Clase 14 Regresión lineal múltiple

OCE 386 - Introducción al análisis de datos con R.

Dr. José A. Gallardo | Pontificia Universidad Católica de Valparaíso

16 November 2021

PLAN DE LA CLASE

1.- Introducción

- Modelo de regresión lineal múltiple.
- ¿Cómo eliminar valores atípicos?
- ¿Cómo seleccionar variables y 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,

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

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.

CASO DE ESTUDIO: GALÁPAGOS

Estudio de diversidad de especies en islas galápagos considera 30 islas y 7 variables aleatorias: 5 de tipo continuo y 2 de tipo discreto.

Extracto del set de datos gala

	Species	Endemics	Area	Elevation	Nearest	Scruz	Adjacent
Baltra	58	23	25.09	346	0.6	0.6	1.84
Bartolome	31	21	1.24	109	0.6	26.3	572.33
Caldwell	3	3	0.21	114	2.8	58.7	0.78
Champion	25	9	0.10	46	1.9	47.4	0.18
Coamano	2	1	0.05	77	1.9	1.9	903.82
Daphne.Major	18	11	0.34	119	8.0	8.0	1.84

DESCRIPCIÓN DE VARIABLES Y OBJETIVO

Objetivo de la RLM

Estimar o predecir el número de *Species* en función de algunas variables independientes o predictoras.

Variable predictora	**Descripción**
Area	Area of the island (km^2)
Elevation	Highest elevation of the island (m)
Nearest	Distance from the nearest island (km)
Scruz	Distance from Santa Cruz island (km)
Adjacent	Area of the adjacent island (square km)

MODELO SIMPLE

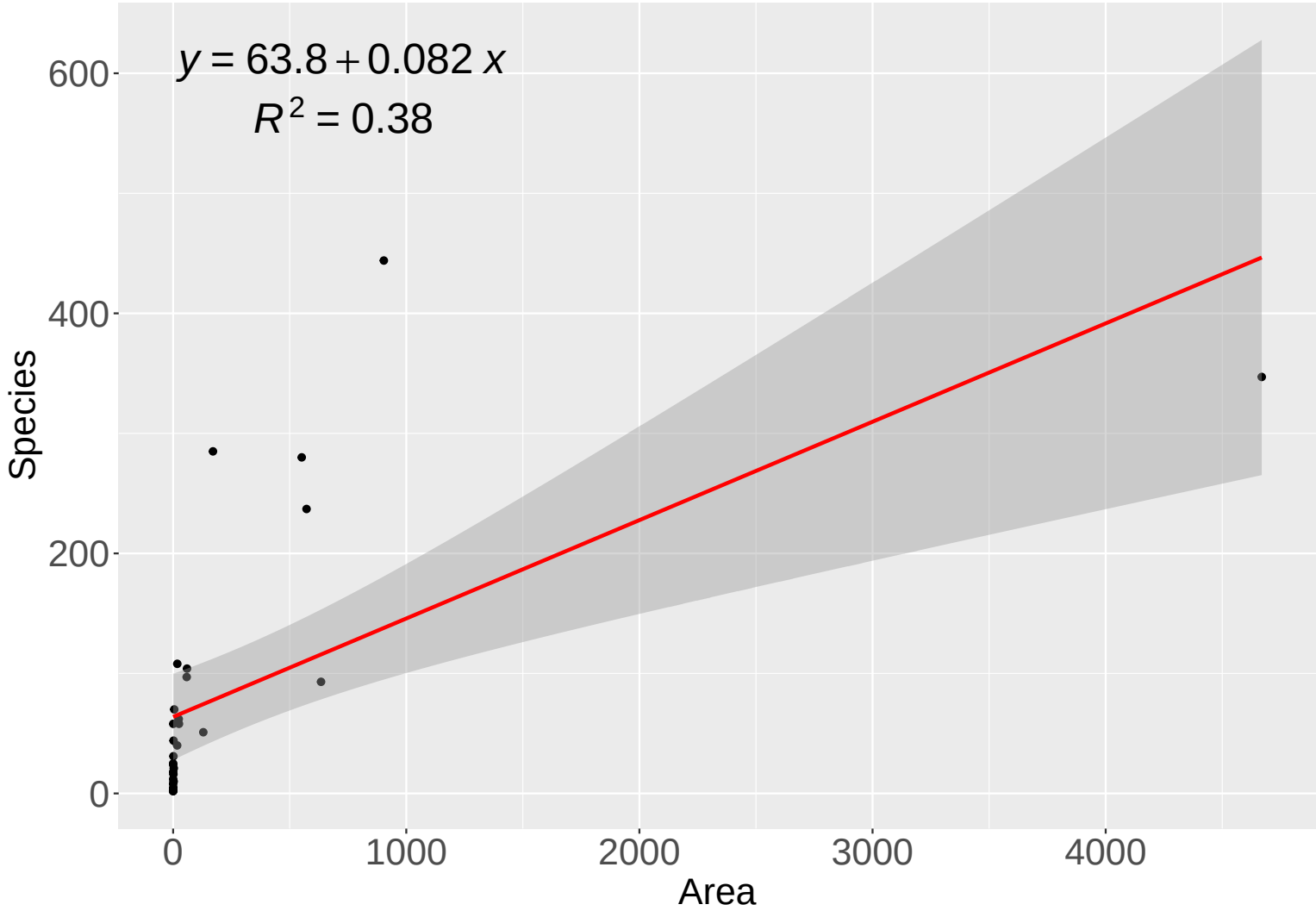
```
lm.1 <- lm (Species ~ Area, data=gala)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	63.78	17.52	3.64	0.001094
Area	0.08196	0.01971	4.158	0.0002748

Table 3: Modelo de regresión lineal simple del set de datos gala

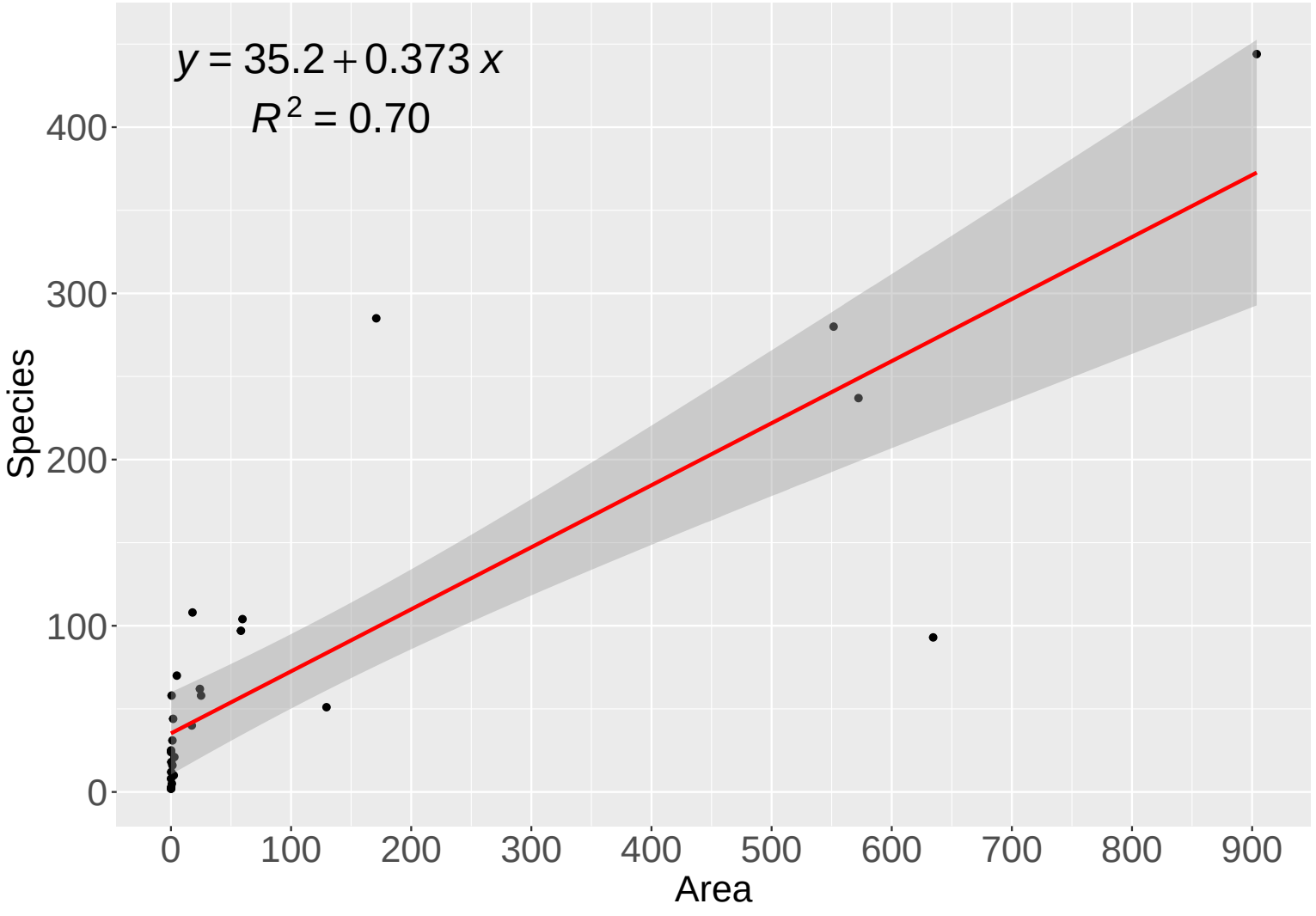
Observations	Residual Std. Error	R^2	Adjusted R^2
30	91.73	0.3817	0.3596

RELACIÓN LINEAL Species ~ Area



RELACIÓN LINEAL Species ~ Area (sin outliers).

```
is.na(gala$Area) <- gala$Area >= 3000
```



MODELO COMPLETO

```
lm.2 <- lm (Species ~ Area + Elevation + Nearest +
                Scruz + Adjacent, data=gala)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	22.59	13.4	1.685	0.1055
Area	0.2957	0.06186	4.781	8.042e-05
Elevation	0.1404	0.0497	2.824	0.009613
Nearest	-0.2552	0.7217	-0.3536	0.7269
Scruz	-0.0901	0.1498	-0.6015	0.5534
Adjacent	-0.06503	0.01223	-5.318	2.124e-05

Table 5: Modelo de regresión múltiple del set de datos gala

Observations	Residual Std. Error	R^2	Adjusted R^2
29	41.65	0.8714	0.8434

MODELO REDUCIDO

```
lm.3 <- lm (Species ~ Area + Elevation +
                Adjacent, data=gala)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.52	12.01	1.375	0.1813
Area	0.309	0.05956	5.189	2.289e-05
Elevation	0.1292	0.04783	2.701	0.01222
Adjacent	-0.06367	0.01159	-5.493	1.047e-05

Table 7: Modelo reducido de regresión múltiple del set de datos gala

Observations	Residual Std. Error	R^2	Adjusted R^2
29	41.01	0.8645	0.8482

COMPARACIÓN DE MODELOS

```
# análisis de residuales
anova(lm.1, lm.3, lm.2) %>% kable()
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
27	93188.71	NA	NA	NA	NA
25	42038.68	2	51150.031	14.7458318	0.0000757
23	39890.96	2	2147.718	0.6191567	0.5471326

```
# Criterio AIC - penaliza el número de variables
AIC(lm.1, lm.3, lm.2) %>% kable()
```

	df	AIC
lm.1	3	322.4759
lm.3	5	303.3909
lm.2	7	305.8701

PRÁCTICA ANÁLISIS DE DATOS

- Guía de trabajo práctico disponible en drive y Rstudio.cloud. **Clase_14**
- El trabajo práctico se realiza en Rstudio.cloud. **Guía 14 Regresión lineal múltiple**

RESUMEN DE LA CLASE

- **Elaborar hipótesis para una regresión lineal múltiple**
- **Realizar análisis de covarianza**
- **Interpretar coeficientes**
- **Comparar modelos**