# Lecture 5: Linear Regressions*

In lecture 2, we introduced stationary linear time series models. In that lecture, we discussed the data generating processes and their characteristics, assuming that we know all parameters (autoregressive or moving average coefficients). However, in empirical studies, we have to specify an econometric model, estimate this model and draw inferences based on the estimates. In this lecture, we will provide an introduction to parametric estimation of a linear model with time series observations. Three commonly used estimation methods are least square estimation (LS), maximum likelihood estimation (MLE) and general method of moments (GMM). In this lecture, we will discuss LS and MLE.

## 1 Least Square Estimation

Least square (LS) estimation is one of the first techniques we learn in econometrics. It is both intuitive and easy to implement, and the famous Gauss-Markov theorem tells that under certain assumptions, ordinary least square (OLS) estimator is the best linear unbiased estimator (BLUE). We will start from review of classical LS estimation and then we will consider estimations with relaxed assumptions.

Below are our notations in this lecture and the basic algebra in LS estimation. Consider the regression

$$y_t = x_t'\beta_0 + u_t, \quad t = 1, \ldots, n \tag{1}$$

where $x_t$ is $k$ by 1 vector and $\beta_0$, also a $k$ by 1 vector is the true parameter. Then the OLS estimator of $\beta_0$, denoted by $\hat{\beta}_n$ is

$$\hat{\beta}_n = \left[\sum_{t=1}^{n} x_t x_t'\right]^{-1} \left[\sum_{t=1}^{n} x_t y_t\right] \tag{2}$$

and the OLS sample residual is

$$\hat{u}_t = y_t - x_t'\hat{\beta}_n.$$

Sometimes, it is more convenient to work in matrix form. Define

$$Y_n = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X_n = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} \quad U_n = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Then the regression can be written as

$$Y_n = X_n\beta_0 + U_n, \tag{3}$$

and the OLS estimator can be written as

$$\hat{\beta}_n = (X_n'X_n)^{-1}X_n'Y_n. \tag{4}$$

Define

$$M_X = I_n - X_n(X_n'X_n)^{-1}X_n'.$$

It is easy to see that $M_x$ is symmetric, idempotent ($M_xM_x = M_x$), and orthogonal to the columns of $X$. Then we have

$$\hat{U}_n = Y_n - X_n\hat{\beta}_n = M_XY_n.$$

To derive the distribution of the estimator $\hat{\beta}_n$,

$$\hat{\beta}_n = (X_n'X_n)^{-1}X_n'Y_n = (X_n'X_n)^{-1}X_n'(X_n\beta_0 + U_n) = \beta_0 + (X_n'X_n)^{-1}X_n'U_n. \tag{5}$$

Therefore, the properties of $\hat{\beta}_n$ depends on $(X_n'X_n)^{-1}X_n'U_n$. For example, if $E[(X_n'X_n)^{-1}X_n'U_n] = 0$, then $\hat{\beta}_n$ is unbiased estimator.

### 1.1 Case 1: OLS with deterministic regressors and *i.i.d.* Gaussian errors

**Assumption 1** *(a) $x_t$ is deterministic; (b) $u_t \sim i.i.d(0, \sigma^2)$; (c) $u_t \sim i.i.d.N(0, \sigma^2)$.*

Under assumption 1 (a) and (b), $E(U_n) = 0$ and $E(U_nU_n') = \sigma^2I_n$. Then from (5) we have

$$E(\hat{\beta}_n) = \beta_0 + (X_n'X_n)^{-1}X_n'E(U_n) = \beta_0,$$

and

$$\begin{aligned} E[(\hat{\beta}_n - \beta_0)(\hat{\beta}_n - \beta_0)'] &= E[(X_n'X_n)^{-1}X_n'U_nU_n'X_n(X_n'X_n)^{-1}] \\ &= (X_n'X_n)^{-1}X_n'E(U_nU_n')X_n(X_n'X_n)^{-1} \\ &= \sigma^2(X_n'X_n)^{-1} \end{aligned}$$

Under these assumptions, Gauss-Markov theorem tells that the OLS estimator $\hat{\beta}_n$ is the best linear unbiased estimator for $\beta_0$. The OLS estimator for $\sigma^2$ is

$$s_n^2 = \hat{U}_n\hat{U}_n'/(n-k) = U_n'M_XM_XU_n/(n-k) = U_n'M_XU_n/(n-k). \tag{6}$$

Since $M_X$ is symmetric, there exists a $n$ by $n$ matrix $P$ such that

$$M_X = P\Lambda P' \quad \text{and} \quad P'P = I_n$$

where $\Lambda$ is a $n$ by $n$ matrix with the eigenvalues of $M_X$ along the principal diagonal and zeros elsewhere. From properties of $M_X$ we can compute that $\Lambda$ contains $k$ zeros and $n-k$ ones along its principal diagonal. Then

$$RSS = U_n'M_XU_n = U_nP\Lambda P'U_n = (P'U_n)\Lambda(P'U_n) = W_n'\lambda W_n = \sum_{t=1}^{n}\lambda_t w_t^2$$

where $W_n = P'U_n$. Then $E(W_nW_n') = P'E(U_nU_n')P = \sigma^2 I_n$, therefore, $w_t$ are uncorrelated with mean 0 and variance $\sigma^2$. Therefore,

$$E(U_n'M_XU_n) = \sum_{t=1}^{n} \lambda_t E(w_t^2) = (n-k)\sigma^2.$$

So the $s_n^2$ defined in (6) is unbiased estimator for $\sigma^2$: $E(s_n^2) = \sigma^2$.
With the Gaussian assumption (c), $\hat{\beta}_n$ is also Gaussian,

$$\hat{\beta}_n \sim N(0, \sigma^2(X_n'X_n)^{-1}).$$

Note that here $\hat{\beta}_n$ is 'exact normal', while many of the estimator in our later discussions are 'asymptotically normal'. Actually, under assumption 1, OLS estimator is optimal. Also, with the Gaussian assumption, $w_t$ is $i.i.d.N(0, \sigma^2)$. Therefore we have

$$U_n'M_XU_n/\sigma^2 \sim \chi^2(n-k).$$

### 1.2 Case 2: OLS with stochastic regressors and *i.i.d.* Gaussian errors

The assumption of deterministic regressors is very strong for empirical studies in economics. Some examples of deterministic regressors are constants and deterministic trend (i.e. $x_t = (1, t, t^2, \ldots)$). However, most data we have for econometric regression are stochastic. Therefore from this subsection, we will allow the regressors to be stochastic. However, in case 2 and case 3, we assume that $x_t$ is independent of errors (leads and lags). This is still too strong in time series, as it rules out many processes including ARMA models.

**Assumption 2** *(a) $x_t$ is stochastic and independent of $u_s$ for all $t, s$; (b) $u_t \sim i.i.d.N(0, \sigma^2)$.*

This assumption can be equivalently written as $U_n|X_n \sim N(0, \sigma^2 I_n)$. Under these assumptions, $\hat{\beta}_n$ is still unbiased:

$$E(\hat{\beta}_n) = \beta_0 + E[(X_n'X_n)^{-1}X_n']E(U_n) = \beta_0.$$

Conditional on $X_n$, $\hat{\beta}_n$ is normal, $\hat{\beta}_n|X_n \sim N(\beta_0, \sigma^2(X_n'X_n)^{-1})$. To get the unconditional probability distribution for $\hat{\beta}_n$, we have to integrate this conditional density over $X$. Therefore, the unconditional distribution of $\hat{\beta}_n$ will depend on the distribution of $X$. However, we still have the unconditional distribution for the estimate of the variance $U_n'M_XU_n/\sigma^2 \sim \chi^2(n-k)$.

### 1.3 Case 3: OLS with stochastic regressors and *i.i.d.* Non-Gaussian errors

Compared to case 2, in this section we let the error terms to follow arbitrary *i.i.d.* distribution with finite fourth moments. Since this is an arbitrary unknown distribution, it is very hard obtain exact distribution (finite sample distribution) for $\hat{\beta}_n$, instead, we will apply asymptotic theory in this problem.

**Assumption 3** *(a) $x_t$ is stochastic and independent of $u_s$ for all $t, s$; (b) $u_t \sim i.i.d.(0, \sigma^2)$, and $E(u_t^4) = \mu_4 < \infty$; (c) $E(x_tx_t') = Q_t$, a positive definite matrix with $(1/n)\sum_{t=1}^{n} Q_t \to Q$, a positive definite matrix; (d) $E(x_{it}x_{jt}x_{kt}x_{lt}) < \infty$ for all $i, j, k, l$ and $t$; (e) $(1/n)\sum_{t=1}^{n}(x_tx_t') \to_p Q$.*

With assumption (a), we still have the $\hat{\beta}_n$ is unbiased estimator for $\beta_0$. The assumption (c) to (e) are restrictions on $x_t$. Basically we want to have $(1/n)\sum_{t=1}^n x_t x_t' \to_p (1/n)\sum_{t=1}^n E(x_t x_t')$.

We have

$$\begin{aligned}\hat{\beta}_n - \beta_0 &= \left[\sum_{t=1}^n x_t x_t'\right]^{-1}\left[\sum_{t=1}^n x_t u_t\right] \\ &= \left[(1/n)\sum_{t=1}^n x_t x_t'\right]^{-1}\left[(1/n)\sum_{t=1}^n x_t u_t\right]\end{aligned}$$

From assumptions and continuous mapping theorem, we have

$$\left[(1/n)\sum_{t=1}^n x_t x_t'\right]^{-1} \to_p Q^{-1}.$$

$x_t u_t$ is a martingale difference sequence with finite variance, then by LLN for mixingales, we have

$$\left[(1/n)\sum_{t=1}^n x_t u_t\right] \to_p 0.$$

Therefore, $\hat{\beta}_n \to_p \beta_0$, so $\hat{\beta}_n$ is a consistent estimator. Next, we will derive the distribution for it. This is the first time we derive asymptotic distribution for an OLS estimator. The routines in deriving asymptotically distribution for $\hat{\beta}_n$ are outlined as follows: first we apply LLN on the term $\sum_{t=1}^n x_t x_t'$, after properly normed (so that the limit is a constant); then apply continuous mapping theorem to get the limit for $[\sum_{t=1}^n x_t x_t']^{-1}$. We already got this in the above proof of consistency for $\hat{\beta}_n$. Then we apply CLT on the term $\sum_{t=1}^n x_t u_t$, also after properly normed (so that the limit is nondegenerate).

Note $E(x_t x_t' u_t^2) = \sigma^2 Q_t$ and $(1/n)\sum_{t=1}^n \sigma^2 Q_t \to \sigma^2 Q$. By CLT for mds, we have

$$\left[(1/\sqrt{n})\sum_{t=1}^n x_t u_t\right] \to N(0, \sigma^2 Q).$$

Therefore,

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n - \beta_0) &= \left[(1/n)\sum_{t=1}^n x_t x_t'\right]^{-1}\left[(1/\sqrt{n})\sum_{t=1}^n x_t u_t\right] \\ &\to N(0, [Q^{-1}(\sigma^2 Q)Q^{-1}]) = N(0, \sigma^2 Q^{-1}).\end{aligned}$$

so the $\hat{\beta}_n$ follows

$$\hat{\beta}_n \approx N\left(\beta_0, \frac{\sigma^2 Q^{-1}}{n}\right).$$

Note that this distribution is not exact, but approximate. So we should read it as 'approximately distributed' as normal.

To compute this variance, we need to know $\sigma^2$. When it is unknown, the OLS estimator $s_n^2$ is still consistent under assumption 3. We have

$$\begin{aligned} u_t^2 &= (y_t - x_t'\beta_0)^2 \\ &= [y_t - x_t'\hat{\beta}_n + x_t'(\hat{\beta}_n - \beta_0)]^2 \\ &= (y_t - x_t'\hat{\beta}_n)^2 + 2(y_t - x_t'\hat{\beta}_n)x_t'(\hat{\beta}_n - \beta_0) + [x_t'(\hat{\beta}_n - \beta_0)]^2 \end{aligned}$$

By LLN, we have $(1/n)\sum_{t=1}^n u_t^2 \to \sigma^2$. There are three terms in the above equation. For the second term, we have

$$(1/n)\sum_{t=1}^n (y_t - x_t'\hat{\beta}_n)x_t'(\hat{\beta}_n - \beta_0) = 0$$

as $(y_t - x_t'\hat{\beta}_n)$ is orthogonal to $x_t$. For the third term,

$$(\hat{\beta}_n - \beta_0)' \left[(1/n)\sum_{t=1}^n x_t'x_t\right](\hat{\beta}_n - \beta_0) \to_p 0$$

as $\hat{\beta}_n - \beta_0$ is $o_p(1)$ and $(1/n)\sum_{t=1}^n x_t'x_t \to Q$. Therefore, we can define

$$\hat{\sigma}_n^2 = (1/n)\sum_{t=1}^n (y_t - x_t'\hat{\beta}_n)^2,$$

and we have

$$\hat{\sigma}_n^2 = (1/n)\sum_{t=1}^n (y_t - x_t'\hat{\beta}_n)^2 = (1/n)\sum_{t=1}^n u_t^2 - (1/n)\sum_{t=1}^n [x_t'(\hat{\beta}_n - \beta_0)]^2 \to \sigma^2.$$

This estimator is only slightly different from $\hat{s}_n^2$ ($\hat{\sigma}_n^2 = (n-k)\hat{s}_n^2/n$). Since $(n-k)/n \to 1$ as $n \to \infty$, if $\hat{\sigma}_n^2$ is consistent, so is $s_n^2$.

Next, to derive the distribution of $\hat{\sigma}_n^2$.

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) = (1/\sqrt{n})\sum_{t=1}^n (u_t^2 - \sigma^2) - \sqrt{n}(\hat{\beta}_n - \beta_0)' \left[(1/n)\sum_{t=1}^n x_t'x_t\right](\hat{\beta}_n - \beta_0).$$

The second term goes to zero as $[(1/n)\sum_{t=1}^n x_t'x_t] \to_p Q$ and $\hat{\beta}_n - \beta_0 \to_p 0$. Define $z_t = u_t^2 - \sigma^2$, then $z_t$ is *i.i.d.* with mean zero and variance $E(u_t^4) - \sigma^4 = \mu_4 - \sigma^4$. Applying CLT, we have

$$(1/\sqrt{n})\sum_{t=1}^n z_t \to_d N(0, \mu_4 - \sigma^4),$$

therefore,

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \to_d N(0, \mu_4 - \sigma^4).$$

The same limit distribution applies for $s_n^2$, since the difference between $\hat{\sigma}_n^2$ and $s_n^2$ is $o_p(n^{-1/2})$.

### 1.4 Case 4: OLS estimation in autoregression with *i.i.d.* error

In an autoregression, say, $x_t = \phi_0 x_{t-1} + \epsilon_t$, where $\epsilon_t$ is *i.i.d.*, the regressors are no longer independent of $\epsilon_t$. In this case, the OLS estimator of $\phi_0$ is biased. However, we will show that under assumption 4, the estimator is consistent.

**Assumption 4** *The regression model is*

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t,$$

*with roots of* $(1 - \phi_1 z - \phi_2 z^2 - ... - \phi_p z^p) = 0$ *outside the unit circle (so* $y_t$ *is stationary) and with* $\epsilon_t$ *i.i.d. with mean zero , variance* $\sigma^2$*, and finite fourth moments* $\mu_4$*.*

Page 215-216 in Hamilton presents the general AR($p$) case with constant. We will use AR(2) as an example, $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$. Let $x_t' = (y_{t-1}, y_{t-2})$, $u_t = \epsilon_t$ and $y_t = x_t'\beta_0 + u_t$ (so $\beta_0' = (\phi_1, \phi_2)$).

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \left[(1/n)\sum_{t=1}^{n} x_t x_t'\right]^{-1} \left[(1/\sqrt{n})\sum_{t=1}^{n} x_t u_t\right] \tag{7}$$

The first term

$$(1/n)\sum_{t=1}^{n} x_t x_t' = (1/n)\begin{bmatrix} \sum_{t=1}^{n} y_{t-1}^2 & \sum_{t=1}^{n} y_{t-1}y_{t-2} \\ \sum_{t=1}^{n} y_{t-1}y_{t-2} & \sum_{t=1}^{n} y_{t-2}^2 \end{bmatrix}$$

In this matrix, first, on the diagonal, $n^{-1}\sum_{t=1}^{n} y_{t-j}^2$ converge to $\gamma_0$. The remaining term $n^{-1}\sum_{t=1}^{n} y_{t-1}y_{t-2}$ converges to $\gamma_1$. Therefore,

$$(1/n)\sum_{t=1}^{n} x_t x_t' \to_p Q = \begin{bmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 \end{bmatrix}.$$

Apply CLT for mds on the second term in (7),

$$\left[(1/\sqrt{n})\sum_{t=1}^{n} x_t u_t\right] \to_d N(0, \sigma^2 Q),$$

therefore,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \to_d N(0, \sigma^2 Q^{-1}).$$

So far we have considered four cases in OLS regressions. The common assumption in all those four cases are *i.i.d.* errors. From next section, we will consider cases where the errors are not *i.i.d.*.

### 1.5 OLS with non-*i.i.d.* errors

When the error $u_t$ is *i.i.d.*, then the variance-covariance matrix $V = E(U_n U_n') = \sigma^2 I_n$. If $V$ is still diagonal but the elements are not equal, for example, the errors on some dates display larger variance and the errors on some dates display smaller variance, then the errors are said to exhibit *heteroskedasticity*. If $V$ is non-diagonal, then the errors are said to be *autocorrelated*. For example, let $u_t = \epsilon_t - \phi\epsilon_{t-1}$ where $\epsilon_t$ is *i.i.d.*, then $u_t$ is serially correlated errors.

Case 5 in Hamilton assumes

**Assumption 5** *(a) $x_t$ is stochastic; (b) conditional on the full matrix $X$, the vector $U \sim N(0, \sigma^2 V)$; (c) $V$ is a known positive matrix.*

Under these assumptions, the exact distribution of $\hat{\beta}_n$ can be derived. However, this is a very strong assumption and it rules out the autoregressive regression. Also, the assumption that $V$ is known rarely holds in applications.

Case 6 in Hamilton assumes uncorrelated but heteroskedastic errors with unknown covariance matrix. Under assumption 6, the OLS estimator is still consistent and asymptotically normal.

**Assumption 6** *(a) $x_t$ stochastic, including perhaps lagged values of $y$; (b) $x_t u_t$ is martingale difference sequence; (c) $E(u_t^2 x_t x_t') = \Omega_t$, a positive definite matrix, with $(1/n)\sum_{t=1}^n \Omega_t \to_p \Omega$ and $(1/n)\sum_{t=1}^n u_t^2 x_t x_t' \to_p \Omega$; (d) $E(u_t^4 x_{it} x_{jt} x_{lt} x_{kt}) < \infty$ for all $i, j, k, l$ and $t$; (e) plims of $(1/n)\sum_{t=1}^n u_t x_{it} x_t x_t'$ and $(1/n)\sum_{t=1}^n x_{it} x_{jt} x_t x_t'$ exist and are finite for all $i$, $j$ and $(1/n)\sum_{t=1}^n x_t' x_t \to_p Q$, a nonsingular matrix.*

Again, write the OLS estimator as

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \left[(1/n)\sum_{t=1}^n x_t x_t'\right]^{-1} \left[(1/\sqrt{n})\sum_{t=1}^n x_t u_t\right]$$

Assumption 6 (e) ensures that

$$\left[(1/n)\sum_{t=1}^n x_t x_t'\right]^{-1} \to_p Q^{-1}.$$

Apply CLT for mds,

$$\left[(1/\sqrt{n})\sum_{t=1}^n x_t u_t\right] \to N(0, \Omega),$$

therefore,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \to N(0, Q^{-1}\Omega Q^{-1}).$$

However, both $Q$ and $\Omega$ are not observable and we need to find consistent estimates for them. White proposes the following estimator $\hat{Q}_n = (1/n)\sum_{t=1}^n x_t x_t'$ and $\hat{\Omega}_n = (1/n)\sum_{t=1}^n \hat{u}_t^2 x_t x_t'$ where $\hat{u}_t$ is the OLS residual $y_t - x_t'\hat{\beta}_n$.

**Proposition 1** *With heteroskedasticity of unknown form satisfying assumption 6, the asymptotic variance-covariance matrix of the OLS coefficient vector can be consistently estimated by*

$$\hat{Q}_n^{-1}\hat{\Omega}_n\hat{Q}_n^{-1} \to_p Q^{-1}\Omega Q^{-1} \tag{8}$$

Proof: Assumption 6 (e) ensures $\hat{Q} \to Q$ and assumption 6 (c) ensures that

$$\tilde{\Omega}_n \equiv (1/n)\sum_{t=1}^n u_t^2 x_t x_t' \to_p \Omega.$$

So to prove (8), we only need to show that

$$\hat{\Omega}_n - \tilde{\Omega}_n = (1/n)\sum_{t=1}^n (\hat{u}_t^2 - u_t^2) x_t x_t' \to 0.$$

The trick here is to make use of a known fact that $\hat{\beta}_n - \beta_0 \to_p 0$. If we could write $\hat{\Omega}_n - \tilde{\Omega}_n$ as sums of some products of $\hat{\beta}_n - \beta_0$ and terms that are bounded, then $\hat{\Omega}_n - \tilde{\Omega}_n \to_p 0$.

$$\begin{aligned}\hat{u}_t^2 - u_t^2 &= (\hat{u}_t + u_t)(\hat{u}_t - u_t) \\ &= [2(y_t - \beta_0' x_t) - (\hat{\beta}_n - \beta_0)' x_t][-(\hat{\beta}_n - \beta_0)' x_t] \\ &= -2u_t(\hat{\beta}_n - \beta_0)' x_t + [(\hat{\beta}_n - \beta_0)' x_t]^2\end{aligned}$$

Then

$$\hat{\Omega}_n - \tilde{\Omega}_n = (-2/n)\sum_{t=1}^{n} u_t(\hat{\beta}_n - \beta_0)' x_t(x_t x_t') + (1/n)\sum_{t=1}^{n}[(\hat{\beta}_n - \beta_0)' x_t]^2(x_t x_t').$$

Write the first term

$$(-2/n)\sum_{t=1}^{n} u_t(\hat{\beta}_n - \beta_0)' x_t(x_t x_t') = -2\sum_{i=1}^{k}(\hat{\beta}_{in} - \beta_{i0})\left[(1/n)\sum_{t=1}^{n} u_t x_{it}(x_t x_t')\right].$$

The term in the bracket has a finite plim by assumption 6 (e) and we have $\hat{\beta}_{in} - \beta_{i0} \to 0$ for each $i$. Then this term converges to zero. (if this looks messy, take $k = 1$, then you can simply move $(\hat{\beta}_n - \beta_0)$ out of the summation. $\hat{\beta}_n - \beta_0 \to_p 0$ and the sum has a finite limit, so the product goes to zero).

Similarly for the second term,

$$(1/n)\sum_{t=1}^{n}[(\hat{\beta}_n - \beta_0)' x_t]^2(x_t x_t') = \sum_{i=1}^{k}\sum_{j=1}^{k}(\hat{\beta}_{in} - \beta_{i0})(\hat{\beta}_{jn} - \beta_{j0})\left[(1/n)\sum_{t=1}^{n} x_{it}x_{jt}(x_t x_t')\right] \to_p 0$$

as the term in bracket has a finite plim. Therefore, $\hat{\Omega}_n - \tilde{\Omega}_n \to 0$.

Define $\hat{V}_n = \hat{Q}_n^{-1}\hat{\Omega}_n\hat{Q}_n^{-1}$, then

$$\hat{\beta}_n \approx N(\beta_0, \hat{V}_n/n),$$

and $V_n/n$ is a heteroskedastic-consistent estimates for the variance-covariance matrix. Newey-West proposes the following estimator for the variance-covariance matrix which is heteroskedastic and autocorrelation consistent (HAC).

$$\hat{V}_n/n = (X_n'X_n)^{-1}\left[\sum_{t=1}^{n}\hat{u}_t^2 x_t x_t' + \sum_{k=1}^{q}\left(1 - \frac{k}{q+1}\right)\sum_{t=k+1}^{n}(x_t\hat{u}_t\hat{u}_{t-k}x_{t-k}' + x_{t-k}\hat{u}_{t-k}\hat{u}_t x_t')\right](X_n'X_n)^{-1}.$$

### 1.6 General least square

General least square (GLS) and feasible general least square (FGLS) is preferred in least square estimation when the errors are heteroskedastic or/and autocorrelated.

Let $x_t$ be stochastic and $U|X \sim N(0, \sigma^2 V)$ where $V$ is known (assumption 5). Since $V$ is symmetric and positive definite, there exists matrix $L$ such that $V^{-1} = L'L$. Premultiply $L$ to our regression and get

$$LY = LX\beta_0 + LU.$$

Then the new error $\tilde{U} = LU$ is *i.i.d.* conditional on $X$,

$$E(\tilde{U}\tilde{U}'|X) = LE(UU'|X)L' = \sigma^2 LVL' = \sigma^2 I_n.$$

Then the estimator

$$\tilde{\beta}_n = (X'L'LX)^{-1}X'L'Ly = (X'V^{-1}X)^{-1}X'V^{-1}y$$

is known as the general least square estimator.

However, as we remarked earlier, in applications, $V$ is rarely known and we have estimate it. The GLS estimator obtained using estimated $V$ is known as feasible GLS estimator. Usually, FGLS require that we specify a parametric model for the error. For example, let the error $u_t$ follow an AR(1) process, $u_t = \rho_0 u_{t-1} + \epsilon_t$ where $\epsilon_t \sim i.i.d.(0, \sigma^2)$. In this case, we can run OLS first and obtain the OLS residual $\hat{u}_t$. Then run OLS estimation for $\rho$ using the $\hat{u}_t$. This estimator, denoted by $\hat{\rho}_n$, is consistent estimator for $\rho$. To show this, write

$$\hat{u}_t = (y_t - \beta_0 x_t + \beta_0 x_t - \hat{\beta}_n x_t) = u_t + (\beta_0 - \hat{\beta}_n)' x_t.$$

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n} \hat{u}_t \hat{u}_{t-1} &= \frac{1}{n}\sum_{t=1}^{n} [u_t + (\beta_0 - \hat{\beta}_n)' x_t][u_{t-1} + (\beta_0 - \hat{\beta}_n)' x_{t-1}] \\
&= \frac{1}{n}\sum_{t=1}^{n} u_t u_{t-1} + (\beta_0 - \hat{\beta}_n)' \frac{1}{n}\sum_{t=1}^{n}(u_t x_{t-1} + u_{t-1} x_t) + (\beta_0 - \hat{\beta}_n)' \left[\frac{1}{n}\sum_{t=1}^{n} x_t x_{t-1}'\right](\beta_0 - \hat{\beta}) \\
&= \frac{1}{n}\sum_{t=1}^{n} u_t u_{t-1} + o_p(1) \\
&= \frac{1}{n}\sum_{t=1}^{n} (\epsilon_t + \rho u_{t-1}) u_{t-1} \\
&\to \rho var(u_t).
\end{aligned}$$

Similarly, we can show that $\frac{1}{n}\sum_{t=1}^{n} \hat{u}_t \hat{u}_t \to_p var(u_t)$, hence $\hat{\rho}_n \to \rho_0$. Still use similar method, we can show that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \hat{u}_t \hat{u}_{t-1} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} u_t u_{t-1} + o_p(1).$$

Hence

$$\sqrt{n}(\hat{\rho} - \rho_0) \to N(0, (1 - \rho_0^2)).$$

Finally the FGLS estimator for $\beta_0$ based on $V(\hat{\rho})$ has the same limit distribution as the GLS estimator based on $V(\rho_0)$ (page 222-225 in Hamilton).

### 1.7 Statistical inference with LS estimation

Some commonly used test statistics for LS estimator are $t$ statistics and $F$ statistics. $t$ statistics is used to test the hypothesis of a single parameter, say $\beta_i = c$. For simplicity, we assume that $c = 0$, so we use $t$ statistics to test if a variable is significant. The $t$ statistics is defined as the ratio

$\hat{\beta}_i/sd(\beta_i)$. Let the estimate of the variance of $\hat{\beta}$ be denoted by $s^2\hat{W}$, then the standard deviation of $\hat{\beta}_i$ is the product of $s$ and the square root of the $i$th element on the diagonal, i.e.,

$$t = \frac{\hat{\beta}_i}{\sqrt{s^2}\sqrt{w_{ii}}}. \tag{9}$$

Recall that if $X/\sigma \sim N(0,1)$, and $Y^2/\sigma^2 \sim \chi^2(m)$, and let $X$ and $Y$ be independent, then

$$t = \frac{X\sqrt{m}}{Y}$$

follows exact student $t$ distribution with $m$ degree of freedom.

$F$-statistics is used to test the hypothesis of $m$ different linear restrictions about $\beta$, say

$$H_0 : R\beta = r,$$

where $R$ is a $m$ by $k$ matrix. The $F$ statistics is then defined as

$$\bar{F} = (R\hat{\beta} - r)'[Var(R\hat{\beta} - r)]^{-1}(R\hat{\beta} - r). \tag{10}$$

This is a Wald statistics. To derive the distribution of the statistics, we will need the following result

**Proposition 2** *If a $k$ by 1 vector $X \sim N(\mu, \Sigma)$, then $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2(k)$.*

Also recall that an exact $F(m,n)$ distribution is defined to be

$$F(m,n) = \frac{\chi^2(m)/m}{\chi^2(n)/n}.$$

With assumption 1 $\hat{W} = (X_n'X_n)^{-1}$, and under the hull hypothesis $\hat{\beta}_i \sim N(0, \sigma^2 w_{ii})$. We can then write

$$t = \frac{\frac{\hat{\beta}_i}{\sqrt{\sigma^2 w_{ii}}}}{\sqrt{\frac{s^2}{\sigma^2}}}.$$

Since the numerator is $N(0,1)$ and the denominator is the square root of $\chi^2(n-k)$ divided by $n-k$ (since $RSS/\sigma^2 \sim \chi^2(n-k)$), and the numerator and denominator are independent, so $t$ statistics (9) under assumption 1 follows exact $t$ distribution.

With assumption 1 and under the null hypothesis, we have

$$R\hat{\beta} - r \sim N(0, \sigma^2 R(X_n'X_n)^{-1}R),$$

then by proposition 2, the $\bar{F}$ statistics defined in (10) under hypothesis $H_0$

$$(R\hat{\beta} - r)'[\sigma^2 R(X_n'X_n)^{-1}R]^{-1}(R\hat{\beta} - r) \sim \chi^2(m).$$

If we replace $\sigma^2$ with $s^2$, and divide it by the number of restrictions $m$, we get the OLS $F$ test of a linear hypothesis

$$\begin{aligned} F &= (R\hat{\beta} - r)'[s^2 R(X_n'X_n)^{-1}R]^{-1}(R\hat{\beta} - r)/m \\ &= \frac{\bar{F}/m}{(RSS/\sigma^2)/(n-k)}, \end{aligned}$$

so $F$ follows a exact $F(m, n-k)$ distribution.

An alternative way to express the $F$ statistics is to compute the estimator without restriction $\hat{\beta}$ and its associated sum of residual $RSS_u$; and the estimator with restriction $\tilde{\beta}$ and its associated sum of residual $RSS_r$, then we can write

$$F = \frac{(RSS_r - RSS_u)/m}{RSS_u/(n-k)}.$$

Now, with assumption 2, $X$ is stochastic and $\hat{\beta}$ is normal conditional on $X$ and $RSS \sim \sigma^2\chi^2(n-k)$ conditional on $X$. This conditional distribution of $RSS$ is the same for all $X$, therefore, the unconditional distribution of $RSS$ is the same as the conditional distribution. The same is true for the $t$ and $F$ statistics. Therefore we have the same results under assumption 2 as that under assumption 1.

From case 3, we no longer have exact distribution for the estimator, and we have to derive the asymptotic distribution for the estimator, so we also use the asymptotic distributions for the test statistics.

$$t_n = \frac{\hat{\beta}_i}{s_n\sqrt{w_{ii}}} = \frac{\sqrt{n}\hat{\beta}_i}{s_n\sqrt{nw_{ii}}}.$$

where $w_{ii}$ is the $i$th element on the diagonal of $\hat{\beta}$'s asymptotic variance $Q^{-1}n^{-1}$. If we let the $i$th element on the diagonal of $Q$ denoted by $q_{ii}$, then we have $\hat{\beta}_i \to_d N(0, \sigma^2 q_{ii})$. Recall that under assumption 3, $s_n \to \sigma$, there we have

$$t_n \to N(0,1).$$

Next, write

$$\begin{aligned} F_n &= (R\hat{\beta} - r)'[s_n^2 R(X_n'X_n)^{-1}R]^{-1}(R\hat{\beta} - r)/m \\ &= \sqrt{n}(R\hat{\beta} - r)'[s_n^2 R(X_n'X_n/n)^{-1}R]^{-1}\sqrt{n}(R\hat{\beta} - r)/m \end{aligned}$$

Now we have $s_n^2 \to_p \sigma^2$, $X_n'X_n/n \to Q$, and under the null,

$$\sqrt{n}(R\hat{\beta} - r) = R\sqrt{n}(\hat{\beta} - \beta_0) \to_d N(0, \sigma^2 RQ^{-1}R').$$

Then by proposition 2, we have

$$mF_n \to \chi^2(m).$$

We can then use similar methods to derive the distribution for other cases. In general if $\hat{\beta} \to_p \beta_0$ and asymptotically normal, $s_n^2 \to \sigma^2$, and we have found a consistent estimate for the variance of $\hat{\beta}$, then the $t$ and $F$ statistics follow asymptotically normal and $\chi^2(m)$ distribution. Actually, under assumption 1 or 2, when the sample size is large, we can also use normal and $\chi^2$ distribution to approximate the exact $t$ and $F$ distribution. Further, since we are using the asymptotic distribution, the Wald test can also be used to test nonlinear restrictions.

# 2 Maximum Likelihood Estimation

## 2.1 Review: maximum likelihood principle and Cramer-Rao lower bound

The basic idea of maximum likelihood principle is to choose the parameter estimates that maximizes the probability of obtaining the observed sample. Consider that we observe a sample $X_n = (x_1, x_2, \ldots, x_n)$ and assume that the sample is drawn from an *i.i.d.* distribution and the associated parameters are denoted by $\theta$. Let $p(x_t; \theta)$ denote the pdf of the $t$th observation. For example, when $x_t \sim i.i.d.N(\mu, \sigma^2)$, then $\theta = (\mu, \sigma^2)$ and

$$p(x_t; \theta) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(x_t - \mu)^2}{2\sigma^2}\right].$$

The likelihood function for the whole sample $X_n$ is

$$L(X_n; \theta) = \prod_{t=1}^{n} p(x_t; \theta)$$

and the log likelihood function is

$$l(X_n; \theta) = \sum_{t=1}^{n} \log p(x_t; \theta).$$

The maximum likelihood estimates for $\theta$ are chosen so that $l(X_n; \theta)$ is maximized. Define the score function $S(\theta) = \partial l(\theta)/\partial\theta$, and the Hessian matrix $H(\theta) = \partial^2 l(\beta)/\partial\theta\partial\theta'$, then the famous Cramer-Rao inequality tells that the lowest bound for the variance of an unbiased estimator of $\theta$ is the inverse of the information matrix $I(\theta_0) = E[S(\theta_0)S(\theta_0)']$, where $\theta_0$ denotes the true value of the parameter. An estimator that have a variance equal to this bound is known as *efficient*. Under some regularity condition which are satisfied for the Gaussian density, we have the following equality

$$I(\theta) = -E[H(\theta)] = -E\left[\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\right].$$

So, if we find an unbiased estimator and its variance achieves the Cramer-Rao lower bound, then we know that this estimator is efficient and there is no other unbiased estimator (linear or nonlinear) that could have smaller variance than this estimator. However, this lower bound is not always achievable. If an estimator does achieve this bound, then this estimator is identical to MLE. Note that Cramer-Rao inequality holds for unbiased estimator while sometimes ML estimators are biased. If the estimator is biased but consistent, and its variance approaches the Cramer-Rao bound asymptotically, then this estimator is known as asymptotically efficient.

**Example 1** (MLE estimation for *i.i.d.* Gaussian distribution) Let $x_t \sim i.i.d.N(\mu, \sigma^2)$, so the parameter $\theta = (\mu, \sigma^2)$. Then we have

$$\begin{aligned} p(x_t; \theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_t - \mu)^2}{2\sigma^2}\right\} \\ l(X_n; \theta) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(x_t - \mu)^2 \end{aligned}$$

$$
\begin{aligned}
S(X_n;\mu) &= \frac{\partial l(X_n;\theta)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{t=1}^{n}(x_t-\mu)^2 \\
S(X_n;\sigma^2) &= \frac{\partial l(X_n;\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{n}(x_t-\mu)^2
\end{aligned}
$$

Set the score functions to zero, we found the MLE estimator for $\theta$ are $\hat{\mu} = \bar{X}_n$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \hat{\mu})^2$. It is easy to verify that $E(\hat{\mu}) = E(\bar{X}_n) = \mu$, so $\hat{\mu}$ is unbiased and its variance $Var(\hat{\mu}) = \sigma^2/n$, while

$$
\begin{aligned}
E\hat{\sigma}^2 &= E\left[\frac{1}{n}\sum_{t=1}^{n}(x_t-\hat{\mu})^2\right] \\
&= E(x_t-\hat{\mu})^2 \\
&= E[(x_t-\mu)+(\mu-\hat{\mu})]^2 \\
&= \sigma^2 - \frac{2}{n}\sigma^2 + \frac{1}{n}\sigma^2 \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}
$$

so $\hat{\sigma}^2$ is biased, but it is consistent as $\hat{\sigma}^2 \to \sigma^2$ as $n \to \infty$. Define $s^2 = \frac{1}{n-1}\sum_{t=1}^{n}(x_t-\hat{\mu})^2$, then $Es^2 = \sigma^2$, and $Var(s^2) = 2\sigma^4/(n-1)$.

We can further compute the Hessian matrix,

$$
H(X_n;\theta) = \begin{bmatrix} \frac{\partial^2 l(X_n;\theta)}{\partial^2 \mu} & \frac{\partial^2 l(X_n;\theta)}{\partial \mu \sigma^2} \\ \frac{\partial^2 l(X_n;\theta)}{\partial \sigma^2 \mu} & \frac{\partial^2 l(X_n;\theta)}{\partial^2 \sigma^2} \end{bmatrix}
$$

where

$$
\begin{aligned}
\frac{\partial l(X_n;\theta)}{\partial^2 \mu} &= -\frac{n}{\sigma^2} \\
\frac{\partial l(X_n;\theta)}{\partial \mu \sigma^2} &= \frac{\partial l(X_n;\theta)}{\partial \sigma^2 \mu} = -\frac{1}{\sigma^4}\sum_{t=1}^{n}(x_t-\mu) \\
\frac{\partial l(X_n;\theta)}{\partial^2 \sigma^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{t=1}^{n}(x_t-\mu)^2
\end{aligned}
$$

We can also compute that

$$
|H(X_n;\theta)|_{\theta=\hat{\theta}} = \frac{n^2}{2\sigma^6} > 0,
$$

so we know that the we have found the maximum (not minimum) of the likelihood function. Next, compute the information matrix,

$$
E_\theta\left[\sum_{t=1}^{n}(x_t-\mu)\right] = 0, \quad E_\theta\left[\sum_{t=1}^{n}(x_t-\mu)^2\right] = n\sigma^2.
$$

therefore the information matrix

$$I(\theta) = E[-H(X_n;\theta)] = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

So the MLE of $\mu$ has achieved the Cramer-Rao lower bound of variance $\frac{\sigma^2}{n}$. Although $s^2$ does not achieve to the lower bound, it turns out it is still the unbiased estimator for $\sigma^2$ with minimum variance.

## 2.2 Asymptotic Normality of MLE

There are a few regularity conditions to ensure that the MLE is consistent. First we assume that the data is strictly stationary and ergodic (for example, *i.i.d.*). Second, we assume that the parameter space $\Theta$ is convex and neither the estimate $\hat{\theta}$ nor the true parameter $\theta_0$ lie on the boundary of $\Theta$. Third, we require that the likelihood function evaluated at $\hat{\theta}$ is different from $\theta_0$, for any $\hat{\theta} \neq \theta_0$ in $\Theta$. This is known as the identification condition. Finally, we assume that $E[\sup_{\theta\in\Theta} |l(X_n;\theta)|] < \infty$. With all those conditions satisfied, the MLE is consistent $\hat{\theta} \to_p \theta_0$.

Next we will discuss the asymptotic results on the score function $S(X_n;\theta)$, the Hessian matrix $H(X_n;\theta)$ and the asymptotic distribution of the MLE estimates $\hat{\theta}$.

First, we want to show that $E[S(X_n,\theta_0)] = 0$ and $E[S(X_n,\theta_0)S(X_n,\theta_0)'] = -E[(H(X_n;\theta_0)]$. Let the integral operator denote integrate over $X_1, X_2, \ldots, X_n$, then we have that

$$\int L(X_n,\theta_0)dX_n = 1.$$

Taking derivative with respect to $\theta$, then we have

$$\int \frac{\partial L(X_n,\theta_0)}{\partial\theta} dX_n = 0.$$

While, we can write

$$\begin{aligned} & \int \frac{\partial L(X_n,\theta_0)}{\partial\theta} dX_n \\ = & \int \frac{1}{L(X_n,\theta_0)} \frac{\partial L(X_n,\theta_0)}{\partial\theta} L(X_n,\theta_0) dX_n \\ = & \int \frac{\partial l(X_n;\theta_0)}{\partial\theta} L(X_n,\theta_0) dX_n \\ = & E[S(X_n,\theta_0)] \end{aligned}$$

So we know that $E[S(X_n,\theta_0)] = 0$. Next, let the integral (which equal to zero) take $\theta'$, it is

$$\int \frac{\partial l(X_n;\theta_0)}{\partial\theta} \frac{\partial L(X_n,\theta_0)}{\partial\theta'} dX_n + \int \frac{\partial^2 l(X_n;\theta_0)}{\partial\theta\partial\theta'} L(X_n,\theta_0) dX_n = 0.$$

The second term is just $E[H(X_n;\theta_0)]$. The first can be written as

$$\begin{aligned} & \int \frac{\partial l(X_n;\theta_0)}{\partial\theta} \left( \frac{1}{L(X_n,\theta_0)} \frac{\partial L(X_n,\theta_0)}{\partial\theta'} \right) L(X_n,\theta_0) dX_n \\ = & \int \frac{\partial l(X_n;\theta_0)}{\partial\theta} \frac{\partial l(X_n;\theta_0)}{\partial\theta'} L(X_n,\theta_0) dX_n \\ = & E[S(X_n,\theta_0)S(X_n,\theta_0)'] \end{aligned}$$

Now, since $E[S(X_n,\theta_0)S(X_n,\theta_0)'] + E[H(X_n;\theta_0)] = 0$, we have that $E[S(X_n,\theta_0)S(X_n,\theta_0)'] = -E[H(X_n;\theta_0)]$.

Next, define that $s(x_t;\theta) = \frac{\partial \log p(x_t;\theta)}{\partial\theta}$, then we write the score function as the sum of $s(x_t;\theta)$, i.e., $S(X_n,\theta) = \sum_{t=1}^n s(x_t;\theta)$. $s(x_t;\theta)$ is *i.i.d.* and we can show that $E[s(x_t;\theta_0)] = 0$ and $E[s(x_t;\theta_0)s(x_t;\theta_0)'] = E[H(x_t;\theta_0)]$. Applying Lindeberg-Levy CLT, we obtain the asymptotic normality of the score function

$$n^{-1/2}S(X_n;\theta_0) \to_d N(0, -\frac{1}{n}E[H(X_n;\theta_0)]).$$

Next, we consider the properties of the Hessian matrix. First we assume that $E[H(X_n;\theta_0)]$ is non-singular. Let $N_\epsilon$ be a neighborhood of $\theta_0$, and

$$E[\sup_{\theta\in N_\epsilon} \|H(X_n;\theta)\|] < \infty,$$

then we have

$$\frac{1}{n}\sum_{t=1}^n H(x_t;\tilde{\theta}) \to E[H(X_n;\theta_0)] \equiv V,$$

where $\tilde{\theta}$ is any consistent estimator for $\theta_0$.

Apply the LLN, we have

$$\frac{1}{n}H(X_n;\theta_0) = \frac{1}{n}\sum_{t=1}^n H(x_t;\theta_0) \to_p E(x_t;\theta_0) = E\left[\frac{1}{n}H(X_n;\theta_0)\right] \equiv -\Sigma.$$

With the notation $\Sigma$, we can write $n^{-1/2}S(X_n;\theta_0) \to_d N(0,\Sigma)$.

**Proposition 3** *(*Asymptotic normality of MLE*) With all the conditions we have outlined above,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, \Sigma^{-1}).$$

Proof: Do a Taylor expansion of $S(X_n;\hat{\theta})$ around $\theta_0$,

$$0 = S(X_n;\hat{\theta}) \approx S(X_n;\theta_0) + (\hat{\theta} - \theta_0)H(X_n;\theta_0).$$

Therefore, we have

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta_0) &= -\sqrt{n}S(X_n;\theta_0)H(X_n;\theta_0) \\
&= \left(\frac{1}{\sqrt{n}}S(X_n;\theta_0)\right)\left(-\frac{1}{n}H(X_n;\theta_0)\right)^{-1} \\
&\to N(0, \Sigma^{-1}\Sigma\Sigma^{-1}) \\
&= N(0,\Sigma^{-1})
\end{aligned}$$

Note that $\Sigma^{-1} = E[\frac{1}{n}H(X_n;\theta_0)]^{-1} = nI(\theta_0)^{-1}$, so the asymptotic distribution of $\hat{\theta}$ can be written as

$$\hat{\theta} \approx N(\theta_0, I(\theta_0)^{-1}).$$

However, $I(\theta_0)$ depends on $\theta_0$ which is unknown. So we need to find a consistent estimator for it, denoted by $\hat{V}$. There are two methods to compute this variance matrix of $\hat{\theta}$. One way is that

we compute the Hessian matrix, and evaluate it at $\theta = \hat{\theta}$, i.e. $\hat{V} = H(X_n; \hat{\theta})$. The second way is to use the outer product estimate, which is

$$\hat{V} = \sum_{t=1}^{n} [S(x_t; \hat{\theta}) S(x_t; \hat{\theta})'].$$

### 2.3 Statistical Inference for MLE

There are three asymptotically equivalent tests for MLE: likelihood ratio (LR) test, Wald test, and Lagrange multiplier (LM) test or score test. You can probably find discussion on these three tests on any graduate text book in econometrics, so we only describe them briefly here.

The likelihood ratio test is based on the difference between the likelihood you computed (maximized) with or without the restriction. Let $l_u$ denote the likelihood without restriction and $l_r$ denote the likelihood with restriction (note that $l_r \leq l_u$). If the restriction is valid, then we expect the $l_r$ should not be too much lower than $l_u$. Therefore, to test if the restriction is valid, the statistics we compute is $2(l_u - l_r)$ which follows a $\chi^2$ distribution with degree of freedom equal to the number of restrictions imposed.

To do LR test, we have to compute the likelihood under both restricted and unrestricted condition. In comparison, the other two tests only use either the estimator without restriction (denoted by $\hat{\theta}$) or the estimator with restriction (denoted by $\tilde{\theta}$).

Let the restriction be $H_0 : R(\theta) = r$, the idea of Wald test is that: if this restriction is valid, then the estimator obtained without restriction $\hat{\theta}$ will make $R(\hat{\theta}) - r$ close to zero. Therefore the Wald statistics is

$$W = (R(\hat{\theta}) - r)'[Var(R(\hat{\theta}) - r)]^{-1}(R(\hat{\theta}) - r),$$

which also follows a $\chi^2$ distribution with degree of freedom equal to the number of restrictions imposed.

To find the ML estimator, we set the score function equal to zero and solve for the estimator, i.e., $S(\hat{\theta}) = 0$. If the restriction is valid, and the estimator we obtained with the restriction is $\tilde{\theta}$, then we expect that $S(\tilde{\theta})$ is close to zero. This idea leads to the LM test or score test. The LM statistics is

$$LM = S(\tilde{\theta})' I(\tilde{\theta})^{-1} S(\tilde{\theta}),$$

which also follows a $\chi^2$ distribution with degree of freedom equal to the number of restrictions imposed.

### 2.4 LS and MLE

In a regression $Y_n = X_n \beta_0 + U_n$ where $U_n | X_n \sim N(0, \sigma^2 I_n)$ (as in assumption 2), the conditional density of $Y$ given $X$ is

$$f(Y|X; \theta) = (2\pi\sigma^2)^{-n/2} \exp\left[ -\frac{1}{2\sigma^2} (Y - X\beta)'(Y - X\beta) \right].$$

The log likelihood function is

$$l(Y|X; \theta) = -\frac{n}{2}\log(2\pi) - -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(X - X\beta)'(X - X\beta)$$

Note that $\hat{\beta}_n$ that maximizes $l$ is the vector that minimizes the sum of squares, therefore, under the assumption 2, the OLS estimator is equivalent to ML estimator for $\hat{\beta}_0$. It can be shown that this estimator is unbiased and achieves the Cramer-Rao lower bound, therefore under assumption 2, the OLS/MLE estimator are efficient (compared to all unbiased linear or nonlinear estimators). Recall that under assumption 1, we have Gauss-Markov theorem to show that OLS estimator is the best linear unbiased estimator. Now, the Cramer-Rao inequality tells the optimality of OLS estimator under assumption2. The ML estimator for $\sigma^2$ is $(Y - X\beta)'(Y - X\beta)/n$. We have introduced this estimator a moment ago and we showed that the difference between $\hat{\sigma}_n^2$ and the OLS estimator $s_n^2$ becomes arbitrarily small as $n \to \infty$.

Next, consider assumption 5, where $U|X \sim N(0, \sigma^2 V)$ and $V$ is known. Then the log likelihood function omitting constant term is

$$l(Y|X, \beta) = -(1/2)\log V - (1/2)(Y - X\beta)'V^{-1}(Y - X\beta).$$

The MLE estimator is

$$\hat{\beta}_n = (X'V^{-1}X)^{-1}X'Y,$$

which is equivalent to the GLS estimator. The score vector is $S_n(\beta) = (Y - X\beta)'V^{-1}X$, the Hessian matrix $H_n(\beta) = X'V^{-1}X$. Therefore, the information matrix is $I(\beta) = X'V^{-1}X$. Therefore, the GLS/MLE estimator is efficient as it achieves the Cramer-Rao lower bound $(X'V^{-1}X)^{-1}$.

When $V$ is unknown, we can parameterize it $V(\psi)$, say, and maximizes the likelihood

$$l(Y|X, \beta, \psi) = -(1/2)\log V(\psi) - (1/2)(Y - X\beta)'V^{-1}(\psi)(Y - X\beta).$$

### 2.5 Example: MLE in autoregressive estimation

In Hamilton's book, you can find many detailed discussions about MLE estimation for an ARMA model in Chapter 5. We will take an AR(1) model as example.

Consider an AR(1) model,

$$x_t = c + \beta x_{t-1} + u_t$$

where $u_t \sim i.i.d.N(0, \sigma^2)$. Let $\theta = (c, \beta, \sigma^2)$ and let the sample size denoted by $n$. There are two ways to construct the likelihood function, and the difference lies in how to specify the initial observation $x_1$. If we let $x_1$ be random, we know that the unconditional distribution of $x_t$ is $N(c/(1-\beta), \sigma^2/(1-\beta^2))$, and this will lead to an exact likelihood function. Alternatively, we can assume that $x_1$ is observable (known) and this will lead to a conditional likelihood function.

We first consider the exact likelihood function. We know that

$$p(x_1; \theta) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(x_1 - c/(1-\beta))^2}{2\sigma^2/(1-\beta^2)}\right].$$

Conditional on $x_1$, the conditional distribution of $x_2$ is $N(c + \beta x_2, \sigma^2)$, then the conditional probability density for the second observation is

$$p(x_2|x_1; \theta) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(x_2 - c - \beta x_1))^2}{2\sigma^2}\right].$$

So the joint probability density for $(x_1, x_2)$ is

$$p(x_1, x_2; \theta) = p(x_2|x_1; \theta)p(x_1; \theta).$$

Similarly, the probability density for the $n$th observation conditional on $x_{n-1}$ is

$$p(x_n|x_{n-1};\theta) = (2\pi\sigma^2)^{-1/2}\exp\left[-\frac{(x_n - c - \beta x_{n-1}))^2}{2\sigma^2}\right].$$

and the density for the joint observation of $X_n = (x_1, x_2, \ldots, x_n)$ is

$$L(X_n;\theta) = p(x_1;\theta)\prod_{t=2}^{n} p(x_t|x_{t-1};\theta).$$

Taking log we get the exact likelihood function (omitting constant terms for simplicity)

$$l(X_n;\theta) = -\frac{1}{2}\log\left(\frac{\sigma^2}{1-\beta^2}\right) - \frac{(x_1 - c/(1-\beta))^2}{2\sigma^2/(1-\beta^2)} - \frac{n-1}{2}\log(\sigma^2) - \sum_{t=2}^{n}\frac{(x_t - c - \beta x_{t-1})^2}{2\sigma^2}. \quad (11)$$

Next, to construct the conditional likelihood, assume that $x_1$ is observable, then the log likelihood function is (again, constant terms are omitted)

$$l(X_n;\theta) = -\frac{n-1}{2}\log(\sigma^2) - \sum_{t=2}^{n}\frac{(x_t - c - \beta x_{t-1})^2}{2\sigma^2}. \quad (12)$$

The maximum likelihood estimates $\hat{c}$ and $\hat{\beta}$ are obtained by maximizing (12), or solving the score function. Note that maximizing (12) with respect to $\beta$ is equivalent to minimizing

$$\sum_{t=1}^{n}(x_t - c - \beta x_{t-1})^2,$$

which is the objective function in OLS.

Compared to the exact likelihood function, we see that the conditional likelihood function is much easier to work with. Actually, when the sample size is large, the first observation becomes negligible to the total likelihood function. When $|\beta| < 1$, the estimator computed from exact likelihood and the estimator from conditional likelihood are asymptotically equivalent.

Finally, if the residual is not Gaussian, and if we estimate the parameter using the conditional Gaussian likelihood as in (12), then the estimate we obtain is known as *quasi-maximum likelihood estimate* (QMLE). QMLE is also very frequently used in empirical estimation. Although we misspecified the density function, in many cases, QMLE is still consistent. For instance, in an AR($p$) process, if the sample second moment converges to the population second moments, then QMLE using (12) is consistent, no matter whether the error is Gaussian or not. However, standard errors for the estimated coefficients that are computed with the Gaussian assumption need not be correct if the true data are not Gaussian (White, 1982).

# 3 Model Selection

In the discussion on estimation above, we assume that the order of the lags is known. However, in empirical estimation, we have to choose a proper order. A larger number of order (parameters) will increase the fitness of the model, therefore we need some criterion to balance the goodness of

fit and model parsimony. There are three commonly used criterion, *Akaike information criterion* (AIC), Schwartz's *Bayesian information criterion* (BIC), and the *posterior information criterion* (PIC) developed by Phillips (1996).

In all these criterion, we specify a maximum order $k_{max}$, and then choose $\hat{k}$ to minimize a criterion equation.

$$\text{AIC} = \log\left(\frac{SSR_k}{n}\right) + \frac{2k}{n} \tag{13}$$

where $n$ is the sample size, $k = 1, 2, \ldots, k_{max}$ is the number of parameters in the model, and $SSR_k$ is the residual from the fitted model. When $k$ increase, the fit increases, so $SSR_k$ decreases, but the second term increases. So this shows a trade off between fit and parsimony. Since the model is estimated using different lags, the sample size also varies. We can either use the different sample size $n - k$, or we can use a fixed sample size $n - k_{max}$. Ng and Perron (2000) has recommended using the fixed sample size and use it to replace $n$ in the criterion. However, the AIC rule is not consistent and tends to overfit the model by choosing larger $k$.

With all other issues similar as in the AIC rule, the BIC rule imposes a larger penalty for increasing number of parameters,

$$\text{BIC} = \log\left(\frac{SSR_k}{n}\right) + \frac{k\log(n)}{n} \tag{14}$$

BIC suggests samller $k$ than AIC and BIC rule is consistent in stationary data, i.e., $\lim_{n\to\infty} \hat{k}_{BIC} = k$. Further, Hannan and Deistler (1988) has shown that $\hat{k}_{BIC}$ is consistent when we set $k_{max} = [c\log(n)]$ (the integer part of $c\log(n)$) for any $c > 0$. Therefore, we can estimate $\hat{k}_{BIC}$ consistently without knowing the upper bound of $k$.

Finally, to present the PIC criterion, let $K = k_{max}$, and let $X(K)$ and $X(k)$ to denote the regressor matrix with $K$ and $k$ parameters respectively. Similar for $\beta$, the parameter vector.

$$\begin{aligned}
Y &= X(K)\beta(K) + error = X(k)\beta(k) + X(*)\beta(*) + error \\
A(*) &= X(*)\beta(*) \\
A(k) &= X(k)\beta(k) \\
A(*,k) &= X(*)X(k) \\
A(**) &= A(*) - A(*,k)A(k)^{-1}A(k,*) \\
\hat{\beta}(*) &= [X(*)'X(*) - X(*)'X(k)(X(k)'X(k))^{-1}X(k)X(*)]^{-1}[X(*)'Y \\
&\quad -X(*)'X(k)(X(k)'X(k))^{-1}X(k)Y] \\
\sigma_K^2 &= SSR_K/(n-K)
\end{aligned}$$

then

$$\text{PIC} = |A(**)/\sigma_k^2|^{-1/2}\exp\left\{\left(\frac{1}{2\sigma_k^2}\right)\hat{\beta}(*)'A(**)\hat{\beta}(*)\right\}.$$

PIC is asymptotically equivalent to the BIC criterion when the data is stationary, and when the data is nonstationary, PIC is still consistent.

Reading: Hamilton, Ch. 5, 8.