Enfoque Estadístico del Aprendizaje

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Lucia Montes Rego y Guadalupe Alonso

Regresión con Componentes Principales (PCR)

Una estrategia posible para abordar la multicolinealidad en Modelos de Regresión Lineal

Motivación

Definición de colinealidad y multicolinealidad. Efectos sobre la Regresión Lineal Múltiple.

Diagnóstico

Diagnóstico informal. Diagnóstico mediante el cálculo de VIF.

Estrategias

Algunas técnicas para abordar el problema de la multicolinealidad

PCR

Descripción del método. Pasos a seguir.

Ejemplo de aplicación

Ejemplo práctico de aplicación de PCR sobre base de datos de NBA

Regresión Lineal Múltiple

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

Suma del cuadrado de los residuos

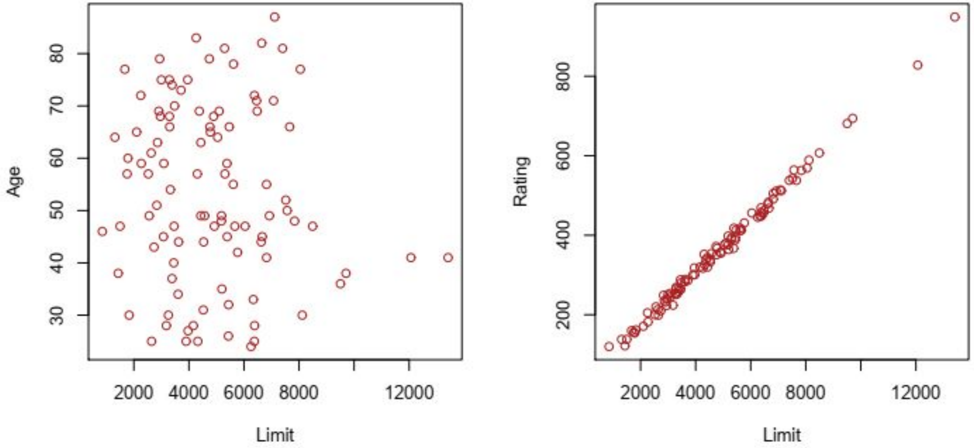
Regresión Lineal Múltiple

Colinealidad

Ejemplo

Base de datos de **crédito** con observaciones *balance, edad, limite y rating*.

Observaciones de *edad, límite y rating*



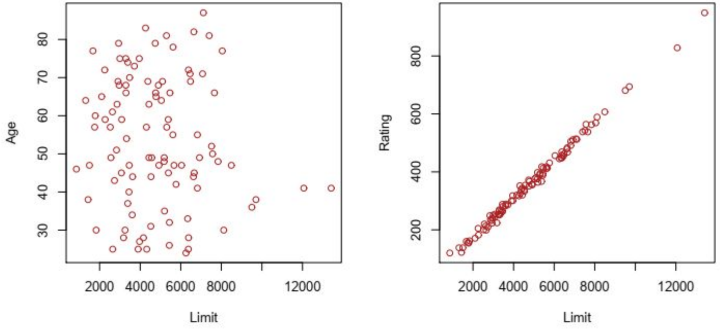
JAMES, Gareth, et al. *An introduction to statistical learning*. New York: springer, 2013.

Regresión Lineal Múltiple

Colinealidad

Ejemplo Base de datos de **crédito** con observaciones *balance, edad, limite y rating*.

Observaciones de *edad, límite y rating*



JAMES, Gareth, et al. *An introduction to statistical learning*. New York: springer, 2013.

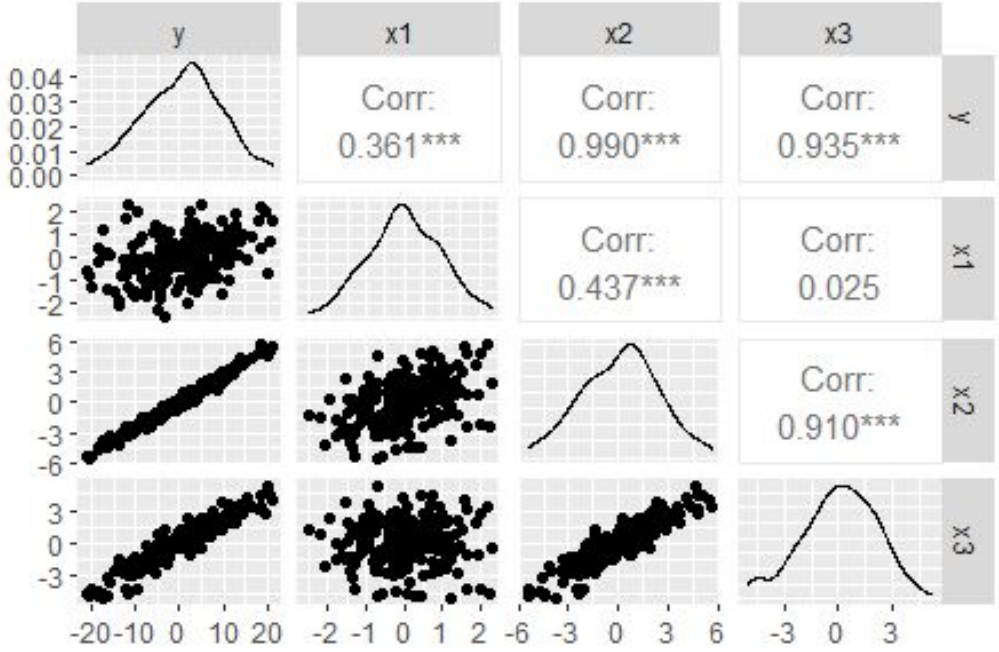
Modelo 1→ *balance ~ age + limit*
Modelo 2→ *balance ~ rating + limit*

		Coefficient	Std. Error	t-statistic	p-value
Model 1	Intercept	-173.411	43.828	-3.957	< 0.0001
	age	-2.292	0.672	-3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	-377.537	45.254	-8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012

Multicolinealidad

Ejemplo de juguete

```
set.seed(31)
n<-200
x1<-rnorm(n)
x3<-rnorm(n,sd=2)
x2<-x1+x3+rnorm(n)/30
y<-2*x1+x2+3*x3+rnorm(n)
```



Multicolinealidad

Ejemplo de juguete

```
set.seed(31)
n<-200
x1<-rnorm(n)
x3<-rnorm(n,sd=2)
x2<-x1+x3+rnorm(n)/30
y<-2*x1+x2+3*x3+rnorm(n)
```

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-2.41431 -0.54805 -0.03971  0.57021  2.72901

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02450    0.07171   0.342    0.733
x1           1.71144    2.20019   0.778    0.438
x2           1.43987    2.19161   0.657    0.512
x3           2.54519    2.19853   1.158    0.248

Residual standard error: 1.012 on 196 degrees of freedom
Multiple R-squared:  0.9873, Adjusted R-squared:  0.9871
F-statistic:  5090 on 3 and 196 DF,  p-value: < 2.2e-16
```

Multicolinealidad

Ejemplo de juguete

```
set.seed(31)
n<-200
x1<-rnorm(n)
x3<-rnorm(n,sd=2)
x2<-x1+x3+rnorm(n)/30
y<-2*x1+x2+3*x3+rnorm(n)
```

```
Call:
lm(formula = y ~ x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-2.45805 -0.54626 -0.04083  0.55585  2.65431

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02227    0.07158   0.311    0.756
x2           3.14364    0.07479  42.031   <2e-16 ***
x3           0.83627    0.08341  10.026   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.011 on 197 degrees of freedom
Multiple R-squared:  0.9873, Adjusted R-squared:  0.9872
F-statistic:  7651 on 2 and 197 DF,  p-value: < 2.2e-16
```

Multicolinealidad

Ejemplo de juguete

```
set.seed(31)
n<-200
x1<-rnorm(n)
x3<-rnorm(n,sd=2)
x2<-x1+x3+rnorm(n)/30
y<-2*x1+x2+3*x3+rnorm(n)
```

```
Call:
lm(formula = y ~ x1 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-2.44458 -0.54240 -0.03136  0.59205  2.79054

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02641    0.07155   0.369    0.712
x1           3.15610    0.07505  42.052   <2e-16 ***
x3           3.98942    0.03460 115.297   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.011 on 197 degrees of freedom
Multiple R-squared:  0.9873, Adjusted R-squared:  0.9872
F-statistic:  7658 on 2 and 197 DF,  p-value: < 2.2e-16
```

Multicolinealidad

Ejemplo

Generan datasets con una variable respuesta y tres predictoras, con diversos grados de multicolinealidad.

VATCHEVA, Kristina P., et al. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale, Calif.)*, 2016, vol. 6, no 2.

Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies

Kristina P. Vatcheva[1], **MinJae Lee**[2], **Joseph B. McCormick**[1], and **Mohammad H. Rahbar**[3,*]

[1]Division of Epidemiology, University of Texas Health Science Center-Houston, School of Public Health, Brownsville Campus, Brownsville, TX

[2]Division of Clinical and Translational Sciences, Department of Internal Medicine, University of Texas Medical School, Biostatistics/Epidemiology/Research Design (BERD) Core, Center for Clinical and Translational Sciences (CCTS), The University of Texas Health Science Center at Houston, Houston, TX

[3]Division of Epidemiology, Human Genetics and Environmental Sciences, University of Texas School of Public Health, Division of Clinical and Translational Sciences, Department of Internal Medicine, University of Texas Medical School at Houston, and Center for Clinical and Translational Sciences at The University of Texas Health Science Center at Houston, Houston, TX

Abstract

The adverse impact of ignoring multicollinearity on findings and data interpretation in regression analysis is very well documented in the statistical literature. The failure to identify and report multicollinearity could result in misleading interpretations of the results. A review of epidemiological literature in PubMed from January 2004 to December 2013, illustrated the need for a greater attention to identifying and minimizing the effect of multicollinearity in analysis of data from epidemiologic studies. We used simulated datasets and real life data from the Cameron County Hispanic Cohort to demonstrate the adverse effects of multicollinearity in the regression analysis and encourage researchers to consider the diagnostic for multicollinearity as one of the steps in regression analysis.

Keywords

Multicollinearity; Regression analysis; Simulation; BMI; Waist circumference

Multicolinealidad

Correlation Scenario (Corr(x1, x2), Corr(x2, x3), Corr(x1, x3))	Predictor Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
1 (.1,.1,.1)	Intercept	70.08	4.34	16.18	<.0001
	x_1	0.3	0.08	3.65	0.0101
	x_2	0.17	0.04	4.68	0.0008
	x_3	0.45	0.04	12.86	<.0001
8 (.85,.5,.1)	Intercept	118.26	3.74	31.68	<.0001
	x_1	2.69	0.2	13.39	<.0001
	x_2	−1.28	0.1	−12.63	<.0001
	x_3	1	0.05	18.74	<.0001

VATCHEVA, Kristina P., et al. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale, Calif.)*, 2016, vol. 6, no 2.

Multicolinealidad

Afecta los resultados obtenidos en Regresión Lineal Múltiple

Multicolinealidad

Diagnóstico informal

Multicolinealidad

Diagnóstico informal

Los **coeficientes de regresión estimados se modifican sustancialmente** al incorporar o quitar variables del modelo

Multicolinealidad

Diagnóstico informal

Los **coeficientes de regresión estimados se modifican sustancialmente** al incorporar o quitar variables del modelo

Test no significativos para los coeficientes asociados a las variables

Multicolinealidad

Diagnóstico informal

Los **coeficientes de regresión estimados se modifican sustancialmente** al incorporar o quitar variables del modelo

Test no significativos para los coeficientes asociados a las variables

Coeficientes estimados con signo contrario al que se espera según consideraciones teóricas

Multicolinealidad

Diagnóstico informal

Los **coeficientes de regresión estimados se modifican sustancialmente** al incorporar o quitar variables del modelo

Test no significativos para los coeficientes asociados a las variables

Coeficientes estimados con signo contrario al que se espera según consideraciones teóricas

Coeficientes de correlación grandes para las predictoras tomadas de a pares

Multicolinealidad

Diagnóstico formal

Multicolinealidad

Diagnóstico formal

SZRETTER NOSTE, María Eugenia. Apunte de Regresión Lineal. *Buenos Aires*, 2013.

VARIANCE INFLATION FACTOR (VIF)

Multicolinealidad

Diagnóstico formal

SZRETTER NOSTE, María Eugenia. Apunte de Regresión Lineal. *Buenos Aires*, 2013.

VARIANCE INFLATION FACTOR (VIF)

$$VIF_k = \frac{1}{1 - R_k^2}, \qquad 1 \leq k \leq p-1,$$

Es un número que **se calcula para cada covariable**

Multicolinealidad

Diagnóstico formal

SZRETTER NOSTE, María Eugenia. Apunte de Regresión Lineal. *Buenos Aires*, 2013.

VARIANCE INFLATION FACTOR (VIF)

$$VIF_k = \frac{1}{1 - R_k^2}, \qquad 1 \leq k \leq p-1,$$

Es un número que **se calcula para cada covariable**

Coeficiente de determinación múltiple de la regresión de X_k sobre el resto de las variables predictoras.

Multicolinealidad

Diagnóstico formal

SZRETTER NOSTE, María Eugenia. Apunte de Regresión Lineal. *Buenos Aires*, 2013.

VARIANCE INFLATION FACTOR (VIF)

$$VIF_k = \frac{1}{1 - R_k^2}, \qquad 1 \le k \le p-1,$$

Es un número que **se calcula para cada covariable**

- $VIF_k = 1$ si la k-ésima covariable no está correlacionada con las restantes variables ($R^2_k = 0$)
- $VIF_k > 1$ si $R^2_k \neq 0$
- Si R^2 está muy cerca de 1, VIF_k se vuelve un número enorme:
  - VIF > 10 → **multicolinealidad**
- Si el promedio de los VIF > 1 → **multicolinealidad**

Multicolinealidad

Diagnóstico

Ejemplo de juguete

VARIANCE INFLATION FACTOR (VIF)

```
set.seed(31)
n<-200
x1<-rnorm(n)
x3<-rnorm(n,sd=2)
x2<-x1+x3+rnorm(n)/30
y<-2*x1+x2+3*x3+rnorm(n)
```

```
library(car)
> vif(ml_132)
       x1        x2        x3
 857.4402 4977.8759 4028.1231
> vif(ml_23)
      x2       x3
5.809174 5.809174
> vif(ml_12)
      x1       x2
1.236561 1.236561
> vif(ml_13)
      x1       x3
1.000632 1.000632
```

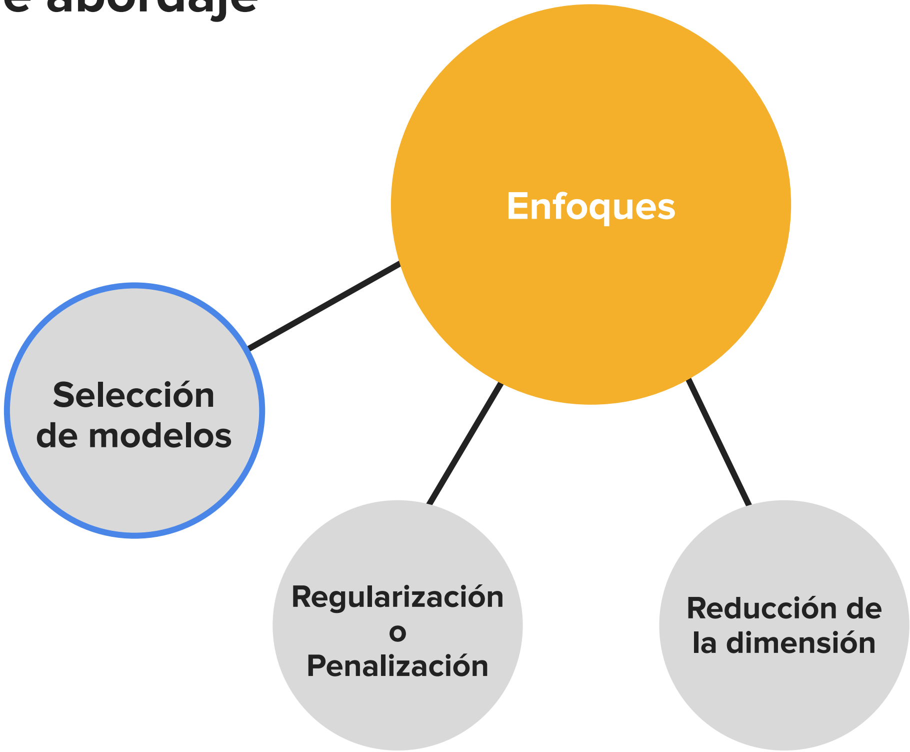
Multicolinealidad

Opciones de abordaje

Multicolinealidad

Opciones de abordaje

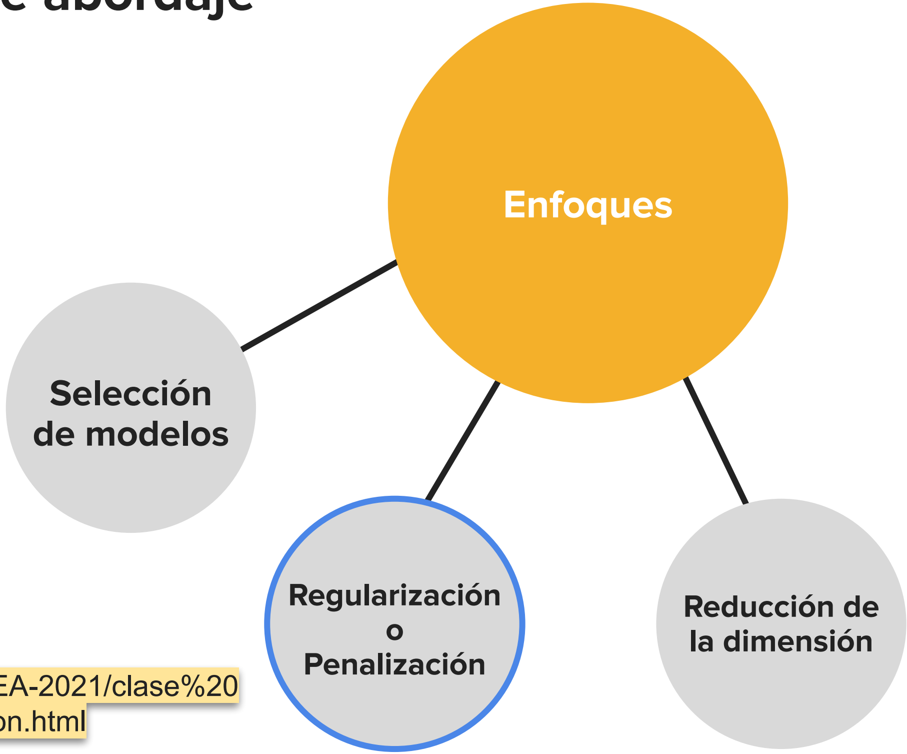
SZRETTER NOSTE, María Eugenia. Apunte de Regresión Lineal. *Buenos Aires*, 2013.



Multicolinealidad

SZRETTER NOSTE, María Eugenia. Apunte de Regresión Lineal. *Buenos Aires*, 2013.

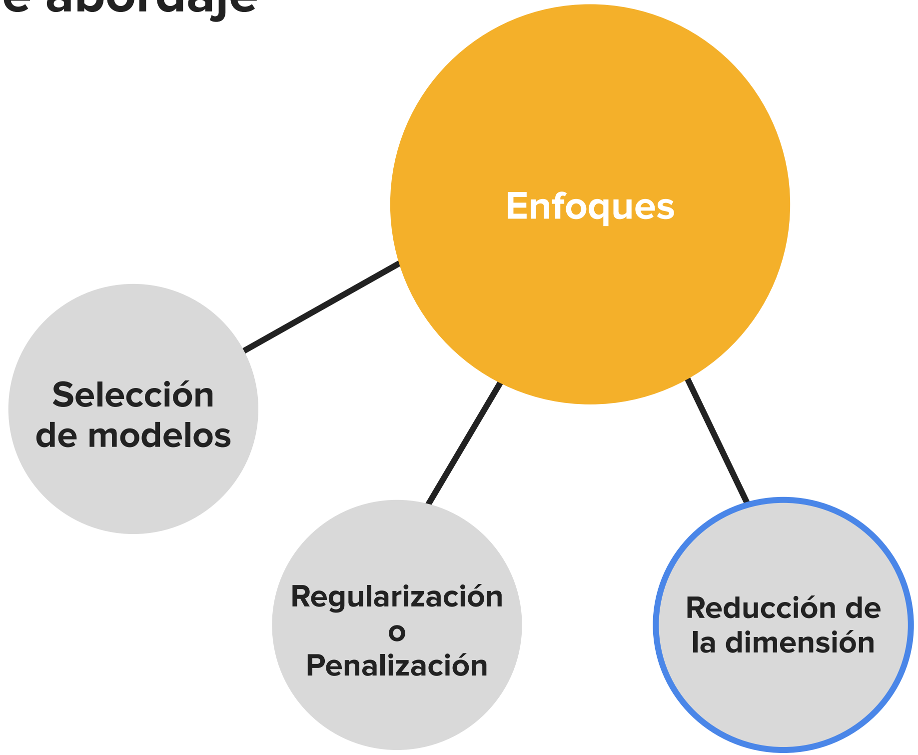
Opciones de abordaje



https://eea-uba.github.io/EEA-2021/clase%2010/regularizacion.html

Multicolinealidad

Opciones de abordaje



JAMES, Gareth, et al. *An introduction to statistical learning*. New York: springer, 2013.

Regresión con Componentes Principales (PCR)

Análisis de Componentes Principales

Breve repaso

- PCA es una técnica para **reducir la dimensión** de una matriz de datos **X** de n×p
- La dirección de PC1 es aquella a lo largo de la cual las observaciones varían más
- PC2 es una combinación lineal de las variables que no están correlacionadas con PC1 y tiene la varianza más grande sujeta a esta restricción (ortogonalidad)
- Las siguientes PC maximizan sucesivamente la varianza, manteniendo la condición de no estar correlacionadas con las PC anteriores
- Se pueden generar tantas componentes principales (PC) como variables existentes

Regresión con Componentes Principales

Estandarización de las variables

Paso 1

Paso 2

Paso 3

Paso 4

Regresión con Componentes Principales

PCA
Construcción de *M* componentes principales $(Z_1,..,Z_M)$

Paso 1

Paso 2

Paso 3

Paso 4

Regresión con Componentes Principales

Selección de los componentes principales (PC) que se utilizarán en la regresión lineal

Paso 1

Paso 2

Paso 3

Paso 4

Regresión con Componentes Principales

Utilización de **componentes principales como predictores** en una regresión lineal por mínimos cuadrados

Paso 1

Paso 2

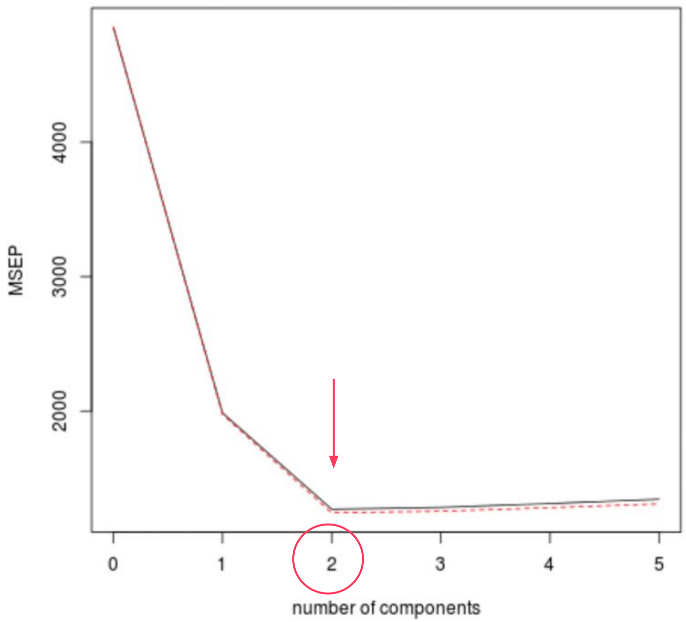
Paso 3

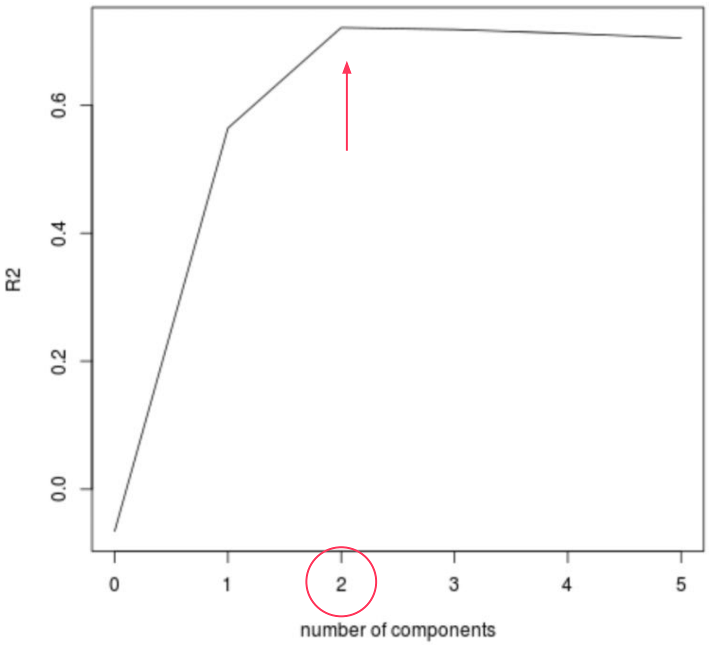
Paso 4

Regresión con Componentes Principales

Selección de Componentes Principales para el Modelo Lineal

Statology
https://www.statology.org/principal-components-regression-in-r/





CASO PRÁCTICO

Regresión con Componentes Principales

Ventajas

- Abordaje del problema de multicolinealidad entre predictores
- Reducción de la dimensionalidad
- Mitigación del sobreajuste

GRACIAS.