Generalized CRM

Devesh Raval

In the Classical Regression Model we assumed that the variance of the error was $\sigma^2 I$: that is, with the same variance across observations and no correlation between observations. In the GCRM we relax this assumption:

$$\begin{aligned} E(Y|X) &= X\beta \\ E(\epsilon|X) &= 0 \\ V(Y|X) &= \Sigma \\ V(\epsilon|X) &= \Sigma \end{aligned}$$

Remember Y is n by 1 so $V(Y|X) = \Sigma$ is n by n- write out the matrix.

Heteroskedasticity- means different diagonal terms.

Autocorrelation/Serial Correlation- non zero off diagonal terms.

First- what are the finite sample properties of the LS Estimator?

Dont want to go immediately to asymptotics b/c in the case of correlated errors- no longer have iid random sampling. So need a new set of asymptotic theorems for some "limited" types of correlation.

LS Estimator:

$$\begin{aligned} b_N &= (X'X)^{-1}X'Y \\ E(b_N|X) &= (X'X)^{-1}X'E(Y|X) \\ &= (X'X)^{-1}X'X\beta = \beta \\ E(b_N) &= \beta \end{aligned}$$

where the last line is LIE.

Thus, OLS is still an unbiased estimator.

The variance of the OLS estimator changes however:

$$\begin{aligned} V(b_N|X) &= (X'X)^{-1}X'V(Y|X)X(X'X)^{-1} \\ &= AV(Y|X)A' \\ &\neq \sigma^2(X'X)^{-1} \end{aligned}$$

Thus, the regular standard errors that we learned earlier are incorrect.

The Gauss Markov Theorem relied on $V(Y|X) = \sigma^2 I$- thus the Gauss Markov Theorem no longer applies. OLS is no longer the min variance linear unbiased estimator.

Two possibilities here:

1. Continue to use the OLS estimator but correct the standard errors.

2. Use information about the new variance matrix to suggest a new estimator.

For right now will cover 2- but will eventually go back to 1 at the end!

Example of 2- Imagine we know the variance of some observations are higher than others- can improve the estimator by giving more weight to more precise estimates- can get a lower variance matrix of the estimates than $(X'X)^{-1}X'\Sigma X(X'X)^{-1}$.

1 Generalized Least Squares Estimator (GLS)

Don't just minimize the sum of squared residuals as we did for OLS.

$$b_N^* = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$$

This is like one of your problem set questions- where Δ replaced Σ^{-1}.

Here Σ^{-1} is positive definite and non stochastic. In the GLS case we know Σ (at least up to a constant of proportionality).

What are the small sample properties of this new estimator?

$$\begin{aligned} E(b_N^*|X) &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}E(Y|X) \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}X\beta \\ &= \beta \end{aligned}$$

$$\begin{aligned} V(b_N^*|X) &= A^* V(Y|X) A^{*'} \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\Sigma\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1} \\ &= (X'\Sigma^{-1}X)^{-1} \end{aligned}$$

Theorem 1. *Aitken's Theorem:*

In the GCRM model, with Σ known, the minimum variance linear unbiased estimator of β is b_N^*.

If $\Sigma = I$, then we have the LS estimator.

Thus, this theorem is a generalization of the Gauss-Markov Theorem.

Proof. Show that the GCRM is equivalent to the CRM on transformed data.

First we can diagonalize the variance matrix Σ (where we use the fact that its positive definite and symmetric):

$$\Sigma = C\Omega C'$$

where C is a matrix containing all n eigenvectors and Ω is a matrix with 0 off diagonals and diagonals of the eigenvalues. Now, we know $CC' = C'C = I$ since the eigenvectors are of unit value $(c_i'c_i) = 1$ and are linearly independent. Then:

$$\begin{aligned}\Sigma^{-1} &= (C\Omega C')^{-1} \\ &= (C')^{-1}\Omega^{-1}C^{-1} \\ &= C\Omega^{-1}C'\end{aligned}$$

Label $P = C\Omega^{-1/2}C'$.
Then:

$$\begin{aligned}PP' &= C\Omega^{-1/2}C'C\Omega^{-1/2}C' \\ &= C\Omega^{-1}C' = \Sigma^{-1}\end{aligned}$$

and

$$\begin{aligned}P\Sigma P' &= C\Omega^{-1/2}C'C\Omega C'C\Omega^{-1/2}C' \\ &= I\end{aligned}$$

Now let us transform the data- weighting Y and X by the inverse of the square root of the variance matrix-

$$\begin{aligned}Y^* &= PY \\ X^* &= PX\end{aligned}$$

Then the transformed data satisfy the assumptions of the CRM:

$$\begin{aligned}E(Y^*|X^*) &= E(PY|X) = PX\beta = X^*\beta \\ V(Y^*|X^*) &= V(PY|X) = PV(Y|X)P' \\ &= P\Sigma P' = I \\ Rank(X^*) &= Rank(X) = k\end{aligned}$$

Thus, the transformed data meets the assumptions of the CRM.
Now, if we regress Y^* on X^*:

$$\begin{aligned}X^{*'}X^* &= X'P'PX = X'\Sigma^{-1}X \\ X^{*'}Y^* &= X'P'PY = X'\Sigma^{-1}Y \\ b_N^{**} &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y\end{aligned}$$

So the OLS estimator in the transformed case is the same as the GLS estimator. If we apply the Gauss-Markov Thm on the transformed data we can see that GLS is now the min variance linear unbiased estimator or BLUE. □

1. Notice here that b_N^* is the solution to the minimization of the sum of squared residuals problem for the transformed data:

$$\begin{array}{rcl} & \min_c & U^{*'}U^* \\ U^* & = & Y^* - X^*c \\ & = & PY - PXc \\ & = & P(Y - Xc) \\ & \min_c & U'\Sigma^{-1}U \\ & \min_c & (Y - Xc)'\Sigma^{-1}(Y - Xc) \end{array}$$

This latter problem is the square of the Mahalanobis distance between Y and Xc. The Mahalanobis distance formula is adjusting for different variances (higher variance observations should be downweighted as they provide less info) and correlations (two highly correlated obs provide less info than 2 uncorrelated obs- if perfect correlation like have 1 obs only). Efficient GMM does the same thing for moments.

2. This applies for Σ known exactly or up to some factor of proportionality.

e.g., $\Sigma = \sigma^2\Omega$, σ^2 is unknown. But if Ω is known, then we can apply GLS.

$$V(Y^*|X^*) = \sigma^2 I$$

just as in the OLS case.

3. Estimation in practice- you can always transform the data and then run an OLS regression.

2 Feasible GLS

If Σ is unknown, what can we do?

Feasible GLS is a two step estimator.

1. Estimate Σ_N with Σ.

2. Use the GLS estimator using Σ_N in place of Σ.

If Σ_N is a consistent estimator of Σ, then FGLS has the same asymptotic properties as GLS.

But need some prior information on the variance matrix.

Need some prior information- b/c for a consistent estimator of variance matrix, need to have more information on each parameter in matrix as sample size increases

If the variance of every observation is different- no more info on any element of matrix as sample size rises- "Incidental Parameters Problem"

One ex- individual specific parameters on slope- goes into error- unless parameterize through some distn.

Some examples we will cover:

1. Pure heteroskedasticity and some info on form (women and men have different errors).

2. Homoskedasticity and off diagonal terms have covariances that depend on a single parameter.

2.1 Pure Heteroskedasticity

Pure heteroskedasticity- examples

y= food consumption, x= income

poor people-only eat fast food

rich people- sometimes fast food, sometimes expensive food

or

y= avg class grade

bigger classes have more obs in average- so lower variance

Basic idea of the GLS estimator:

Provide more weight to more precise observations, so optimally reweight data.

Sometimes called "Weighted Least Squares".

We have the following regression model:

$$\begin{aligned} E(Y|X) &= X\beta \\ V(Y_i|X_i) &= \sigma_i^2 \\ Rank(X) &= k \end{aligned}$$

Write out Σ- different diagonals and zero off diagonals.

P n by n matrix- diagonal matrix with $\frac{1}{\sigma_i}$ on the diagonal. (Write it out).

Now- verify that P satisfies the earlier properties:

$$\begin{aligned} PP' &= \Sigma^{-1} \\ P\Sigma P' &= I \end{aligned}$$

Three approaches:

1. Run OLS and adjust the standard errors to match arbitrary heteroskedasticity- White std. errors-will cover this case later.

2. Suppose Σ is known. Implement the GLS estimator.

$$\begin{aligned} b_N^* &= (X'\Sigma X)^{-1}(X'\Sigma^{-1}Y) \\ Y^* &= PY \\ X^* &= PX \end{aligned}$$

What are these P matrices doing?

$$\begin{aligned} Y_i^* &= Y_i(1/\sigma_i) = Y_i/\sigma_i \\ X_i^* &= X_i/\sigma_i \end{aligned}$$

Thus, more weight is given to observations with smaller variances.

Then regress Y^* on X^*- get b_N^*.

$$\begin{aligned} E(Y|X) &= X\beta = \beta_1 * 1 + \beta_2 * X_2 + \beta_3 * X_3 + ... + \beta_k X_k \\ \hat{y}^* &= b_1^* X_1 + b_2^* X_2^* + ... + b_k^* X_k^* \end{aligned}$$

Here $X_1 = 1/\sigma_i$- so don't put a constant/intercept in the new regression!!

If iid sampling, can apply asymptotic results and get consistency.

3. Σ is unknown- so we need a consistent estimator of it.

One example- White Standard error variance matrix- will cover later.

Or have some prior info.

Then we have the following two step procedure:

1. Consistently estimate Σ.

$$\Sigma_N \quad \rightarrow_p \quad \Sigma$$

2. Treat Σ_N as if it were Σ and implement GLS.

What prior information?

One example- homoskedasticity- its just the identity matrix, up to a constant of proportionality.

Another example-

Split the n observations as follows- the first set are men and second set women.

Assume men all have the same variance and women all have the same variance- but men and women can have different variances.

Thus- instead of having n unknown parameters- just have 2 unknown parameters.

Notice- as N increases- as long as getting more men and more women- will have more info to estimate both parameters.

If we can reduce the number of unknown parameters such that Σ can be consistently estimated using some function of e, then we can use FGLS.

Continuing example-

Observations of men have σ_m^2, observations of women have σ_w^2.

Then we have two unknowns (σ_m^2, σ_w^2) and we know $\sigma_m^2 = V(\epsilon|male = 1)$, $\sigma_w^2 = V(\epsilon|male = 0)$.

Order male observations first (WLOG)

The FGLS strategy would be:

1. Run LS regression of Y on X

$$e = [e_m e_w]$$

2. Calculate the variance matrix- in this case variance of men and variance of women:

$$\begin{aligned} \sigma^2_{mN} &= \frac{e'_m e_m}{N_m} \\ \sigma^2_{wN} &= \frac{e'_w e_w}{N_w} \end{aligned}$$

3. Transform Y and X by dividing the male observations by σ_{mN} and female observations by σ_{wN}.

4. Run LS regression on transformed variables.

Another ex- school size- if sample mean them know variance of mean is σ^2/N- so variances of observations depend on N- so here only one parameter to estimate.

2.2 Autocorrelation

Models of dependence in data- most common application is time, but can also have spatial data.

For example: Expect crime rate in urban core to influence crime rate in expanding spatial rings around it, but this effect will decay with distance.

Another example: Macro model with shocks and impulse responses.

Both models we will study are stationary- a series is (weakly) stationary if the mean and variance exist and unconditional variances and covariances do not depend on the time t, i.e.

$$E(\epsilon_t \epsilon_{t-j})$$

does not depend on t (though it can depend on the lag j!)

First Model: AR(1) Model

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_t + \epsilon_t \\ \epsilon_t &= \rho \epsilon_{t-1} + u_t \end{aligned}$$

Today's error depends on yesterday's error, downweighted, plus a random shock.

Here u_t is iid, so

$$\begin{aligned} Cov(u_t, u_{t-j}) &= 0 \\ Var(u_t) &= \sigma^2 \perp t \end{aligned}$$

For "stationarity"- define $|\rho| < 1$. This means that dependence decays over time.

What are the properties of ϵ_t? Of the variance matrix?

It turns out that we can write ϵ_t as an infinite sum of all of the u_t:

$$\begin{aligned} \epsilon_t &= \rho\epsilon_{t-1} + u_t \\ &= \rho(\rho\epsilon_{t-2} + u_{t-1}) + u_t \\ &= \rho^2\epsilon_{t-2} + \rho u_{t-1} + u_t \\ &= \rho^2(\rho\epsilon_{t-3} + u_{t-2}) + \rho u_{t-1} + u_t \\ &= (\rho^3\epsilon_{t-3} + \rho^2 u_{t-2}) + \rho u_{t-1} + u_t \\ &= \rho^l \epsilon_{t-l} + \sum_{j=0}^{l-1} \rho^j u_{t-j} \end{aligned}$$

As long as $|\rho| < 1$, we know that

$$\begin{aligned} \lim_{l\to\infty} \rho^l \epsilon_{t-l} &= 0 \\ \epsilon_t &= \sum_{j=0}^{\infty} \rho^j u_{t-j} \end{aligned}$$

thus ϵ_t is an infinite sum of discounted iid shocks (instead of one iid shock as before).

Then we have that:

$$\begin{aligned} Cov(\epsilon_{t-1}, u_t) = Cov(\epsilon_t, u_{t+1}) &= Cov(\sum_{j=0}^{\infty} \rho^j u_{t-j}, u_{t+1}) \\ &= 0 \end{aligned}$$

since the $\{u_t\}$ are iid.

Since the $\{u_t\}$ are iid the variance of ϵ_t should not depend on t (stationarity!!)

$$\begin{aligned} Var(\epsilon_t) &= Var(\rho\epsilon_{t-1} + u_t) \\ \sigma_\epsilon^2 &= \rho^2\sigma_\epsilon^2 + \sigma_u^2 + 2\rho Cov(\epsilon_{t-1}, u_t) \\ &= \rho^2\sigma_\epsilon^2 + \sigma_u^2 \\ \sigma_\epsilon^2(1-\rho^2) &= \sigma_u^2 \\ \sigma_\epsilon^2 &= \frac{\sigma_u^2}{1-\rho^2} \end{aligned}$$

So the variance is homoskedastic- doesnt depend on time i.e. on observation. We can also calculate the covariances:

$$\begin{aligned} Cov(\epsilon_t, \epsilon_{t-1}) &= Cov(\rho\epsilon_{t-1} + u_t, \epsilon_{t-1}) \\ &= \rho Var(\epsilon_{t-1}) + Cov(u_t, \epsilon_{t-1}) \\ &= \rho Var(\epsilon_{t-1}) + 0 \\ &= \rho\sigma_\epsilon^2 \end{aligned}$$

$$\begin{aligned} Cov(\epsilon_t, \epsilon_{t-2}) &= Cov(\rho\epsilon_{t-1} + u_t, \epsilon_{t-2}) \\ &= \rho Cov(\epsilon_{t-1}, \epsilon_{t-2}) + Cov(u_t, \epsilon_{t-2}) \\ &= \rho Cov(\epsilon_{t-1}, \epsilon_{t-2}) + 0 \\ &= \rho(\rho\sigma_\epsilon^2) = \rho^2\sigma_\epsilon^2 \end{aligned}$$

Notice nothing here depends on t- only on the lag. If we continue this, we get the following covariance between t and its jth lag:

$$Cov(\epsilon_t, \epsilon_{t-j}) = \rho^j\sigma_\epsilon^2$$

The dependence between terms slowly dies out as the lag gets further out. Notice the entire variance matrix can be described by two parameters- σ_ϵ^2 and ρ!!

Write out the entire variance matrix- diagonals are $\frac{\sigma_u^2}{1-\rho^2}$, off diagonals $\rho^j * \frac{\sigma_u^2}{1-\rho^2}$ so multiply $\frac{\sigma_u^2}{1-\rho^2}$ by matrix.

The other most common Correlation process- Moving Average Process:

$$\begin{aligned} \epsilon_t &= \lambda u_{t-1} + u_t \\ var(\epsilon_t) &= \lambda^2 Var(u_{t-1}) + Var(u_t) \\ &= (1+\lambda^2)\sigma_u^2 \\ Cov(\epsilon_t, \epsilon_{t-1}) &= Cov(\lambda u_{t-1} + u_t, \lambda u_{t-2} + u_{t-1}) \\ &= \lambda\sigma_u^2 \\ Cov(\epsilon_t, \epsilon_{t-2}) &= Cov(\lambda u_{t-1} + u_t, \lambda u_{t-3} + u_{t-2}) \\ &= 0 \end{aligned}$$

Can generalize these two:

$$\epsilon_t = \rho\epsilon_{t-1} + u_t + \lambda u_{t-1}$$

ARMA(1,1)- 1s are the order of the AR and MA process- can have higher orders.

I will go the basic GLS/FGLS approach to the AR(1) process.

Can think of this in two ways- difference out the previous time period so the only error left is u_t:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_t + \epsilon_t \\ \rho y_{t-1} &= \rho\beta_0 + \beta_1 \rho x_{t-1} + \rho\epsilon_{t-1} \\ (y_t - \rho y_{t-1}) &= (1-\rho)\beta_0 + \beta_1(x_t - \rho x_{t-1}) + u_t \\ y_t^* &= (1-\rho)\beta_0 + \beta_1 x_t^* + u_t \end{aligned}$$

where:

$$\begin{aligned} y_t^* &= y_t - \rho y_{t-1} \\ x_t^* &= x_t - \rho x_{t-1} \end{aligned}$$

This procedure is differencing out the dependency- so now the errors are iid as before!

What do we do for $t = 1$? In that case there is no x_0. Instead, can not include first observation, or weight down by the variance of the first observation-

$$Var(\epsilon_1) = \frac{\sigma_u^2}{1-\rho^2}$$

so weight first observation by $\sqrt{1-\rho^2}$.

This procedure is equivalent to formally diagonalizing the variance matrix and computing P.

P will become:

$$\begin{bmatrix} \sqrt{1-\rho^2} & & & & \\ -\rho & 1 & & & \\ & -\rho & 1 & & \\ & & -\rho & 1 & \\ & & & -\rho & 1 \end{bmatrix}$$

Remember up to a constant of proportionality, which is σ_u^2.

GLS case- we know ρ- so estimate this transformed regression.

FGLS- we will need to compute ρ using $\hat{\rho}$.

If ρ is unknown, what is it in the model?

$$\begin{aligned} Cov(\epsilon_t, \epsilon_{t-1}) &= \rho\sigma_\epsilon^2 \\ \frac{Cov(\epsilon_t, \epsilon_{t-1})}{Var(\epsilon_{t-1})} &= \rho \end{aligned}$$

This is just the bivariate slope coefficient of a regression of ϵ_t on ϵ_{t-1}.
So we have the following strategy:
1. Run OLS of Y on X- get residuals e (n by 1 vector).
2. Regress e_t on e_{t-1} (obv dropping one observation).
Get the slope coefficient

$$\hat{\rho} = \frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=2}^{T} e_{t-1}^2}$$

3. Transform the data using $\hat{\rho}$ in place of ρ.

2.3 HACC Estimation and Clustering

(This section is incomplete).

The other approach we can take is to estimate a variance matrix that is robust to arbitrary levels of heteroskedasticity or autocorrelation while still using OLS- this is the typical approach used now.

Remember that the variance of the OLS estimator under the GCRM is:

$$V(b_N|X) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$$

Since we know $(X'X)^{-1}$ will consistently estimate its population counterpart the only challenge is to find a consistent estimator V_N for $V = \frac{1}{n}X'\Sigma X \rightarrow_p E[X'\Sigma X]$. The challenge again is lack of knowledge of Σ.

First- let us write out what V will look like:

$$\begin{array}{cccc} & & & \\ X_1 & X_2 & \ldots & X_n \\ & & & \end{array} \begin{array}{cccc} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & & \\ & & & \\ \sigma_{n1} & & & \sigma_{nn} \end{array} \begin{array}{c} X_1' \\ X_2' \\ \ldots \\ X_n' \end{array}$$

Remember each X_i here is k by 1- multiplying the first two we get:

$$\begin{array}{cccc} \sum_j X_j\sigma_{1j} & \sum_j X_j\sigma_{2j} & \ldots & \sum_j X_j\sigma_{nj} \end{array} \begin{array}{c} X_1' \\ X_2' \\ \ldots \\ X_n' \end{array}$$

LHS is k by n, RHS is n by k

Multiplying it all out we get:

$$V = \frac{1}{N} X'\Sigma X \quad = \quad \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_{ij} X_i X_j'$$

If we separate the variance and covariance terms in this gigantic sum, we get:

$$\begin{aligned} V = \frac{1}{N} X'\Sigma X &= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_{ij} X_i X_j' \\ &= \frac{1}{N} \sum_{i=1}^{N} \sigma_{ii} X_i X_i' + \frac{1}{N} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \sigma_{ij} (X_j X_{j-i}' + X_{j-i} X_j') \end{aligned}$$

I will proceed in a series of cases:
Pure heteroskedasticity:

$$\begin{aligned} E(\epsilon_i \epsilon_j | X_i) &= 0 \forall i \neq j \\ V(\epsilon_i | X_i) &= \sigma_{ii} \end{aligned}$$

Here if we knew Σ, since all off-diagonal terms are zeroed-out, we will have:

$$\begin{aligned} V &= X'\Sigma X \\ &= \frac{1}{N} \sum_{i=1}^{N} \sigma_{ii} X_i' X_i \end{aligned}$$

Since we dont know σ_{ii}, we will replace with the sample analog from the residuals.

$$V_N \quad = \quad \frac{1}{N} \sum_{i=1}^{N} e_i^2 X_i' X_i$$

Here $e_i^2 = (Y_i - X_i b_N)^2$ or is the squared residual from OLS. Is this estimator consistent?

Asymptotics are more complicated as e is a function of the OLS estimator- so its value is changing as N changes as well as terms being added to the sum.

Under some regularity conditions (so we have Uniform Law of Large Numbers- won't cover these in class):

$$\begin{aligned} V_N &= \frac{1}{N}\sum_{i=1}^{N} e_i^2 X_i' X_i \\ &\to_p E(\epsilon_i^2 X_i' X_i) \\ &= E(E(\epsilon_i^2|X_i) X_i' X_i) \\ &= E(\sigma_{ii} X_i' X_i) \end{aligned}$$

The last line is by LIE.

Thus, V_N converges to the right variance matrix- these are the White Std errors. They work for arbitrary degrees of heteroskedaticity.

Autocorrelation:

Here we know that we can not just zero out the correlation terms- what to do?

Let's assume regularity conditions so asymptotics hold under dependent sampling- one general case is time series mixing, so we have approximate independence as the distance between observations gets large.

Lets first take the MA(1) case- so we know there is correlation at the first lag only- but now want to relax the weak stationary assumption. Then V is:

$$V = \frac{1}{N} X' \Sigma X \qquad = \frac{1}{N}\sum_{i=1}^{N} \sigma^2 X_i X_i' + \frac{1}{N}\sum_{i=2}^{N} \sigma_{i,i-1}(X_i X_{i-1}' + X_{i-1} X_i')$$

We would then use the sample analog of $\sigma_{i,i-1}$ - $e_i e_{i-1}$- so our estimator is:

$$V_N = \frac{1}{N} X' \Sigma X \qquad = \frac{1}{N}\sum_{i=1}^{N} e_i^2 X_i X_i' + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N} e_i e_j 1(|i-j| \leq 1)(X_j X_{j-i}' + X_{j-i} X_j')$$

We can also write this as:

$$V_N = \frac{1}{N} X' \Sigma X \qquad = \frac{1}{N}\sum_{i=1}^{N} e_i^2 X_i X_i' + \frac{1}{N}\sum_{i=2}^{N} e_i e_{i-1}(X_i X_{i-1}' + X_{i-1} X_i')$$

By similar asymptotics as before will converge to $E(\epsilon_i \epsilon_{i-1}(X_i X_{i-1}' + X_{i-1} X_i'))$- then LIE will give $E(\sigma_{i,i-1}(X_i X_{i-1}' + X_{i-1} X_i'))$.

If we have a MA(q): correlation for q periods- then will have q such lags- as follows:

$$V_N = \frac{1}{N} X'\Sigma X \qquad = \frac{1}{N}\sum_{i=1}^{N} e_i^2 X_i X_i' + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N} e_i e_j 1(|i-j| \leq q)(X_j X_{j-i}' + X_{j-i} X_j')$$

in the general case you would have:

$$V_N = \frac{1}{N} X'\Sigma X \quad = \quad \frac{1}{N}\sum_{i=1}^{N} \sigma^2 X_i X_i' + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N} \lambda_T(l) e_i e_j (X_j X_{j-i}' + X_{j-i} X_j')$$

Why is $\lambda_T(l)$ needed? At very large lags there are few observations- so the variance of those parts of the variance matrix is huge- hence λ is a weighting function to weight them down, or zero them out. There is a bias here from excluding parts so there is a bias-variance tradeoff. Example- $\lambda = 1$ if below some lag cutoff. Triangle weights also possible.

2.4 Clustering

Clustering addresses a similar problem- correlations between observations in the same group- the only difference is group is not a lag in time series, can be anything.

Ex. state or Dominicks store or product

One ex- observe wages and schooling- but state level shocks in wage errors (demand shocks across states)- so error in wages will be correlated within state- less indep observations

The underlying regression model is now:

$$Y_{ig} \quad = \quad X_{ig}'\beta + u_{ig}$$

where g is the group or cluster- we want arbitrary correlation within this group. One type of error that could generate this would be:

$$u_{ig} \quad = \quad \alpha_g + \nu_{ig}$$

α_g is a random component common to everyone in the group, while ν_{ig} is iid.

In the general case, the V matrix now becomes:

$$V_N = \frac{1}{N} X'\Sigma X \qquad = \frac{1}{N}\sum_{i=1}^{N} e_i^2 X_i X_i' + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N} e_i e_j 1(i \text{ and } j \in g)(X_j X_{j-i}' + X_{j-i} X_j')$$

Thus, we only include the covariance terms for same group observations- using the OLS residual correlation to proxy for the true correlation. This procedure is then robust to arbitrary forms of correlation within the group!

Clustering now is very common in empirical work.

This will work as long as have enough clusters- can't have just one cluster and so arbitrary correlation across all observations (same as in time series case- why need weighting function). In general need at least 50 clusters- asymptotics become as number of clusters goes to infinity.