Workshop3. Assumptions Under Linear Regression & Multicollinearity

UNAM – FE   Econometrics I

Esp. Humberto Acevedo
Assistant. Emilio Sandoval

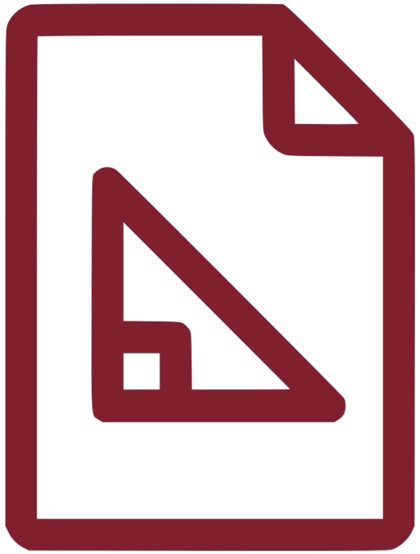
March 1, 2022

Student Learning Outcomes (SLOS)

- **a**. Understand BLUE Assumption
- **b**. Understand the concept of multicollinearity
- **c**. Analyze the maths behind multicollinearity
- **d**. Recognize ways to detect multicollinearity
- **e**. Understand VIF test.
- **f**. Apply VIF Test

BLUE

Best
Linear
Unbiased
Estimator



Given assumptions of linear regression model, estimation of Least Squares own optimal properties which are referred to Gauss-Markov Theorem

Blue

Best

Linear

Unbiased

Estimator

Linear function of a random variable

Best

Linear

Unbiased

Estimator

Expected value $E(\beta^2)$ equal to true value β

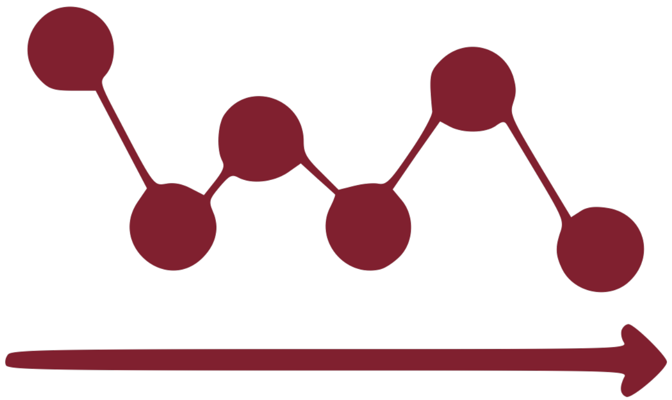
Blue

Best

Linear

Unbiased

Estimator



Minimum variance

To **remember**

Prefix “co” referring to linear movement in tandem (correlation)

Suffix meaning the quality or state of

Multi – col – linearity – ity

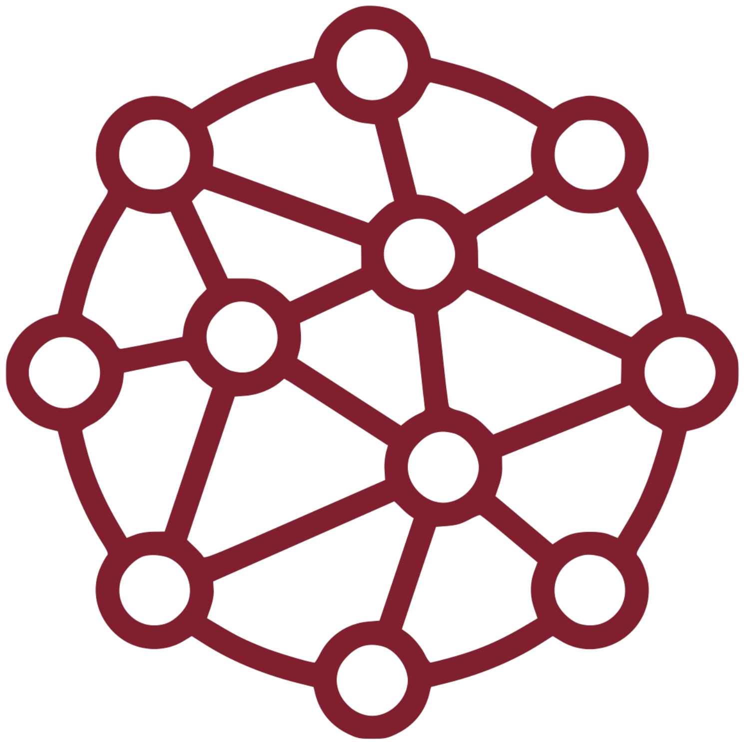
Multiple independent variables within multiple regression

Ocurring whithin a linear equation

Taken from Andrew Ozbun (2021)

Empirical studies have shown that finding correlation levels between independent variables is quite usual.

"Linear 'perfect' relationship among some or all independent variables from a regression model."

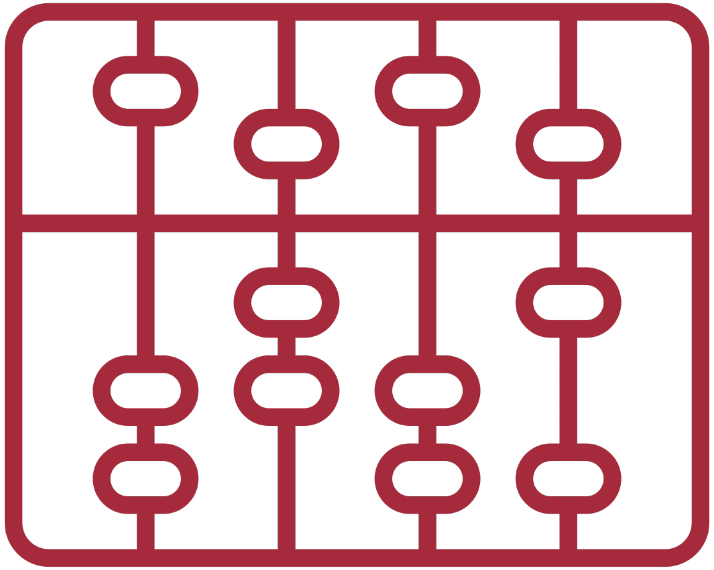


The existence of any level of association among them has effects when estimating parameters and their variances.

To the exist multicollinearity it is not feasible to separate into neatly, the effects on the dependent variable of each of the explanatory variables

One of the basic assumptions of the general linear model states that the explanatory variables are linearly independent

When independent variables are correlated in such a way that any of the columns of the matrix explanatory variables can be written as a linear combination of the others it is not possible to get the matrix inverse of $(X'X)1$



There is perfect multicollinearity when columns from matrix X are linearly dependent.

$$X = \begin{pmatrix} 1 & 3 & 7 \\ 1 & 2 & 5 \\ 1 & 4 & 9 \end{pmatrix}$$

Do you notice something?

Third column is obtained by taking first and add two times the second column

Sometimes columns from matrix X are almost linearly dependent

$$\lambda_1 X_1 + \lambda_1 X_1 + \cdots + \lambda_k X_k \approx 0$$

- Matrix X has a rank equal to k
- Matrix $X'X$ is not singular
- OLS can be calculated

There is approximated multicollinearity when columns from matrix X are almost linearly dependent

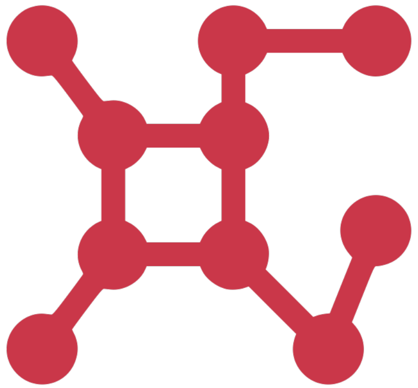
For example, columns from matrix X are almost linearly dependent.

$$X = \begin{pmatrix} 1 & 3 & 7.01 \\ 1 & 2 & 5 \\ 1 & 4 & 9 \end{pmatrix}$$

$$|X'X| = 0.02$$

Presence of approximated multicollinearity allows a better coefficient estimate but variance will be higher

Structural multicollinearity

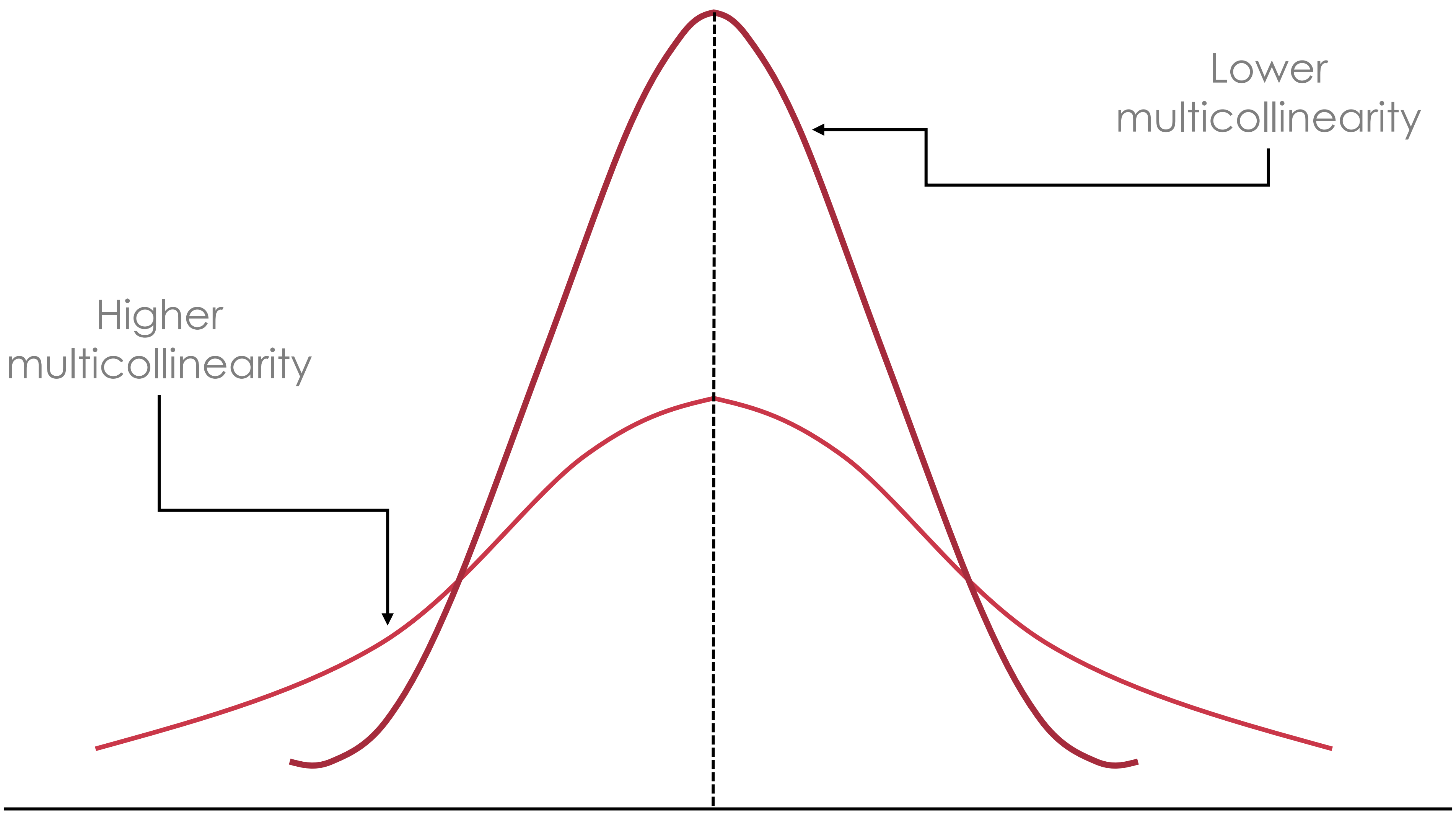


Mathematical artifact caused by creating new predictors from other predictors

Data-Based Multicollinearity

Results from a poorly designed experiment, reliance on purely observational data, or the inability to manipulate the system on which the data are collected.

Properties of **estimators**



How to detect multicollinearity?

1. A first glance, we can obtain **simple sampling correlation coefficients** for each **pair** of independent variables, then we check if the **degree of correlation among them is high**

2. The other way around is to run a **regression for each independent variable** over the rest and then analyze **coefficients of determination** of each regression

Do not worry if you do not understand, it will be crystal clear with practicing!

Super note!

If any of this coefficients of determination is high, it would indicate possible presence of multicollinearity.

Open presion.dta

It contains information about 20 persons that have high blood pressure (arterial hypertension). Researchers were keen on determining relationship between blood pressure, age, weight, body surface, duration, pulse, and levels of stress.

We check the scatter matrix and correlation matrix.

To notice the effects that this correlation has over variance we need to see a single case that has no multicollinearity

STATA COMMANDS

1. `graph matrix bp age weight bsa dur pulse strees, half`
2. `corr bp age weight bsa dur pulse stress`

Open nomulti.dta

In this dataset regressors have a correlation equals to zero.

We check the scatter matrix and correlation matrix.

To notice the effects that this correlation has over variance we need to see a single case that has no multicollinearity.

Then we execute a series of regressions to analyze the information from this dataset.

Look that we store results.

STATA COMMANDS

```
3. graph matrix y x1 x2, half
4. corr x1 x2
5. reg y x1
6. est store yvsx1
7. reg y x2
8. est store yvsx2
9. reg y x1 x2
10.est store yvsx1x2
11.reg y x2 x1
12.est store yvsx2x1
```

Now we are going to integrate all this results in a table (this is very common and requested in any analytical job)

STATA COMMANDS

```
13. estimates table yvsx1 yvsx2 yvsx1x2 yvsx2x1, b(%9.2f) se(%9.2f)
14.anova y x1 x2
15.anova y x2 x1
```

From here, we will review two cases:

when regressors are slightly correlated

when regressors are highly correlated

What happens when regressors are slightly correlated?

Return to presion.dta database.

We may focus on relationship between *BP* and regressors *bsa* and *stress*.

To visualize this, we calculate correlation matrix.

STATA COMMANDS

```
16.Graph matrix bp age stress, half

17.Corr bp age stress
```

Again, we run multiple regressions.

STATA COMMANDS

18. reg bp stress
19. Est store bpstress
20. Reg bp age
21. Est store bpage
22. Reg bp stress age
23. Est store stressage
24. Reg bp age stress
25. Est store agestress

Finally, we create a comparative table and analyze variance.

STATA COMMANDS

```
26. estimates table bpstress bpage stressage
    agestress, b(%92.f) se(%9.2f)
27. anova bp stress age
28. anova bp age stress
```

STATA COMMANDS

Now, what happens when regressors are highly correlated?

Return to presion.dta database.

When regressor is correlated, the estimated coefficient will depend on variations from another regressor on which the former maintains that relationship.

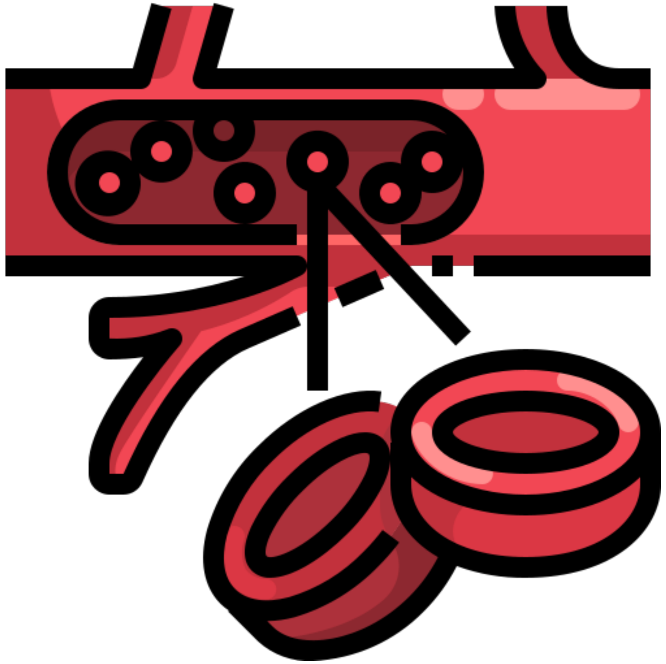
```
29. corr bp age weight bsa dur pulse stress
```

Again, we generate regressions

STATA COMMANDS

```
30. reg bp bsa
31. est store bsa
32. reg bp weight
33. est store weight
34. reg bp bsa weight
35. est store bsa weight
36. estimates table bsa weight bsa weight b(%9.2f)
```

Conclussion



If BSA is the only regressor, we can say that for each additional square metre in body surface (*bsa*), blood pressure increases in 34.4 mm Hg.

If we include *weight* and *bsa* in the model, it is possible to point out that for each additional square metre in body surface (bsa) keeping weight constantly, then blood pressure increases only in 5.83 mm. Hg.

Conclussion

We can observe that variable *BSA* is meaningful in simple regression

Weight ceases to be significant in the regression where it appears

This may be contradictory due to the conclussion that blood pressure is related with body surface

VIF quantifies how big is the variance over estimator.

The closer R^2 gets to 1 or the higher the colinearity of variable X_j with the rest of variables, the greater the value of VIF and the larger the variance of estimated coefficient turns .

$$VIF = \frac{1}{(1 - R_j^2)}$$

$$TOL = \frac{1}{(VIF)}$$

Multicollinearity inflate variance.

If $VIF_j > 10$ then conclude that collinearity of variable X_j regarding with the rest of variables is high.

If $TOL < 0.1$ there is collinearity.

Open elemapi2.dta

This dataset contains information about academic performance from elementary education.

Let's prove that academic performance (*api00*) depends on the percentage of students that receive free meals (*meals*), that are learning English (*ell*), on percentage of teacher with new accreditations (*emer*), and if parents have any college degree (*some_col*)

STATA COMMANDS

```
37. reg api00 meals ell emer some_col
38. vif
```

Let's run a second estimation now adding the following variables:

Grad_sch: Parents' educational level.

Col_grad: Number of parents with college degree.

Avg_ed: Parent's educatonal level average.

STATA COMMANDS

```
39. reg api00 meals ell emer some_col avg_ed
    grad_sch col_grad
40. vif
```

Solutions

Drop explicative variables: it is possible there is a problem due to specificiation error (omission of any relevant variable.)

Transform data: with cross-sectional data it is adsivable to use variables quotients, such as:

$$\frac{Y_i}{X_{3i}} = \beta_1 \frac{1}{X_{3i}} + \beta_2 \frac{1}{X_{3i}} + \beta_3 \frac{1}{X_{3i}}$$

Note that with time series, using data in first differentiation is recommended

$$\Delta Y_t = \beta_2 \Delta X_{2t} + \beta_3 \Delta X_{3t} + e_t$$

Having said dropping variables, we run a regression with some variables.

We drop `avg_ed`

STATA COMMANDS

```
41. reg api00 meals ell emer some_col grad_sch
    col_grad
42. vif
```

What happens with VIF test?

References

- **Salvatore, D., & Sarmiento, J. C.** (1983). *Econometría* (No. HB141 S39). McGraw-Hill.
- **Kumari K., J. Pract Cardiovasc, Wooldridge, J.** (2020). *Introductory econometrics : a modern approach*. Boston, MA: Cengage. Gujarati, D. & Porter, D. (2009). *Basic econometrics*. Boston: McGraw-Hill Irwin.
- **Gujarati, D. N.** (2009). Basic econometrics. Tata McGraw-Hill Education.
- PennState Eberly College of Science, *Reducing Structural Multicollinearity*, from https://online.stat.psu.edu/stat462/node/182/
- PennState Eberly College of Science, *Reducing Data-Based Multicollinearity*, from https://online.stat.psu.edu/stat462/node/181/