Clase 16 Regresión lineal múltiple y supuestos.

Diplomado en Análisis de Datos con R e Investigación reproducible para Biociencias.

Dr. José Gallardo Matus & Dra. Angélica Rueda Calderón

Pontificia Universidad Católica de Valparaíso

18 October 2022

PLAN DE LA CLASE

1.- Introducción

- Modelo de regresión lineal múltiple (MRLM).
- Estudio de caso: transformación de variables predictoras.
- Pruebas de hipótesis.
- Supuestos de MRLM
- 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 html.

REGRESIÓN LINEAL MÚLTIPLE

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

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

$\beta_0 =$ Intercepto. $\beta_1, \beta_2, ..., \beta_p =$ Coeficientes de regresión.

ESTUDIO DE RENDIMIENTO EN MAÍZ

En este estudio de caso trabajaremos con un subset de datos de maíz (*corn*) del paquete de R **agridat** (n=162).

heady.fertilizer {agridat}

Table 1: Tabla de datos de maíz

Variable	Descripción	Tipo de efecto/ variable
Crop	Cultivo se seleccionó maíz	Criterio de clasificación/ Factor de tratamiento cualitativo
P	Mediciones de fósforo	Variable regresora númerica
N	Mediciones de nitrógeno	Variable regresora númerica
yield	Rendimiento (g).	Variable respuesta/ Cuantitativa continua

PRUEBAS DE HIPÓTESIS REGRESIÓN LINEAL MÚLTIPLE

- **Intercepto**
  Igual que en regresión lineal simple.
- **Modelo completo**
  Igual que en regresión lineal simple.
- **Coeficientes**
  Uno para cada variable y para cada factor de una variable de clasificación.

ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE: PROBLEMAS

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

Problemas

- ¿Qué hacer si las variables predictoras están correlacionadas?.
- ¿Cómo seleccionar variables para incluir en el modelo?.
- ¿Qué hacemos con las variables que no tienen efecto sobre la variable respuesta?.
- Dado N modelos ¿Cómo compararlos?, ¿Cuál es mejor?.

SUPUESTOS DE MODELO DE REGRESIÓN LÍNEAL MÚLTIPLE

¿Cuales son los supuestos?

- Independencia.
- Linealidad entre variable cada variable independiente y dependiente.
- Homocedasticidad.
- Normalidad.
- Multicolinealidad.

MODELO DE REGRESIÓN LÍNEAL MÚLTIPLE

```
# Crea modelo de regresión múltiple con lm()
m1 <- lm(yield ~ N + P + sqrt(N) + sqrt(P) + sqrt(N*P),
         data=d2)
# Imprime resultado con función summary()
summary(m1)$coefficients%>%kable()
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5.6944239	6.6272936	-0.8592382	0.3921123
N	-0.3162168	0.0399606	-7.9132187	0.0000000
P	-0.4174864	0.0399606	-10.4474565	0.0000000
sqrt(N)	6.3532022	0.8681460	7.3181263	0.0000000
sqrt(P)	8.5176589	0.8681460	9.8113206	0.0000000
sqrt(N * P)	0.3409584	0.0385388	8.8471577	0.0000000

$R^2 = 0.92$, $p\text{-}val = 6.836811 \times 10^{-35}$

SUPUESTO DE INDEPENDENCIA: PRUEBA DE DURBIN-WATSON

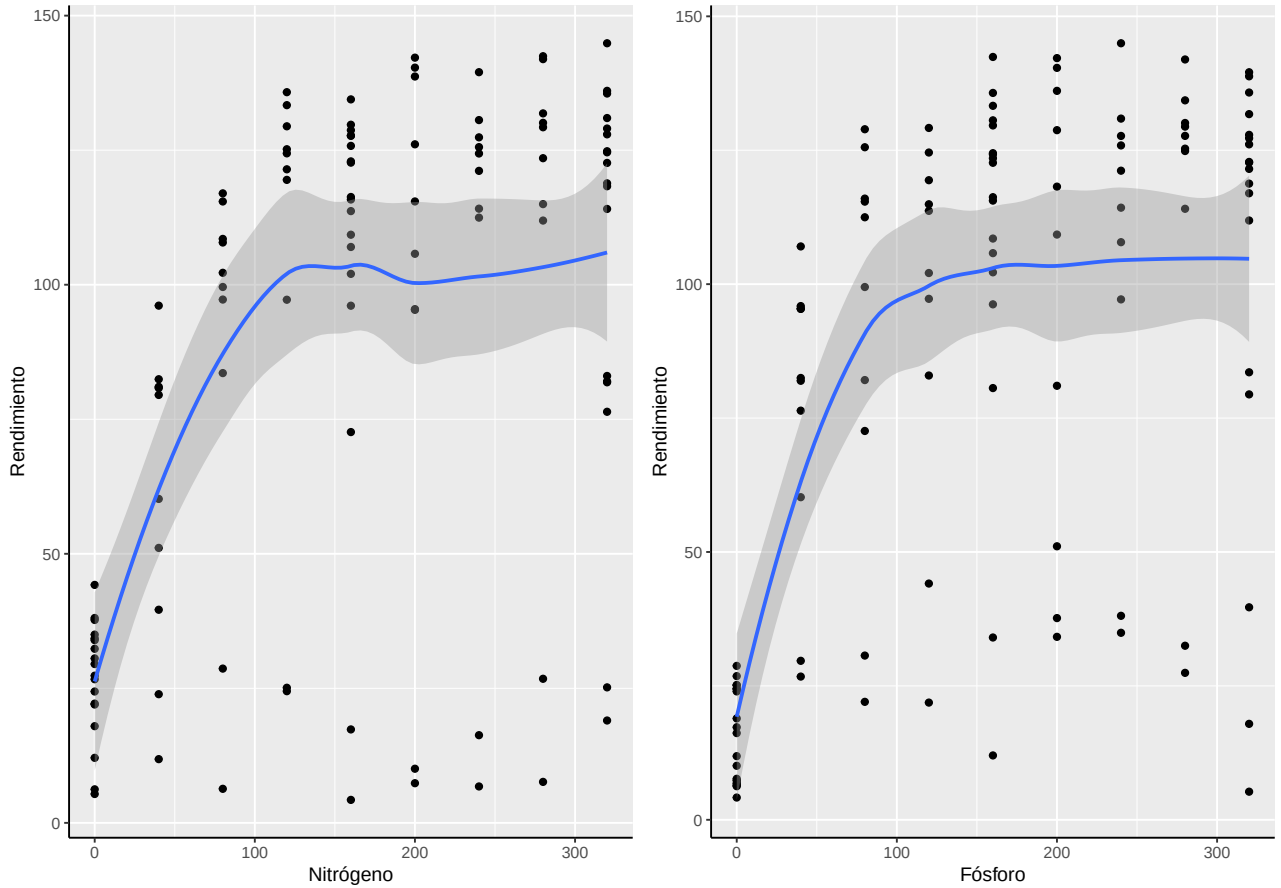
$\mathbf{H_0}$: Los residuos son independientes entre sí.

```
dwtest(yield ~ N + P + sqrt(N) + sqrt(P) + sqrt(N*P),
       data=d2,
       alternative = c("two.sided"),
       iterations = 15)
```

```
##
##  Durbin-Watson test
##
## data:  yield ~ N + P + sqrt(N) + sqrt(P) + sqrt(N * P)
## DW = 1.8923, p-value = 0.3301
## alternative hypothesis: true autocorrelation is not 0
```

SUPUESTO DE LINEALIDAD: MÉTODO GRÁFICO

H_0: Hay relación lineal entre cada variable predictora y la variable respuesta.



SUPUESTO DE HOMOGENEIDAD DE VARIANZAS: PRUEBA DE BREUSCH-PAGAN

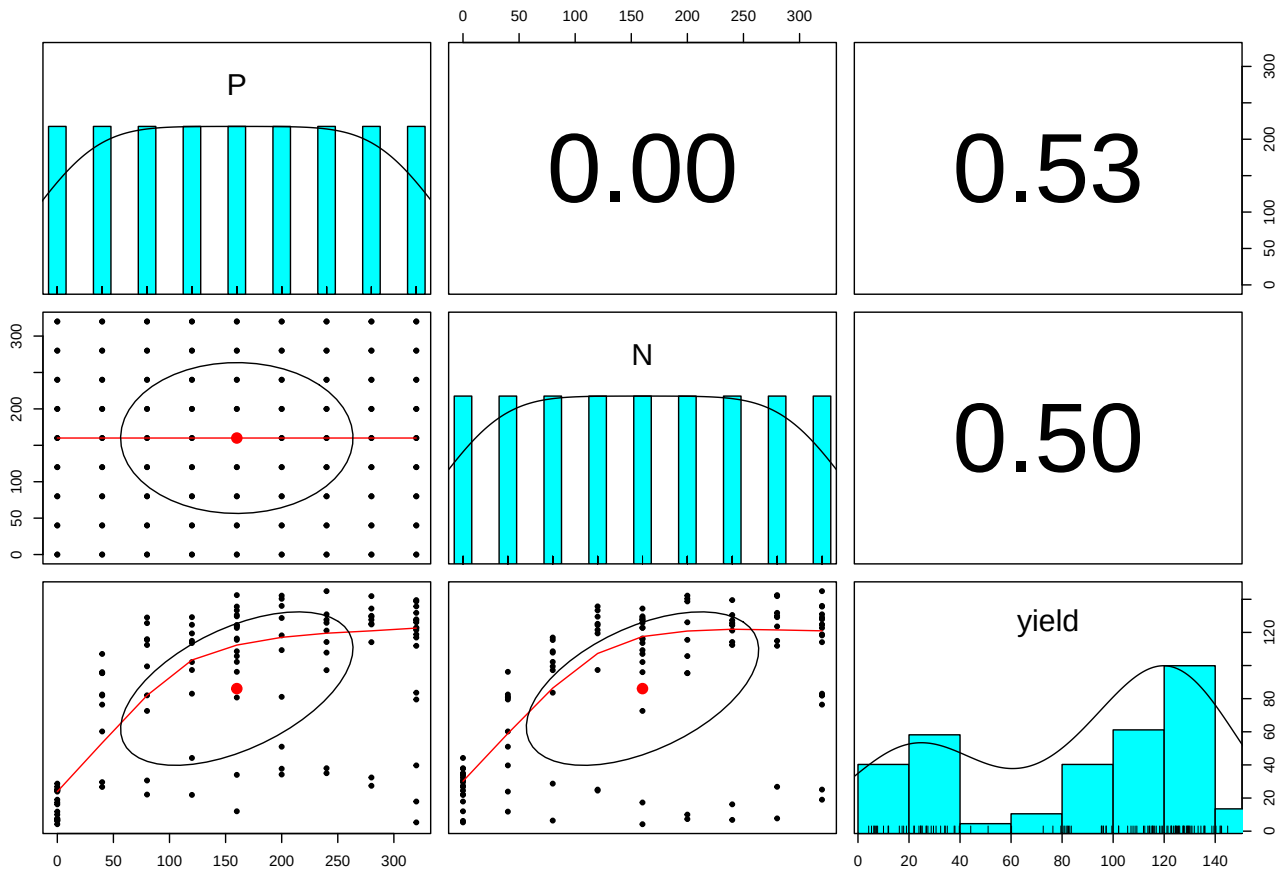
$\mathbf{H_0}$: La varianza de los residuos es constante.

```
bptest(m1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m1
## BP = 0.21033, df = 5, p-value = 0.999
```

SUPUESTO DE MULTICOLINEALIDAD

Correlaciones >0,80 es problema.



FACTOR DE INFLACIÓN DE LA 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.
- Mayor a 10 es evidencia de alta multicolinealidad.

```
m1 <- lm(yield ~ N + P + sqrt(N) + sqrt(P) + sqrt(N*P),
         data=d2)
vif(m1) %>% kable(digits=2, col.names = c("VIF"))
```

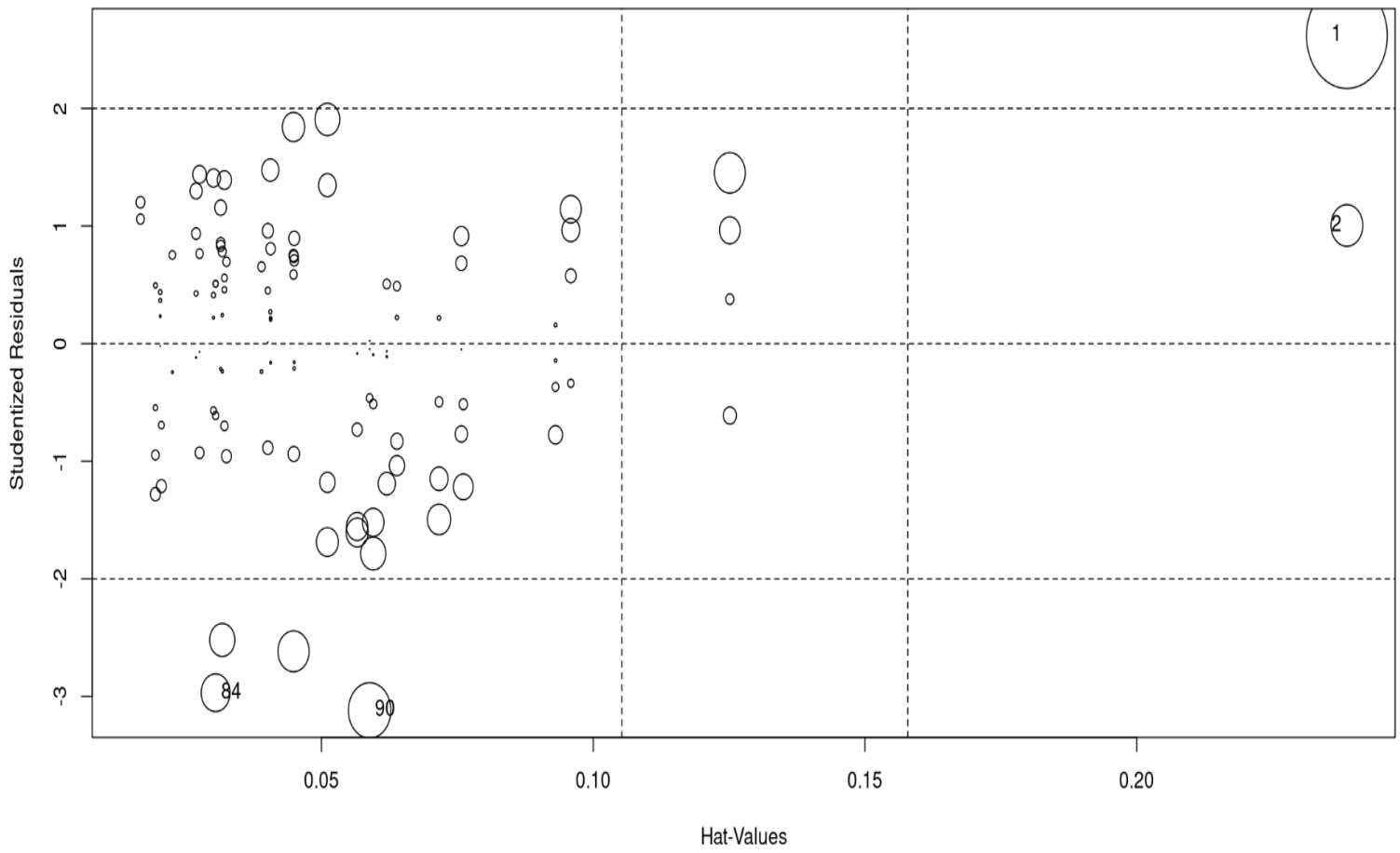
	VIF
N	11.89
P	11.89
sqrt(N)	16.25
sqrt(P)	16.25
sqrt(N * P)	9.29

¿CÓMO RESOLVEMOS MULTICOLINEALIDAD?

- Eliminar variables correlacionadas, pero podríamos eliminar una variable causal.
- Transformar una de las variables: log u otra.
- Reemplazar por variables ortogonales: Una solución simple y elegante son los componentes principales (ACP).

DATOS INFLUYENTES

influencePlot(m1)



SUPUESTO DE NORMALIDAD: PRUEBA DE SHAPIRO-WILKS

$\mathbf{H_0}$: Los residuos tienen distribución normal.

```
shapiro.test(x= rstudent(m1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(m1)
## W = 0.97683, p-value = 0.04511
```

```
shapiro.test(x= rstudent(m1)[-90])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(m1)[-90]
## W = 0.97784, p-value = 0.05735
```

COMPARACIÓN: MODELO COMPLETO 0

```
# Crea modelo de regresión múltiple
lm0<- lm(yield ~ N + P + sqrt(N) + sqrt(P) + sqrt(N*P),
         data=d2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5.6944239	6.6272936	-0.8592382	0.3921123
N	-0.3162168	0.0399606	-7.9132187	0.0000000
P	-0.4174864	0.0399606	-10.4474565	0.0000000
sqrt(N)	6.3532022	0.8681460	7.3181263	0.0000000
sqrt(P)	8.5176589	0.8681460	9.8113206	0.0000000
sqrt(N * P)	0.3409584	0.0385388	8.8471577	0.0000000

$R^2 = 0.92$, $p\text{-}val = 6.836811 \times 10^{-35}$

COMPARACIÓN: MODELO REDUCIDO 1

```
lm1<- lm(yield ~ N + P + N:P ,
         data=d2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.1649969	9.1398529	4.066258	0.0000900
N	0.0860205	0.0477341	1.802077	0.0742720
P	0.0964640	0.0477341	2.020861	0.0457231
N:P	0.0007631	0.0002478	3.079794	0.0026165

$R^2 = 0.56$, $p\text{-}val = 1.4988588 \times 10^{-12}$

COMPARACIÓN: MODELO REDUCIDO 2

```
lm2<- lm(yield ~ sqrt(N) + sqrt(P) + sqrt(N*P),
         data=d2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.8198018	9.5995947	2.6896762	0.0082665
sqrt(N)	0.5255024	0.7623597	0.6893102	0.4920796
sqrt(P)	0.9082304	0.7623597	1.1913410	0.2360839
sqrt(N * P)	0.3524855	0.0602904	5.8464635	0.0000001

$R^2 = 0.8$, $p\text{-}val = 4.1336333 \times 10^{-25}$

CRITERIOS PARA COMPARAR MODELOS

Existen diferentes criterios para comparar modelos.

- Anova de residuales (RSS).
- Criterios que penalizan incrementar el número de parámetros estimados (más variables predictoras):
  - **a)** Akaike Information Criterion (AIC).
  - **b)** Bayesian Information Criterion (BIC).
- En todos los casos mientras menor es el valor de RSS, AIC o BIC mejor es el modelo.
- No necesariamente los resultados son equivalentes entre criterios.

COMPARACIÓN USANDO RESIDUALES

```
anova(lm0, lm1, lm2) %>% kable()
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
108	19873.76	NA	NA	NA	NA
110	106457.61	-2	-86583.85	235.2614	0
110	49567.37	0	56890.24	NA	NA

COMPARACIÓN USANDO AIC Y BIC

AIC $= -2 * log - likelihood + 2 * K$
BIC $= -2 * log - likelihood + log(n) * K$
K= número de parámetros a estimar.

	df	AIC
lm0	7	925.8671
lm1	5	1113.1986
lm2	5	1026.0554

	df	BIC
lm0	7	945.0205
lm1	5	1126.8796
lm2	5	1039.7364

PRÁCTICA ANÁLISIS DE DATOS

- El trabajo práctico se realiza en Rstudio.cloud.

Guía 16 Regresión lineal multiple.

RESUMEN DE LA CLASE

- Elaborar hipótesis para una regresión lineal múltiple.
- Interpretar coeficientes.
- Evaluar supuestos.
- Comparar modelos: residuales, AIC, BIC.