Panel data analysis: clustered estimator, FGLS and PCSE
Day 1.

David Pupovac

Longitudinal data

- Panel data analysis?

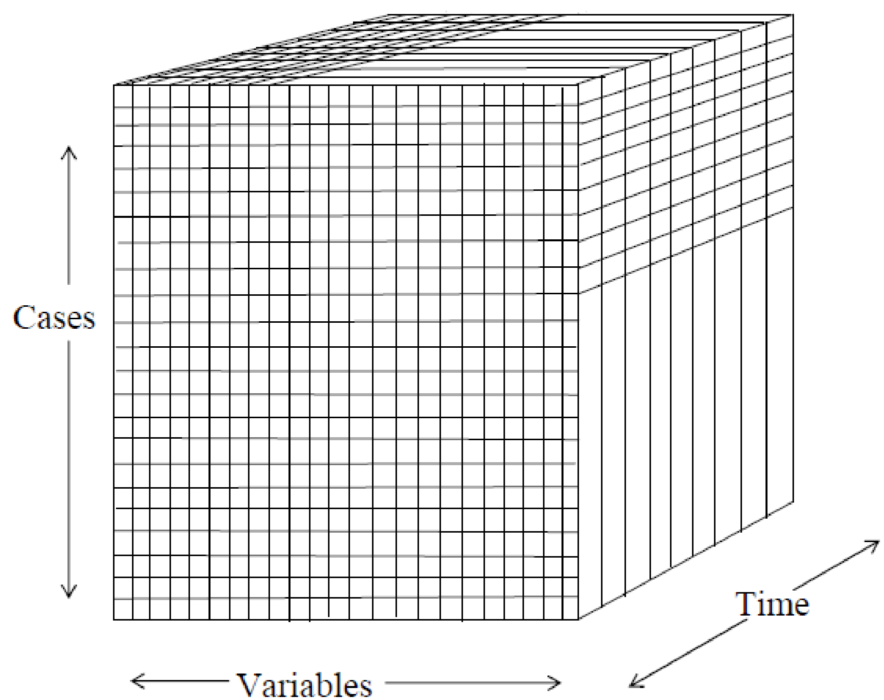
- Longitudinal data is data observed over time as well as over space – i.e. repeated observations on units observed over time
- *Therefore, longitudinal data is a collection of the time series (TS) across cross-sections (CS)*

- **Relation to multi-level modeling:**

  - **Longitudinal data can be perceived as a subset of hierarchical data where observations are correlated because they belong to the same unit**

e.g. educational studies: we observe student i in school u (presumably there is some connection between the observations in the same school)

The structure of longitudinal data



The structure of longitudinal data

country	year	Y	X1	X2	X3
1	2000	6.0	7.8	5.8	1.3
1	2001	4.6	0.6	7.9	7.8
1	2002	9.4	2.1	5.4	1.1
2	2000	9.1	1.3	6.7	4.1
2	2001	8.3	0.9	6.6	5.0
2	2002	0.6	9.8	0.4	7.2
3	2000	9.1	0.2	2.6	6.4
3	2001	4.8	5.9	3.2	6.4
3	2002	9.1	5.2	6.9	2.1

Balance vs. unbalanced data

- A balanced data set is a set that contains all elements observed at all time units. Whereas unbalanced data is a set of data where at certain time units, the data category is not observed - this difference was more relevant in past. If there are not major asymmetries, then there is not a big problem. If a huge problem - imputation.

Types of longitudinal data

- **Two main types of data we will deal with:**
  - **Panel study (cross-section time series - CSTS)**

  (data sets: NES, Congressional election outcomes by CD and year)
  - **Time series cross-section (TSCS)**

  (data sets: political economy data on 15 OECD nations observed annually)

    ❖ sometimes these are referred to as **short panels** $(N > T)$ and **long panels** $(N < T)$

- **Types of longitudinal data:**
  - Event history data
  - Dyad year design in international relations
  - Data combining different surveys taken at different times; rolling cross-section (data sets: Canadian Election Study)
  - Pseudo panel - (respondents grouped by cohort) based on repeated cross section data" (e.g. Family Expenditure Surveys)

Panels vs. TSCS data

- Panel data have large number of cross-sections (big N) with each unit observed only a few times (small T)

  (panel data usually have a larger number of data points)

- TSCS data has reasonable sized T and not very large N.

- **The main consequence of this difference:**
  - **For panel data, asymptotics are in N, while T is fixed.**
  - **For TSCS data, asymptotics in T, while N is fixed.**

- This results in major differences. With small T there is no option of inferring anything about the time series structure of the data, while with bigger T there is. With TSCS data we are concerned with the units (states, firms,…), while we usually do not care about the units in panel models. Furthermore, in panel data the units are typically sampled, while the number of waves is fixed. In TSCS units are fixed but we can contemplate what happens as T approaches infinity.

- **Consequently, we will have to use different methods in analyzing these types of data.**

Advantages of TSCS/CSTS data

- TSCS/CSTS data usually give the researcher a larger number of data points, increasing the degrees of freedom, hence improving the efficiency of estimates.

- TSCS/CSTS data allows making inferences about the dynamics.

- The TSCS/CSTS data allow us to construct and test assertion that the real reason one finds (or does not find) certain effects is the presence of omitted (mismeasured or unobserved) variables that are correlated with explanatory variables.

- Consider the estimation of a distributed-lag model: multicollinearities appear among $h + 1$ lagged variables, x_t, $x_{t-1} \ldots x_{t-h}$. If TSCS/CSTS data are available, we can utilize the interindividual differences in x values to reduce the problem of collinearity

- **However, despite the advantages, the complicated structure of the data present major challenges for modeling**

- As we are going to stick to regression setting, let's first look at the OLS assumptions

Linear regression

- Terminology: regressing dependent (y) on independent (x)

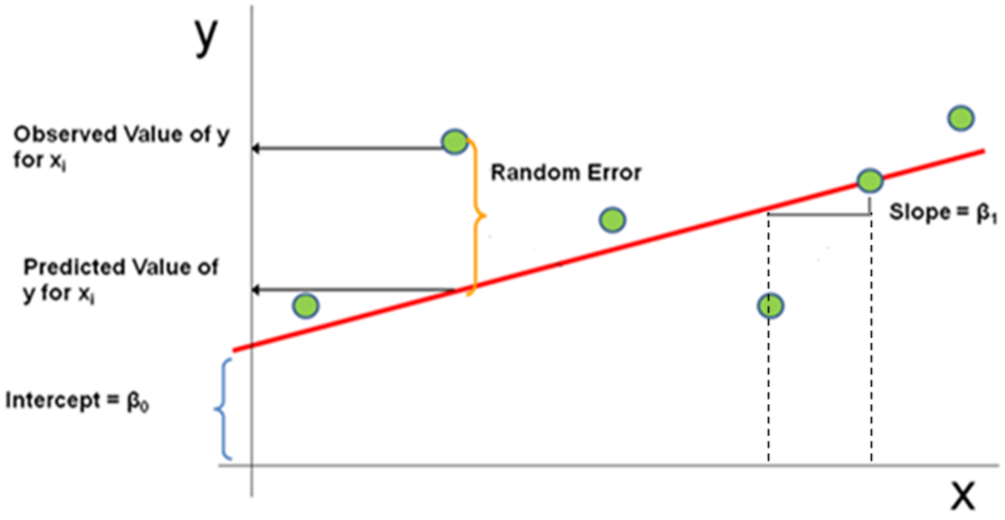
$$\boldsymbol{y = \beta_0 + \beta_1 x + \varepsilon}$$

- Elements of the equation:

  intercept , β_0

  slope, β_1

  random error, ε

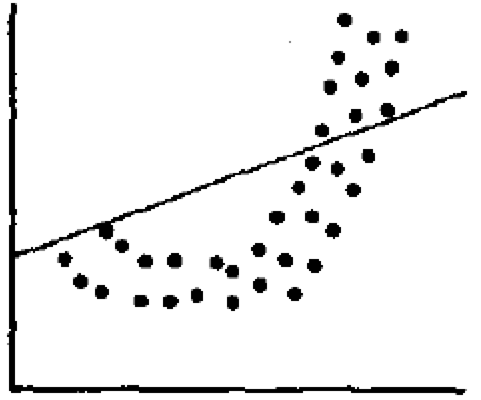


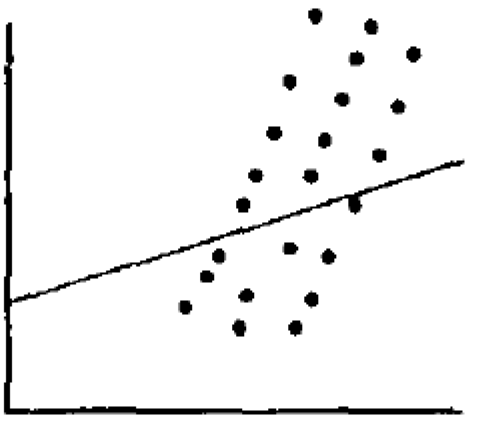
- Multiple regression:

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon}$$

Assumptions: correct model specification

- Linear relation (Gauss-Markov)





- No omitted variables (causes bias)
- No irrelevant variables

From: Lewis-Beck (1980) Applied Regression: An Introduction

Assumptions (Gauss-Markov): exogeneity - mean of error is zero

$$E[\varepsilon] = 0$$

- The sum of the positive errors should be equal to the sum of the negative errors.

$$E[\varepsilon_i | x_i] = 0 \ \ \forall\, i$$

- **Weak exogeneity** - the errors should have conditional mean zero (contemporaneous exogeneity)

$$E[\varepsilon_t | x_s] = 0 \ \ \forall\, s$$

- **Strict exogeneity** - the errors should have mean zero for all x, that is even in case if $t \neq s$

- ***The main consequence of no endogeneity assumption is lack of <u>correlation</u> between disturbances and explanatory variables***

Assumptions (Gauss-Markov) : no multicollinearity

- $\boldsymbol{X}$ is an $n \times k$ matrix of full rank - the columns of $\boldsymbol{X}$ are linearly independent

- Multicollinearity typically arises when two variables that measure the same thing are both included in a multiple regression model.

- The major undesirable consequence of multicollinearity is that the variances of the OLS estimates of the parameters of the collinear variables are quite large.

Assumptions: normality

$$E[\varepsilon|X] \sim N[0, \sigma^2 I]$$

- For any given X, the distribution of errors is normal.
- Also, Y is distributed normally at each value of X
- These assumptions are not actually required for the Gauss-Markov Theorem. However, we often assume them to make hypothesis testing easier.

Assumptions:

Spherical disturbances **(Gauss-Markov)**

- Constant error variance : $E[\varepsilon_i|X] = \sigma^2 \; \forall \; i$

No heteroscedasticity

- Independent errors : $E[\varepsilon_i \varepsilon_j|X] = 0 \; \forall \; i \neq j$

No autocorrelation

Pooled data

- General model:

$$y_{i,t} = x_{i,t}\beta + \varepsilon_{i,t}$$

- $x_{i,t}$ is a vector of exogenous variables observations indexed by both unit (i) and time (t).

  where $i = 1, \ldots, s$ is the number of units (or panels) and $t = 1, \ldots, j$ is the number of observations for panel i.

Can we use OLS in analyzing the pooled TSCS/CSTS data?

- This depends on several preconditions including “pooling” and violation of Gauss-Markov assumptions.

OLS on pooled data

- “Pooling” assumes that all units are characterized by the same regression equation at all points in time – in other words, it requires identical parameters for all cross-sectional and/or time units. Thus, pooling the data requires homogeneity assumption.

  – Are the intercepts same for each unit?
  – Are the slopes same for each unit?
  – Are the parameters constant over time?
  – Are there no effects of units on each other?
  – Is there any temporal dependence and does it varies across units?

- How to model homogeneity?
  – Complete heterogeneity
  – Complete homogeneity
- Typically, neither is justified

- **We will start addressing these topics in the next class.**

Pooling and errors

- Even if we assume that all units are fit by the same model there is a question of spherical errors

- However, as we have a more complicated data structure, the violation of the assumption of spherical errors may come from various sources

- <u>Also, keep in mind that some violations can be perceived from both the viewpoint of cross-sectional and time units</u>. This also is also applicable to the issues related to homogeneity (e.g. homogeneity of intercepts and slopes my be perceived from the viewpoint of cross-sections and time)

- **Today we will only deal with the issues of spherical errors**

(keep in mind that this is somewhat artificial division because in research situations you will most likely have to combine the methods of modeling heterogeneity and the methods of dealing with non-spherical errors as nuisance)

OLS matrices

- ***Projection P matrix:*** $X(X^TX)^{-1}X^T$ *also called influence or hat (H) matrix*
- ***Residual maker:*** $I - X(X^TX)^{-1}X^T$

- $\hat{\beta}_{OLS} = (X^TX)^{-1}X^Ty$ *this minimizes the sum of squared residuals*
- $y - X\hat{\beta}_{OLS} = (I - X(X^TX)^{-1}X^T)y = (I - P)y$

- $Var\_cor(\hat{\beta}) = \sigma^2(X^TX)^{-1}$

- The square roots of the variances on the diagonal of $Var\_cor(\hat{\beta}_{OLS})$ are **standard errors**

Spherical errors - OLS

- Under OLS assumptions variance covariance matrix of β is $(X'X)^{-1}\{X'\Omega X\}(X'X)^{-1}$, where assuming that $\Omega = \sigma^2 I$ we get:

$$\Omega = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

- The assumption of $\sigma^2 I$ reflects the assumption of spherical disturbances where:

$$\hat{\sigma}^2 = \frac{\varepsilon'\varepsilon}{n-k}$$

- So we finally we get:

$$\widehat{Cov}(\hat{\beta}) = \sigma^2 I\ (X^T X)^{-1}$$

- Commonly, the violation of this assumption may lead to:
  - **Heteroscedasticity** - diagonal elements of matrix not are equal
  - **Autocorrelation** - off-diagonal elements of matrix are not zero

- Analogous problems, but with more complicated structure, may be inherent to TSCS/CSTS data

Spherical errors

- Assuming that the errors are spherical (i.e., satisfy the Gauss-Markov assumptions) and the model is appropriately specified, OLS is optimal.

- Thus, if all errors are independent and identically distributed,

$$E\left(\varepsilon_{i,t}\varepsilon_{j,s}\right) = \begin{Bmatrix} \sigma^2 \; if \; i = j \; and \; s = t \\ 0 \; otherwise \end{Bmatrix}$$

this assumes that Ω matrix for TSCS/CSTS looks like this:

$$\Omega = \begin{bmatrix} \sigma^2 I & 0 & \dots & 0 \\ 0 & \sigma^2 I & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 I \end{bmatrix}$$

- In case of TSCS/CSTS data OLS estimates this by:

$$\widehat{Cov}\left(\hat{\beta}\right) = (X'X)^{-1}X'\left(\frac{\sum_i \sum_t \varepsilon_{i,t}^2}{NT-k}\right)X(X'X)^{-1}$$

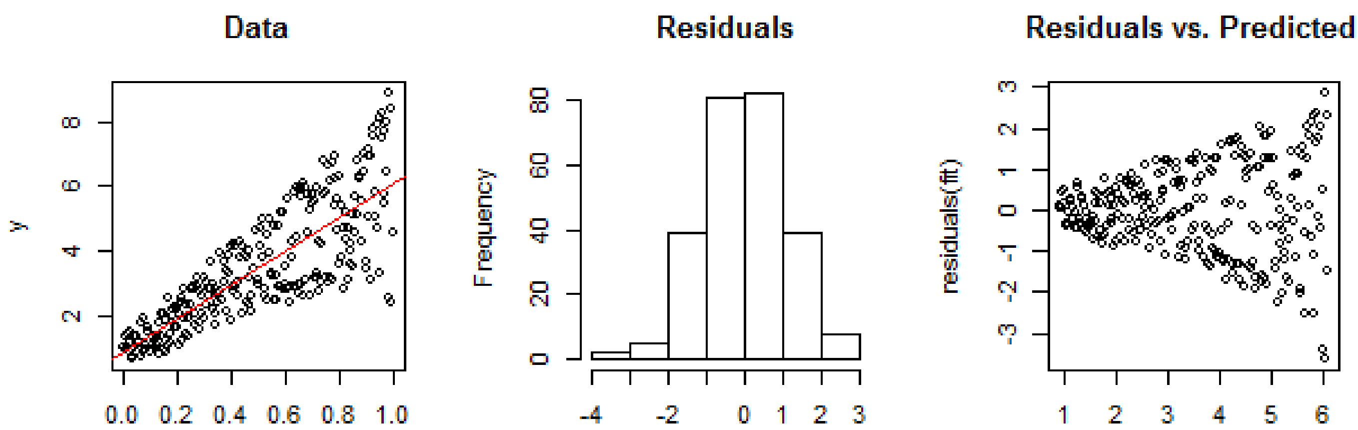
Non-spherical errors

- **The OLS errors for TSCS/CSTS data will be wrong if there are:**
  - panel heteroskedasticity
  - contemporaneous correlation of the errors
  - serial correlation of errors

(usual) Heteroskedasticity

- **In standard (non-TSCS/CSTS) setting heteroskedasticity implies:**

$$E[\varepsilon_i|X] = \sigma^2 \; \forall \; i$$



- <u>Detection</u>: visualization, Breusch-Pagan, White test
- <u>Remedy</u>: robust standard errors (Huber-White), WLS

Non-spherical errors - heteroskedasticity

- In TSCS/CSTS setting this implies that the variance for each of the unit differ. This concept differs from simple heteroskedasticity in that error variances are constant within a unit. This results in following Ω matrix:

$$\Omega = \begin{bmatrix} \sigma_1^2 I & 0 & 0 & 0 \\ 0 & \sigma_2^2 I & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_m^2 I \end{bmatrix}$$

- However, keep in mind that you can define heteroskedasticity in terms of $E(\varepsilon_i) = \sigma_i^2$, $E(\varepsilon_t) = \sigma_t^2$ or $E(\varepsilon_{i,t}) = \sigma_{i,t}^2$.

- Thus, in addition to panel heteroskedasticity you may have regular heteroskedasticity – however, as mentioned above we typically assume that the error variances within each unit do not differ over time

Non-spherical errors – cross-sectional correlation

Correlation across panels (cross-sectional, spatial correlation)

- This violation assumes that the unit specific error terms are correlated through time. For example, the large errors for $unit\ i$ at time t will often be associated with large errors for $unit\ j$ at time t.

- Thus, we may wish to assume that the error terms of panels are correlated, in addition to having different variances. This will result in following Ω matrix:

$$\Omega = \begin{bmatrix} \sigma_1^2 I & \sigma_{1,2} I & \cdots & \sigma_{1,m} I \\ \sigma_{2,1} I & \sigma_2^2 I & \cdots & \sigma_{2,m} I \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m,1} I & \sigma_{m,2} I & \cdots & \sigma_m^2 I \end{bmatrix}$$

- Analysts typically assume that all spatial correlation is both contemporary and does not vary with time.

Non-spherical errors – cross-sectional correlation

Examples:

- Economies of Belgium and Netherlands may be associated and, consequently, their errors may be correlated.

- These contemporaneous correlations may differ by unit. The errors in the Scandinavian economies may be linked together, but remain independent of errors in the southern European countries.

- Because we must estimate cross-sectional correlation in this model, the panels must be balanced (and $T \geq N$ for valid results).

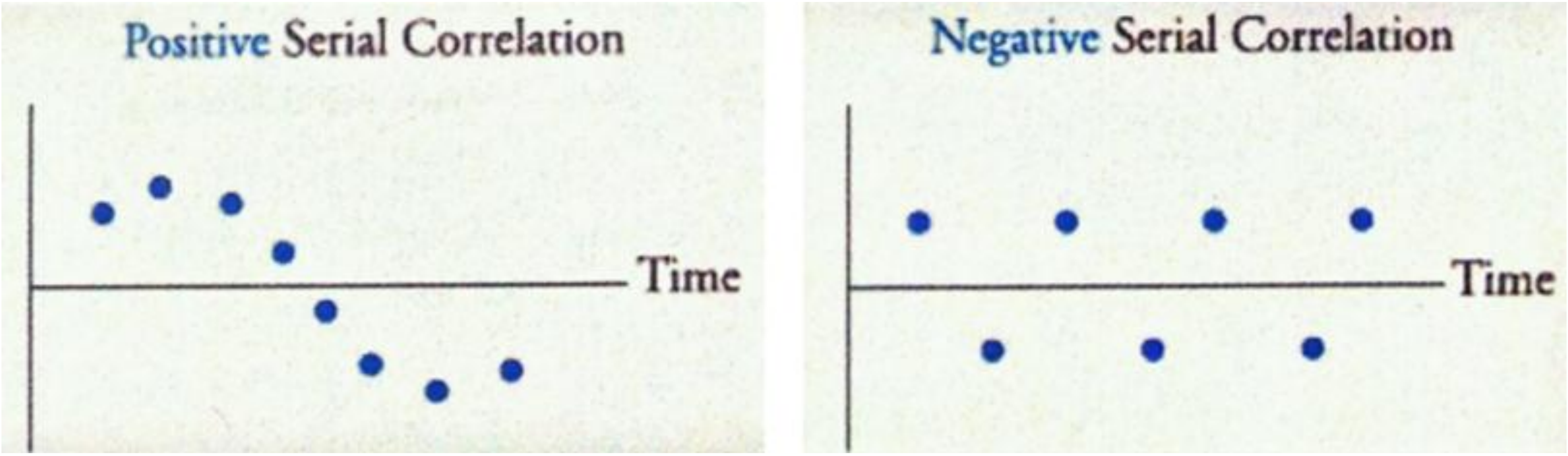
Non-spherical errors – autocorrelation

- In time series terms , if $y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} \ldots + \beta_n x_{nt} + \varepsilon_t$,the first-order autocorrelation $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$
  - where: ε_t = the error term of the equation in question
  - ρ = the first-order autocorrelation coefficient
  - u_t = a classical (not serially correlated) error term
- **The magnitude of $\boldsymbol{\rho}$ indicates the strength of the serial correlation:**
  - If ρ is zero, there is no serial correlation
  - As ρ approaches one in absolute value, the previous observation of the error term becomes more important in determining the current value of ε_t and a high degree of serial correlation exists
- **Autocorrelation ranges:**

$$-\mathbf{1} < \boldsymbol{\rho} < +\mathbf{1}$$

- **Positive**:
  - implies that the error term tends to have the **same sign** from one time period to the next
- **Negative**:
  - implies that the error term has a tendency to **switch signs** from negative to positive and back again in consecutive observations

Positive and negative autocorrelation



Positive autocorrelation; $\rho > 0$ very common and consequential, understates standard errors – there will be a tendency to reject the null hypothesis when it should not be rejected.

Negative autocorrelation; $\rho < 0$ – less common and less consequential.

Detection of autocorrelation – single series

- **Visualization (plot residuals)**
- **Statistical tests:**
- No dependent lags - **Durbin Watson d**: $d = \frac{\sum_t (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_t \varepsilon_t^2}$

  Runs from 0 to 4. No autocorrelation is 2.0, let's say 1.4-2.6 (in smaller samples N=25 just below 2.0) – positive autocorrelation if closer to 0, negative autocorrelation if closer to 4

- **Portmanteau test (Ljung-Box)**

$$Q^* = n(n+2)\sum_{k=1}^{h} \frac{r_k^2}{(n-k)}$$

  where n is the length of the residual time series, r_k is the kth autocorrelation coefficient of the residuals, and h is the number of lags to test. Large values of Q^* indicate that there are significant autocorrelation

- **Breusch-Godfrey test** – allows for multiple lags

  – Breusch-Godfrey is a Lagrange Multiplier (LM) test where we regress residual on X variables and lagged residual(s) and test for significance of latter. For testing we can use either F test or LM statistics and χ^2 distribution

Non-spherical errors - autocorrelation

Autocorrelation within panels

- The individual identity matrices along the diagonal of Ω may be replaced with more general structures to allow for serial correlation.

- In repeated samples these poor estimates will average out, since we are as likely to start with a negative error as with a positive one, leaving the OLS estimator unbiased, but the high variation in these estimates will cause the variance of β_{OLS} to be greater than it would have been had the errors been distributed randomly.

- The temporal dependence exhibited by the errors is also assumed to be time-invariant and it is often assumed be invariant across units. Some of these assumptions may be relaxed.

Dealing with non-spherical errors - clustered sandwich estimator

- In OLS setting we were dealing with heteroskedasticity using Huber/White sandwich estimator or some of its variants

- White showed that $X^T \varepsilon \varepsilon^T X = \sum_{i=1}^{N} \varepsilon_i^2 \, x_i x_i^T$ is a good estimator of the corresponding expectation term $X^T u u^T X$

- Heteroskedasticity consistent estimators are constructed by plugging an estimate of type $\widehat{\Omega} = diag(u_1, \ldots, u_n)$ into central term (i.e. $X^T \widehat{\Omega} X_W$) of $(X^T X)^{-1} X^T \Omega \, X \, (X^T X)^{-1}$

- This can be generalized to the TSCS/CSTS data

Dealing with non-spherical errors - clustered sandwich estimator

- The clustered sandwich estimator applies the logic of White estimator to the data collected across cross-sections and time. The formula for the clustered estimator is simply that of the robust (unclustered) estimator with the individual $\varepsilon_i \times x_i$ replaced by their sums over each cluster

- Clustered sandwich estimator allows for <u>independence across groups (clusters, units) but not necessarily within groups</u>. Clustered sandwich estimator specifies that the standard errors allow for <u>intragroup correlation</u>, relaxing the usual requirement that the observations are independent

- However, clustered sandwich estimator ignores the fact that we assume there is a common variance structure within a cluster and that the intercorrelation across units follows a very specific pattern. **Panel corrected standard errors fix this.**

Clustered sandwich estimator – remarks

- **The cluster–robust estimator requires that there are many clusters.**

- If you have a very small number of clusters compared to your overall sample size it is possible that the standard errors could be larger than the OLS results. Thus, if the number of clusters is small, clustered sandwich estimator is not an improvement over the non-robust option, and in fact it can be substantially worse.

- Experts disagree about just how small is small for these purposes (most would agree that $N > 10 - 15$)

- In addition, you might think your data correlates in more than one way (include fixed-effects for one group and clustered estimator for the other; there are also multi-way clustering options as well - see Cameron, Gelbach and Miller, 2006)

Dealing with non-spherical errors - generalized least squares

- GLS transforms the model in order to remove heteroskedasticity, contemporaneous correlation or autocorrelation from the estimates. It requires the knowledge of the structure of heteroskedasticity, contemporaneous correlation or autocorrelation in order to propose the transformed model

- For **autocorrelation** start with OLS estimation

$$\boldsymbol{Y_t = \beta_0 + \beta_1 X_{1t} + \epsilon_t}$$

where $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$ (due to serial correlation),

- Then lag the new equation by one period and multiply by ρ, obtaining:

$$\boldsymbol{\rho Y_{t-1} = \rho\beta_0 + \rho\beta_1 X_{1t-1} + \rho\epsilon_{t-1}}$$

- Next, subtract the later equation from the former:

$$\boldsymbol{Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_{1t} - \rho X_{1t-1}) + u_t}$$

- GLS is only theoretical, we do not typically know the value of $\boldsymbol{\rho}$ - that is why we use **feasible/estimated GLS**

Feasible generalized least squares – autocorrelation

- ***Cochrane-Orcutt feasible GLS:***

1. Regress $\boldsymbol{Y}$ on $\boldsymbol{X}$
2. Generate residuals, i.e., $\epsilon = Y - \beta_0 - \beta_1 X$
3. Regress current residuals on lagged residuals to produce estimate of $\boldsymbol{\rho}$; $\epsilon_t = \rho\epsilon_{t-1} + u$
4. Transform $\boldsymbol{Y}$ and $\boldsymbol{X}$ using $\boldsymbol{\rho}$ *:* $Y_t^* = Y_t - \rho Y_{t-1}$ and $X_t^* = X_t - \rho X_{t-1}$
5. Regress $\boldsymbol{Y_t^*}$ on $\boldsymbol{X_t^*}$
6. Return to 2 and continue to cycle until convergence (there is little change in $\boldsymbol{\rho}$)

Prais-Winsten approach reintroduces first case with imputation

Generalized least squares – heteroskedasticity

- Let's assume we have an issue of heteroscedasticity. A common way of dealing with this problem is <u>**weighted least square**</u> (WLS)
  - We can think of the heteroscedastic data as being divided in two regions one with high and the other with low heteroscedasticity. The idea of weighted least square is to give more weight to cases coming from low variance and less weight to the cases coming from high variance (i.e. a large deviation in a low variance region is more important).
- So in the case of heteroscedasticity in $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ we know that: $Var(\epsilon_i|X_i) = \sigma^2 X_i$. If we take the equation and divide it with the $\sqrt{X_i}$ we get:

$$\frac{Y_i}{\sqrt{X_i}} = \frac{\beta_0}{\sqrt{X_i}} + \beta_1\sqrt{X_i} + \frac{\epsilon_i}{\sqrt{X_i}}$$

As $Var(cons \times X_i) = cons^2 \times Var(X_i)$ we can only take $\frac{1}{\sqrt{\mathrm{X_i}}}$ out:

$$Var\left(\frac{\epsilon_i}{\sqrt{X_i}}|X_i\right) = \frac{1}{X_i}Var(\epsilon_i|X_i) = \frac{1}{X_i}\sigma^2\ X_i = \sigma^2$$

- So, the transform model is homoscedastic. Similarly, to the case above, we do not know variance σ^2, and we have to use feasible GLS

Parks-Kmenta - feasible generalized least squares in TSCS data

- Similarly to heteroscedasticity and serial correlation, the same logic is applicable to contemporaneously correlated errors

- **Parks-Kmenta** is FGLS for TSCS with <u>serial correlation</u> and <u>contemporaneously correlated</u> errors. The procedure was first described by Parks, and then popularized by Kmenta (hence, Parks-Kmenta methodology). The method was popular in the second half of the 20th century but it is largely considered obsolete now. **This model is not appropriate for <u>TSCS</u> data.**

- The Parks method consists of two sequential FGLS transformations, first estimating serial correlation of the errors and then eliminating contemporaneous correlation.

  1. First you start with OLS, and use residuals to estimate the unit-specific serial correlation of the errors, which is used to transform the model into one with serially independent errors.

  2. Residuals from this estimation are then use to estimate the contemporaneous correlation of the errors and the data is once again transformed to allow for OLS estimation with now spherical errors.

Generalized least squares - properties

- Feasible generalized least squares perform well in large samples. In the limit, it is equivalent to full maximum likelihood, so it has all the optimal asymptotic properties of maximum likelihood.

- GLS estimator can be shown to be the BLUE. Instead of minimizing the sum of squared residuals (OLS estimation), an appropriately weighted sum of squared residuals is minimized. As consequence, the GLS procedure produces a more efficient estimator.

- However, we do not know much about finite sample properties of FGLS. It actually may be biases (and sometimes less efficient that OLS) in finite samples.

Problems with Parks-Kmenta and TSCS

- FGLS underestimates sampling variability (for normal errors) This problem is not severe if there are only a small number of parameters in the variance-covariance matrix to be estimated (as in Cochrane-Orcutt for time series) but the problem is severe if there are a lot of parameters relative to the amount of data.

- Parks-Kmenta method requires estimating an enormous number of parameters for the error covariances. Beck and Katz (1995) showed that properties of this estimator for typical T in TSCS are very bad, and they can underestimate variability by 50%-200%. So, standard errors are way too optimistic – so they recommend not to use this model for TSCS data

(An additional problem with this procedure is that it is basically weighting units by how well they fit the underlying regression)

Panel Corrected Standard Errors

- Panel corrected standard errors correctly measure the sampling variability of the OLS estimates even with panel heteroskedastic and contemporaneously correlated errors.

- **In normal setting any serial correlation must be removed before the panel corrected standard errors are calculated**

  Stata allows you to avoid this (the disturbances may be assumed to be autocorrelated within panel, and the autocorrelation parameter may be constant across panels or different for each panel) – more on this later

- Note that this is a “sandwich estimator” in the style of the Huber-White robust estimator. The only difference is that Ω in $(X^T X)^{-1} X^T \Omega\, X\, (X^T X)^{-1}$ is estimated differently.

PCSE - estimation

- TSCS models with contemporaneously correlated and panel-heteroscedastitic errors have the covariance matrix of the errors, Ω as an $NT \times NT$ block diagonal matrix with an $N \times N$ matrix of contemporaneous covariances along the block diagonal Σ

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \cdots & \sigma_{1,N} \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} & \cdots & \sigma_{2,N} \\ & & & \vdots & \\ \sigma_{1,N} & \sigma_{2,N} & \sigma_{3,N} & \cdots & \sigma_N^2 \end{pmatrix} \qquad \Omega = \begin{pmatrix} \Sigma & 0 & 0 & \ldots & 0 \\ 0 & \Sigma & 0 & \ldots & 0 \\ 0 & 0 & \Sigma & \ldots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \ldots & \Sigma \end{pmatrix}$$

- Σ is the panel-by-panel covariance matrix. We thus have T replications of the error that can be used to estimate this block; for large T, this estimate is quite good.

PCSE - estimation

- We need to estimate these diagonals, Σ. As OLS is consistent we can use its residual. Let $e_{i,t}$ be the OLS residuals for unit i and time t. We can estimate typical element of Σ as:

$$\hat{\Sigma}_{i,j} = \frac{\sum_{t=1}^{T} e_{i,t} e_{j,t}}{T}$$

- To calculate PCSE, organize the residuals from the fitted model according to the group. Letting E denote the $T \times N$ matrix of the OLS residuals we can estimate Ω by:

$$\Omega = \frac{(E'E)}{T} \otimes I_t$$

- Panel corrected standard errors are then estimated by taking square root of the diagonal elements:

$$(X'X)^{-1} X' \left(\frac{E'E}{T} \otimes I_t \right) X (X'X)^{-1}$$

PCSE

- Unlike White errors, which only accounts for regular heteroskedactity, this estimator accounts for panel heteroskedactity and contemporaneous correlation of errors.

- **Panel-corrected standard error (PCSE) estimates are estimated by either OLS or Prais–Winsten regression (FGLS).**

- **Prais–Winsten estimates are used when autocorrelation is specified!!!**

(Obviously, it would be good to model variables that lead to the error complications, but in practice this can be hard to do, so OLS with PCSE is a practical solution)

PCSE vs. FGLS

- Panel-corrected standard error is an alternative to feasible generalized least squares (FGLS)
- The full FGLS variance–covariance estimates are typically unacceptably optimistic (anticonservative) when used with the type of TSCD data analyzed by most social scientists — $10–20$ panels with $10–40$ periods per panel.
- Beck and Katz (1995) show that the OLS or Prais–Winsten estimates with PCSEs have coverage probabilities that are closer to nominal.
- The correction works only because we have repeated information on the contemporaneous errors. **Thus, the method works only in TSCS context.**