Recitation 5: Dependent dataset structure

Seung-hun Lee

Columbia University
Introduction to Econometrics II Recitation

February 7th, 2022

Standard errors in a dependent dataset

Datasets are not always IID!

- In a time series setting past shocks can carry over to future periods.
- The data may be clustered: There are groups within the observations whose members show correlated patterns
- In such case, there may be a dependence structure across error terms - so we look at clustered standard errors and (not today) Newey standard errors

IID features to set a yardstick

- Location: Assume strict exogeneity, or $E[e_i|X_1,..,X_n] = 0 \; \forall i$. This implies

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

- 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.

IID is especially crucial in characterizing dispersion

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

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

- IID allows us to focus on homoskedasticity and rule out dependence across e_i, e_j
- If homoskedastic, we can assume that $D = \sigma^2 I_n$. Then the variance can be written as

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

- We still can get White standard errors $\left(\sqrt{\widehat{V}_{\hat{\beta}}}\right)$ if homoskedasticity does not hold

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

WLLN and CLT hold nicely in IID

- Define our moment condition $g_i = X_i e_i = X' e$ and $\bar{g}(\beta) = \frac{1}{n}\sum_{i=1}^{n} X_i e_i$
- 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 asymptotic distribution for $\hat{\beta}$

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

Variances are different in homoskedasticity vs heteroskedasticity

- Under homoskedasticity, write $\Omega_1 = \sigma^2 Q_{XX}$ and the asymptotic variance is $\sigma^2 Q_{XX}^{-1}$.
- Otherwise, we work with $\Omega_2 = E[x_i x_i' e_i^2]$ which is

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

- 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
(we also need different wald statistic!)
- If homoskedascitity is the correct assumption but we stick with Ω_2, we end up with a slightly larger standard error
(consistent in asymptotics and the wald statistic converges to chi-squared distribution)

Detecting heteroskedasticity: White test

- The idea is to test whether the difference between Ω_1 and Ω_2 is large or not
- Original idea is to use

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

- h_i: Stacked up vector of all nonzero $x_i x_i'$ element (*vec* operator)
- 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

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

where m is the number of all the nonzero elements in $x_i x_i'$

An implementable version of White test

- Regress $\hat{e}_i^2$ onto the nonzero elements of $x_i x_i'$, which leads to

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

  Obtain the R^2 and use nR^2 as our test statistic against the distribution χ_m^2
- 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

Clustered samples

- 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).
- For DGP $y = X\beta + e$, the individual and cluster level regression is

$$y_{ig} = x_{ig}'\beta + e_{ig}, \quad y_g = X_g\beta + e_g$$

- OLS: $\left(\sum_{g=1}^{G} X_g' X_g\right)^{-1} \left(\sum_{g=1}^{G} X_g' y_g\right) = \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} x_{ig} x_{ig}'\right)^{-1} \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} x_{ig} y_{ig}\right)$

Bias and variance for clustered samples

Assumptions for the clustered samples

A1 Clusters are known to researchers

A2 (y_g, X_g) are independent across clusters

A3 $E[e_g|X_g] = 0$ ($E[e_{ig}|X_g] = 0$ for $i \in \{1, .., n_g\}$) - within-cluster correlation allowed

- OLS is still unbiased
- Conditional variance takes this form:

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

  Hence, $var(\hat{\beta}|X) = (X'X)^{-1}\Omega_n(X'X)^{-1}$
- Estimate with $\widehat{var}(\hat{\beta}|X) = (X'X)^{-1}\widehat{\Omega}_n(X'X)^{-1}$, where $\widehat{\Omega}_n = \sum_{g=1}^{G} X_g' \hat{e}_g \hat{e}_g' X_g$

Time series data

Time series is not likely to be IID

- When there is a time structure (or fixed ordering) with the dataset, the data is unlikely to satisfy the independence assumptions
- Serial correlation (autocorrelation) is likely an issue
- Most shocks that affect many macroeconomic variables - GDP, unemployment to name a few - carry over to the next period
- Our asymptotic theories involved IID assumptions, so we need new tools to analyze key features in time series regressions
- We need concepts that allows us to pin the distribution to some constant parameter to obtain means, variances, and covariances

Dependence between y_t and y_{t-k}

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

$$\gamma_t(k) = cov(y_t, y_{t-k}),\ \gamma_t(0) = cov(y_t, y_t) = var(y_t)$$

- **Autocorrelation** between y_t and y_{t-k} is defined as

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

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

Covariance stationarity: Weakest version of 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

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

- Mean can be characterized without t
- Autocovariance and autocorrelation only depends on the relative difference (k) between the sequences and not t
- This implies that $\gamma(k) = \gamma(-k)$ (The multivariate equivalent is $\Gamma(k) = \Gamma(-k)'$)

Strict stationarity: All features of the distribution is time-invariant

- 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.
- This means that the distribution of $\{y_t\}$ remains stable over time
- If $\{y_t\}$ satisfies strict stationarity, a linear transformation of that process $x_t = \phi(y_t, ..., y_{t-k})$ is also strict stationary.

Ergodicity: Sample of $\{y_t\}$ captures all features of the full observation

- Stationary process may show limited time series variation (being stuck at a particular area), resulting in a sample average representing a local area but not 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 $k \to \infty$.
- **Ergodic stationarity**: We say $\{y_t\}$ satisfies ergodic stationarity if distribution of $(y_t, ...y_{t-k})$ is strictly stationary and ergodic.
- $x_t = \phi(y_t, ..., y_{t-k})$ is also ergodic stationary if $(y_t, ..., y_{t-k})$ is ergodic stationary.

Ergodic theorem: LLN for dependent dataset

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

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

IID version of CLT may fail even with white noise error

- **White noise error (WN)**: an error e_t satisfying $E[e_t] = 0$ and $\gamma(k) = 0 \ \forall k \neq 0$
- Let $y_t = \mu + e_t$ where e_t is WN. 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

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

- The central limit theorem would yield $\sqrt{T}\frac{\bar{y}-\mu}{\sigma} \sim N(0, 1)$

...because long-run variances differ from simple short-run version

- This is because $E[e_t e_j]$ is nonzero!
- The version of CLT with this in mind is the following

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

Lags of dependent variable as independent variables

- **AR(p)** process: Including p lags of dependent variable as independent variable

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

- Lag operator L: $Ly_t = y_{t-1}, .., L^p y_t = y_{t-p}$.
- AR(1): Can be written as

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

where α is our parameter of interest

Rewriting AR(1) with just e_t's

- Take the process one period back and write $y_{t-1} = \alpha y_{t-2} + e_{t-1}$.
- Then plug this into $y_t = \alpha y_{t-1} + e_t$ to 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

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

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

Use of AR(1): How long does a random shock persist?

- Impulse response function: Useful for shock analysis
- Suppose that a random shock happens at $t = 1$ (so $e_1 \neq 0$) and nowhere else ($y_0 = 0$)
- Then we can write

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

- α represents the persistence of a shock
  - α close to 0 implies that the shock is transient
  - $|\alpha| < 1$ that is close to 1 is a persistent shock that dies away eventually
  - If not in this range, divergent shock that does not die out!