Clase 13 Introducción a los modelos lineales

Curso Introducción al Análisis de datos con R para la acuicultura.

Dr. José A. Gallardo y Dra. María Angélica Rueda.
jose.gallardo@pucv.cl | Pontificia Universidad Católica de Valparaíso

27 July 2021

PLAN DE LA CLASE

1.- Introducción

-Modelo de regresión lineal múltiple.

-El problema de la multicolinealidad.

-¿Cómo seleccionar variables?.

-¿Cómo comparar modelos?.

-Interpretación regresión lineal múltiple con R.

2.- Práctica con R y Rstudio cloud

-Realizar análisis de regresión lineal múltiple.

-Realizar gráficas avanzadas con ggplot2.

-Elaborar un reporte dinámico en formato pdf.

REGRESIÓN LINEAL MÚLTIPLE

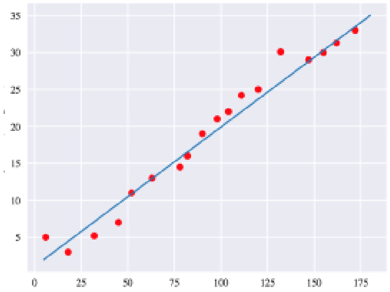
Sea Y una variable respuesta continua y X_1, X_2,. . . ,X_p variables predictoras, un modelo de regresión lineal múltiple se puede representar como,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i$$

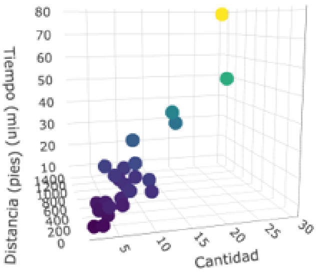
donde β_0 es el intercepto y β_1, β_2,. . . , β_p representan los coeficientes de regresión estandarizados.

COMPARACIÓN MODELO LINEAL SIMPLE Y MÚLTIPLE

2 dimensiones



3 o más dimensiones



COMPARACIÓN MODELOS LINEAL SIMPLE Y MÚLTIPLE 2

Minimizar suma de cuadrados

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\sum_{i=1}^{n}\left(y_i - (\beta_0 + \sum_{j=1}^{p-1}\beta_j x_{ij})\right)^2$$

Predecir observaciones

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p-1}\hat{\beta}_j x_{ij}.$$

Calcular residuos

$$\hat{\epsilon}_i = y_i - \hat{y}_i.$$

$$\hat{\epsilon}_i = y_i - \hat{y}_i.$$

PRUEBAS DE HIPÓTESIS EN RLM

Igual que RL simple

- ***Hipótesis intercepto***

$H_0 : \beta_0 = 0$

$H_1 : \beta_0 \neq 0$

Igual que RL simple, pero...

- ***Hipótesis coef. regresión***

$H_0 : \beta_{i1}, ..., \beta_{ip} = 0$

$H_1 : \beta_{i1}, ..., \beta_{ip} \neq 0$

Igual que RL simple

- ***Hipótesis modelo de regresión***

$H_0 : \beta_j = 0 \ (j = 1, 2, ..., k)$

$H_1 : \beta_j \neq 0$

PROBLEMAS A RESOLVER CON LOS MODELOS DE REGRESIÓN MÚLTIPLE

Para p variables predictoras existen N modelos diferentes que pueden usarse para estimar, modelar o predecir la variable respuesta.

Problemas

1). ¿Qué hacer si las variables predictoras están correlacionadas?

2). ¿Cómo seleccionar variables para incluir en el modelo?

3). ¿Qué hacemos con las variables que no tienen efecto sobre la variable respuesta?

4). Dado N modelos ¿Cómo compararlos?, ¿Cuál es mejor?

ESTUDIO DE CASO - REGRESIÓN LINEAL MÚLTIPLE

Origen de los datos: Simulación de una variable respuesta Y y dos variables predictoras X_1 y X_2.

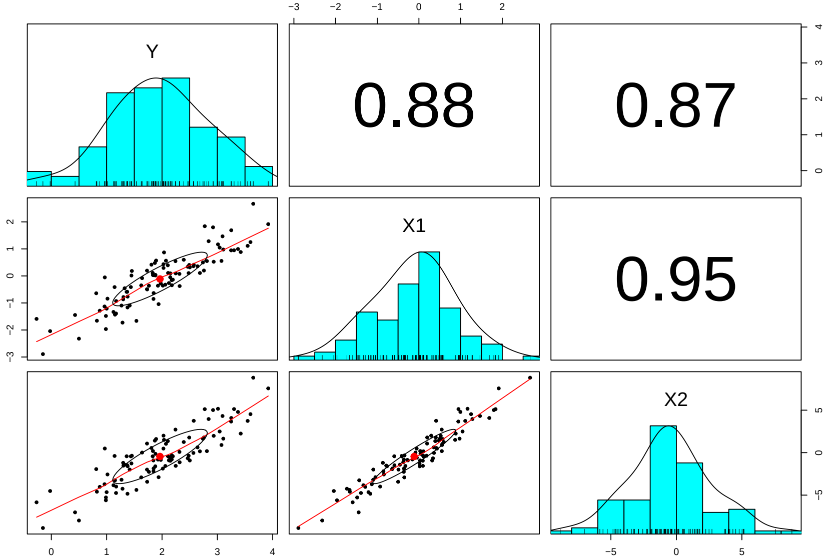
Modelo lineal lm1<- lm(Y~X1+X2)

Table 1: Tabla de datos

Y	X1	X2
2.811	0.5497	0.1796
1.015	-0.8416	-2.566
1.836	0.033	0.1865
2.934	0.5241	1.979
1.287	-1.728	-4.251
1.978	-0.2779	-0.857

SUPUESTO 1: MULTICOLINEALIDAD

Gráfica de correlaciones (>0,80 es problema)



SUPUESTO 1: MULTICOLINEALIDAD

Factor de inflación-varianza (VIF)

VIF: Es una medida del grado en que la varianza del estimador de mínimos cuadrados incrementa, por la colinealidad entre las variables predictoras.

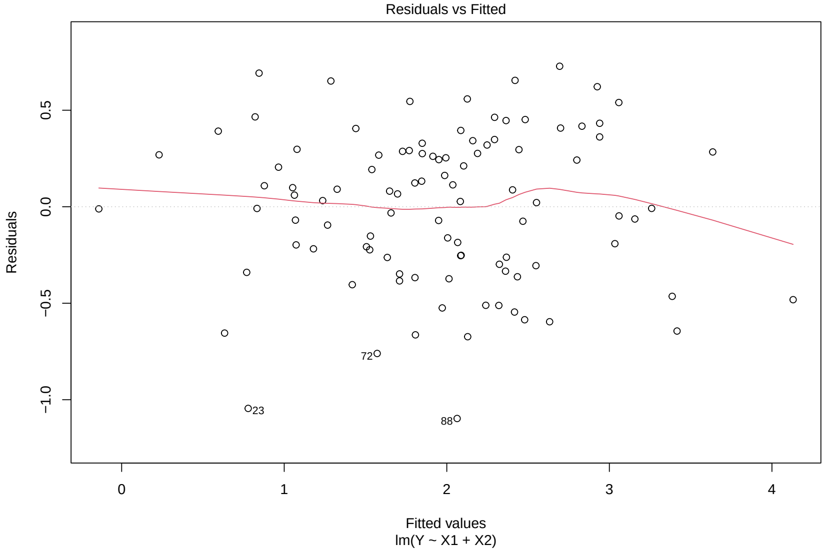
$$VIF = \frac{1}{1 - R_i^2}$$

R_i^2 es el coeficiente de determinación de la ecuación de regresión de Y_i como variable respuesta en función del resto de variables predictoras. $\mathbf{VIF} > \mathbf{10}$ es evidencia de alta multicolinealidad.

X1	X2
10.6	10.6

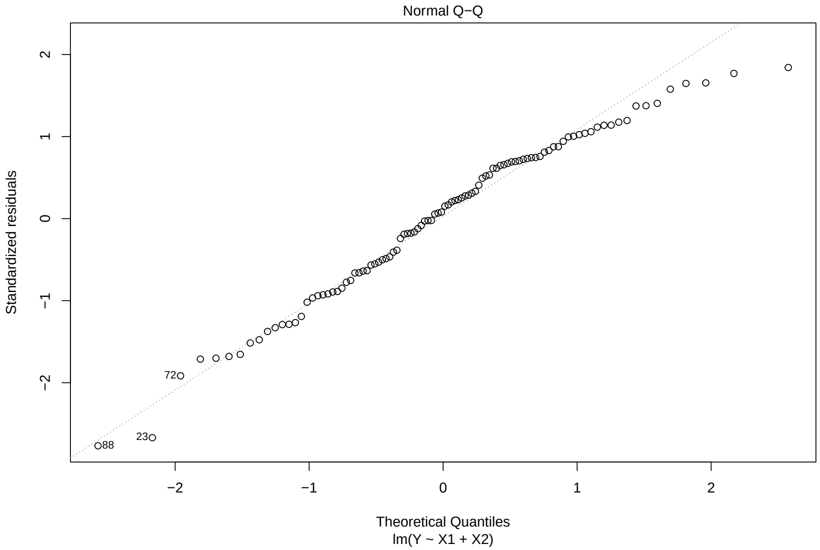
SUPUESTO 2: HOMOGENEIDAD DE VARIANZAS

```
plot(lm1, which = 1)
```



SUPUESTO 3: NORMALIDAD

```
plot(lm1, which = 2)
```



REGRESIÓN LINEAL MÚLTIPLE

Modelo lineal

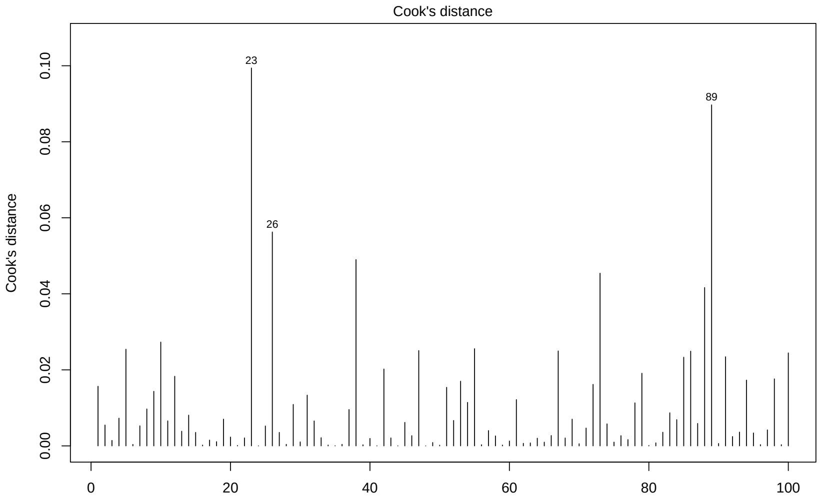
```
summary(lm1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.09720 -0.27163  0.04586  0.29220  0.72779
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.05696    0.04044  50.865  < 2e-16 ***
## X1           0.53563    0.13172   4.067 9.71e-05 ***
## X2           0.07307    0.04087   1.788   0.0769 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

IDENTIFICAR VALORES ATÍPICOS "OUTLIERS": DISTANCIA DE COOK

La distancia de Cook es un criterio para identificar datos atípicos.

```
plot(lm1, which = 4)
```



ELIMINACIÓN DE DATOS ATÍPICOS

datos_new <- sim_dat[-c(23,89,26),]

```
sim_dat<-as.data.frame(cbind(Y,X1,X2))
str(sim_dat)
```

```
## 'data.frame':    100 obs. of  3 variables:
##  $ Y : num  2.81 1.01 1.84 2.93 1.29 ...
##  $ X1: num  0.55 -0.842 0.033 0.524 -1.728 ...
##  $ X2: num  0.18 -2.566 0.187 1.979 -4.251 ...
```

```
datos_new <- sim_dat[-c(23,89,26),]
str(datos_new)
```

```
## 'data.frame':    97 obs. of  3 variables:
##  $ Y : num  2.81 1.01 1.84 2.93 1.29 ...
##  $ X1: num  0.55 -0.842 0.033 0.524 -1.728 ...
##  $ X2: num  0.18 -2.566 0.187 1.979 -4.251 ...
```

¿CÓMO RESOLVEMOS MULTICOLINEALIDAD?

1). Eliminar variables correlacionadas: pero podríamos estar generando el problema de las variables omitidas.

2). Transformar una de las variables: log u otra.

3). Reemplazar por variables ortogonales: Una solución simple y elegante son los componentes principales.

REGRESIÓN LINEAL MÚLTIPLE

Modelo lineal

```
summary(lm2)
```

```
##
## Call:
## lm(formula = Y ~ X1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10803 -0.24008  0.05222  0.26130  0.75213
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.04930    0.04066   50.40   <2e-16 ***
## X1           0.75974    0.04090   18.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
```

¿CUÁL ES EL MEJOR MODELO LINEAL?

Comparación por criterios

Akaike Information Criterion **(AIC)** y Bayesian Information Criterion **(BIC)** ambos criterios penalizan la complejidad del modelo. Al igual que RSS mientras menor su valor, mejor es el modelo.

AIC

```
aic <- AIC(lm1, lm2)
```

Table 4: Comparación modelos usando AIC

	df	AIC
lm1	4	105.2
lm2	3	106.5

¿CUÁL ES EL MEJOR MODELO LINEAL?

Comparación por criterios

BIC

```
bic <- BIC(lm1, lm2)
```

Table 5: Comparación modelos usando BIC

	df	BIC
lm1	4	115.6
lm2	3	114.3

RESUMEN DE LA CLASE

1). Revisión de conceptos de pruebas de hipótesis y modelos lineales.

2). Elaborar y evaluar modelos lineales simples.