Clase 4. Regresiones (en Rmarkdown)

Demian Zayat

6/6/2019

Trabajaremos nosotros en grupos de no más de tres personas, con regresiones, la proxima es Geo y la otra ya la entrega del examen. Vamos a trabajar en Rmarkdown o Rnotebook

Lo que está entre `---` son *chunks*. Lo que viene escrito es un ejemplo, hay que borrar todo antes de empezar.

Desde el menu "insert" se puede insertar un chunk. Puede ser un chunk de Python, R, etc. Hay que tener instalado todo lo necesario para correr el chunk. Tambien con `control + Alt + I`. Hay cosas en las que R es mejor (gráfico por ejemplo) y otras en lo que Python es mejor (scrapping por ejemplo).

En general el primer chunk está dedicada a la carga de librerías. Lo podemos denominar `setup`.

Markdown es un texto al que le agregamos código y vemos el resultado del código.

La Regresión es un modelo que tiene la forma matemática `y = a + bx`. `y` es el resultado o variable dependiente. Depende del resto de la ecuación. `a` es la ordenada en origen o intercepción. `b` es la pendiente o coeficiente y `x` es la variable independiente, o input o variante explicativa.

Por ejemplo, puedo tener la regresión entre años de estudio e ingresos. Años de estudio en el eje x e ingresos en el eje y. Es un plot de puntos. La regresión es la línea que mejor junta los promedios, elevado al cuadrado para perder el signo.

El intercepto es donde arranca la curva. Cuanto vale el valor de y cuando x es cero. Es teórico muchas veces. el Coeficiente es cuanto varía la variable dependeinte por una unidad más de la variable independiente.

Estoy buscando una regresión lineal. Si la nube no está correlacionada, va a intentar lo mismo. No será adecuado. Se puede buscar otro tipo de modelo.

Pueden ser variables numéricas. También categóricas, y sigue siendo lineal. La variable explicativa puede ser categórica (sexo, por ejemplo). Si la variable dependiente no fuera numérica (muerto/vivo), es una regresión logística o multinominal. No buscas linealidad.

Puede ser multilineal $x = a + b1x1 + b2x2$. Hay correlaciones positivas (linea ascendente) y negativas (linea descendente).

En la multilineal, beta (b) es la modificación de y con cada aumento de una unidad de x, manteniendo constante el resto. Quizas hay otra cosa que explica la relación que se ve. La otra variable afecta la ecuación dejando el resto igual (variación neta).

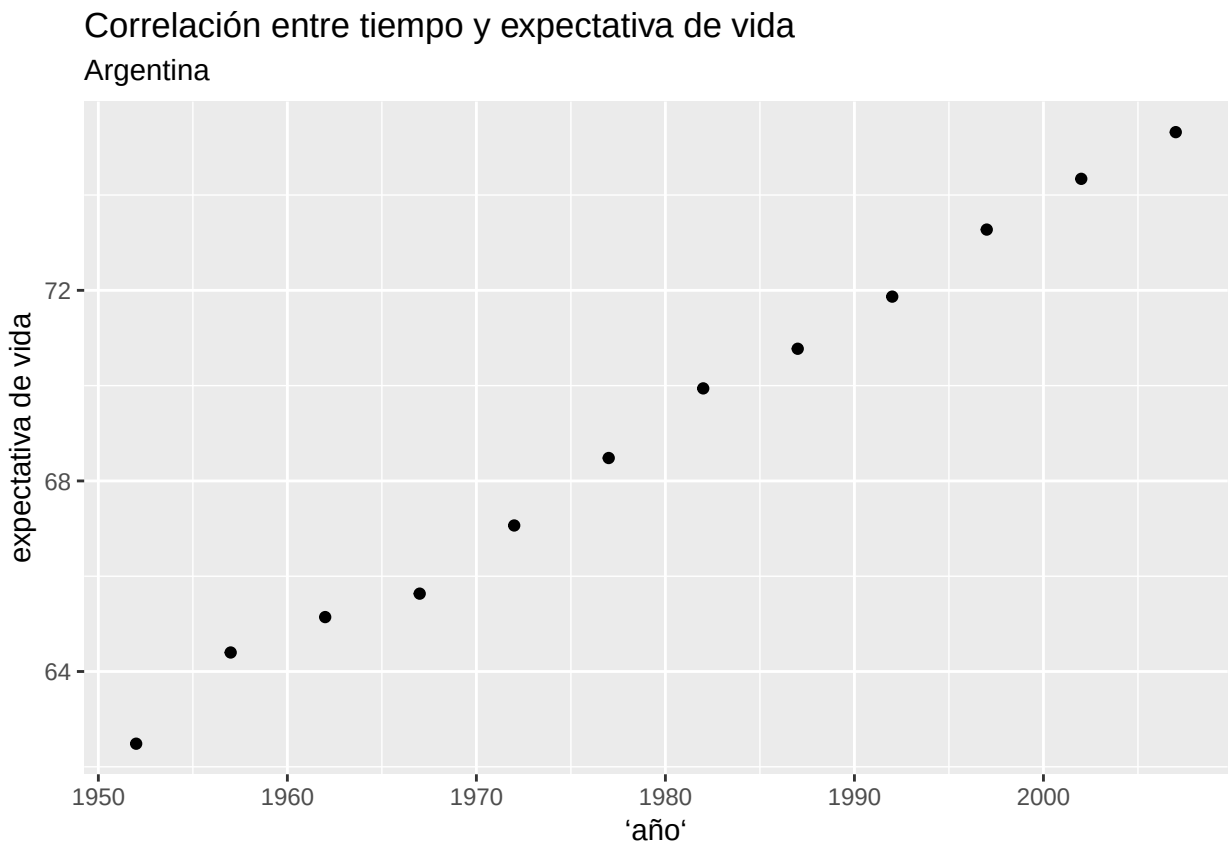
No es causalidad. Es asociación! Ningún modelo explica causalidad, todas muestran asociación o correlación.

La regresión puede tener dos fines: predictivos o explicativos. Los predictivos nos permiten saber cuánto ganará una persona según sus años de estudios. O explicativos, donde interesa cómo se construye la fórmula: importa el género o la raza en el resultado del juicio? Hay modelos mejores que la regresión lineal en cuanto predicción. No hay muchos mejores en términos explicativos. Mientras más predictivo, menos intelegible queda. Interpretar eso es muy dificil.

Como saber si la variable independiente sea realmente una variable influyente? Si la probabilidad es `p < 0.05` es estadísticamente significativo. Si no, no se puede descartar que sea por azar. Una cosa es la fuerza, que está dado por la beta, y otra la significancia estadística de ese valor.

```
data_mundial <- read.csv("https://bitsandbricks.github.io/data/gapminder.csv", encoding = "UTF-8")
data_arg <- data_mundial %>%
  filter(pais == "Argentina")
```

```
ggplot(data = data_arg) +
    geom_point(aes(x = año, y = expVida)) +
    labs(title = "Correlación entre tiempo y expectativa de vida",
         subtitle = "Argentina",
         y = "expectativa de vida")
```



La correlación es lineal perfecta.

```
cor(data_arg$año, data_arg$expVida)
```

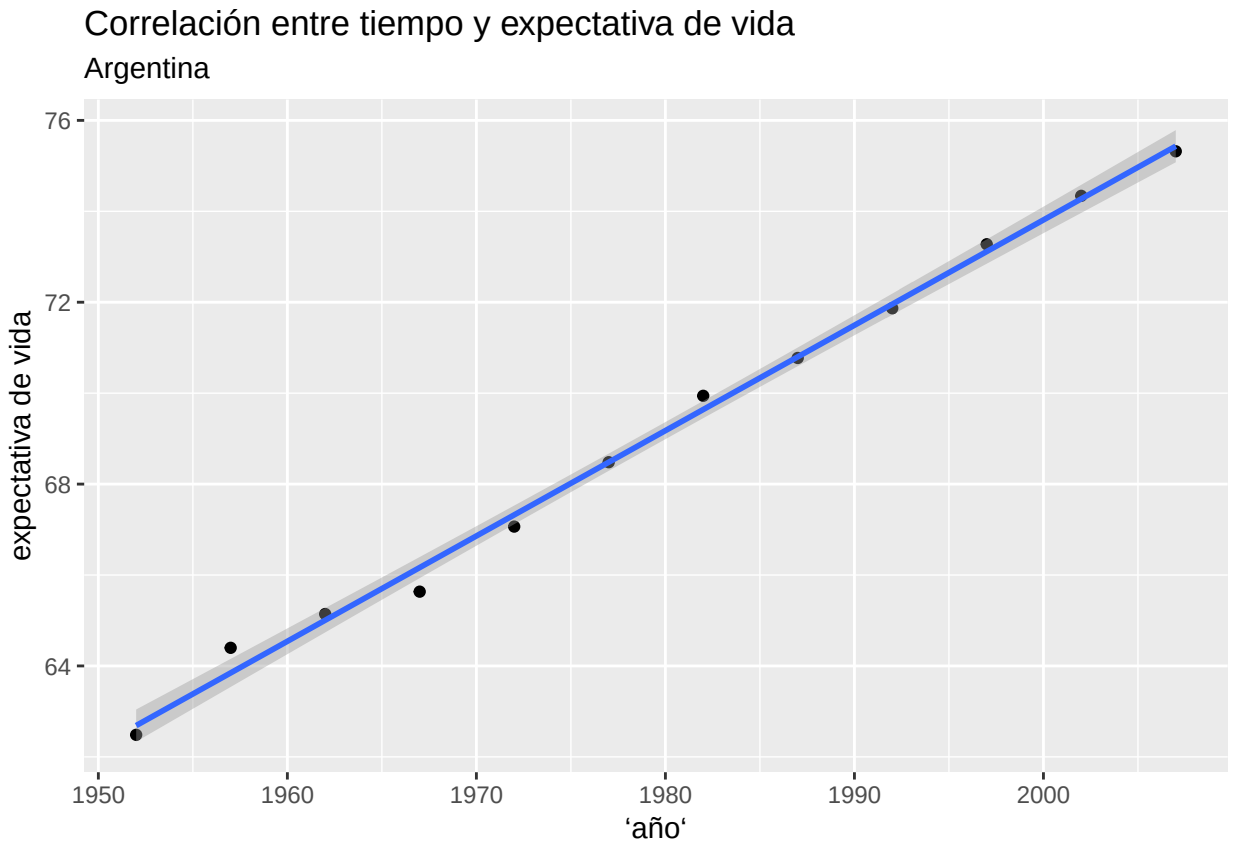
```
## [1] 0.9977816
```

```
modelo_exp <- lm(expVida ~ año, data = data_arg)
```

`lm` es lineal model, la función para las regresiones lineales. Primero variable dependiente, ~ según la variable independiente, y luego la data de dónde tomarlas. Es una única línea de código.

Por cada año calendario aumenta un 0,23 la expectativa de vida en argentina.

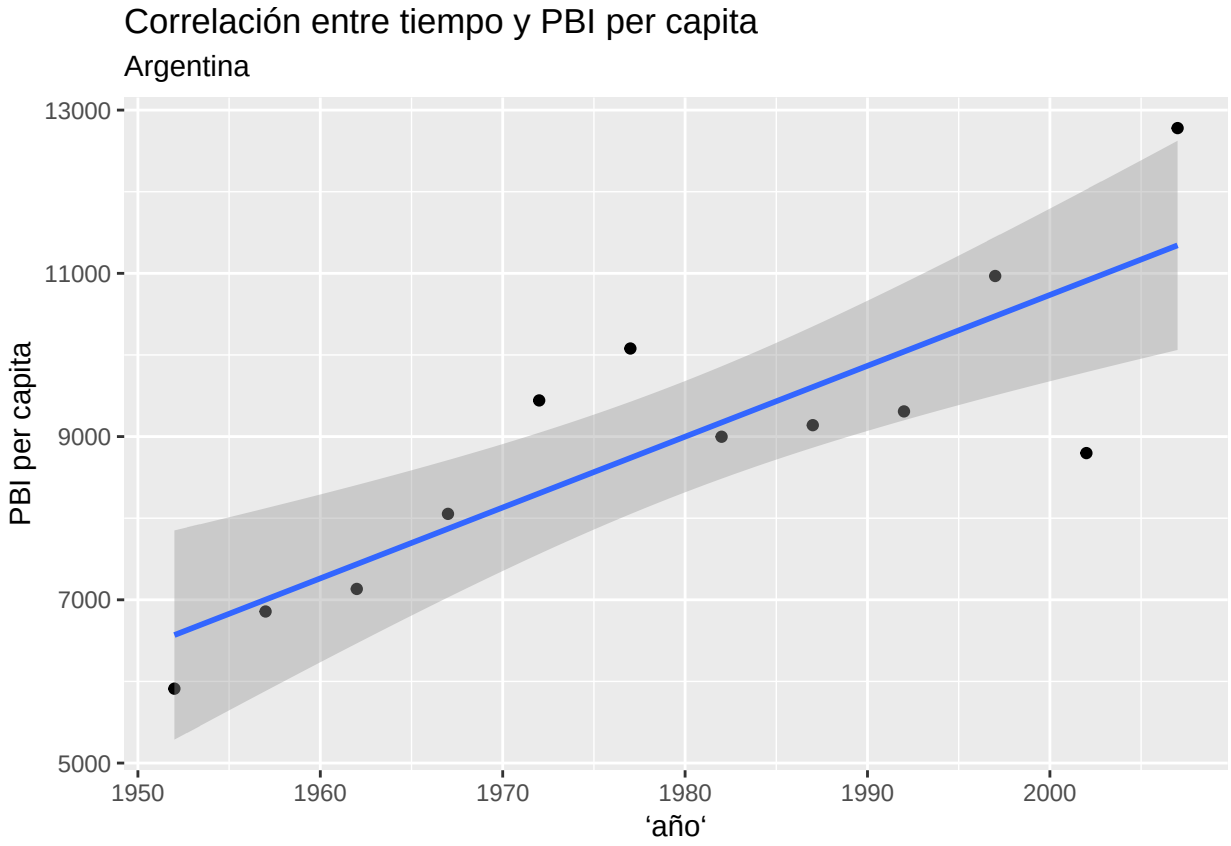
```
ggplot(data = data_arg) +
    geom_point(aes(x = año, y = expVida)) +
    labs(title = "Correlación entre tiempo y expectativa de vida",
         subtitle = "Argentina",
         y = "expectativa de vida")+
   geom_smooth(aes(x = año, y = expVida), method = "lm") #para graficar la regresion
```



Con el `geom_smooth` no es necesario antes calcular la regresión, la grafica directamente. El sombreado es el intervalo de confianza del 0.95 de probabilidad.

Como es una muestra y no el total, hay incerteza, que se representa en el intervalo de confianza. Se puede hacer con medias, o con regresiones. y entonces el Beta tendrá un intervalo de confianza. Un beta de cero significa que no hay asociación. No es significativo. En cambio, si el beta no es cero en el intervalo de confianza, algo modifica a la dependiente. Desvío estandar, como diferencia de cada valor a la media, sin signo.

```
ggplot(data = data_arg) +
    geom_point(aes(x = año, y = PBI_PC)) +
    labs(title = "Correlación entre tiempo y PBI per capita",
         subtitle = "Argentina",
         y = "PBI per capita")+
   geom_smooth(aes(x = año, y = PBI_PC), method = "lm") #para graficar la regresion
```


Correlación entre tiempo y PBI per capita
Argentina
PBI per capita
13000
11000
9000
7000
5000
1950
1960
1970
1980
1990
2000
‘año‘