

Clase 16 Regresión lineal múltiple.

Diplomado en 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

09 November 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, \dots, X_p variables predictoras, un modelo de regresión lineal múltiple se puede representar como,

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

β_0 = Intercepto. $\beta_1 X_{i1}, \beta_2 X_{i2}, \beta_p X_{ip}$ = Coeficientes de regresión estandarizados.

Si $p = 1$, el modelo es una regresión lineal simple.

Si $p > 1$, el modelo es una regresión lineal múltiple.

Si $p > 1$ y alguna variable predictora es Categórica, el modelo se denomina ANCOVA.

ESTUDIO DE CASO DIETAS MITILIDOS

Dieta microencapsulada en mitilidos.

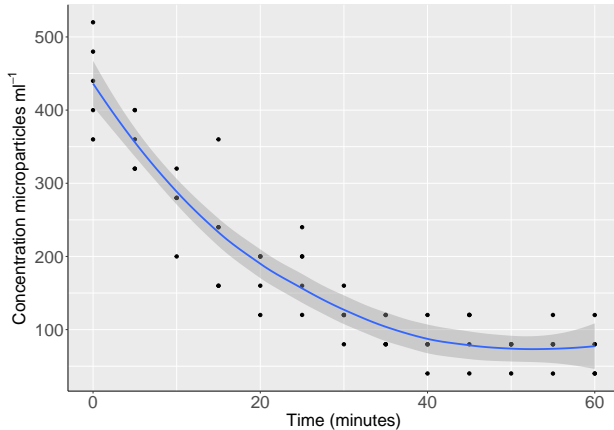
Fuente: Willer and Aldridge 2017

time	sample	replicate	particle concentration
0	mussel	a	400
5	mussel	a	320
10	mussel	a	280
...
0	control	a	160
5	Control	a	120
10	Control	a	120

TASA DE ACLARACIÓN MUSSEL.

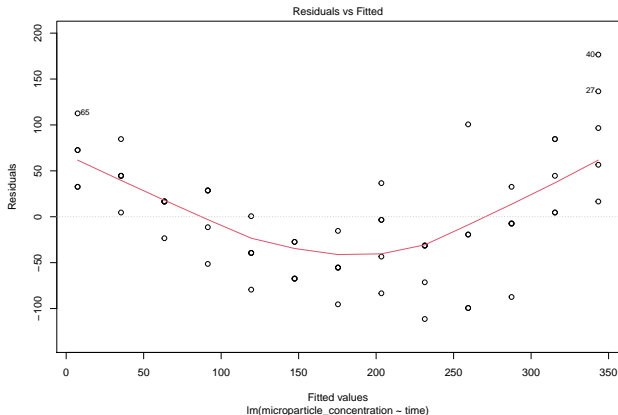
Problemas: La concentración es discreta y la relación es no lineal.

Tips: `stat_smooth(method='loess', formula=y~x, se=T)`



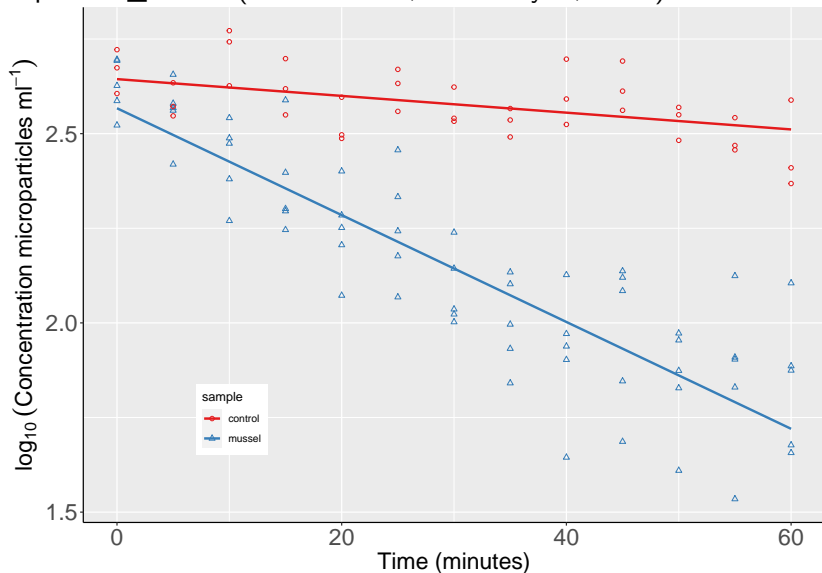
EVALUACION SUPUESTOS.

```
reg_mussel <- lm(microparticle_concentration ~ time,  
                  data=mussel)  
plot(reg_mussel, which = 1)
```



REGRESIÓN LINEAL SOBRE LOG10(TASA ACLARACIÓN)

Tips: `stat_smooth(method='lm', formula=y~x, se=F)`



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.

REGRESIÓN LINEAL MULTIPLE

```
lm.full <- lm(log_microparticle_concentration  
  ~ time*sample + time + sample,  
  data = clearance)  
  
summary(lm.full)$coef %>% kable()
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.6440298	0.0355452	74.385053	0.0000000
time	-0.0022153	0.0010054	-2.203443	0.0298584
samplemussel	-0.0769430	0.0449615	-1.711309	0.0901242
time:samplemussel	-0.0119008	0.0012717	-9.358133	0.0000000

$$R^2 = 0.87, p\text{-val} = 1.0691926 \times 10^{-28}$$

ANCOVA

```
anova(lm.full) %>% kable()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
time	1	3.391944	3.391944	245.84687	0
sample	1	4.590457	4.590457	332.71466	0
time:sample	1	1.208266	1.208266	87.57466	0
Residuals	100	1.379698	0.013797	NA	NA

REGRESIÓN LINEAL SIMPLE: MUSSEL

```
reg_mussel <- lm(log_microparticle_concentration  
                  ~ time, data=mussel)  
summary(reg_mussel)$coef %>% kable()
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.567087	0.0333508	76.97221	0
time	-0.014116	0.0009433	-14.96447	0

$$R^2 = 0.78, p\text{-val} = 2.0490325 \times 10^{-22}$$

REGRESIÓN LINEAL SIMPLE: CONTROL

```
reg_control <- lm(log_microparticle_concentration  
                  ~ time, data=control)  
summary(reg_control)$coef %>% kable()
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.6440298	0.0160715	164.516921	0.00e+00
time	-0.0022153	0.0004546	-4.873339	2.08e-05

$$R^2 = 0.39, p\text{-val} = 2.0849643 \times 10^{-5}$$

PROBLEMAS CON LOS ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

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?.

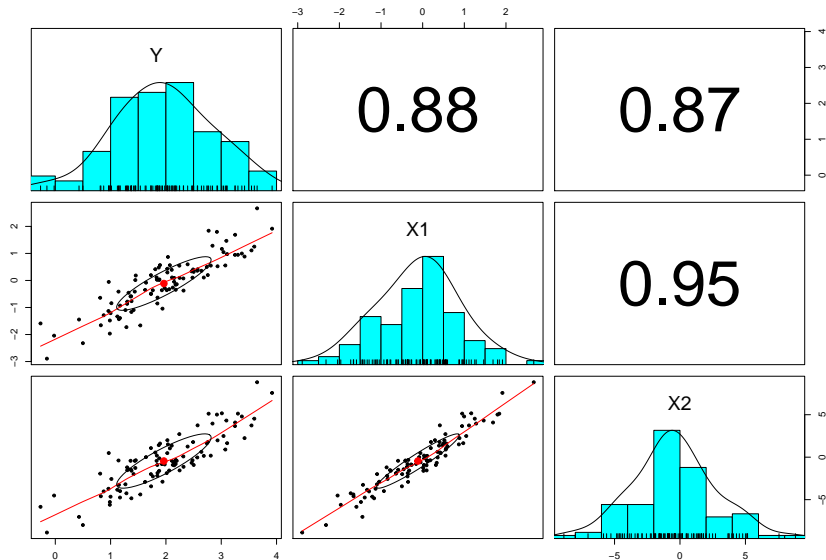
DATOS SIMULADOS PARA REG. LINEAL MÚLTIPLE

100 datos simulados de 3 variables cuantitativas continuas.

Y	X1	X2
2.81	0.55	0.18
1.01	-0.84	-2.57
1.84	0.03	0.19
2.93	0.52	1.98
1.29	-1.73	-4.25
1.98	-0.28	-0.86

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

```
lm1<- lm(Y~X1+X2)
vif(lm1) %>%
  kable(digits=2, col.names = c("VIF"))
```

	VIF
X1	10.6
X2	10.6

¿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).

COMPARACIÓN DE MODELOS: MODELO COMPLETO

```
lm1 <- lm(Y~X1+X2)
```

```
summary(lm1)$coef %>% kable()
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.0569644	0.0404396	50.865151	0.0000000
X1	0.5356269	0.1317168	4.066505	0.0000971
X2	0.0730690	0.0408696	1.787858	0.0769216

$$R^2 = 0.79, p\text{-val} = 4.4295606 \times 10^{-34}$$

COMPARACIÓN DE MODELOS: MODELO REDUCIDO

```
lm<- lm(Y~X1)
```

```
summary(lm)$coef %>% kable()
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.049298	0.0406597	50.40121	0
X1	0.759739	0.0408995	18.57574	0

$$R^2 = 0.78, p\text{-val} = 7.108665 \times 10^{-34}$$

COMPARACIÓN DE MODELOS

```
# analisis de residuales
```

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

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
97	15.48007	NA	NA	NA	NA
98	15.99018	-1	-0.5101139	3.196436	0.0769216

```
# Criterio AIC - penaliza el número de variables
```

```
AIC(lm1, lm2) %>% kable()
```

	df	AIC
lm1	4	105.2260
lm2	3	106.4682

PRÁCTICA ANÁLISIS DE DATOS

- ▶ Guía de trabajo práctico disponible en drive y Rstudio.cloud.

Clase_15

- ▶ El trabajo práctico se realiza en Rstudio.cloud.

Guía 15 Regresión lineal

RESUMEN DE LA CLASE

- ▶ **Elaborar hipótesis para una regresión lineal múltiple**
- ▶ **Realizar análisis de covarianza**
- ▶ **Interpretar coeficientes**
- ▶ **Evaluar supuestos: multicolinealidad**
- ▶ **Comparar modelos**