthweight, the estimated coefficient for smoking is about 0.92 which can be interpret as if a person smokes, it will increase the log odd ratio by about 0.92 or it will increase the odd ratio by about 2.5 (exp(0.92)).

6.1 Marginal effect

The marginal effect of a logit model is $(\frac{\partial F(y)}{\partial x_i})$ just the partial derivative for $F(y)$ with respect to a regressor x_i. In the example of the low birthweight, the marginal effect or partial derivative for $F(y)$ with respect to a regressor is evaluated at the mean observation. For example, if a person age is about 23.2381, their probability of getting a baby with low birthweight will decrease about 1%. Except for the age and weight at last menstrual period regressors, all the other regressors are binary variables (0 and 1) and the interpretation of these marginal effects makes limited sense.

6.2 Link function

A link function is a function linking the actual Y to the estimated Y in an econometric model. For example, as seen above, the logit link function

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1,$$

and the probit link function is

$$\Phi^{-1}(Y) = \beta_0 + \beta_1 x_1,$$

Therefore, logit and probit models differs in their link function. In the example of the low birthweight, the probit model estimated coefficients have similar signs to the estimated logit model coefficients. Thus, this

implies the results are robust.