

# Tema 7. Multicolinealidad

Gustavo A. García

[ggarci24@eafit.edu.co](mailto:ggarci24@eafit.edu.co)

Econometría para la Toma de Decisiones

Maestría en Economía Aplicada

Escuela de Finanzas, Economía y Gobierno

Universidad EAFIT

Link slides en formato **html**

Link slides en formato **PDF**

## En este tema

- Naturaleza del problema
- Multicolinealidad perfecta
- Ausencia total de multicolinealidad (ortogonalidad)
- Multicolinealidad imperfecta
- Consecuencias
- Criterios para detectar multicolinealidad
- Medidas correctivas
- Ejercicio aplicado en R

# Lecturas

- Wooldridge, Jeffrey (2013). *Introducción a la econometría*. 5a edición, Cengage Learning. [Cap. 3.4](#)
- Gujarati, D. y Porter, D. (2010). *Econometría*. 5a edición, Mc Graw Hill. [Cap. 10](#)

# Naturaleza del problema

- La multicolinealidad es un problema de **dependencia lineal entre las variables explicatorias que obstaculizan el poder aislar el efecto de cada una por separado**
- Su principal efecto serán varianzas excesivamente grandes para cada estimador, y esto llevará a que cada parámetro pueda estar mal estimado
- No obstante:
  - alta dependencia no necesariamente lleva a molestias
  - en presencia de molestias para cada parámetro, una combinación lineal puede estar bien estimada
  - puede haber una contradicción estadística: **por separado ningún parámetro significativo pero de conjunto si lo son**

# Naturaleza del problema

Lo primero que hay que tener en cuenta es que es un problema muestral, al ser consecuencia de excesiva dependencia entre las  $X$

Se pueden distinguir tres situaciones:

- Multicolinealidad perfecta
- Ausencia total de multicolinealidad (Ortogonalidad)
- multicolinealidad imperfecta

# Multicolinealidad perfecta

Representa la violación del supuesto de rango completo de la matriz  $\mathbf{X}_{n \times k}$

$$\text{Si } \rho(\mathbf{X}_{n \times k}) < k$$

$$|\mathbf{X}'\mathbf{X}| = 0$$

$(\mathbf{X}'\mathbf{X})$  es matriz singular

$$(\mathbf{X}'\mathbf{X})^{-1} \text{ no existe}$$

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \text{ no se puede calcular}$$

$$\text{Cov}(\hat{\mathbf{B}}) = \hat{\sigma}_u^2(\mathbf{X}'\mathbf{X})^{-1} \text{ no se puede calcular}$$

La trampa de las variables binarias es un ejemplo de multicolinealidad perfecta, en este caso por una incorrecta especificación del modelo

# Multicolinealidad perfecta

Supóngase un modelo de 2 variables

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Así que la matriz  $\mathbf{X}'\mathbf{X}$  será

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum X_{2i} & \sum X_{3i} \\ \sum X_{2i} & \sum X_{2i}^2 & \sum X_{2i}X_{3i} \\ \sum X_{3i} & \sum X_{2i}X_{3i} & \sum X_{3i}^2 \end{pmatrix}$$

Si se diese el caso que  $X_{3i} = qX_{2i}$  entonces el  $\rho(\mathbf{X}_{n \times 3}) = 2$ , entonces se tendría

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum X_{2i} & q \sum X_{2i} \\ \sum X_{2i} & \sum X_{2i}^2 & q \sum X_{2i}^2 \\ q \sum X_{2i} & q \sum X_{2i}^2 & q^2 \sum X_{2i}^2 \end{pmatrix}$$

Se observa que la 3a columna de  $\mathbf{X}'\mathbf{X}$  es  $q$  veces la 2a, por lo tanto los datos muestrales no permiten estimar el modelo planteado



# Multicolinealidad perfecta

En efecto

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 q X_{2i} + u_i$$

$$Y_i = \beta_1 + (\beta_2 + q\beta_3) X_{2i} + u_i$$

La muestra contiene información para estimar  $\beta_1$  y  $\beta_2 + q\beta_3$ , pero no hay suficiente información para estimar  $\beta_2$  y  $\beta_3$  por separado

En este caso extremo de multicolinealidad perfecta el modelo  $\mathbf{Y} = \mathbf{XB} + \mathbf{u}$  no puede ser estimado. Si es consecuencia de una incorrecta especificación el único camino es re-especificar el modelo

# Ausencia total de multicolinealidad (ortogonalidad)

En el otro extremo está la ausencia total de interdependencia entre regresores. Es el caso de la **ortogonalidad**: como la variable  $X_j$  es totalmente independiente de otra  $X_m$ , su producto vectorial será nulo  $X_j'X_m = 0$ . Este es el mejor caso para estimar  $\beta_j$ , pero rara vez sucede

Si la correlación entre los regresores es 0, el estimador MCO,  $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ , equivale a esta expresión:

$$\begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} \frac{\sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})}{\sum (X_{2i} - \bar{X}_2)^2} \\ \frac{\sum (X_{3i} - \bar{X}_3)(Y_i - \bar{Y})}{\sum (X_{3i} - \bar{X}_3)^2} \end{bmatrix}$$

- La matriz  $(\mathbf{X}'\mathbf{X})$  se convierte en una matriz diagonal, se pierde la componente de covarianza entre los regresores por la incorrelación los mismos
- Las estimaciones del modelo múltiple coinciden con la de dos modelos simples por separado
- Los valores de la varianza si varían puesto que  $\hat{\sigma}_u^2$  depende del número de regresores utilizados:  $\hat{\sigma}_u^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{N-k}$  y  $Cov(\hat{\mathbf{B}}) = \hat{\sigma}_u^2(\mathbf{X}'\mathbf{X})^{-1}$
- Difícilmente se da en la práctica

# Multicolinealidad imperfecta

En este caso existe una altísima dependencia entre regresores mas no es una relación exacta

En algunos casos el origen está en el movimiento conjunto de algunas variables. Es el caso de ciertas series temporales macroeconómicas. En otros casos, el origen puede estar en la construcción del modelo: si se incluyen cuadrados y productos cruzados, además de los regresores lineales, como en la función translog o en la ecuación de mincer

Es la situación más habitual en la práctica. Se cumple la condición de rango:

$$\begin{aligned} \rho(\mathbf{X}_{n \times k}) &= k \\ |\mathbf{X}'\mathbf{X}| &\neq 0 \\ (\mathbf{X}'\mathbf{X}) &\text{ es una matriz no singular} \\ (\mathbf{X}'\mathbf{X})^{-1} &\text{ existe} \end{aligned}$$

| $y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + u_i$ |           |           |           | $r_{x_2 x_3} = 0$ | $r_{x_2 x_3} = 0,9$ | $r_{x_2 x_3} = 0,99$ |             |
|---|-----------|-----------|-----------|-------------------|---------------------|----------------------|-------------|
| $r_{x_2 x_3}$   | $(X'X)$   |           |           | $ X'X $           | $(X'X)^{-1}$        |                      |             |
| 0   | 120       | 61,004614 | 66,708938 | 1                 | 6,220273            | -6,603373            | -5,135692   |
|   | 61,004614 | 31,090011 | 33,912942 |                   | -6,603373           | 12,989269            | 0           |
|   | 66,708938 | 33,912942 | 37,192264 |                   | -5,135692           | 0                    | 9,238395    |
| 0,9   | 120       | 61,004614 | 57,283422 | 0,13888           | 3,377475            | -5,782412            | -0,899793   |
|   | 61,004614 | 31,090011 | 29,191517 |                   | -5,782412           | 68,364564            | -60,692330  |
|   | 57,283422 | 29,191517 | 27,424042 |                   | -0,899793           | -60,692330           | 66,519899   |
| 0,99  | 120       | 61,004614 | 58,487073 | 0,01354           | 3,381193            | -9,791567            | 3,292823    |
|   | 61,004614 | 31,090011 | 29,807718 |                   | -9,791567           | 652,724630           | -660,731100 |
|   | 58,487073 | 29,807718 | 28,579785 |                   | 3,292823            | -660,731100          | 682,415830  |

# Consecuencias de la multicolinealidad imperfecta

1. Hay pérdida de precisión en la estimación individual de los parámetros
2. Como consecuencia de la mayor varianza habrán intervalos de confianza más amplios y más tendencias al no rechazo de  $H_0 : \beta_j = 0$ . No obstante la prueba en su conjunto no se verá afectada
3. Si como consecuencia del no rechazo de  $\beta_j = 0$  se elimina  $X_{ji}$ , puede ocurrir un sesgo en la estimación del resto de parámetros
4. Excesiva sensibilidad muestral

# Criterios para detectar multicolinealidad

- La multicolinealidad es una cuestión de grado y no de clase. La distinción importante no es entre presencia o ausencia de multicolinealidad, sino entre sus diferentes grados
- Como la multicolinealidad se refiere a la condición de las variables explicativas que son no estocásticas por supuestos, es una característica de la muestra y no de la población

Por consiguiente, no es necesario "llevar a cabo pruebas sobre multicolinealidad", pero, si se desea, es posible medir su grado en cualquier muestra determinada

No existe un método único para detectar o medir la fuerza de la multicolinealidad. Lo que se tiene son ciertas reglas prácticas, algunas informales y otras formales, pero todas reglas prácticas

# Criterios para detectar multicolinealidad

- Diagrama de dispersión: permite observar cómo se relacionan las diversas variables de un modelo de regresión
- Un  $R^2$  elevado o significancia conjunta del modelo y pocas razones  $t$  significativas
- Altas correlaciones entre parejas de regresores
- Regresiones auxiliares y calculo de  $R_j^2$
- La regla práctica de Klein: la multicolinealidad puede ser un problema complicado solamente si el  $R^2$  de una regresión auxiliar es mayor que el  $R^2$  de la regresión de  $Y$  sobre todos los regresores
- Factor inflacionario de la varianza (FIV)

$$Var(\hat{\beta}_j) = \frac{\hat{\sigma}_u^2}{\sum x_{ji}^2(1-R_j^2)} = \frac{\hat{\sigma}_u^2}{\sum x_{ji}^2} FIV_j, \text{ donde } FIV_j = \frac{1}{1-R_j^2}$$

La regla práctica es: si el  $FIV$  de una variable es superior a 10 (esto sucede si  $R_j^2$  excede 0.90), se dice que esa variable es muy colineal

# Medidas correctivas

- La multicolinealidad es en esencia un problema de deficiencia de datos, y en algunas ocasiones no hay opción respecto de los datos disponibles para el análisis empírico
- Procedimiento de reglas prácticas
  - Información *a priori*
  - Combinación de información de corte transversal y de series de tiempo
  - Eliminación de una(s) variable(s) y el sesgo de especificación
  - Transformación de variables
  - Datos nuevos o adicionales

# Ejercicio aplicado en R

Los datos para este ejercicio fueron extraídos de *the 1974 Motor Trend US magazine*, y contiene información sobre consumo de gasolina y 10 características del diseño y desempeño de 32 automóviles.

Se dispone de las siguientes variables:

- mpg: Miles/(US) gallon
- disp: Displacement (cu.in.)
- hp: Gross horsepower
- wt: Weight (1000 lbs)
- qsec: 1/4 mile time

La idea es analizar los determinantes del desempeño de los automóviles y se estima la siguiente ecuación:

$$mpg_i = \beta_1 + \beta_2 disp_i + \beta_3 hp_i + \beta_4 wt + \beta_5 qsec + u_i$$



# Ejercicio aplicado en R

```
library(Hmisc); library(corrplot); library(olsrr); library(mctest)
data(mtcars)
model <- lm(mpg ~ disp + hp + wt + qsec, data = mtcars)
summary(model)
```

Call:

```
lm(formula = mpg ~ disp + hp + wt + qsec, data = mtcars)
```

Residuals:

|  | Min     | 1Q      | Median  | 3Q     | Max    |
|--|---------|---------|---------|--------|--------|
|  | -3.8664 | -1.5819 | -0.3788 | 1.1712 | 5.6468 |

Coefficients:

|             | Estimate  | Std. Error | t value | Pr(> t )   |
|-------------|-----------|------------|---------|------------|
| (Intercept) | 27.329638 | 8.639032   | 3.164   | 0.00383 ** |
| disp        | 0.002666  | 0.010738   | 0.248   | 0.80576    |
| hp          | -0.018666 | 0.015613   | -1.196  | 0.24227    |
| wt          | -4.609123 | 1.265851   | -3.641  | 0.00113 ** |
| qsec        | 0.544160  | 0.466493   | 1.166   | 0.25362    |

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.622 on 27 degrees of freedom

Multiple R-squared: 0.8351, Adjusted R-squared: 0.8107

F-statistic: 34.19 on 4 and 27 DF, p-value: 3.311e-10

```
# Matriz de correlaciones parciales y su significancia estadística
subdata <- mtcars[,c("disp", "hp", "wt", "qsec")]
cor(subdata)
```

|      | disp       | hp         | wt         | qsec       |
|------|------------|------------|------------|------------|
| disp | 1.0000000  | 0.7909486  | 0.8879799  | -0.4336979 |
| hp   | 0.7909486  | 1.0000000  | 0.6587479  | -0.7082234 |
| wt   | 0.8879799  | 0.6587479  | 1.0000000  | -0.1747159 |
| qsec | -0.4336979 | -0.7082234 | -0.1747159 | 1.0000000  |

```
cor <- cor(mtcars[,c("disp", "hp", "wt", "qsec")])
cor
```

|      | disp       | hp         | wt         | qsec       |
|------|------------|------------|------------|------------|
| disp | 1.0000000  | 0.7909486  | 0.8879799  | -0.4336979 |
| hp   | 0.7909486  | 1.0000000  | 0.6587479  | -0.7082234 |
| wt   | 0.8879799  | 0.6587479  | 1.0000000  | -0.1747159 |
| qsec | -0.4336979 | -0.7082234 | -0.1747159 | 1.0000000  |

```
# Correlaciones con significancia estadística
rcorr(as.matrix(subdata))
```

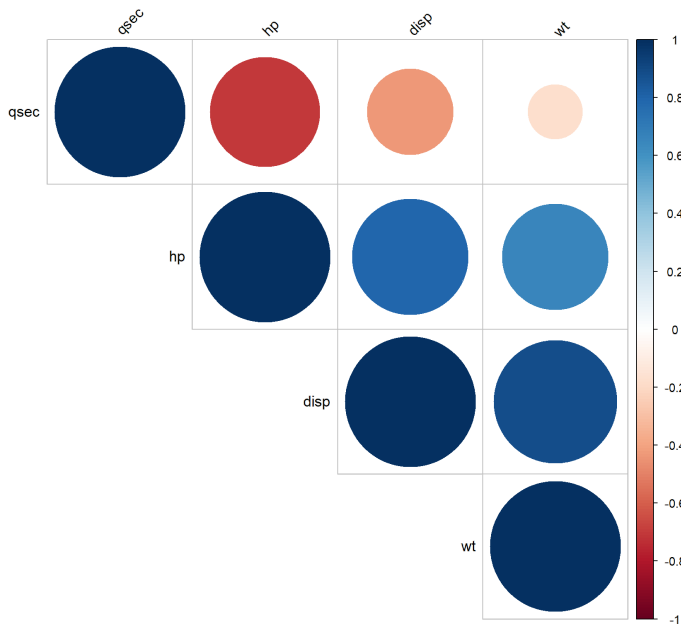
|      | disp  | hp    | wt    | qsec  |
|------|-------|-------|-------|-------|
| disp | 1.00  | 0.79  | 0.89  | -0.43 |
| hp   | 0.79  | 1.00  | 0.66  | -0.71 |
| wt   | 0.89  | 0.66  | 1.00  | -0.17 |
| qsec | -0.43 | -0.71 | -0.17 | 1.00  |

n= 32

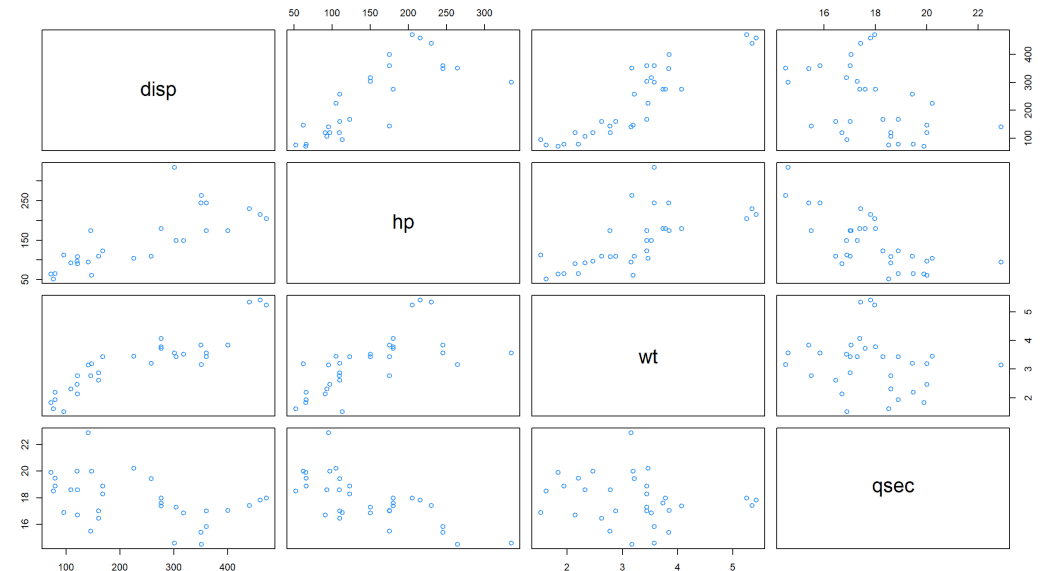
| P    | disp   | hp     | wt     | qsec   |
|------|--------|--------|--------|--------|
| disp |        | 0.0000 | 0.0000 | 0.0131 |
| hp   | 0.0000 |        | 0.0000 | 0.0000 |
| wt   | 0.0000 | 0.0000 |        | 0.3389 |

# Ejercicio aplicado en R

```
# Corrplot function  
corrplot(cor, type="upper", order = "hclust", tl.col = "black",  
         tl.srt = 45, sig.level = 0.05, insig = "blank")
```



```
# Diagrama de dispersión  
plot(subdata, col = "dodgerblue")
```



# Ejercicio aplicado en R

```
# FIV
# vif: factor de inflación de la varianza (FIV). VIF mayores a 10 son preocupantes
# la tolerancia = 1/vif. Tolerancia más bajo que 0.1 es comparable a un VIF alto

ols_vif_tol(model)
```

|   | Variables | Tolerance | VIF      |
|---|-----------|-----------|----------|
| 1 | disp      | 0.1252279 | 7.985439 |
| 2 | hp        | 0.1935450 | 5.166758 |
| 3 | wt        | 0.1445726 | 6.916942 |
| 4 | qsec      | 0.3191708 | 3.133119 |

```
ols_coll_diag(model)
```

## Tolerance and Variance Inflation Factor

|   | Variables | Tolerance | VIF      |
|---|-----------|-----------|----------|
| 1 | disp      | 0.1252279 | 7.985439 |
| 2 | hp        | 0.1935450 | 5.166758 |
| 3 | wt        | 0.1445726 | 6.916942 |
| 4 | qsec      | 0.3191708 | 3.133119 |

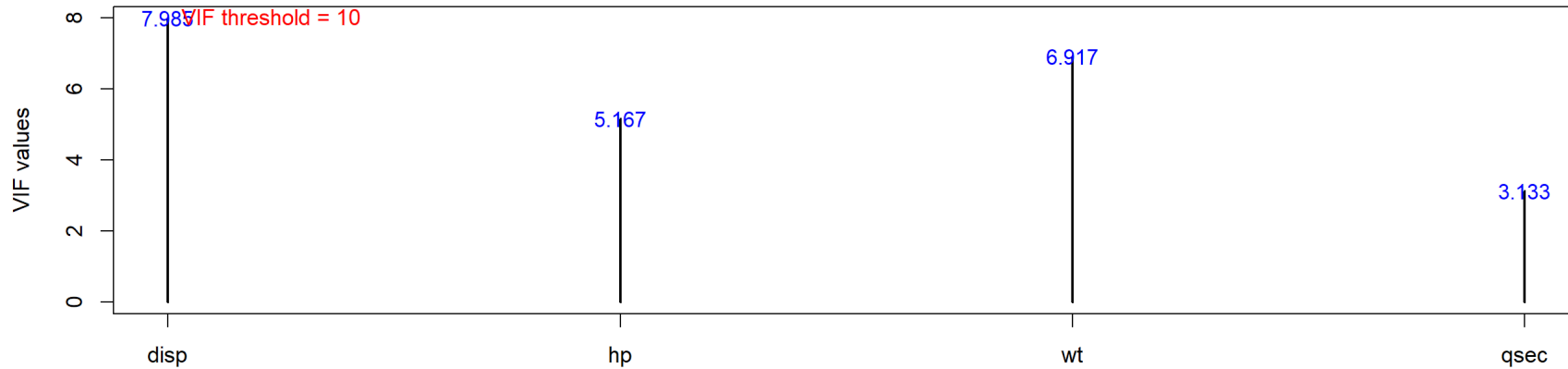
## Eigenvalue and Condition Index

|   | Eigenvalue   | Condition Index | intercept   | disp        | hp          |
|---|--------------|-----------------|-------------|-------------|-------------|
| 1 | 4.721487187  | 1.000000        | 0.000123237 | 0.001132468 | 0.001413094 |
| 2 | 0.216562203  | 4.669260        | 0.002617424 | 0.036811051 | 0.027751289 |
| 3 | 0.050416837  | 9.677242        | 0.001656551 | 0.120881424 | 0.392366164 |
| 4 | 0.010104757  | 21.616057       | 0.025805998 | 0.777260487 | 0.059594623 |
| 5 | 0.001429017  | 57.480524       | 0.969796790 | 0.063914571 | 0.518874831 |
|   | wt           | qsec            |             |             |             |
| 1 | 0.0005253393 | 0.0001277169    |             |             |             |
| 2 | 0.0002096014 | 0.0046789491    |             |             |             |
| 3 | 0.0377028008 | 0.0001952599    |             |             |             |
| 4 | 0.7017528428 | 0.0024577686    |             |             |             |
| 5 | 0.2598094157 | 0.9925403056    |             |             |             |

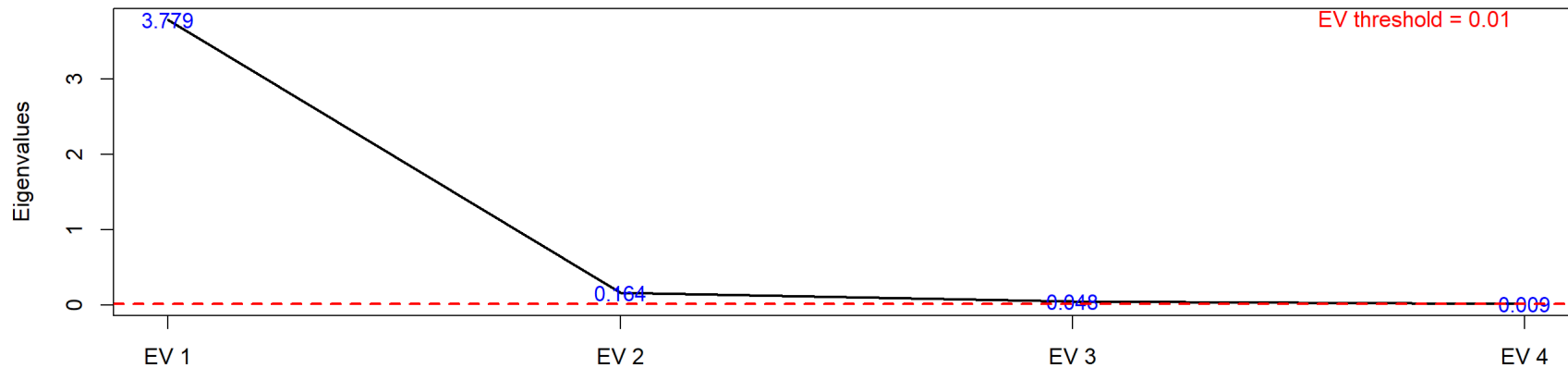
# Ejercicio aplicado en R

```
# Otro paquete para hacer diagnóstico de colinealidad es mctest  
# Leer https://journal.r-project.org/archive/2016/RJ-2016-062/RJ-2016-062.pdf y https://cran.r-project.org/web/packages/mctest/mctest.pdf para entender  
mc.plot(model)
```

VIF Plot



Eigenvalues Plot



# Ejercicio aplicado en R

```
imcdiag(model)
```

Call:  
imcdiag(mod = model)

All Individual Multicollinearity Diagnostics Result

|      | VIF    | TOL    | Wi      | Fi       | Leamer | CVIF    | Klein | IND1   |
|------|--------|--------|---------|----------|--------|---------|-------|--------|
| disp | 7.9854 | 0.1252 | 65.1974 | 101.2889 | 0.3539 | -1.0541 | 1     | 0.0134 |
| hp   | 5.1668 | 0.1935 | 38.8897 | 60.4180  | 0.4399 | -0.6820 | 0     | 0.0207 |
| wt   | 6.9169 | 0.1446 | 55.2248 | 85.7957  | 0.3802 | -0.9130 | 1     | 0.0155 |
| qsec | 3.1331 | 0.3192 | 19.9091 | 30.9302  | 0.5650 | -0.4136 | 0     | 0.0342 |

IND2

|      |        |
|------|--------|
| disp | 1.0875 |
| hp   | 1.0026 |
| wt   | 1.0635 |
| qsec | 0.8464 |

1 --> COLLINEARITY is detected by the test  
0 --> COLLINEARITY is not detected by the test

disp , hp , qsec , coefficient(s) are non-significant may be due to multicollinearity

R-square of y on all x: 0.8351

use method argument to check which regressors may be the reason of collinearity

=====

```
imcdiag(model, method = "Klein")
```

Call:  
imcdiag(mod = model, method = "Klein")

Klein Multicollinearity Diagnostics

|      | R2j    | R2(overall) | Difference | detection |
|------|--------|-------------|------------|-----------|
| disp | 0.8748 | 0.8351      | 0.0396     | 1         |
| hp   | 0.8065 | 0.8351      | -0.0287    | 0         |
| wt   | 0.8554 | 0.8351      | 0.0203     | 1         |
| qsec | 0.6808 | 0.8351      | -0.1543    | 0         |

Multicollinearity may be due to disp wt regressors

1 --> COLLINEARITY is detected by the test  
0 --> COLLINEARITY is not detected by the test

=====

# Ejercicio aplicado en R

```
# Solución: eliminar disp
model1 <- lm(mpg ~ hp + wt + qsec, data = mtcars)
summary(model1)
```

Call:  
lm(formula = mpg ~ hp + wt + qsec, data = mtcars)

Residuals:

|  | Min     | 1Q      | Median  | 3Q     | Max    |
|--|---------|---------|---------|--------|--------|
|  | -3.8591 | -1.6418 | -0.4636 | 1.1940 | 5.6092 |

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t )     |
|-------------|----------|------------|---------|--------------|
| (Intercept) | 27.61053 | 8.41993    | 3.279   | 0.00278 **   |
| hp          | -0.01782 | 0.01498    | -1.190  | 0.24418      |
| wt          | -4.35880 | 0.75270    | -5.791  | 3.22e-06 *** |
| qsec        | 0.51083  | 0.43922    | 1.163   | 0.25463      |

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.578 on 28 degrees of freedom  
Multiple R-squared: 0.8348, Adjusted R-squared: 0.8171  
F-statistic: 47.15 on 3 and 28 DF, p-value: 4.506e-11

```
imcdiag(model1, corr=T)
```

Call:  
imcdiag(mod = model1, corr = T)

## All Individual Multicollinearity Diagnostics Result

|      | VIF    | TOL    | Wi      | Fi       | Leamer | CVIF    | Klein | IND1   | IND2 |
|------|--------|--------|---------|----------|--------|---------|-------|--------|------|
| hp   | 4.9220 | 0.2032 | 56.8684 | 117.6587 | 0.4507 | -1.5328 | 0     | 0.0140 |      |
| wt   | 2.5304 | 0.3952 | 22.1914 | 45.9133  | 0.6286 | -0.7880 | 0     | 0.0273 |      |
| qsec | 2.8738 | 0.3480 | 27.1702 | 56.2141  | 0.5899 | -0.8950 | 0     | 0.0240 |      |

```
# hp y qsec tienen alta correlación, se podría eliminar qsec
model2 <- lm(mpg ~ hp + wt, data = mtcars)
summary(model2)
```

Call:  
lm(formula = mpg ~ hp + wt, data = mtcars)

Residuals:

|  | Min    | 1Q     | Median | 3Q    | Max   |
|--|--------|--------|--------|-------|-------|
|  | -3.941 | -1.600 | -0.182 | 1.050 | 5.854 |

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t )     |
|-------------|----------|------------|---------|--------------|
| (Intercept) | 37.22727 | 1.59879    | 23.285  | < 2e-16 ***  |
| hp          | -0.03177 | 0.00903    | -3.519  | 0.00145 **   |
| wt          | -3.87783 | 0.63273    | -6.129  | 1.12e-06 *** |

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.593 on 29 degrees of freedom  
Multiple R-squared: 0.8268, Adjusted R-squared: 0.8148  
F-statistic: 69.21 on 2 and 29 DF, p-value: 9.109e-12

```
imcdiag(model2, corr=T)
```

Call:  
imcdiag(mod = model2, corr = T)

## All Individual Multicollinearity Diagnostics Result

|    | VIF    | TOL    | Wi      | Fi  | Leamer | CVIF    | Klein | IND1   | IND2 |
|----|--------|--------|---------|-----|--------|---------|-------|--------|------|
| hp | 1.7666 | 0.5661 | 22.9987 | Inf | 0.7524 | -0.8613 | 0     | 0.0189 | 1    |
| wt | 1.7666 | 0.5661 | 22.9987 | Inf | 0.7524 | -0.8613 | 0     | 0.0189 | 1    |

1 --> COLLINEARITY is detected by the test  
0 --> COLLINEARITY is not detected by the test