Multicollinearity

EC 339

Marcio Santetti

Fall 2022

Motivation

Linear relationships

Let us recall **CLRM Assumption VI**:

> *No explanatory variable is a perfect linear function of any other explanatory variable.*

This assumption implies a **deterministic** relationship between two independent variables.

$$x_1 = \alpha_0 + \alpha_1 x_3$$

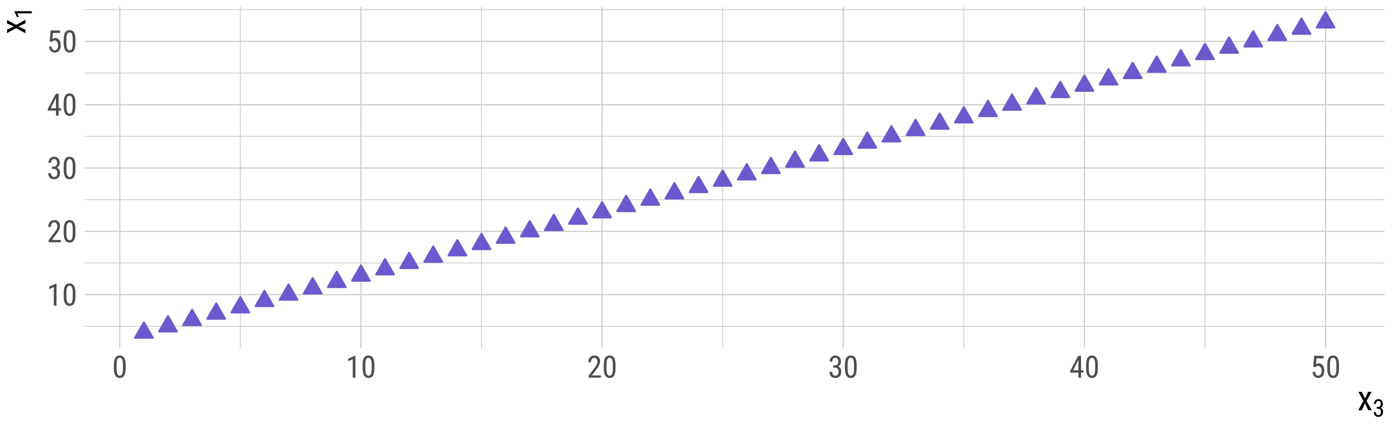
However, in practice we should worry more about strong **stochastic** relationships between two independent variables.

$$x_1 = \alpha_0 + \alpha_1 x_3 + \epsilon_i$$

Linear relationships

What does a linear relationship between two independent variables mean in practice?

- If two variables (say, x_1 and x_3) move **together**, then how can OLS **distinguish** between the effects of these two on y?
  - It **cannot**!



Perfect multicollinearity

Perfect multicollinearity

CLRM Assumption VI only refers to **perfect** multicollinearity.

With its presence, OLS estimation is **indeterminate**.

- Why?

How to disentangle the effect of each independent variable on y?

The *ceteris paribus* assumption no longer holds.

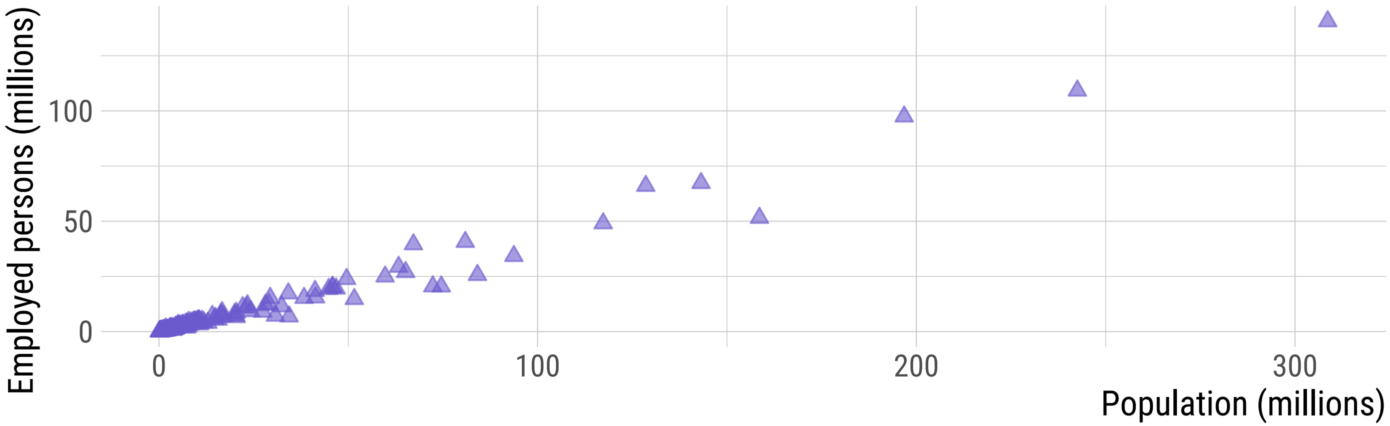
- **Good news**: *rare* to occur in practice.

Imperfect multicollinearity

Imperfect multicollinearity

Even though CLRM Assumption VI **does not** contemplate this version of multicollinearity, it is an actual problem within OLS estimation.

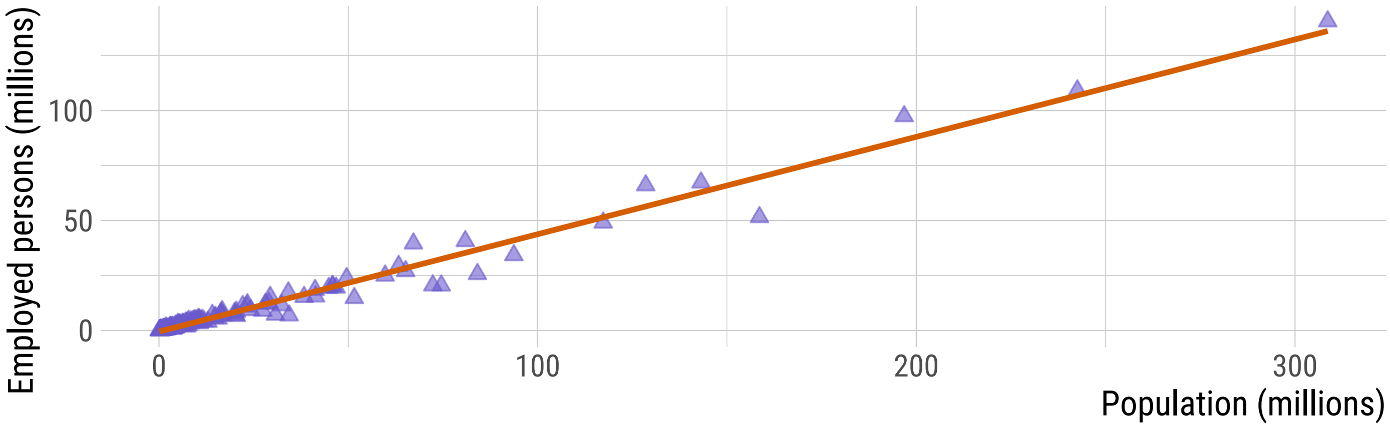
Strong **stochastic** relationships imply strong **correlation coefficients** between two independent variables.



Imperfect multicollinearity

Even though CLRM Assumption VI **does not** contemplate this version of multicollinearity, it is an actual problem within OLS estimation.

Strong **stochastic** relationships imply strong **correlation coefficients** between two independent variables.



Consequences of multicollinearity

Consequences of multicollinearity

By itself, multicollinearity **does not** cause **bias** to OLS β coefficients.

However, it affects OLS **standard errors**.

Recall that standard errors are part of the **t-test formula**:

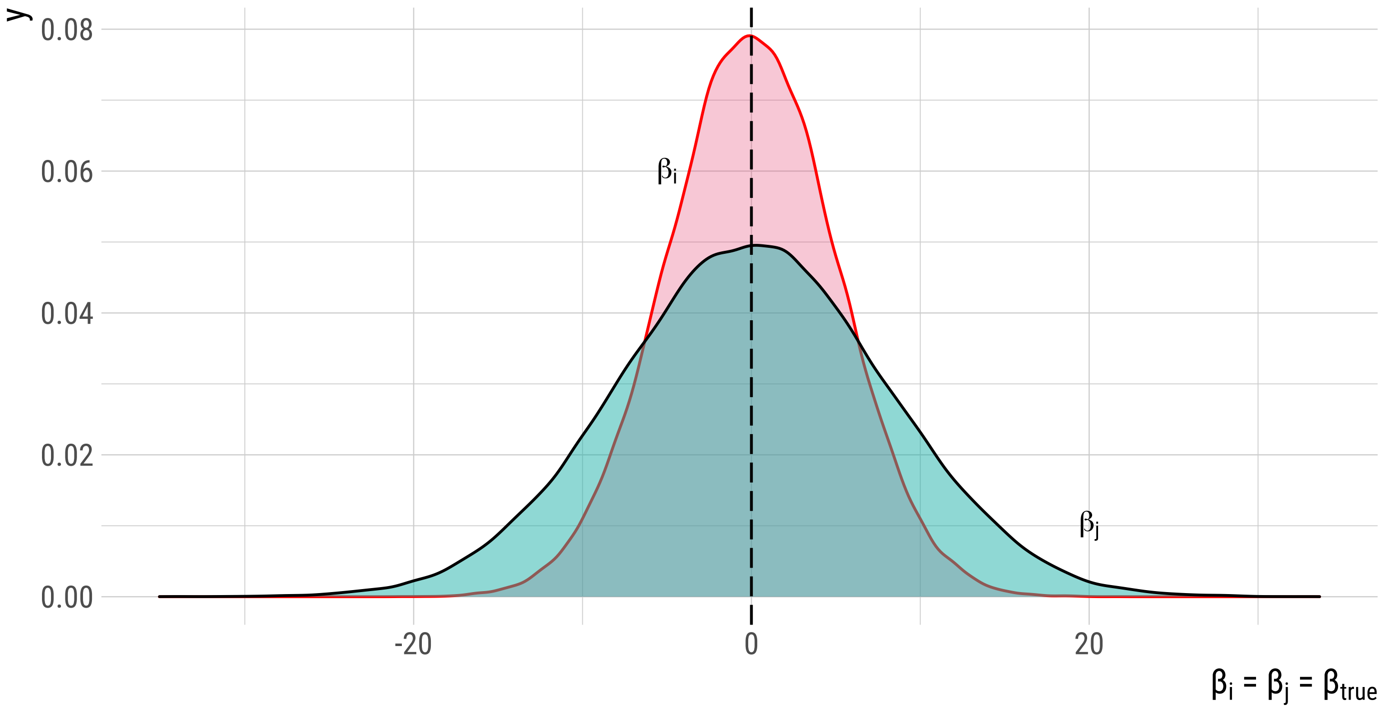
$$t = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)}$$

Therefore, it affects OLS **inference**.

Consequences of multicollinearity

Visually:

- Which estimate is *relatively more efficient*?



Dealing with multicollinearity

Dealing with multicollinearity

Consider the following model:

$$log(rgdpna_i) = \beta_0 + \beta_1 pop_i + \beta_2 emp_i + \beta_3 ck_i + \beta_4 ccon_i + u_i$$

where (for each country *i*):

- `rgdpna`: real GDP (millions 2011 USD)
- `pop`: population (millions)
- `emp`: number of employed persons (millions)
- `ck`: capital services levels (index, USA = 1)
- `ccon`: real consumption (households and government)

Dealing with multicollinearity

```
                        #>
#> ===============================================
#>                         Dependent variable:
#>                     ---------------------------
#>                             log(rgdpna)
#> -----------------------------------------------
#> pop                          0.050***
#>                               (0.018)
#> emp                           -0.069
#>                               (0.042)
#> ck                           26.632***
#>                               (6.518)
#> ccon                        -0.00000***
#>                              (0.00000)
#> Constant                     10.785***
#>                               (0.145)
#> -----------------------------------------------
#> Observations                    130
#> R2                             0.478
#> Adjusted R2                    0.461
#> Residual Std. Error      1.404 (df = 125)
#> F Statistic          28.605*** (df = 4; 125)
#> ===============================================
#> Note:               *p<0.1; **p<0.05; ***p<0.01
```

Dealing with multicollinearity

A little modification:

$$log(rgdpna_i) = \beta_0 + \beta_1 log(emp_i) + \beta_3 ck_i + \beta_4 log(ccon_i) + u_i$$

Dealing with multicollinearity

```
                        #>
#> ===============================================
#>                         Dependent variable:
#>                     ---------------------------
#>                             log(rgdpna)
#> -----------------------------------------------
#> log(emp)                     -0.059**
#>                               (0.029)
#> ck                            -0.206
#>                               (0.288)
#> log(ccon)                    1.076***
#>                               (0.027)
#> Constant                      -0.487*
#>                               (0.275)
#> -----------------------------------------------
#> Observations                    130
#> R2                             0.979
#> Adjusted R2                    0.979
#> Residual Std. Error      0.277 (df = 126)
#> F Statistic         2,001.826*** (df = 3; 126)
#> ===============================================
#> Note:               *p<0.1; **p<0.05; ***p<0.01
```

Dealing with multicollinearity

Checking **correlation** coefficients:

- $Corr(pop_i, emp_i) = 0.987$
- $Corr(ccon_i, emp_i) = 0.980$

- $Corr(log(ccon_i), emp_i) = 0.584$

Dealing with multicollinearity

A recommended procedure is to always check out the **correlation coefficient** among the chosen independent variables.

- In addition, we can calculate **Variance Inflation Factors** (VIFs):

$$VIF(\hat{\beta}_i) = \frac{1}{(1 - R_i^2)}$$

where R_i^2 is the coefficient of determination of the *auxiliary regression* models.

- The procedure is to estimate one auxiliary regression model for *each* independent variable.
- Then, store the R^2 for each regression.
- A *VIF* greater than 5 is already sifficient to imply high multicollinearity.

Dealing with multicollinearity

In R...

```
model_1 %>%
  vif()
```

```
#>      pop      emp       ck     ccon
#> 42.68883 48.52425 30.43790 27.30301
```

```
model_2 %>%
  vif()
```

```
#>  log(emp)        ck log(ccon)
#>  3.717818  1.516566  4.236570
```

- What do we conclude?

Dealing with multicollinearity

In Stata…

```
reg lrdgpna pop emp ck ccon

vif

    Variable |       VIF       1/VIF
-------------+----------------------
         emp |     48.52    0.020608
         pop |     42.69    0.023425
          ck |     30.44    0.032854
        ccon |     27.30    0.036626
-------------+----------------------
    Mean VIF |     37.24
```

- What do we conclude?

Dealing with multicollinearity

In `Stata`...

```
reg lrdgpna lemp ck lccon

vif

    Variable |       VIF       1/VIF
-------------+----------------------
       lccon |      4.24    0.236040
        lemp |      3.72    0.268975
          ck |      1.52    0.659385
-------------+----------------------
    Mean VIF |      3.16
```

- What do we conclude?

Next time: Multicollinearity in practice