AP: Duflo (2001)

Alessandro Caggia

June 2025

Abstract

1. Robust Standard Errors

- A typical strategy in applied work is to report **heteroskedasticity-robust standard errors** for OLS regressions ("robust" option in Stata).
- Estimator for the variance of the asymptotic distribution of the OLS estimator that is **(asymptotically) consistent both under homo- and heteroskedasticity.**

Under homoskedasticity:

- OLS standard errors are not only **consistent**, but also the **best linear unbiased estimate** (BLUE, most efficient).
- Therefore, it is not recommended to use the "robust" ones.

Under heteroskedasticity:

- OLS estimator of standard errors is **inconsistent**, but the "robust" one is consistent.
- When the regression approximates a non-linear CEF, we have heteroskedastic errors in general, so better to use the robust one. Eg As soon as the dependent variable is not a continuous variable (e.g. binary), the CEF is not linear (easy proof). Then you will have, **by construction**, heteroskedastic errors.

- Since we do not know the CEF or form of heteroskedasticity, **in large samples it is safer to use robust standard errors.**
- But, both the OLS and the robust estimators are **biased in small samples**, and the robust ones can be **more biased**.
- **Conservative strategy:** use the larger of OLS (conventional) and robust standard errors, typically the robust ones.

2. Clustering and the Moulton Factor

- With large samples, we do not need to care too much about heteroskedasticity.
- However, **clustering (group structures) can affect standard errors substantially**, even with large samples.
  - *E.g., we randomize at the school level, but outcomes are measured at the student level.*
  - *Clustering impacts the off-diagonal terms in the variance-covariance matrix of errors.* Basically you have a block diagonal matrix: **correlation within groups**, but **no correlation across groups**! This affecs the error term. Obs are no longer iid. You udnerestimate standard errors. yo uoverestiamte the precision of your esitmator (the idea is that the true variabilty in rela workd is higer, and it is as if you have lower than relanumber of observatiosn to learn from the data)
- **Failure to control for within-cluster error correlation** can lead to:
  - Misleadingly small standard errors,
  - Inflated t-statistics,
  - Falsely small p-values.
- If we assume that **non-diagonal terms in the variance-covariance matrix are constant**, we can compute the exact bias and derive a correction: the **Moulton factor**.

Model for Clustered Observations

$$Y_{ig} = \beta_0 + \beta_1 X_g + e_{ig}$$

- $g = 1, \dots, G$ groups. Covariates X vary at group level.
- Test scores in the same school are correlated ⇒ errors e_{ig} are correlated.
- $\mathbb{E}(e_{ig}e_{jg}) = \rho_e\sigma^2 \neq 0$, about two observations i and j within the same group g.
- where considering a **Random effects model:** $e_{ig} = \upsilon_g + \eta_{ig}$, so we can decopmose the observation i error as the sum of a group level error and an individual level error. The idea is putting all of the correlation among individuals beonging to the same clusetr in this term and assume η_{ig} is hmoskedastic and uncorrelated across individauls, if yo udo dthis you find β_1
- **Intraclass correlation coefficient:**

$$\rho_e = \frac{\sigma_\upsilon^2}{\sigma_\upsilon^2 + \sigma_\eta^2}$$

  This measure how similar are observation within a class.
- Interpretation
  - If $\sigma_\upsilon^2 = 0 \Rightarrow \rho_e = 0$
    All variance is from individual noise ⇒ No within-group correlation.
  - If $\sigma_\eta^2 = 0 \Rightarrow \rho_e = 1$
    All individuals in a cluster have the same error ⇒ Perfect intra-cluster correlation.
- Implications for Standard Errors
  - When $\rho_e > 0$, individuals in the same group are not independent.
  - If you ignore this and assume i.i.d. errors:
    - Your standard errors will be too small.
    - You overstate precision, leading to over-rejection of null hypotheses.

Moulton Factor

$$\sqrt{\frac{V(\hat{\beta}_1)}{V_c(\hat{\beta}_1)}} = \sqrt{1 + (n-1)\rho_e}$$

- $V(\hat{\beta}_1)$: correct variance, $V_c(\hat{\beta}_1)$: conventional variance ignoring clustering.
- n: average group size.
- **Bias increases with group size n and ρ_e.**
- as soon as rho is hreater than zero and rgoup have mroe than 2 observations yo uhaev true standard errors > those asusmed udner homoskedasticityò.
- if rho is equal to zero true se are euql to hooskeedastij ones
- When $\rho_e = 1$, bias factor $= \sqrt{n}$. With rho $= 1$, **the bias is of size n.** No additional info from more observations in same group.So, n observations within a cluster are not adding any independent

information. But if you mistakenly treat them as n independent observations, you overestimate the information by a factor of n. All errors within group are exactly the same, thus their outcome should be the same, and more observations do not provide any additional information! you get nor info by observing dierent usnits within the same group, you ahve actually n oabservaiton = n number of hroup.

General Moulton Factor The covariates play a key role in the generalized Moulton correction because the bias from clustering depends not only on the correlation of the errors within groups, but also on the correlation of the covariates within groups.

$$\sqrt{\frac{V(\hat{\beta}_1)}{V_c(\hat{\beta}_1)}} = \sqrt{1 + \left(\frac{V(n_g)}{E(n_g)} + E(n_g) - 1\right)\rho_e\rho_X}$$

- ρ_X: intraclass correlation of X.
- Formula adapts if n_g not constant (different group sizes).

Correcting Clustered SE: 5 Approaches

1. **Parametric**: use Moulton formula. Mouoto correciton is a pramatric rule to corect for itnraclass correlaiton. THE CORREITON IS TRIVIAL, SVEGLIA GUARDA LA OFMULA SU ED APPLICALA!
2. **Cluster SE**: flexible. Consistent if G (number of groups) is large.
3. **Group Averages**: most conservatve startegy: constraint myself to running a reg at the groups levels (throuw away all incdividual level ifnormaiton). ISSUE, SCREENSHOT!
4. **GLS or MLE**: One would estimate the parameters of the error correlation model, and estimate the original model by feasible generalized least squares. Not robust to wrong specifications.
5. **Block Bootstrap**: resample clusters, compute $\hat{\beta}_1^{(b)}$, and estimate variance across B bootstrap samples. Need to do bootstrap with blocks to keep in all bootstrap sampels the group structure!

When to Use Cluster Option in Stata?

- Must have a random sample of clusters from a super-population with large G (BASICALLY, LERGE NUMEBR OF GROUPS).
- With regressor randomized **within clusters** ⇒ clustering not needed.
- With regressor randomized **at cluster level** ⇒ **must cluster**.

3. Bootstrapping

1. **Block (or pairs) Bootstrap**: re-sample entire clusters instead of individuals. Do the following steps B times (recommended $B \geq 400$):
   - 0.1 Form G clusters based on (y_i, x_i) by resampling with replacement G times from the original sample of clusters.
   - 0.2 Compute B estimates, one $\hat{\beta}_i$, for each bootstrap sample. Compute the variance of the B estimates.
2. **Problem**: does not have much better properties than cluster option. But, you can use it for methods when that option is not available.

Resampling technique that can be used as an alternative to inference based on asymptotic formulas.
Draw repeatedly from your sample of N observations, as if it were the population.
Get the bootstrap sampling distribution as the distribution of the estimator across the different draws.
The bootstrap can be useful when we do not have asymptotic formulas for the estimator, or for asymptotic refinements (i.e. reduction in finite sample
bias for consistent estimators).

4. Serial Correlation

4.1. Serial Correlation in Panels and DID Models

Suppose the treatment d_{ts} varies at the state level and over time, and we have individual-level data:

$$y_{its} = \gamma d_{ts} + x'_{its}\delta + \alpha_s + \delta_t + u_{its}$$

- **Key concern:** with serial correlation in errors *over time within states*, we must adjust standard errors for:
  - group correlation (individuals in the same state),
  - and serial correlation (within a state over time).

Cases

1. **Serial correlation but no group correlation**:
   Use robust SEs or ARIMA-based methods to model serial correlation directly.
2. **No serial correlation, but group correlation within states** (e.g., shocks independent over time but correlated within state):
   Cluster SEs at the group*time level.
3. **Both serial correlation (within state) and group correlation:**
   Cluster at the **group level** (e.g., state), as a conservative and recommended approach. Because serial correlation cannot be ignored, you must collapse to the state level, which gives you fewer clusters (bigger clusters less fun).

Important: We need a *large number of groups* to estimate intracluster correlations over time reliably.

4.2. What to Cluster Over?

- Individuals often reside in subregions within larger regions.
- **Recommendation:** cluster at the **region** level (Cameron and Miller, 2015).
- Clustering at subregion level is incorrect if regressors or errors are correlated across subregions within the same region.
- **Practical strategy:** start clustering at the lowest level, try higher levels, and stop when standard errors stabilize.
- *Note:* larger clusters tend to have higher standard errors due to greater average n.

4.3. Small Number of Clusters

- Clustered SEs: t-statistic converges to $N(0,1)$ only as $G \to \infty$.
- With finite G: inference should be based on a t-distribution with $G-1$ degrees of freedom.
- **Donald and Lang (2007)**: if there are K regressors invariant within cluster (take same values), use t-distribution with $G-K$ degrees of freedom.
- With few clusters, SEs are downward biased even with clustering → over-rejection of the null.
- Empirical rule: clustering at group level is typically safe if $G \geq 50$ and cluster sizes are similar.

4.4. Fixes for Few Clusters

What if less than 50 clusters?

1. Bias-Corrected Cluster-Robust Variance

Cluster generalization of HC2/HC3 heteroskedasticity-robust formulas. Helps, but doesnt fully eliminate bias or over-rejection.

2. Cluster Bootstrap with Asymptotic Refinement (e.g., Wild Cluster Bootstrap)

1. Estimate unrestricted model, get $\hat{\beta}_k$ and t-stat.
2. Re-estimate under $\hat{\beta}_k = 0$ get restricted $\tilde{\beta}$, residuals $\tilde{u}$.
3. Do B bootstrap replications:

$$y_{ig}^b = X_{ig}\tilde{\beta} + \tilde{u} \cdot v_g^b$$

where $v_g^b = \pm 1$, randomly assigned per cluster. boh.

4. Estimate t-statistic for each bootstrap draw.
5. Compute p-value = share of draws with bootstrapped t-stat > original.
6. In Stata: use `cgmwildboot` or `boottest`.

Randomization Inference

- Instead of drawing samples from a super population, we can draw from repeatedly sampling the treatment assignment.
- Define your statistic (e.g. difference in means between treatment and control) and derive its exact distribution by calculating the test statistic under each possible permutation of treatment allocation.
- Assume you have a sample with 18 units, 9 assigned to treatment, you have $\binom{18}{9} = \frac{18!}{9!\cdot 9!} = 48{,}620$ different possible assignment vectors, get one statistic for each.
- We ask how likely it is to observe a value of the test statistic as large as the one observed: count how many of the 48,620 test statistics have a value larger than the one observed.
- Lets say it is true for 1,500 of them, then your Exact p-value is $\frac{1500}{48620} = 0.03$, and you can reject the null of no difference in means at the 97% level.
- If the number of permutations is very large, you can calculate the statistic for a random sub-sample of permutations and get close to exact p-values.

Stata Implementation: `ritest`

- `ritest` (Hess, 2017 Stata Journal), `ssc install ritest`
- `ritest treatment _b[treatment], reps(1000) seed(1111) cluster(id): reg Y treatment, cluster(id)`