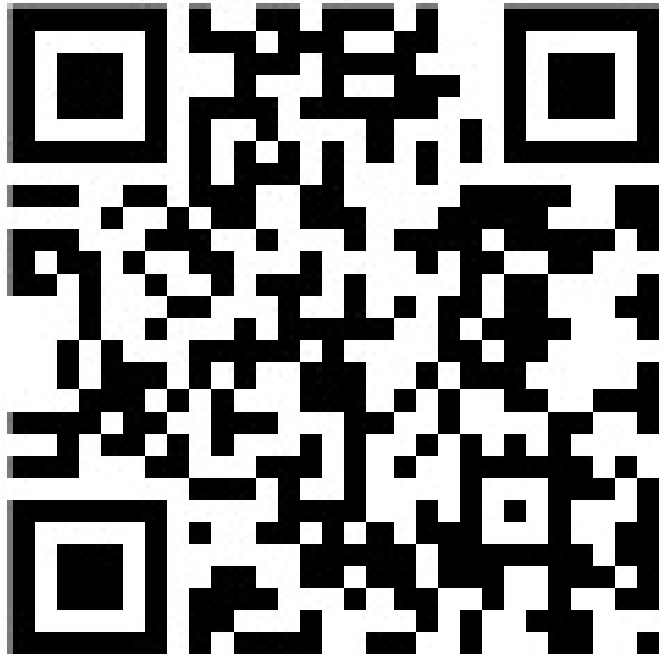
R Markdown, investigación reproducible y aplicaciones educativas

Lino AA Notarantonio (lino@tec.mx)

12/16/2019

Presentación

La pueden bajar en el repositorio de Github. Sigan `Jump to file` en la página.



Investigación reproducible: Antecedentes

Los métodos estadísticos juegan un papel fundamental en las ciencias empíricas con baja relaciòn señal-ruido.

En los últimos años se han presentados problemas de reproducibilidad de muchos resultados en ciencias sociales y ciencias de la salud. Entre otros, mencionamos

- Reproducibility Project in Psychology (Science, 2015). Una de las conclusiones es una evidencia débil para los resultados en los artículos originales, aún con datos proporcionados por los mismos autores.

Investigación reproducible: Antecedentes

- Social Sciences Replication Project (2018) De los 21 experimentos considerados (todos publicados en *Nature* o *Science*, en el periodo 2010-2015), sólo 13 pudieron ser replicados. Además, el tamaño del efecto en el estudio es aproximadamente la mitad de los que se publicó.

Problemas adicionales

- Limpieza de datos: como manejar valores ausentes; outliers, etc.
- Datos no públicos
- Dragado de datos (*data dredging*) (típico en Big Data)
- Datos perdidos: L. Weitzman

Investigación reproducible

Investigación que permite la presentación de los resultados con la manipulación necesaria y transparente de los datos correspondientes.

Beneficios

- Aplicaciones educativas: presentación del material y de la manipulación de los datos asociados de una manera coordinada y clara.
- Investigación: facilidad de réplica de resultados empíricos.

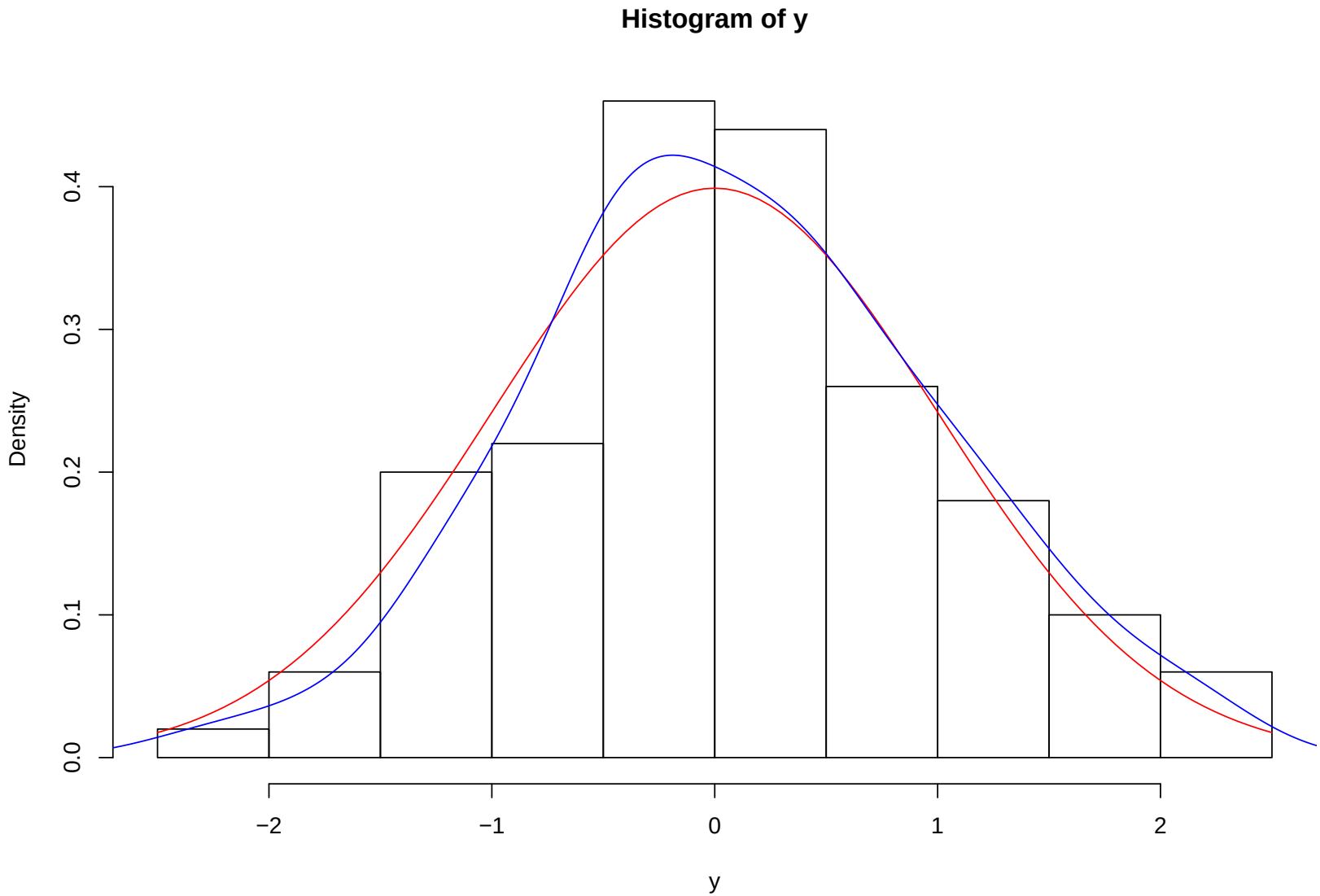
Ejemplo 1

Simulación de 100 muestras tomadas de una población normal estándar con histograma y la denstidad de probabilidad teórica correspondiente.

```
set.seed(123)
y <- rnorm(100)
hist(y, probability = TRUE)
curve(dnorm(x), col = "red", add = TRUE)
d <- density(y)
lines(d, col = "blue", add = TRUE)
```

Ejemplo 1

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): '
## graphical parameter
```



Ejemplo 2: Manipulación transparente de datos

- Los datos a continuación se bajaron del sitio NOAA (National Oceanic and Atmospheric Administration) de EEUU.
- Los datos abarcan observaciones de CO2 en el periodo marzo 1958 a marzo 2017.
- La columna (variable) *average* contiene las medias mensuales de CO2 (ppm, partes per millones).
- Datos ausentes para la variale *average* se denotan con -99.99; para la variable *no. days* con -1.

Ejemplo 2: Manipulación transparente de datos

Se usará la variable *average*. El valor -99.99 indica valores ausentes en el dataset.

Ejemplo 2: Manipulación transparente de datos

```
##      V1 V2         V3     V4     V5     V6 V7
## 1 1958  3 1958.208 315.71 315.71 314.62 -1
## 2 1958  4 1958.292 317.45 317.45 315.29 -1
## 3 1958  5 1958.375 317.50 317.50 314.71 -1
## 4 1958  6 1958.458 -99.99 317.10 314.85 -1
```

Ejemplo 2: Manipulación transparente de datos

Cambiamos el valor numérico -99.99 a *NA*, sin cambiar el dataset original:

```
##      V1 V2         V3     V4     V5     V6 V7
## 1 1958  3 1958.208 315.71 315.71 314.62 -1
## 2 1958  4 1958.292 317.45 317.45 315.29 -1
## 3 1958  5 1958.375 317.50 317.50 314.71 -1
## 4 1958  6 1958.458     NA 317.10 314.85 -1
```

Ejemplo 2: Manipulación transparente de datos

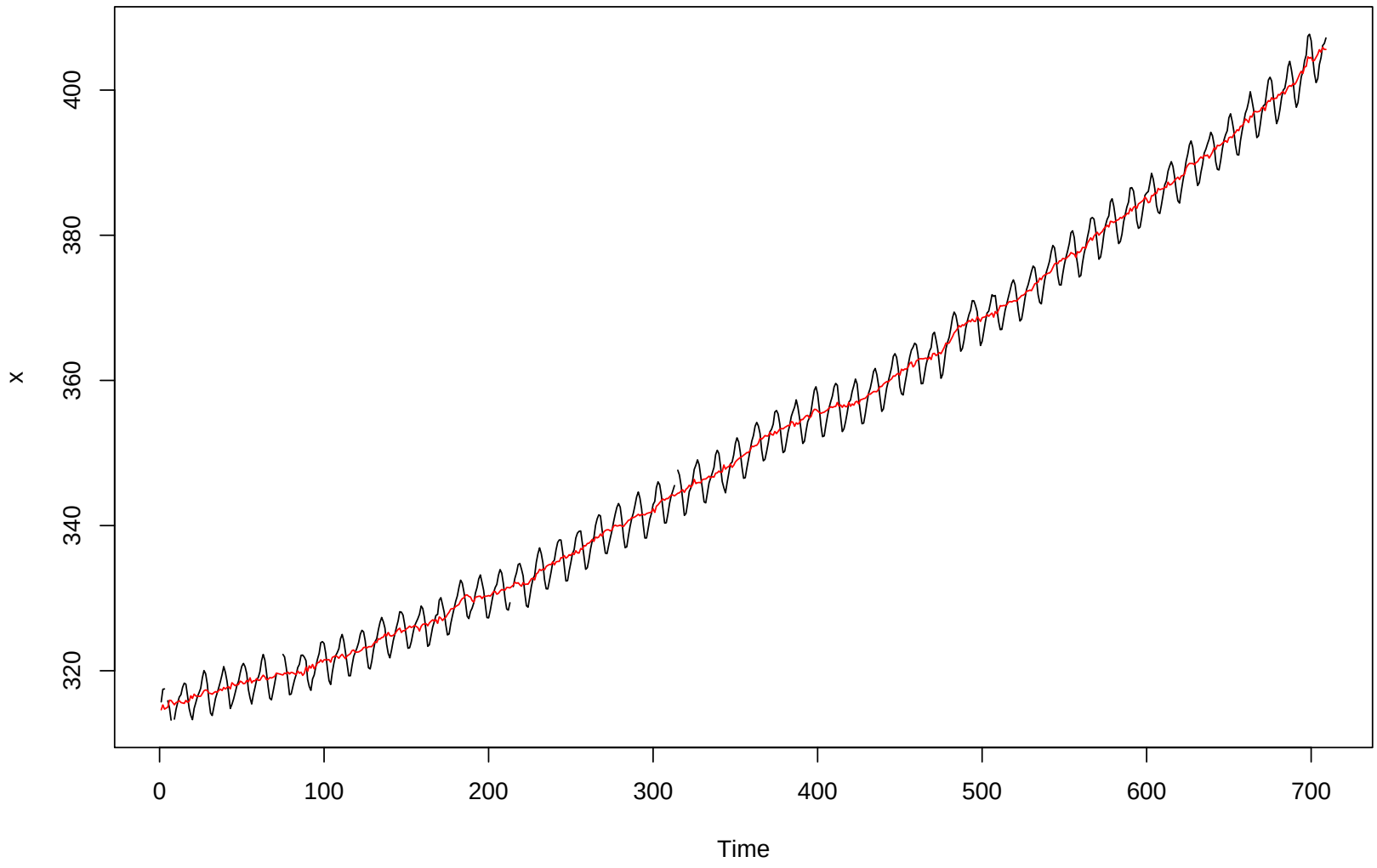
Cálculo de estadísticos descriptivos:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   313.2   328.5   350.2   352.5   373.7   407.7       7
```

Ejemplo 2: Manipulación transparente de datos

Ploteo de la variable *average* como series de tiempo, incluyendo la tendencia:

Ejemplo 2: Manipulación transparente de datos



Ejemplo 3

Se bajan los datos, en formato de variables separadas por comas, del Institute for Digital Research and Education de la UCLA.

La base de datos contiene las variables *admit*, *gre*, *gpa*, *rank*.

- *admit* es una variable dummy (dicotómica) que vale 1 si el candidato es admitido a un posgrado (y vale cero, de otra manera);
- las variables *gre*, *gpa* denotan el promedio del GRE y GPA, respectivamente y se consideran variables continuas;
- la variable *rank* es una variable ordinal, con valores de 1 a 4: $rank = 1$ indica el ranqueo máximo; $rank = 4$ el menor.

Ejemplo 3

Estimamos un modelo de probabilidad lineal

$$admit = \beta_0 + \beta_1 gre + \beta_2 gpa + \beta_3 rank + u$$

```
mydata <-
  read.csv("https://stats.idre.ucla.edu
           /stat/data/binary.csv")
prob.lineal <- lm(admit ~ gre + gpa + factor(rank),
    data = mydata)
prob.lineal.summary <- summary(prob.lineal)
prob.lineal.summary$coef[,1:3]
```

Ejemplo 3

Los coeficientes estimados, con valor-p asociado, se dan a continuación:

```
##                     Estimate   Std. Error
## (Intercept)    -0.258910210 0.2159903968
## gre             0.000429572 0.0002107347
## gpa             0.155535025 0.0639618357
## factor(rank)2 -0.162365349 0.0677144786
## factor(rank)3 -0.290570480 0.0702453273
## factor(rank)4 -0.323026366 0.0793163565
```

Ejemplo 3

```
##                     Estimate   t value     Pr(>|t|)
## (Intercept)    -0.258910210 -1.198712 2.313606e-01
## gre             0.000429572  2.038449 4.217224e-02
## gpa             0.155535025  2.431685 1.547363e-02
## factor(rank)2 -0.162365349 -2.397794 1.695874e-02
## factor(rank)3 -0.290570480 -4.136510 4.313004e-05
## factor(rank)4 -0.323026366 -4.072632 5.621191e-05
```