Clase 14 Regresión lineal múltiple.

Diplomado en Análisis de datos con R para la Acuicultura.

Dr. José Gallardo Matus

Pontificia Universidad Católica de Valparaíso

24 May 2022

PLAN DE LA CLASE

1.- Introducción

- Modelo de regresión lineal múltiple.
- Estudio de caso: transformación de variable respuesta.
- Pruebas de hipótesis.
- 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 $X_1, \ldots, X_p$ 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. $\beta_1 X_{i1}, \beta_2 X_{i2}, \beta_p X_{ip} =$ 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 ALIMENTACION MOLUSCOS FILTRADORES

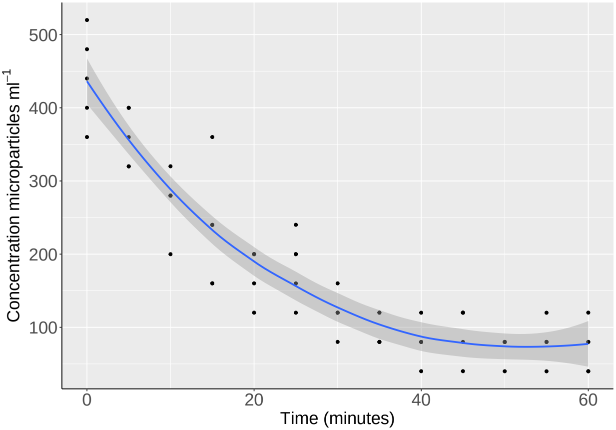
Dieta microencapsulada en mitilidos.

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

Fuente: Willer and Aldridge 2017

TASA DE ACLARACIÓN (PROXY DE CONSUMO DE PARTÍCULAS).

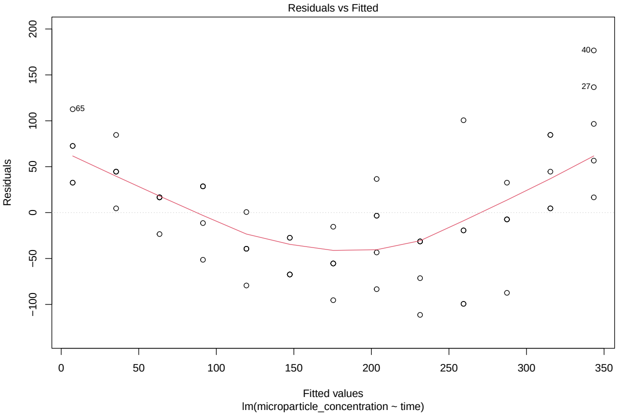
Problemas: Concentración es discreta y relación es no lineal.



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

EVALUACION SUPUESTOS.

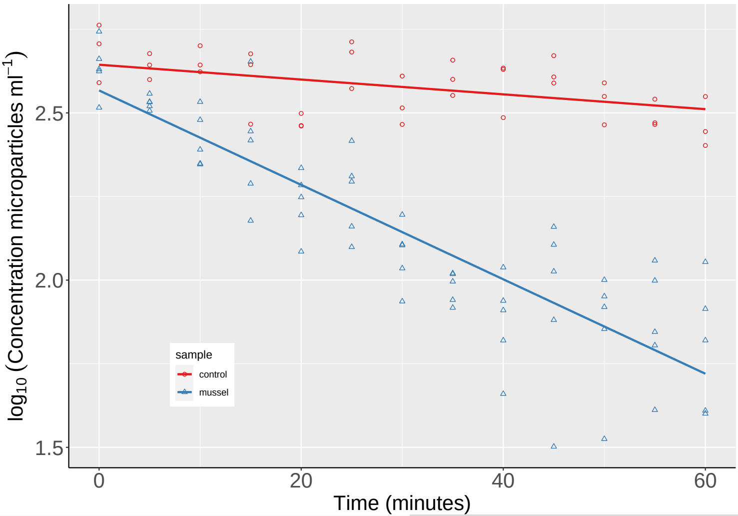
```
reg_mussel <- lm(microparticle_concentration ~ time,
                 data=mussel)
plot(reg_mussel, which = 1)
```



TRANSFORMACIÓN DE VARIABLE RESPUESTA.

Regresión lineal sobre Log10(Tasa de 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.

INTERPRETACIÓN COMO REGRESIÓN LINEAL MULTIPLE

```r
# Crea modelo de regresión múltiple (RM) con lm()
lm.full <- lm(log_microparticle_concentration
              ~ time*sample + time + sample,
              data = clearance)

# Imprime resultado RM con función summary()
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

INTERPRETACIÓN COMO 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

COMPARACIÓN CON REGRESIONES LINEALES SIMPLES

```r
# Crea dos modelos de regresión lineal simple
reg_mussel <- lm(log_microparticle_concentration
                  ~ time, data=mussel)

reg_control <- lm(log_microparticle_concentration
                   ~ time, data=control)
```

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

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

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

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?.

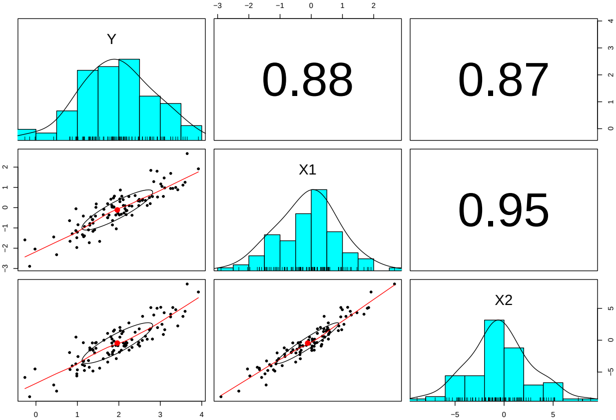
DATOS SIMULADOS PARA REG. LINEAL MÚLTIPLE

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 0

```r
# Crea modelo de regresión múltiple
lm0<- lm(Y~X1+X2)
```

	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 1

```r
# Crea modelo de regresión simple variable X1
lm1<- lm(Y~X1)
```

	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: MODELO REDUCIDO 2

```r
# Crea modelo de regresión simple variable X2
lm2<- lm(Y~X2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.0678250	0.0434322	47.61041	0
X2	0.2312349	0.0135089	17.11726	0

$R^2 = 0.75$, $p\text{-}val = 3.3098905 \times 10^{-31}$

CRITERIOS PARA COMPARAR MODELOS.

Existen diferentes criterios para comparar modelos.
- Anova de residuales (RSS).
- Criterios que penalizan número de variables:

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.

COMPARACIÓN DE MODELOS USANDO RESIDUALES.

```r
anova(lm0, 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
98	18.11910	0	-2.1289130	NA	NA

COMPARACIÓN DE MODELOS USANDO AIC Y BIC.

```r
AIC <- AIC(lm0, lm1, lm2)
AIC <- BIC(lm0, lm1, lm2)
```

	df	AIC
lm0	4	105.2260
lm1	3	106.4682
lm2	3	118.9673

	df	BIC
lm0	4	115.6467
lm1	3	114.2837
lm2	3	126.7828

PRÁCTICA ANÁLISIS DE DATOS

- El trabajo práctico se realiza en Rstudio.cloud.
  Guía 14 Regresión lineal multiple

RESUMEN DE LA CLASE

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