Adjusting Standard Errors, Bootstrap Multicolinearity Diagnosis

Yabin YIN

Division of Social Science, HKUST

Feb 19, 2021

Outline

Regression with Matrix Algebra

For i.i.d. random vectors (Y_1, X_1), (Y_2, X_2), ..., (Y_n, X_n),the matrix version residuals of the OLS regression can be write as

$$\begin{pmatrix} e_0 \\ e_1 \\ e_2 \\ \vdots \\ e_K \end{pmatrix} = \begin{pmatrix} Y_0 \\ Y_1 \\ Y_2 \\ \vdots \\ Y_K \end{pmatrix} - \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{Kn} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_K \end{pmatrix}$$

Finding the OLS regression estimator is equivalent to the following minimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_b \sum_{i=1}^{n} e_i^2 = \arg\min_b (\mathbf{e}^T \mathbf{e}) = \arg\min_b (\mathbf{Y} - \mathbb{X}\mathbf{b})^T (\mathbf{Y} - \mathbb{X}\mathbf{b}).$$

Regression with Matrix Algebra

The first-order condition (that is, setting the derivative of the sum of squared residuals with respect to the coefficients equal to 0) yields,

$$-2\mathbb{X}^T(\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}) = 0$$

Solving for $\hat{\boldsymbol{\beta}}$

$$\begin{aligned}
-\mathbb{X}^T(\mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}) &= 0 \\
-\mathbb{X}^T\mathbf{Y} + \mathbb{X}^T\mathbb{X}\hat{\boldsymbol{\beta}} &= 0 \\
\mathbb{X}^T\mathbb{X}\hat{\boldsymbol{\beta}} &= \mathbb{X}^T\mathbf{Y} \\
(\mathbb{X}^T\mathbb{X})^{-1}(\mathbb{X}^T\mathbb{X})\hat{\boldsymbol{\beta}} &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y} \\
\hat{\boldsymbol{\beta}} &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}
\end{aligned}$$

Standard Errors

Define the vector of errors ϵ as the differences between the observed values of $\hat{Y}$ and the (true) value of Y:

$$\boldsymbol{\epsilon} = \mathbf{Y} - \mathbb{X}\boldsymbol{\beta}$$

We can decompose $\hat{\boldsymbol{\beta}}$ as

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y} \\
&= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T(\mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\boldsymbol{\beta} + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\epsilon} \\
&= \boldsymbol{\beta} + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\epsilon}
\end{aligned}$$

Then, the variance of $\hat{\boldsymbol{\beta}}$ is:

$$\boldsymbol{Var(\hat{\beta})} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Var(\epsilon)\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}$$

Variance-Covariance Structure

To estimate $\boldsymbol{Var(\hat{\beta})}$, we need to estimate $Var(\epsilon)$

$$Var(\epsilon) = E(\epsilon\epsilon^T) = \begin{bmatrix} var(\epsilon_1) & cov(\epsilon_1\epsilon_2) & \cdots & cov(\epsilon_1\epsilon_n) \\ cov(\epsilon_2\epsilon_1) & var(\epsilon_2) & \cdots & cov(\epsilon_2\epsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\epsilon_n\epsilon_1) & cov(\epsilon_n\epsilon_2) & \cdots & var(\epsilon_n) \end{bmatrix}$$

General case:

$$Var(\epsilon) = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_2^2 & \cdots & \sigma_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots & \sigma_n^2 \end{bmatrix}$$

Variance-Covariance Structure

Homoskedastic case: identical variance

$$Var(\epsilon) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

Heteroskedastic case: variance is not identical, should use robust standard errors

$$Var(\epsilon) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Cluster Standard Errors

Cameron and Miller, 2015, JHR

Suppose that we have G groups (e.g., G villages) in the sample, and the error terms are correlated within each group (no correlation between groups). For instance, individuals within a village would behave in a similar manner.

Stacking all observations in the g^{th} cluster, the model can be rewrite as

$$\mathbf{Y}_g = \mathbb{X}_g \boldsymbol{\beta} + \boldsymbol{\epsilon}_g, g = 1, ..., G.$$

The OLS estimator is

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{Y} = (\sum_{g=1}^{G} \mathbb{X}_g^T \mathbb{X}_g)^{-1} \sum_{g=1}^{G} \mathbb{X}_g^T \mathbf{Y}_g$$

Cluster Standard Errors

In general, the variance matrix is

$$\boldsymbol{Var(\hat{\beta})} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Var(\epsilon)\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}$$

Given error independence across clusters, $V[\epsilon]$ has a block-diagonal structure

$$\boldsymbol{Var(\hat{\beta})} = (\sum_{g=1}^{G}\mathbb{X}_g^T\mathbb{X}_g)^{-1}\sum_{g=1}^{G}\mathbb{X}_g^T E[\epsilon_g\epsilon_g^T]\mathbb{X}_g(\sum_{g=1}^{G}\mathbb{X}_g^T\mathbb{X}_g)^{-1}$$

At individual level, we can rewrite

$$B_{clu} = \sum_{g=1}^{G}\mathbb{X}_g^T E[\epsilon_g\epsilon_g^T]\mathbb{X}_g = \sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} x_{ig}x_{jg}^T E(\epsilon_{ig}\epsilon_{jg})$$

$E(\epsilon_{ig}\epsilon_{jg})$ is the error covariance for the i^{th} and j^{th} observations in the same group g.

Cross group variance covariance structure:

$$Var(\epsilon) = \begin{bmatrix} E(\epsilon_{g1}\epsilon_{g1}^T) & E(\epsilon_{g1}\epsilon_{g2}^T) & \cdots & E(\epsilon_{g1}\epsilon_{gG}^T) \\ E(\epsilon_{g2}\epsilon_{g1}^T) & E(\epsilon_{g2}\epsilon_{g2}^T) & \cdots & E(\epsilon_{g2}\epsilon_{gG}^T) \\ \vdots & \vdots & \ddots & \vdots \\ E(\epsilon_{gG}\epsilon_{g1}^T) & E(\epsilon_{gG}\epsilon_{g2}^T) & \cdots & E(\epsilon_{gG}\epsilon_{gG}^T) \end{bmatrix}$$

$$= \begin{bmatrix} E(\epsilon_{g1}\epsilon_{g1}^T) & 0 & \cdots & 0 \\ 0 & E(\epsilon_{g2}\epsilon_{g2}^T) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & E(\epsilon_{gG}\epsilon_{gG}^T) \end{bmatrix}$$

$$= \begin{bmatrix} var(\epsilon_{g1}) & 0 & \cdots & 0 \\ 0 & var(\epsilon_{g2}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & var(\epsilon_{gG}) \end{bmatrix}$$

Angrist and Pischke, 2008, MHE

within cluster variance covariance structure:

$$E(\epsilon_{ig}\epsilon_{jg}) = \begin{bmatrix} E(\epsilon_{1g}\epsilon_{1g}^T) & E(\epsilon_{1g}\epsilon_{2g}^T) & \cdots & E(\epsilon_{1g}\epsilon_{ng}^T) \\ E(\epsilon_{2g}\epsilon_{1g}^T) & E(\epsilon_{2g}\epsilon_{2g}^T) & \cdots & E(\epsilon_{2g}\epsilon_{ng}^T) \\ \vdots & \vdots & \ddots & \vdots \\ E(\epsilon_{ng}\epsilon_{1g}^T) & E(\epsilon_{ng}\epsilon_{2g}^T) & \cdots & E(\epsilon_{ng}\epsilon_{ng}^T) \end{bmatrix}$$

$$= \begin{bmatrix} var(\epsilon_{1g}) & cov(\epsilon_{1g}\epsilon_{2g}^T) & \cdots & cov(\epsilon_{1g}\epsilon_{ng}^T) \\ cov(\epsilon_{2g}\epsilon_{1g}^T) & var(\epsilon_{2g}) & \cdots & cov(\epsilon_{2g}\epsilon_{ng}^T) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\epsilon_{ng}\epsilon_{1g}^T) & cov(\epsilon_{ng}\epsilon_{2g}^T) & \cdots & var(\epsilon_{ng}) \end{bmatrix}$$

Inference

The variance of coef. ($\text{Var}(\beta)$) will be bigger when:

- Regressors within cluster are correlated
- Errors within cluster are correlated
- N_g is large

Bootstrap

See from *SOSC 5340 Tutorial 1.rmd*

Multicolinearity

Collinearity implies two variables are near perfect linear combinations of one another. Multicollinearity involves more than two variables. In the presence of multicollinearity, regression estimates are unstable and have high standard errors.

We can use Variance inflation factors (VIF) to dignose multicollinearity. VIF measures the inflation in the variances of the parameter estimates due to collinearities that exist among the predictors.

VIF

Steps to calculate VIF:

- Regress the k^{th} predictor on rest of the predictors in the model.
- Compute the R_k^2
- $VIF = \frac{1}{1-R_k^2} = \frac{1}{Tolerance}$
- Tolerance: Percent of variance in the predictor that cannot be accounted for by other predictors

VIF $= 1$: there is no correlation among the k^{th} predictor and the remaining predictor variables

VIF

Steps to calculate VIF:

- Regress the k^{th} predictor on rest of the predictors in the model.
- Compute the R_k^2
- $VIF = \frac{1}{1-R_k^2} = \frac{1}{Tolerance}$
- Tolerance: Percent of variance in the predictor that cannot be accounted for by other predictors

Rule of thumb:
VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

- Abadie, Athey, Imbens, and Wooldridge, 2020, ECMA, Sampling-Based versus Design-Based Uncertainty in Regression Analysis.
- Angrist, J. D., Pischke, J. S. (2008). Mostly harmless econometrics: An empiricist's companion. Princeton university press.
- Aronow, P. M., Miller, B. T. (2019). Foundations of agnostic statistics. Cambridge University Press.
- Beck and Katz, 2011, ARPS, Modeling Dynamics in Time-Series-Cross-Section Political Economy Data.
- Cameron, Gelbach, and Miller, 2008, REStat, Bootstrap-based improvements for inference with clustered errors.
- Cameron and Miller, 2015, JHR, A Practitioner's Guide to Cluster-Robust Inference.

- Driscoll and Kraay, 1998, REStat, Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data.
- Newey and West, 1987, ECMA, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix.
- Peterson, 2009, RFS, Estimating Standard Errors in Finance Panel Data Sets Comparing Approaches.
- Thompson, 2011, JFE, Simple formulas for standard errors that cluster by both firm and time.
- White, 1980, ECMA, A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.