Taller R - Sesión 7: Reportería con Rmarkdown

José Daniel Conejeros

Julio 2020

El **objetivo** de esta sesión es introducir a la reportería de análisis estadísticos con Rmarkdown.

## I. ¿Qué es R Markdown?

Es un espacio para elaborar documentos en PDF, HTML, MS-WORD (entre muchas otras) con una sintaxis de formato simple. Dado que R es un entorno de Desarrollo Integrado, R Markdown proporciona un marco unificado para la ciencia de datos combinando código, resultados y texto con el fin de crear documentos para que puedan ser reproducibles.

Los principales elementos son:

1. **Encabezado YAML: configuraciones preliminares al documento.** Va en el encabezado del Rmarkdown con especificaciones tipo:

* Título.
* Autor.
* Fecha.
* Formato de salida.
* Otros (infinitas configuraciones adicionales).

1. **Trozos/Bloques de códigos rodeados por: ```{r} ```**

Esto es un **chunk**:

vector <- c(1:10)   
vector

## [1] 1 2 3 4 5 6 7 8 9 10

Los argumentos nos permite especificar la salida de los **chunk** los más genéricos son los siguientes:

* inclue = FALSE. Los códigos y los resultados no van aparecer en el documento final.
* echo = FALSE. Los códigos no aparecen pero sí los resultados.
* message = FALSE. No aparecen mensajes.
* warning = FALSE. No aparecen mensajes tipo warnings.
* fig.cap="Título plot". Agrega un título a los resultados gráficos.

1. **Texto con formato simple.** Todo lo que no está como chunk o en el YAML es texto. A este texto yo le puedo agregar código en línea. Por ejemplo, puedo decir que la media del vector es: mean(vector) y arroja como resultado 5.5.

## II. Flujo en Rmarkdown

Al realizar click en la opción **Knit** (tejer), automáticamente se generará un documento que incluye tanto el contenido como la salida de cualquier fragmento de código en R. El flujo de trabajo en R es el siguiente:



Flujo en Rmarkdown

## III. Ejemplo: Datos PSU

### a. Código - Resultado

# Utilizamos este comando para desactivar notación científica   
options(scipen=999)  
  
# Abrimos nuestra base de datos  
library(readr)  
psu <- read\_csv("psu\_sample.csv")   
  
# Exploramos  
dim(psu)

## [1] 9623 13

colnames(psu)

## [1] "sexo" "estado\_civil" "grupo\_depend" "x\_nem" "leng"   
## [6] "mate" "cien" "beca" "edad" "jefe\_familia"  
## [11] "nem" "educpadre" "educmadre"

head(psu)

## # A tibble: 6 x 13  
## sexo estado\_civil grupo\_depend x\_nem leng mate cien beca edad  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Masc… Soltero Municipal (… 50 353 499 461 0 18  
## 2 Masc… Soltero Particular … 58 611 656 608 0 19  
## 3 Feme… Soltero Particular … 59 685 569 626 0 19  
## 4 Feme… Soltero Particular … 50 413 391 347 0 18  
## 5 Masc… Soltero Particular … 57 622 644 579 0 19  
## 6 Feme… Soltero Municipal (… 60 506 448 448 0 19  
## # … with 4 more variables: jefe\_familia <chr>, nem <dbl>, educpadre <dbl>,  
## # educmadre <dbl>

### b. Descriptivos

# Podemos utilizar el clásico  
summary(psu)

## sexo estado\_civil grupo\_depend x\_nem   
## Length:9623 Length:9623 Length:9623 Min. :41.00   
## Class :character Class :character Class :character 1st Qu.:54.00   
## Mode :character Mode :character Mode :character Median :57.00   
## Mean :57.41   
## 3rd Qu.:61.00   
## Max. :70.00   
## leng mate cien beca   
## Min. :197.0 Min. :179.0 Min. :216.0 Min. :0.0000   
## 1st Qu.:440.0 1st Qu.:459.0 1st Qu.:431.0 1st Qu.:0.0000   
## Median :516.0 Median :539.0 Median :499.0 Median :0.0000   
## Mean :516.1 Mean :534.2 Mean :501.7 Mean :0.1137   
## 3rd Qu.:589.0 3rd Qu.:609.0 3rd Qu.:575.0 3rd Qu.:0.0000   
## Max. :841.0 Max. :850.0 Max. :850.0 Max. :1.0000   
## edad jefe\_familia nem educpadre   
## Min. :17.00 Length:9623 Min. :4.100 Min. : 1.000   
## 1st Qu.:19.00 Class :character 1st Qu.:5.400 1st Qu.: 4.000   
## Median :19.00 Mode :character Median :5.700 Median : 5.000   
## Mean :18.84 Mean :5.741 Mean : 6.958   
## 3rd Qu.:19.00 3rd Qu.:6.100 3rd Qu.:11.000   
## Max. :22.00 Max. :7.000 Max. :14.000   
## educmadre   
## Min. : 1.000   
## 1st Qu.: 4.000   
## Median : 5.000   
## Mean : 6.691   
## 3rd Qu.:11.000   
## Max. :14.000

### c. Expresiones matemáticas y código en línea

En la tabla anterior podemos ver la estimación de dos estadísticos:

* **El promedio**. Por ejemplo el puntaje promedio de la prueba de matemáticas es de 534.2417126
* **La desviación estándar**. Por ejemplo la desviación estándar del puntaje de la prueba de matemáticas es de 110.3027465.

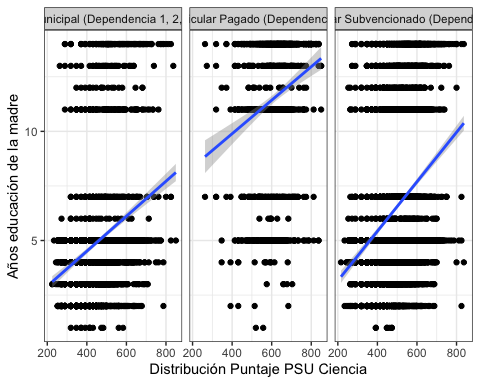
### d. Tablas

# Tabla con estadísticos descriptivos  
library(dplyr)  
tabla <- psu %>%  
 dplyr::summarize(n = n(),  
 mean = mean(mate, na.rm = T),  
 sd = sd(mate, na.rm = T),   
 min = min(mate, na.rm = T),  
 q25 = quantile(mate, probs = 0.25, na.rm = T),  
 median = median(mate, na.rm = T),  
 q75 = quantile(mate, probs = 0.75, na.rm = T),  
 max = max(mate, na.rm = T)  
 ) %>%  
 rename(Frecuencia=n,  
 Promedio=mean,  
 Std = sd,  
 Min = min,  
 Q25 = q25,  
 Q50 = median,  
 Q75 = q75,  
 Max = max  
 )   
  
# Podemos guardar como una data y transformar en tabla:  
library(knitr)  
kable(tabla)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Frecuencia | Promedio | Std | Min | Q25 | Q50 | Q75 | Max |
| 9623 | 534.2417 | 110.3027 | 179 | 459 | 539 | 609 | 850 |

### e. Gráficos

library(ggplot2)  
ggplot(psu, aes(x=cien, y=educmadre)) +   
 geom\_point()+  
 stat\_smooth(method="lm", se=T, formula=y ~ x) +   
 xlab("Distribución Puntaje PSU Ciencia") +   
 ylab("Años educación de la madre") +   
 theme\_bw() +   
 facet\_wrap(~grupo\_depend)



Mi primer gráfico en RMarkdown

### f. Ejemplo de un modelo de regresión

# Estimo los modelos  
m1 <- lm(mate ~ sexo, data=psu)  
m2 <- lm(mate ~ sexo + grupo\_depend, data=psu)  
m3 <- lm(mate ~ sexo + grupo\_depend + educpadre + educmadre, data=psu)  
m4 <- lm(mate ~ sexo + grupo\_depend + educpadre + educmadre + factor(beca), data=psu)  
  
# Genero una tabla con mis modelos  
library(texreg)  
texreg(l=list(m1, m2, m3, m4),   
caption="Modelos de regresión lineal", float.pos="h" , digits = 3,single.row = FALSE,  
custom.model.names = c("Modelo 1","Modelo 2","Modelo 3","Modelo 4"),  
caption.above = TRUE, include.ci = FALSE, stars = c(0.01, 0.05, 0.1), fsingle.row = T,  
custom.coef.names=c("Intercepto", "Masculino", "Part. Pagado", "Part. Subvencionado",   
 "Educación Padre", "Educación Madre","Becado"))

## Referencias

Podemos incluir links o accesos directos:

[Página oficial de Rmarkdown](http://rmarkdown.rstudio.com): Aquí puedes todo lo referido a esta herramienta para generar reportes. Hay galerías, formatos, documentos de replicación u otros. Solo tienes que explorar.

[R para Ciencia de Datos. Capítulo 5](https://r4ds.had.co.nz/r-markdown.html): En este capítulo se detallan los principales comandos y opciones para operar con las funciones de Rmarkdown.

[Video introductorio para construir presentaciones en Markdown](https://www.youtube.com/watch?v=i0TIYtqPMNQ&feature=youtu.be). Recomiendo revisar los últimos 20 minutos para automatizar reportes.