

Clase 13 Regresión lineal simple

**DBT 845 - Investigación reproducible y análisis de datos
biotecnológicos con R.**

Dr. José Gallardo Matus

Pontificia Universidad Católica de Valparaíso

23 May 2022

PLAN DE LA CLASE

1.- Introducción

- ▶ ¿Qué son los modelos lineales?
- ▶ ¿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 html.

MODELOS LINEALES

Los modelos lineales se usan para explicar, modelar o predecir la relación lineal de una variable respuesta Y con una o más P variables predictoras X_1, X_2, \dots, X_P .

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_P X_P + \epsilon_i$$

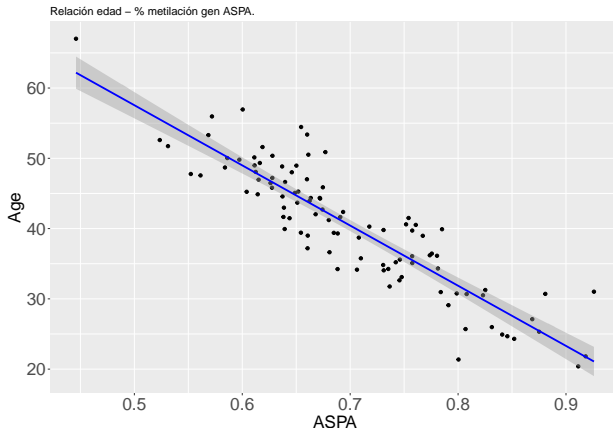
Si $p=1$, regresión lineal simple.

Si $= >1$ regresión lineal múltiple.

Si $p > 1$ y existe una variable categórica, se llama anova.

REGRESIÓN LINEAL SIMPLE

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



Fuente: Adaptado de Huang et al. 2015

REGRESIÓN LINEAL: PREDICCIÓN

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

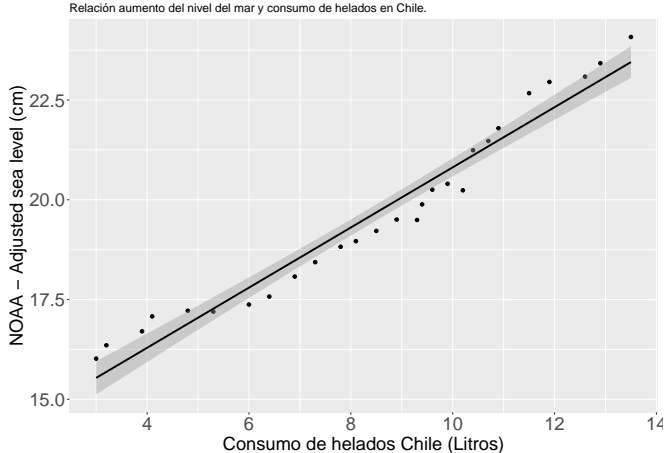
$$Y = a + \beta_1 X_1$$

Predicción de edad cuando la metilación es 0,5%, 1,0%, 1,5%.

```
reg <- lm(Age ~ ASPA, data = age.aspa)
predict.lm(reg,
            newdata=data.frame(ASPA=c(0.5, 1, 1.5)),
            interval="confidence")
```

	fit	lwr	upr
## 1	57.55590	55.68747	59.42433
## 2	14.74239	12.06592	17.41886
## 3	-28.07112	-34.92509	-21.21714

CORRELACIÓN NO IMPLICA CAUSALIDAD



¿Si dejamos de tomar helados disminuirá el aumento nivel del mar?
¿Qué factor “z” puede explicar la correlación entre consumo de helados y nivel del mar?

¿Cómo probar que existe una relación causal entre X e Y?

1. Temporalidad: La causa X debe preceder al efecto Y.
2. Dirección: La relación va desde la causa X al efecto Y.
3. Asociación (regresión): Debe ser distinta de cero.

La asociación es lo único que puedo probar con un análisis de regresión.

ECUACIÓN DE REGRESIÓN LINEAL: BETAS

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

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

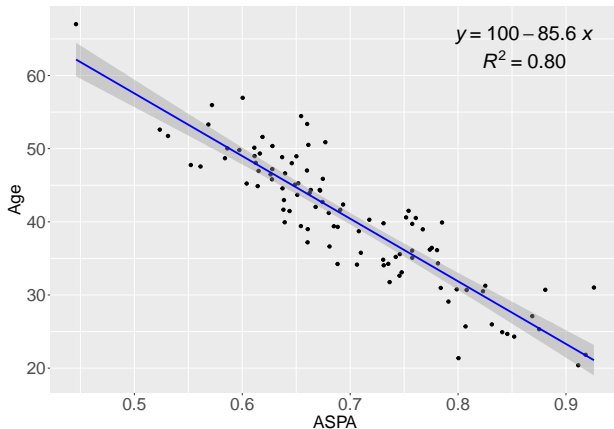
β_0 = Intercepto = valor que toma “y” cuando $x = 0$.

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

ϵ = mide la variabilidad de la variable respuesta que no es explicada por la recta de regresión.

LINEA DE REGRESIÓN

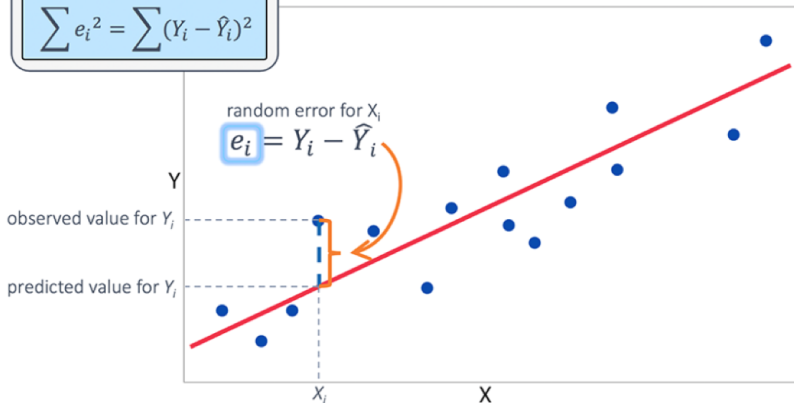
Línea de regresión: Corresponde a los valores “ajustados” o estimados de “y” en función de “x”. Se calcula con los estimadores de *mínimos cuadrados* de β_0 y β_1 .



RESIDUOS Y MÉTODOS DE MÍNIMOS CUADRADOS

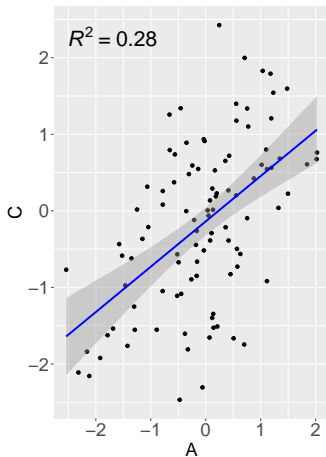
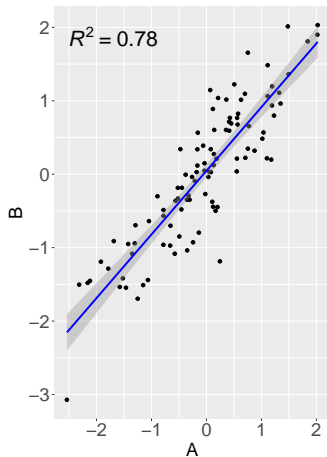
Method of Least Squares

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



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.



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.

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

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

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

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

Un Beta significativo indica que X esta correlacionado con Y, pero no necesariamente es un indicador de causalidad.

REGRESIÓN LINEAL CON R: COEFICIENTES

```
reg <- lm(Age~ ASPA, < data = age.aspa) summary(reg)
```

Call:

```
lm(formula = Age ~ ASPA, data = age.aspa)
```

Residuals:

Min	1Q	Median	3Q	Max
-10.4653	-3.1157	-0.0222	2.0904	10.1301

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	100.369	3.022	33.22	<2e-16 ***
ASPA	-85.627	4.284	-19.99	<2e-16 ***

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

Residual standard error: 3.978 on 98 degrees of freedom

Multiple R-squared: 0.803, Adjusted R-squared: 0.801

F-statistic: 399.5 on 1 and 98 DF, p-value: < 2.2e-16

REGRESIÓN LINEAL CON R: PRUEBA DE F

Anova de la regresión.

```
anova(reg)
```

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

```
##
```

```
## Response: Age
```

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

```
## ASPA      1 6319.8   6319.8   399.46 < 2.2e-16 ***
```

```
## Residuals 98 1550.4     15.8
```

```
## ---
```

```
## 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.      M  
## -10.46534  -3.11570  -0.02219   0.00000   2.09042  10.13
```

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

```
## [1] 3.977514
```

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

```
## [1] 0.8030012
```

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

```
## [1] 0.800991
```

PREDICCIÓN LINEAL DE LA EDAD

Predicción de la Edad con 0,25 - 0,50 y 0,75% de metilación del gen ASPA

```
predict.lm(reg,  
            newdata=data.frame(ASPA=c(0.25,0.50,0.75)),  
            interval="confidence")
```

	fit	lwr	upr
## 1	78.96265	75.06294	82.86236
## 2	57.55590	55.68747	59.42433
## 3	36.14914	35.24935	37.04893

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.

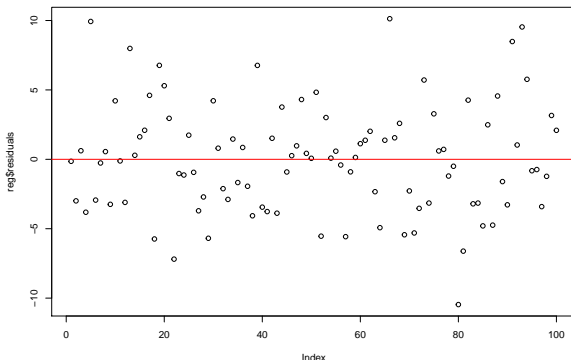
 - En caso de incumplimiento se pueden transformar datos o elaborar otros modelos (Regresión logística).

INDEPENDENCIA: MÉTODO GRÁFICO

H_0 : Los residuos son independientes entre sí.

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

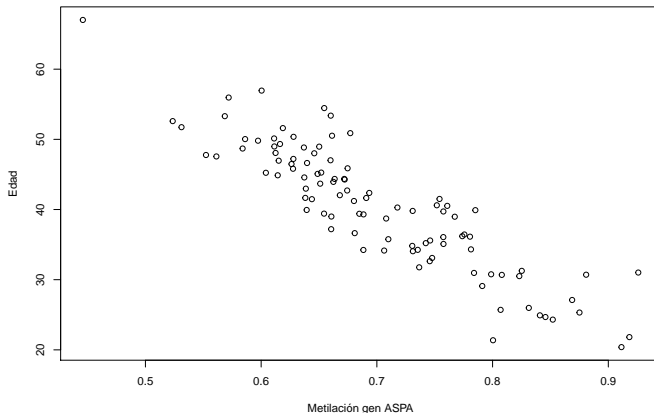
```
plot(reg$residuals)  
abline(h=0, col="red")
```



LINEALIDAD: MÉTODO GRÁFICO

H₀: 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.

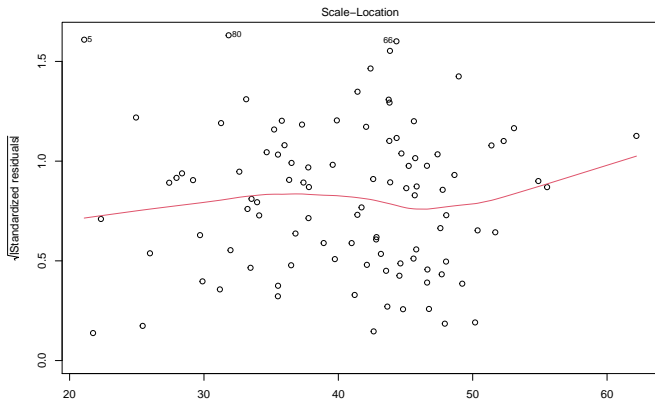


HOMOGENEIDAD DE VARIANZAS: MÉTODO GRÁFICO

H_0 : La varianza de los residuos es constante.

H_A : La varianza de los residuos no es constante.

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

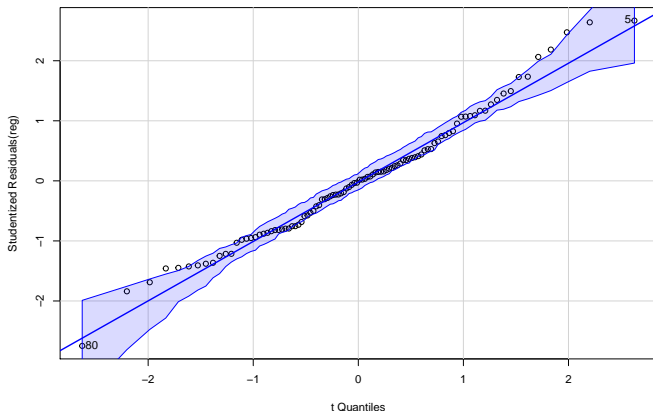


NORMALIDAD: GRÁFICO DE CUANTILES

H_0 : Los residuos tienen distribución normal.

H_A : Los residuos no tienen distribución normal.

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

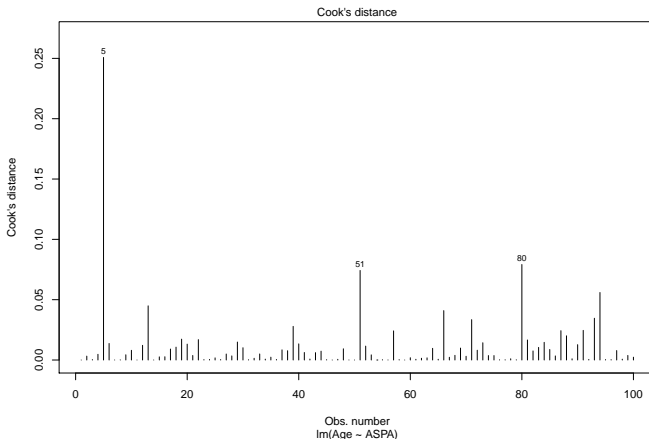


```
## [1] 5 80
```

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.