Introduction to Econometrics II: Recitation 5

Seung-hun Lee[*]

February 21th, 2022

1 Standard errors in a non-IID setting

Up until this point, the assumption of identical and independent distribution (IID) has been left untouched. In a time series setting, however, this is going to be unlikely to be true as past shocks can carry over to future periods. In such case, there may be a dependence structure across error terms - one of them being a **serial correlation** (or **autocorrelation**) problem. Before going into this, we review the variance structure in the IID, finite sample setting and then relax the IID assumption later on.

1.1 Review of the finite sample IID world

Assume the following data generating process

$$y = X\beta + e$$

with $X \in \mathbb{R}^{n\times k}$. Recall that for an OLS estimator, $\hat{\beta} - \beta = (X'X)^{-1}X'e$. There are two features of importance in the finite sample

- Location: This would be represented by checking for the bias. For this, we need strict exogeneity condition, where $E[e_i|X_1, .., X_n] = 0\ \forall i$. This implies

$$E[e|X] = \begin{pmatrix} E[e_1|X] \\ ... \\ E[e_n|X] \end{pmatrix} = \begin{pmatrix} E[e_1|X_1, .., X_n] \\ ... \\ E[e_n|X_1, .., X_n] \end{pmatrix} = 0$$

[*] Contact me at sl4436@columbia.edu if you spot any errors or have suggestions on improving this note.

and the bias can be calculated as

$$E[\hat{\beta}-\beta|X] = E[(X'X)^{-1}X'e|X] = (X'X)^{-1}X'E[e|X] = 0$$

Here, IID is used in the sense that $E[e_i|X_1,..,X_n]$ holds identically for all values of i.

- Dispersion: This relates to the variance of the $\hat{\beta}$ estimates. We write

$$\begin{aligned} var(\hat{\beta}|X) &= E[(\hat{\beta}-E[\hat{\beta}])(\hat{\beta}-E[\hat{\beta}])'|X] \\ &= E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X] \\ &= (X'X)^{-1}X'E[ee'|X]X(X'X)^{-1} \end{aligned}$$

where the final form of this depends on what assumptions we set on $D = E[ee'|X]$. We will rule out a very general structure with cross dependence across errors for the moment and focus on homoskedastic and heteroskedastic cases

If homoskedastic, we can assume that $D = \sigma^2 I_n$. Then the variance can be written as

$$var(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$$

and the distribution of $\hat{\beta}$ can be characterized as

$$\hat{\beta}-\beta|X \sim (0, \sigma^2(X'X)^{-1})$$

Since we are in a finite sample world, we cannot invoke the central limit theorem to claim normality. The distribution depends on the small sample distribution of e_i.

If not, the D matrix is a diagonal matrix whose diagonal entries are $E[e_1^2|X], ..., E[e_n^2|X]$. Since we do not know the true e_i^2, we need to estimate this with $\hat{e}_i^2$ and use $\widetilde{D}$. Then the variance becomes

$$var(\hat{\beta}|X) = \widehat{V}_{\hat{\beta}} = (X'X)^{-1}\left(\sum_{i=1}^{n} x_i x_i' \hat{e}_i^2\right)(X'X)^{-1}$$

and $\sqrt{\widehat{V}_{\hat{\beta}}}$ is what is known as **White standard errors**.

Another way to fix the standard errors is to use GLS, which minimizes $(y-X\beta)'D^{-1}(y-X\beta)$ with weights given by σ_i^{-1}. Unlike White standard errors, this is a two-step procedure in that we need to specify each σ_i.

1.2 Review of the asymptotic IID world

Here, we can appeal to the central limit theorem by taking $n \to \infty$. We define our moment condition $g_i = X_i e_i = X'e$ and the sample analogue $\bar{g}(\beta) = \frac{1}{n}\sum_{i=1}^n X_i e_i$. The weak law of large numbers and the central limit theorem implies

$$\bar{g}(\beta) \xrightarrow{p} 0, \ \sqrt{n}\bar{g}(\beta) \xrightarrow{d} N(0, \Omega)$$

where $\Omega = E[g_i g_i'] = E[x_i x_i' e_i^2]$. We write $Q_{XX} = E[x_i x_i'] = E[X'X]$ and obtain

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, Q_{XX}^{-1}\Omega Q_{XX}^{-1})$$

If we assume homoskedasticity, we can write $\Omega_1 = \sigma^2 Q_{XX}$ and the asymptotic variance becomes $\sigma^2 Q_{XX}^{-1}$. If otherwise, we work with $\Omega_2 = E[x_i x_i' e_i^2]$. Note that this can alternatively be written as

$$E[x_i x_i' e_i^2] = E[x_i x_i' E[e_i^2|X]] + E[x_i x_i'(e_i^2 - E[e_i^2|X])]$$

where the second term on the right hand side can be shown to be close to zero (Hansen 7.7).

The takeaway from the above equation is that If homoskedasticity is the wrong assumption, the we are missing out the $E[x_i x_i'(e_i^2 - E[e_i^2|X])]$ part and get the wrong estimator. Moreover, the wald statistic

$$n(\hat{\beta} - \beta)'(\hat{\sigma}^2(X'X/n)^{-1})^{-1}(\hat{\beta} - \beta)$$

does not converge to a chi-squared distribution since we are using an estimator of a variance that is inconsistent for the true variance. If homoskedascitity is the correct assumption but we stick with Ω_2, the cost for this is having a slightly larger standard error, which is still consistent in asymptotics and the wald statistic converges to chi-squared distribution.

So how do you detect heteroskedasticity? One such test designed to do this is the White's test of heteroskedasticity. The idea is to test whether the difference between Ω_1 and Ω_2 is large or not. If homoskedasticity is true, the distance between the two should not be large. Specifically, we set up as

$$\widehat{\Omega}_2 = \sum_{i=1}^n x_i x_i' \hat{e}_i^2, \ \widehat{\Omega}_1 = s^2 \frac{1}{n}\sum_{i=1}^n x_i x_i', \ \widehat{\Omega}_2 - \widehat{\Omega}_1 = \frac{1}{n}\sum_{i=1}^n x_i x_i'(\hat{e}_i^2 - s^2)$$

where $s^2 = \frac{1}{n-k}\sum_{i=1}^n \hat{e}_i^2$ and $\hat{e}_i$ is OLS residual. Then define h_i as a column vector that stacks up all the nonzero elements in $x_i x_i'$ matrix. Then, define $c_n = \frac{1}{n}\sum_{i=1}^n (\hat{e}_i^2 - s^2)h_i$ and we test using a the following test statistic

$$c_n'(var(c_n))^{-1}c_n \xrightarrow{d} \chi^2_m$$

where m is the number of all the nonzero elements in $x_i x_i'$. Because of the computational burden, we use an alternative approach by regressing $\hat{e}_i^2$ onto the nonzero elements of $x_i x_i'$, which leads to

$$\hat{e}_i^2 = \delta_0 + \delta_1 x_1 + ... + \delta_k x_k + \delta_{k+1} x_1^2 + ... + \delta_{2k} x_k^2 + \delta_{2k+1} x_1 x_2 + ... + \delta_{2k+[(k-1)k]/2} x_{k-1} x_k + \epsilon_i$$

Then, we obtain the R^2 and use nR^2 as our test statistic against the distribution χ^2_m, where m is the total number of regressors in the above equation. A shorter version that does not subtract too much from the degrees of freedom is to regress the squared of residual onto the constant and the squared of the regressors.

1.3 Clustered sampling

Now we move on to the clustered samples, where the total of n observations can be split into G separate groups. Let $g \in \{1, .., G\}$ index the group and $i \in \{1, .., n_g\}$ index individual in group g which has n_g observations (so $n = \sum_{g=1}^{G} n_g$). The individual level observation for individual i in group g will be written as (y_{ig}, x_{ig}). The cluster level observations will be denoted as (y_g, X_g), where $y_g = (y_{1g}, ..., y_{n_g g})' \in \mathbb{R}^{n_g}$ and $X_g \in \mathbb{R}^{n_g \times k}$ can be defined similarly. The population level variables will be written as (y, X).

Let our data generating process be $y = X\beta + e$. The individual level regression for this is

$$y_{ig} = x_{ig}'\beta + e_{ig}$$

which can be stacked to a cluster level regression as

$$y_g = X_g\beta + e_g$$

In this setup, the OLS can be written equivalently in these ways

$$\hat{\beta} = (X'X)^{-1}X'y = \left(\sum_{g=1}^{G} X_g' X_g\right)^{-1}\left(\sum_{g=1}^{G} X_g' y_g\right) = \left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} x_{ig}x_{ig}'\right)^{-1}\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} x_{ig}y_{ig}\right)$$

To analyze the features of the estimators in a clustered samples, we need to set these assuptions

Assumption 1.1 (Assumptions for Clustered samples)**.** *We assume the following for the clustered samples*

A1 *Clusters are known to researchers*

A2 (y_g, X_g) *are independent across clusters*

A3 $E[e_g|X_g] = 0$ *(same as* $E[e_{ig}|X_g] = 0$ *for* $i \in \{1, .., n_g\}$*)*

However, we do allow for the correlation of observations within the same cluster.

With this, we can show that the OLS is unbiased.

$$\begin{aligned} E[\hat{\beta} - \beta|X] &= E\left[\left(\sum_{g=1}^{G} X_g' X_g\right)^{-1} \left(\sum_{g=1}^{G} X_g' e_g\right) |X\right] \\ &= \left(\sum_{g=1}^{G} X_g' X_g\right)^{-1} \left(\sum_{g=1}^{G} X_g' E[e_g|X]\right) \\ &= \left(\sum_{g=1}^{G} X_g' X_g\right)^{-1} \left(\sum_{g=1}^{G} X_g' E[e_g|X_g]\right) = 0 \end{aligned}$$

The conditional variance-covariance matrix of $\hat{\beta}$ is as follows. Define $D_g = E[e_g e_g' | X_g]$. Then, we can write

$$\begin{aligned} var\left[\sum_{g=1}^{G} X_g' e_g |X\right] &= \sum_{g=1}^{G} var(X_g' e_g|X_g) \\ &= \sum_{g=1}^{G} X_g' E[e_g e_g'|X_g] X_g \\ &= \sum_{g=1}^{G} X_g' D_g X_g = \Omega_n \end{aligned}$$

Hence, $var(\hat{\beta}|X) = (X'X)^{-1}\Omega_n(X'X)^{-1}$. We estimate this using $\widehat{\Omega}_n = \sum_{g=1}^{G} X_g' \hat{e}_g \hat{e}_g' X_g$ and thus, $\widehat{var}(\hat{\beta}|X) = (X'X)^{-1}\widehat{\Omega}_n(X'X)^{-1}$.

Note that the above example uses OLS estimates. You can extend this to other estimators.

2 Dependent data and time series

We now move on to the times series data, where we no longer assume that the data satisfies IID assumption. When there is a time structure (or fixed ordering) with the dataset, the data is unlikely to satisfy the independence assumptions. There are likely to be correlation of observations across time - **serial correlation** (or **autocorrelation**, I prefer to use this term). For instance, most shocks that affect many macroeconomic variables - GDP, unemployment to name a few - carry over to the next period. Since most of our asymptotic theories involved IID assumptions implicitly, we need new tools to analyze key features in time series regressions.

2.1 Concepts of stationarity

In the cross-sectional observations, we treat a sequence $(y_1, ..., y_n)$ as a random draws from an underlying population larger than n. Since our data now has an order structure, we should treat this differently. In times series observations, we should see $(y_1, ..., y_T)$ as a subset of a dependent stochastic processes. We are still interested in deriving the mean, variance and covariances of the distribution, which is obtainable only after pinning the distribution to some constant parameter. This is where stationarity kicks in. Here are some key definitions to go through before getting there though.

Definition 2.1 (Autocovariance)**.** Let $(y_t, ..., y_{t-k})$ be part of a ordered sequence. **Autocovariance** between y_t and y_{t-k} is defined as

$$\gamma_t(k) = cov(y_t, y_{t-k})$$

Similarly, we can characterize the variance of y_t as

$$\gamma_t(0) = cov(y_t, y_t) = var(y_t)$$

Definition 2.2 (Autocorrelation)**. Autocorrelation** between y_t and y_{t-k} is defined as

$$\rho_t(k) = \frac{cov(y_t, y_{t-k})}{\sqrt{var(y_t)var(y_{t-k})}}$$

So autocovariances capture the linear dependence between y_t and its lags. Autocorrelation is a scale-free version of this measure.

We are now ready to characterize the various concepts of stationarity. We start with the version which has the weakest requirements.

Definition 2.3 (Covariance stationarity, or weak stationarity)**.** Let $\{y_t\}$ be a sequence of observations. We say $\{y_t\}$ satisfies covariance stationarity if the first and second moments are time-invariant. Or

$$E[y_t] = \mu_t = \mu, \;\; \gamma_t(k) = \gamma(k) \text{ and } \rho_t(k) = \rho(k)\ \forall k,$$

What this implies is that the mean can be characterized without t and that autocovariance and autocorrelation only depends on the relative difference (which is our k) between the sequences and not t. This is a nice feature as it implies that $\gamma(k) = \gamma(-k)$ (The multivariate equivalent is $\Gamma(k) = \Gamma(-k)'$)

A stronger version of stationarity requires that the distribution of $\{y_t\}$ remains stable over time. More formally,

Definition 2.4 (Strict stationarity)**.** Let $\{y_t\}$ be a sequence of observations. We say $\{y_t\}$ satisfies **strict stationarity** if distribution of $(y_t, ...y_{t-k})$ is independent of t for all k.

If $\{y_t\}$ satisfies strict stationarity, a linear transformation of that process $x_t = \phi(y_t, ..., y_{t-k})$ is also strict stationary.

Unfortunately, strict stationarity is not enough for us to characterize the law of large numbers. One pitfall is that a stationary process may show limited time series variation - i.e. being stuck at a particular area. This may result in a sample average which only represents a local area but not the full distribution. The intuitive definition of ergodicity is that the process should move around the average and take all values over its support. A sufficient condition for this to happen is that the the process is asymptotically independent - or that any two random variables far apart in the sequence is almost independently distributed. or that $\gamma(k) \to 0$ as $k \to \infty$. Then we can define ergodic stationarity as follows.

Definition 2.5 (Ergodic stationarity)**.** Let $\{y_t\}$ be a sequence of observations. We say $\{y_t\}$ satisfies **ergodic stationarity** if distribution of $(y_t, ...y_{t-k})$ is strictly stationary and ergodic.

Likewise, $x_t = \phi(y_t, ..., y_{t-k})$ is also ergodic stationary if $(y_t, ..., y_{t-k})$ is ergodic stationary.

2.2 Ergodic theorem and central limit theorem

We are now ready to characterize the weak law of large numbers that works in the time series context.

Theorem 2.1 (Ergodic theorem). *Let $\{y_t\}$ be strictly stationary and ergodic, $E|y_t| = \mu < \infty$. Then as $T \to \infty$,*

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \xrightarrow{p} E[y_t] = \mu$$

We can analyze the consistency of the time series estimators with this. This also implies that the sample autocovariances converge to its true value

$$\widehat{\gamma}_t(j) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y}) \xrightarrow{p} \gamma(k)$$

We now present the central limit theorem that works on the dependent data. For this, we define the white noise error e_t as and error satisfying $E[e_t] = 0$ and $\gamma(k) = 0\ \forall k \neq 0$. To motivate the why IID CLT does not work, we work with a process $y_t = \mu + e_t$ where e_t is white noise. The sample variance $var(\bar{y})$ can be calculated as

$$var(\bar{y}) = var\left(\frac{1}{T}\sum_{t=1}^{T} y_t\right) = \frac{1}{T^2} var\left(\sum_{t=1}^{T} y_t\right)$$

In the case where y_t satisfies IID(μ, σ^2), with $E[y_t y_s] = 0$ for $t \neq s$, calculating the sample variance is simple and becomes

$$var(\bar{y}) = \frac{1}{T^2}\sum_{t=1}^{T} var(y_t) = \frac{\sigma^2}{T}$$

and the central limit theorem would yield

$$\sqrt{T}\frac{\bar{y} - \mu}{\sigma} \sim N(0,1)$$

However, we are working with a dependent dataset where the short-run variance σ^2 may not equal long-run variance due to the dependence nature. Instead, we need a CLT that uses a long-run variance.

Theorem 2.2 (Central limit theorem for dependent data)**.** *Let $\{y_t\}$ come from a dependent data satisfying strict stationarity and ergodicity. Specifically, let $y_t = \mu + e_t$ with e_t being a white noise. then, as $T \to \infty$, the following holds,*

$$\sqrt{T}\frac{\bar{y}-\mu}{\omega} \sim N(0,1)$$

where $\omega^2 = \sum_{j=-\infty}^{\infty} \gamma(j) = \gamma(0) + 2\sum_{j=1}^{\infty} \gamma(j)$

2.3 AR(p) processes

One of the most used time series process is a regression where the independent variables are composed of the lags of the dependent variable. That process is called an **AR(p)** process where p is the number of lags of dependent variable included as independent variable. It takes a following form

$$y_t = \alpha_1 y_{t-1} + ... + \alpha_p y_{t-p} + e_t$$

Another way to characterize this is to use the lag operator L where $Ly_t = y_{t-1}, .., L^p y_t = y_{t-p}$.

We will focus on cases where $p = 1$. So our process has this form

$$y_t = \alpha y_{t-1} + e_t \iff (1 - \alpha L) y_t = e_t$$

This process can be written with many lags of e_t instead. Note that we can take the process one period back and write $y_{t-1} = \alpha y_{t-2} + e_{t-1}$. Then we get

$$y_t = \alpha(\alpha y_{t-2} + e_{t-1}) + e_t$$

Repeat this for all the possible lags of y variable. This would leave us with

$$y_t = e_t + \alpha e_{t-1} + \alpha^2 e_{t-2} + ... + \alpha^t e_0$$

If we work with infinite periods of time, we can write

$$y_t = \sum_{j=0}^{\infty} (\alpha L)^j e_t = \sum_{j=0}^{\infty} \alpha^j e_{t-j}$$

However, for this summation to not diverge, we need that $|\alpha| < 1$. Otherwise, y_t diverges.

An interesting feature of this process is that we can capture how persistent a random shock at a certain period is. This is what an **impulse response function** is. Suppose that a random shock happens at $t = 1$ (so $e_1 \neq 0$) and nowhere else, with $y_0 = 0$. Then we can write

$$\begin{aligned} y_0 &= 0 \\ y_1 &= \alpha y_0 + e_1 = e_1 \\ y_2 &= \alpha y_1 + e_2 = \alpha(\alpha y_0 + e_1) + e_2 = \alpha e_1 + e_2 = \alpha e_1 \\ y_3 &= \alpha^2 e_1 + \alpha e_2 + e_3 = \alpha^2 e_1 \\ &\dots \end{aligned}$$

Key takeaway here is that α represents the persistence of a shock. α close to 0 implies that the shock is transient. On the other hand, $|\alpha| < 1$ that is close to 1 is a persistent shock that dies away eventually. Any α that deviates away from aforementioned values is a divergent shock that does not die out.