

Clase 13 Regresión lineal simple

Diplomado en Análisis de datos con R para la Acuicultura.

Dr. José Gallardo Matus

Pontificia Universidad Católica de Valparaíso

17 May 2022

PLAN DE LA CLASE

1.- Introducción

- ▶ ¿Qué es y para qué sirve una Regresión lineal?
- ▶ 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).

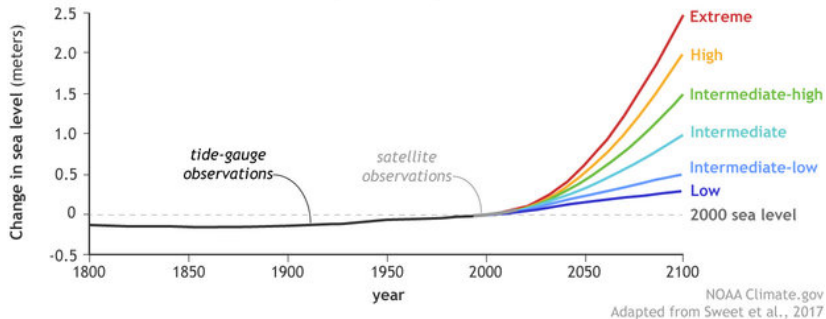
Nivel del mar en función del tiempo. Fuente: epa.gov



REGRESIÓN LINEAL: PREDICCIÓN

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

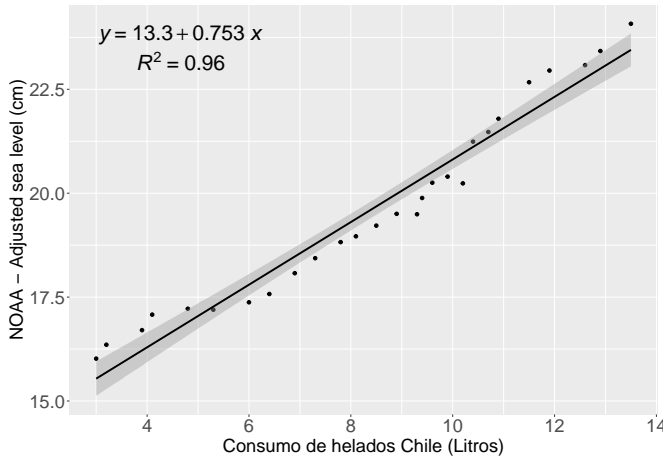
Possible future sea levels for different greenhouse gas pathways



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: BETAS

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

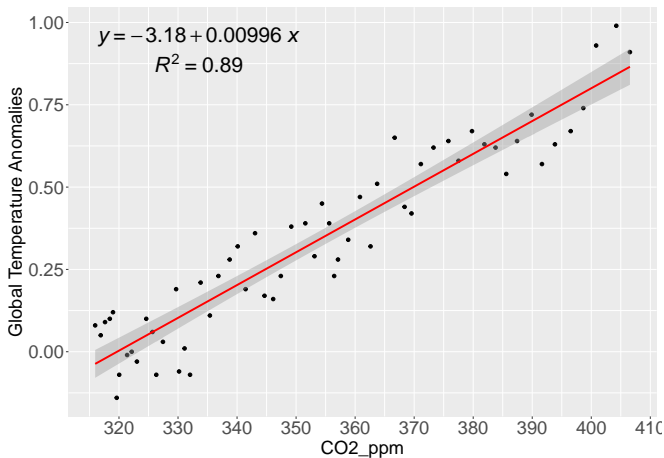
$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

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

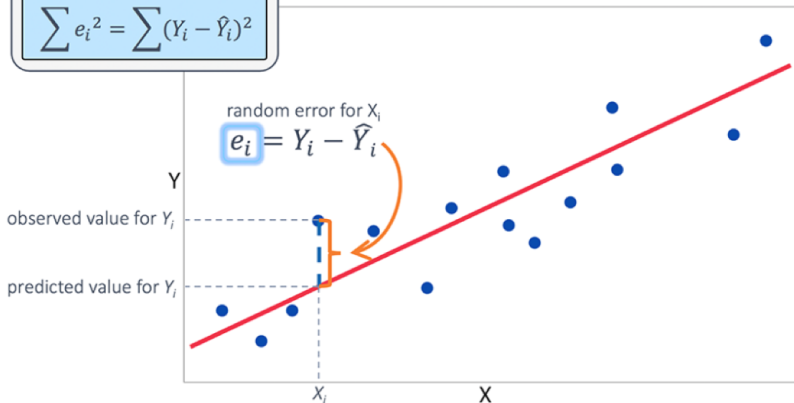
Línea de regresión: Corresponde a los valores “ajustados” o estimados de “y” en función de “x”. Se calcula con los estimadores de *mínimos cuadrados* de β_0 y β_1 .



RESIDUOS Y MÉTODOS DE MÍNIMOS CUADRADOS

Method of Least Squares

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



COEFICIENTE 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)
```

Call:

```
lm(formula = `Global Temperature Anomalies` ~ CO2_ppm, data = Global_warming)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.193189	-0.075250	0.000167	0.079210	0.181541

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.1844662	0.1629649	-19.54	<2e-16 ***
CO2_ppm	0.0099616	0.0004597	21.67	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09404 on 57 degrees of freedom

Multiple R-squared: 0.8917, Adjusted R-squared: 0.8898

F-statistic: 469.5 on 1 and 57 DF, p-value: < 2.2e-16

REGRESIÓN LINEAL CON R: PRUEBA DE F

Anova de la regresión.

```
anova(reg)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: Global Temperature Anomalies
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## CO2_ppm     1 4.1526   4.1526  469.53 < 2.2e-16 ***
```

```
## Residuals  57 0.5041   0.0088
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

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
```

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.

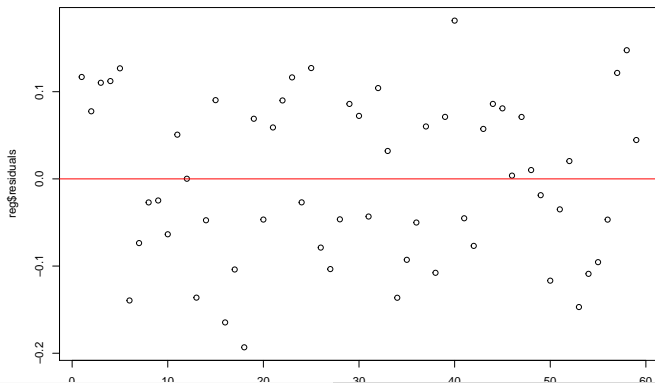
 - En caso de incumplimiento se pueden transformar datos o elaborar otros modelos (Regresión logística).

INDEPENDENCIA: MÉTODO GRÁFICO

H_0 : Los residuos son independientes entre sí.

H_A : Los residuos no son independientes entre sí (existe autocorrelación).

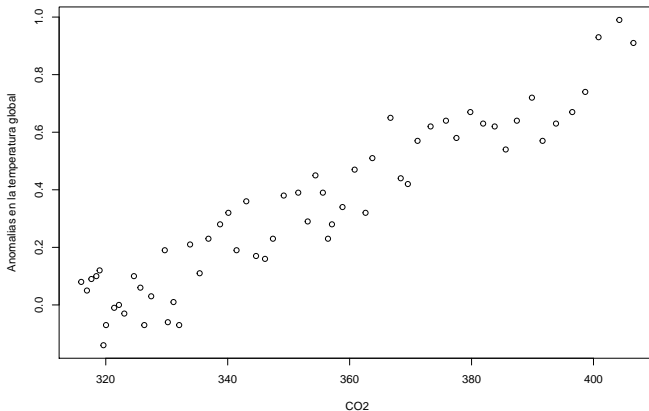
```
plot(reg$residuals)  
abline(h=0, col="red")
```



LINEALIDAD: MÉTODO GRÁFICO

H_0 : Hay relación lineal entre la variable regresora y la variable predictora.

H_A : No hay relación lineal entre la variable regresora y la variable predictora.

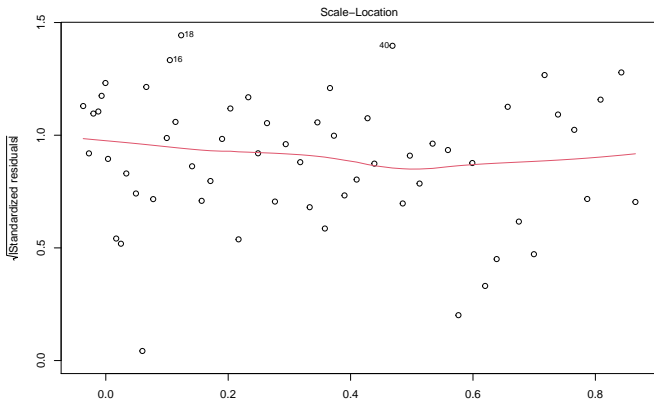


HOMOGENEIDAD DE VARIANZAS: MÉTODO GRÁFICO

H_0 : La varianza de los residuos es constante.

H_A : La varianza de los residuos no es constante.

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

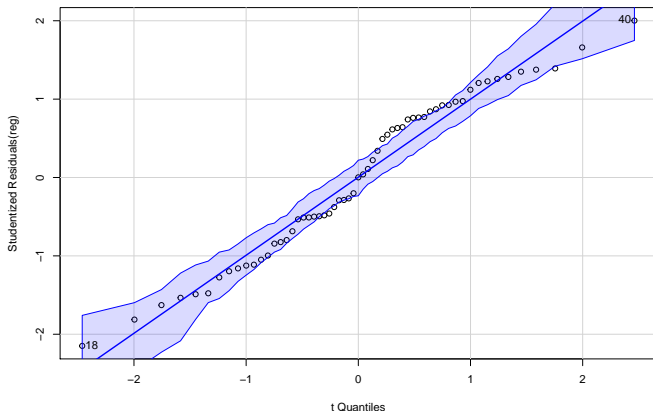


NORMALIDAD: GRÁFICO DE CUANTILES

H_0 : Los residuos tienen distribución normal.

H_A : Los residuos no tienen distribución normal.

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

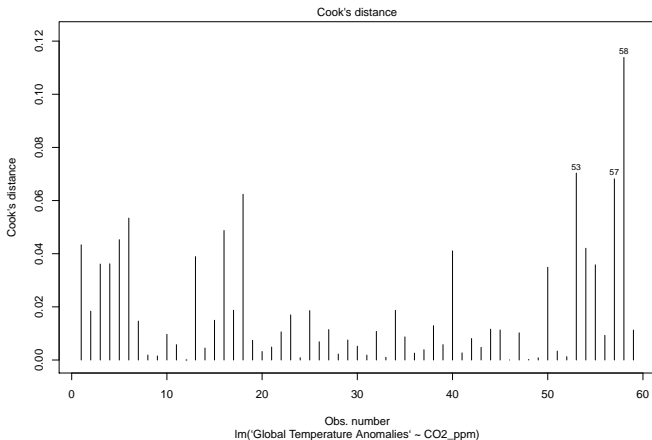


```
## [1] 18 40
```

VALORES ATÍPICOS

Una observación se puede considerar influyente (valor atípico) si tiene un valor de distancia de Cook mayor a 1.

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



PRÁCTICA ANÁLISIS DE DATOS

- ▶ Guía de trabajo práctico disponible en [Rstudio.cloud](https://rstudio.cloud).

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.