

Clase 15 Regresión lineal

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

06 November 2021

PLAN DE LA CLASE

1.- Introducción

- ▶ Regresión lineal ¿Qué es y para qué sirve?
- ▶ Correlación v/s causalidad.
- ▶ Repaso ecuación de regresión lineal.
- ▶ Repaso betas y causalidad.
- ▶ Interpretación Regresión lineal con R.
- ▶ Evaluación de supuestos.

2.- Práctica con R y Rstudio cloud

- ▶ Realizar análisis de regresión lineal.
- ▶ Realizar gráficas avanzadas con ggplot2.
- ▶ Elaborar un reporte dinámico en formato pdf.

INTRODUCCIÓN REGRESIÓN LINEAL

Herramienta estadística que permite determinar si existe una relación (asociación) entre una variable predictora (independiente) y la variable respuesta (dependiente).

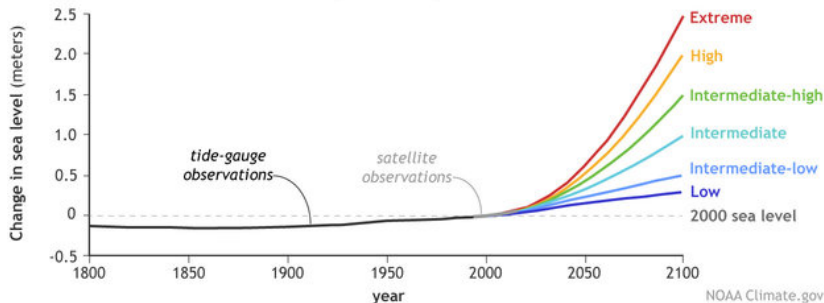
Nivel del mar en función del tiempo. Fuente: epa.gov



REGRESIÓN LINEAL: PREDICCIÓN.

La ecuación de la regresión permite, bajo ciertos supuestos, predecir el valor de una variable respuesta “y” a partir de una o más variables predictoras “x”.

Possible future sea levels for different greenhouse gas pathways

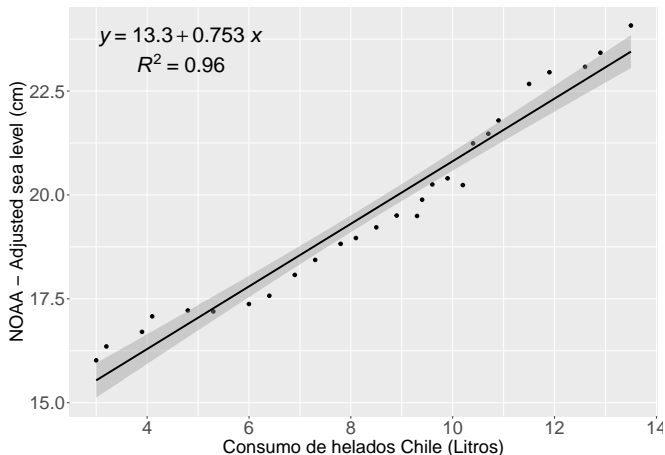


NOAA Climate.gov
Adapted from Sweet et al., 2017

CORRELACIÓN NO IMPLICA CAUSALIDAD

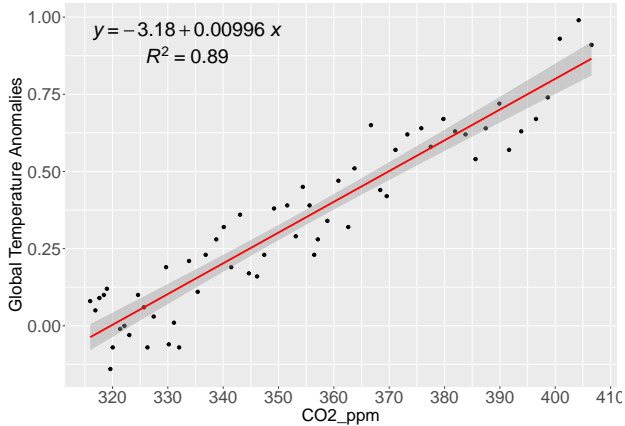
¿Si dejamos de tomar helados disminuirá el nivel del mar?

¿Qué factor “z” puede explicar la correlación entre consumo de helados y nivel del mar?



REGRESIÓN LINEAL CAUSALIDAD

Calentamiento global Fuente: climate.gov



REGRESIÓN LINEAL: BETAS

Betas miden la influencia del intercepto y la pendiente sobre la variable Y .

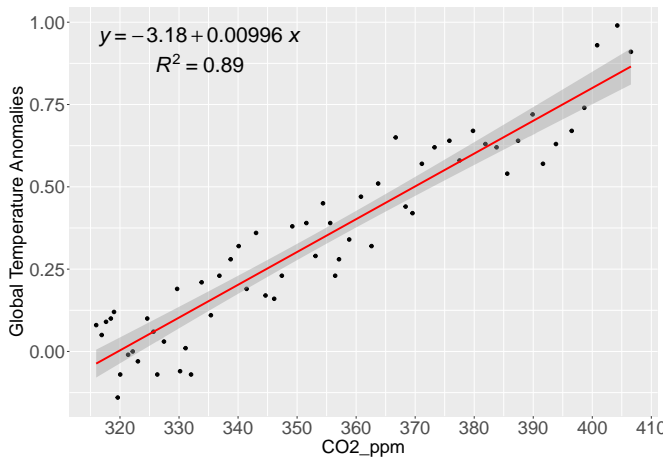
$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

β_0 = Intercepto = valor que toma “y” cuando $x = 0$.

β_1 = Pendiente = Cambio promedio de “y” cuando “x” cambia en una unidad.

LINEA DE REGRESIÓN.

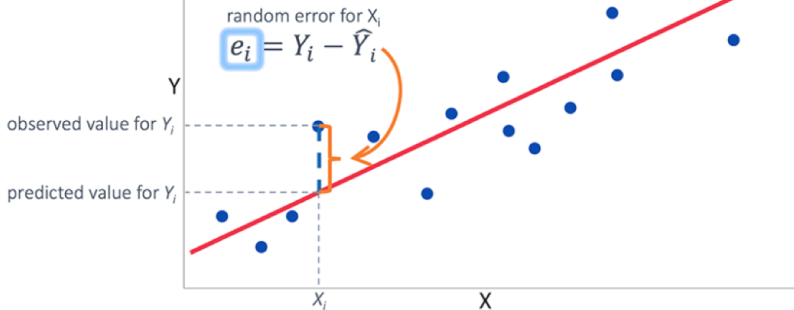
Línea de regresión: Corresponde a los valores “ajustados” o estimados de “y” en función de “x”. Se calcula con los estimadores de *mínimos cuadrados* de β_0 y β_1



RESIDUOS Y MÉTODOS DE MÍNIMOS CUADRADOS

Method of Least Squares

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$



COCIENTE DE DETERMINACIÓN

R^2 mide la proporción de la variación muestral de “y” que es explicada por x (varía entre 0-1). Se calcula como el cuadrado del coeficiente de correlación de pearson.

R^2_{ajust} nos dice qué porcentaje de la variación de la variable dependiente es explicado por la o las variables independientes de manera conjunta.

$$R^2_{ajust} = 1 - (1 - R^2) \frac{n-1}{n-p-1}$$

donde:

n = tamaño de la muestra

p = cantidad de variables predictoras en el modelo

PRUEBAS DE HIPÓTESIS

Prueba de hipótesis del coeficiente de regresión y el intercepto Tipo de prueba: Prueba de t – student

La hipótesis nula en ambos casos es que los coeficiente (β_0) y (β_1) son iguales a 0, es decir sin asociación entre las variables.

$$H_0 : \beta_0 = 0 \text{ y } H_0 : \beta_1 = 0$$

Prueba de hipótesis del modelo completo Tipo de prueba: Prueba de F.

La hipótesis nula es que los coeficientes son iguales a 0.

$$H_0 : \beta_j = 0 ; j = 1, 2, \dots, k$$

REGRESIÓN LINEAL CON R: COEFICIENTES

```
reg <- lm(`Global Temperature Anomalies` ~ CO2_ppm,  
          data = Global_warming)  
# summary(reg)
```

Coeficientes

	Estimate	Std. Error	t value	Pr(>
Intercepto	-3.18	0.1629	-19.54	<2e-16 ***
CO2	0.0099	0.0004	21.67	<2e-16 ***

REGRESIÓN LINEAL CON R: PRUEBA DE F

Anova de la regresión.

```
anova(reg) %>% kable()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
CO2_ppm	1	4.1525733	4.1525733	469.5292	0
Residuals	57	0.5041149	0.0088441	NA	NA

EXTRAER INFORMACIÓN DE LA REGRESIÓN LINEAL

```
summary(reg$residuals)
```

```
##           Min.        1st Qu.         Median         Mean         3rd Qu.
## -0.1931891 -0.0752495  0.0001668  0.0000000  0.0792104
```

```
summary(reg)$sigma
```

```
## [1] 0.09404319
```

```
summary(reg)$r.squared
```

```
## [1] 0.8917439
```

```
summary(reg)$adj.r.squared
```

```
## [1] 0.8898447
```

PREDICCIÓN LINEAL DEL NIVEL DEL MAR

Predicción de la anomalía próximos años

```
predict.lm(reg, newdata=data.frame(CO2_ppm=c(410,420,430)),  
           interval="confidence")
```

	fit	lwr	upr
## 1	0.8997967	0.8422796	0.9573139
## 2	0.9994129	0.9334513	1.0653745
## 3	1.0990291	1.0244426	1.1736155

PREDICCIÓN LINEAL FUERA DEL RANGO OBSERVADO

Predicción de aumento de temperatura con 10 veces más CO2 que ahora.

¿Por qué esta predicción es inadecuada?

```
predict.lm(reg, newdata=data.frame(CO2_ppm=c(4100,4200,4300),  
                                     interval="prediction")
```

##		fit	lwr	upr
## 1	37.65816	34.20395	41.11237	
## 2	38.65432	35.10819	42.20046	
## 3	39.65049	36.01242	43.28855	

SUPUESTOS DE LA REGRESIÓN LINEAL SIMPLE

- ▶ ¿Cuales son los supuestos?

Independencia.

Linealidad entre variable independiente y dependiente.

Homocedasticidad.

Normalidad.

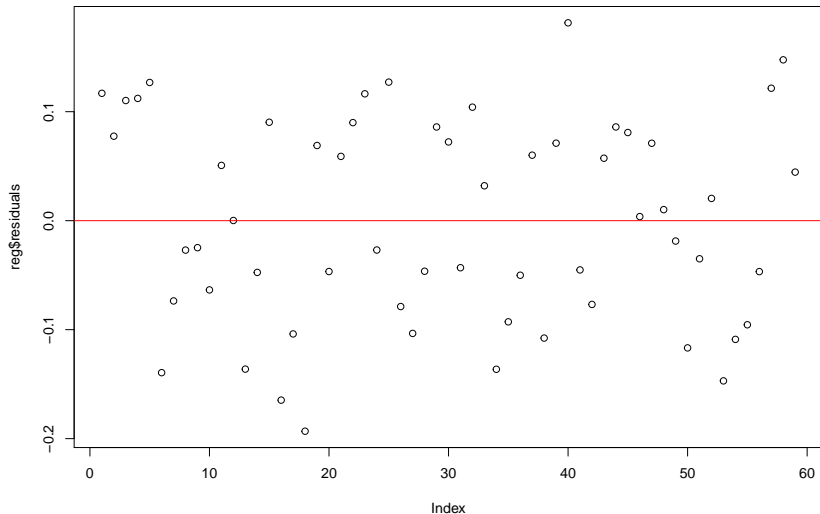
- ▶ ¿Por qué son importantes?

Para validar el resultado obtenido.

En caso de incumplimiento se pueden transformar datos o elaborar otros modelos (Regresión logística).

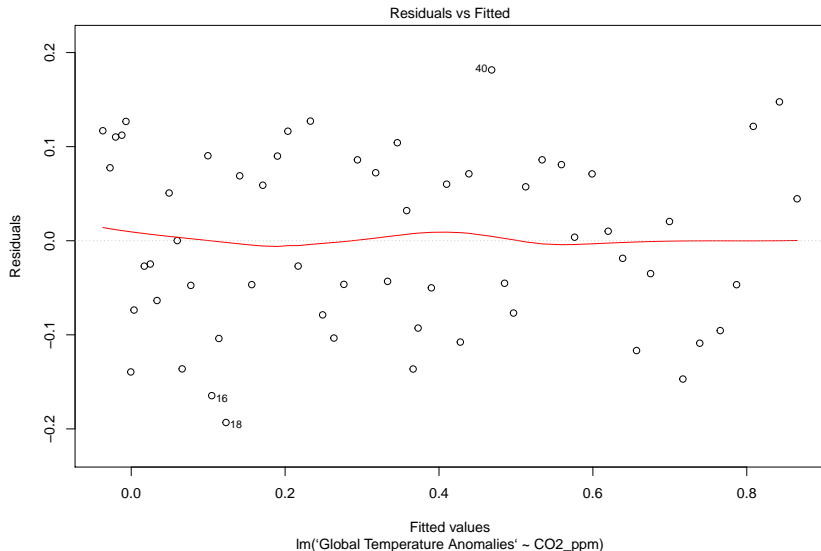
INDEPENDENCIA: MÉTODO GRÁFICO

```
plot(reg$residuals)  
abline(h=0, col="red")
```



LINEALIDAD: MÉTODO GRÁFICO

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



INDEPENDENCIA: DURBIN WATSON

H_0 : No existe autocorrelación entre los datos (lo que deseamos).

H_0 : Existe autocorrelación entre los datos.

Dado que $p < 0,05$ se rechaza independencia.

```
# durbin watson test  
durbinWatsonTest(reg)
```

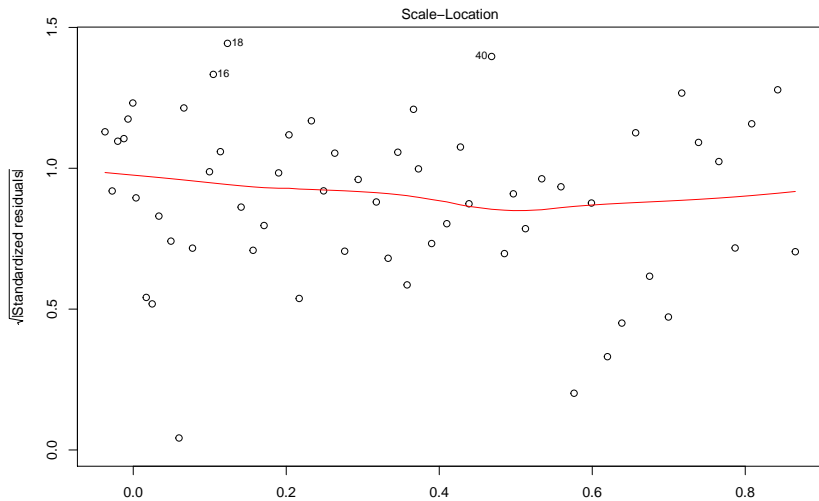
```
## lag Autocorrelation D-W Statistic p-value  
## 1 0.2338696 1.501215 0.048  
## Alternative hypothesis: rho != 0
```

HOMOGENEIDAD DE VARIANZAS: MÉTODO GRÁFICO

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_A: \sigma_1^2 \neq \sigma_2^2$$

```
plot(reg, which=3)
```



HOMOGENEIDAD DE VARIANZAS: PRUEBAS ESTADÍSTICAS

$p > .05$, No tenemos evidencias para rechazar que nuestros datos son homocedasticos.

```
ncvTest(reg) # library(car)
```

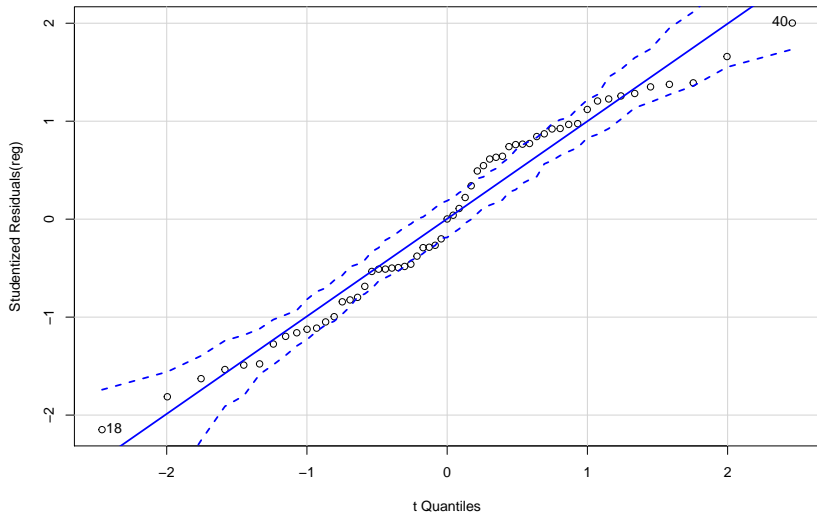
```
## Non-constant Variance Score Test  
## Variance formula: ~ fitted.values  
## Chisquare = 0.1563339, Df = 1, p = 0.69255
```

```
bptest(reg) # library(lmtest)
```

```
##  
## studentized Breusch-Pagan test  
##  
## data: reg  
## BP = 0.34329, df = 1, p-value = 0.5579
```

NORMALIDAD: GRÁFICO DE CUANTILES

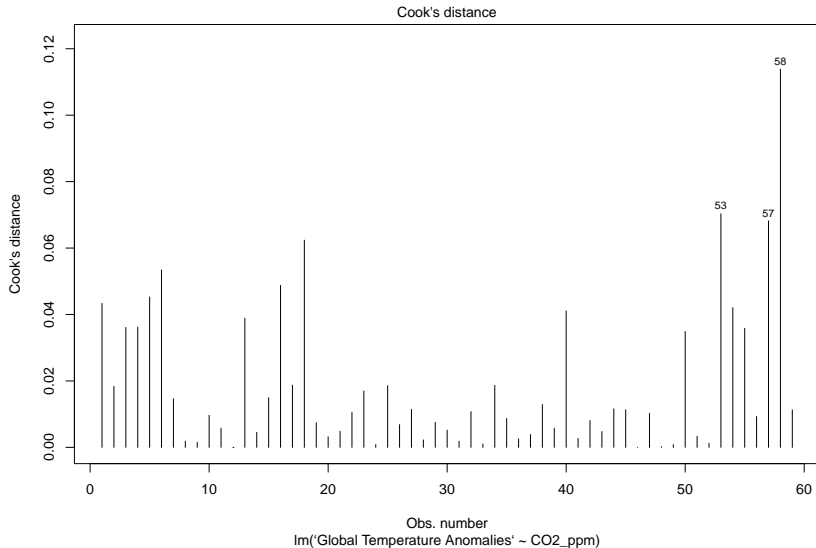
```
qqPlot(reg) # library(car)
```



```
## [1] 18 40
```

VALORES ATÍPICOS

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



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**
- ▶ **Realizar análisis de regresión lineal simple**
- ▶ **Interpretar coeficientes y realizar predicciones**
- ▶ **Evaluar supuestos de los análisis de regresión**