Aprendiendo a Computar vs. Computar para Aprender

Hacia una nueva forma de enseñar y aprender estadística

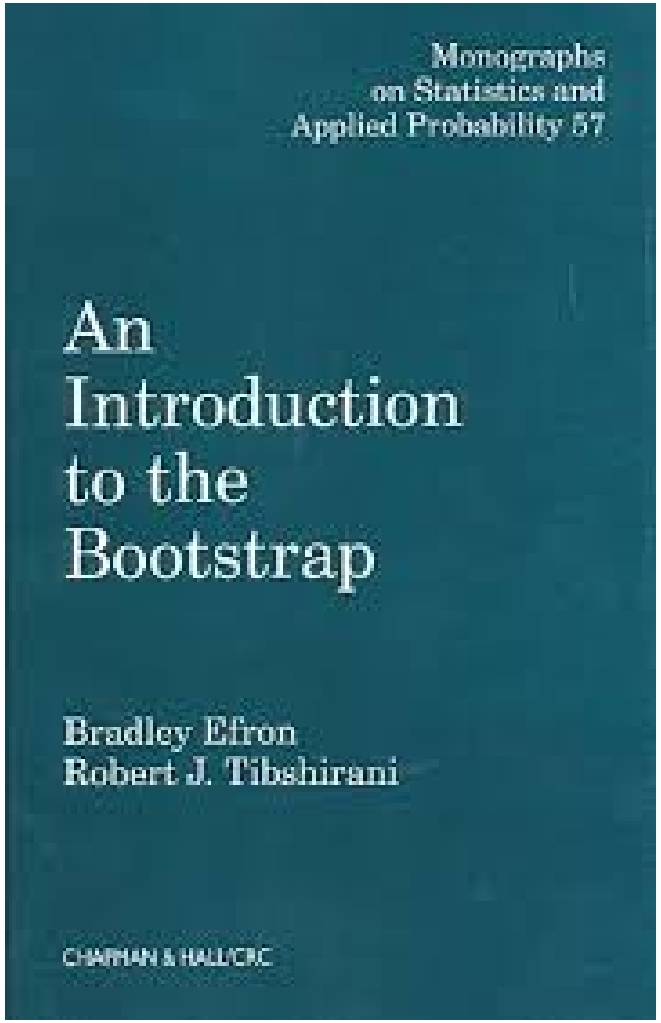
Walter Sosa Escudero

Universidad de San Andrés y CONICET

LatinR. Agosto de 2018. Buenos Aires

product names mentioned herein are trademarks or registered trademarks of their respective companies.
Copyright 1997 MathSoft, Inc. All rights reserved. S-PLUS is a registered trademark of MathSo
Explore
Analyze
Discover
Visualize
MathSoft
RELEASE 2
S-PLUS 4
S-Plus 4

1993...


Monographs
on Statistics and
Applied Probability 57
An
Introduction
to the
Bootstrap
Bradley Efron
Robert J. Tibshirani
CHAPMAN & HALL/CRC

1993...

(1) How should I collect my data?

(2) How should I analyze and summarize the data that I've collected?

(3) How accurate are my data summaries?

Question 3 constitutes part of the process known as statistical inference. The bootstrap is a recently developed technique for making certain kinds of statistical inferences. It is only recently developed because it requires modern computer power to simplify the often intricate calculations of traditional statistical theory.

The explanations that we will give for the bootstrap, and other

2018...

Walter Sosa Escudero @wsosaescudero · 37 min

Una encuestita mas (plis RT). Si tomaste por lo menos un curso de estadistica universitaria, ¿sabrias que es el boostrap?

74% No

26% Si

215 votos • Falta(n) 23 horas

Walter Sosa Escudero @wsosaescudero · 38 min

Denme una manito (respondan y RT), gracias. Si tomaste por lo menos un curso de estadistica a nivel de grado, ¿aprendiste boostrap en el primer curso que tomaste?

92% No

8% Si

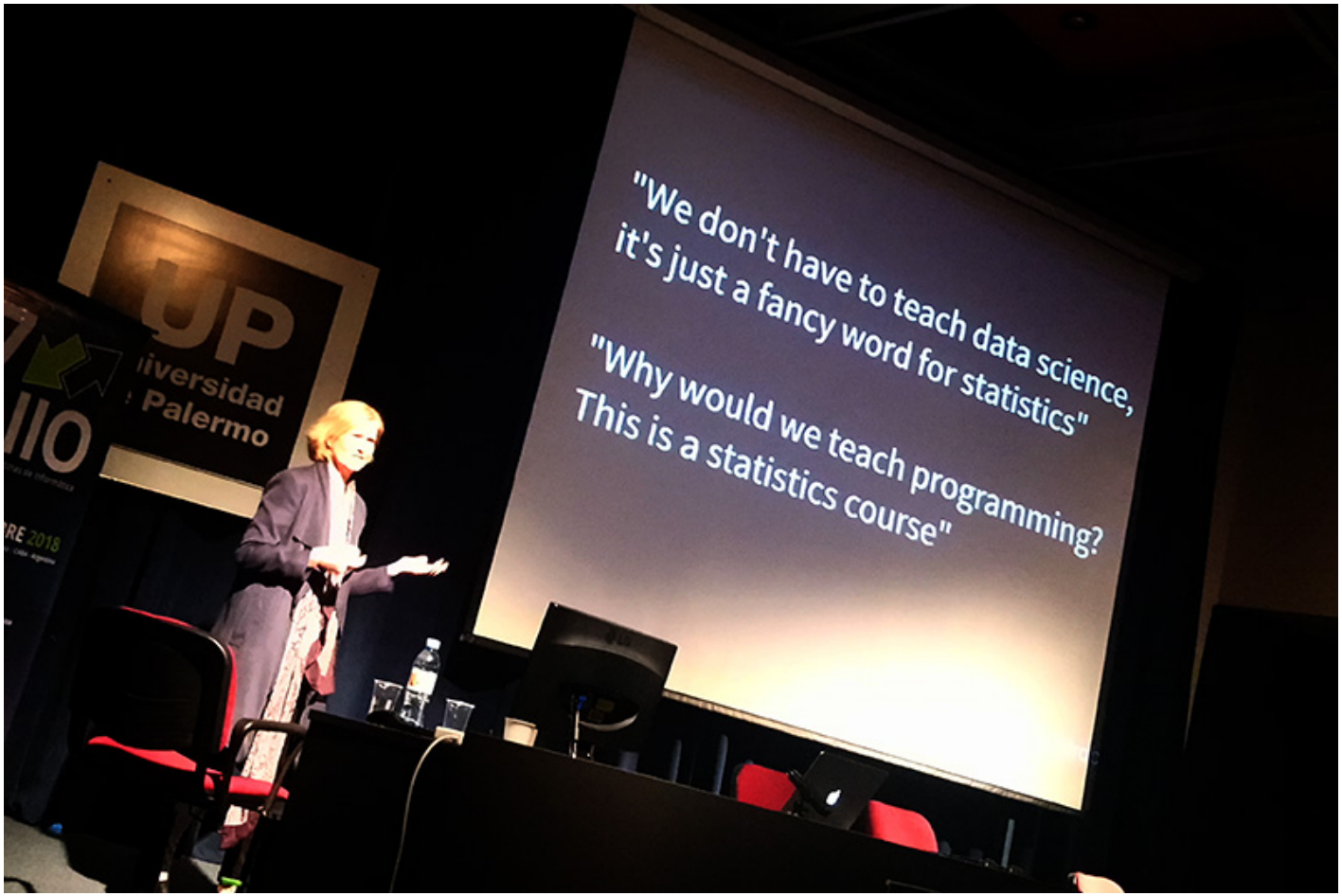
270 votos • Falta(n) 23 horas

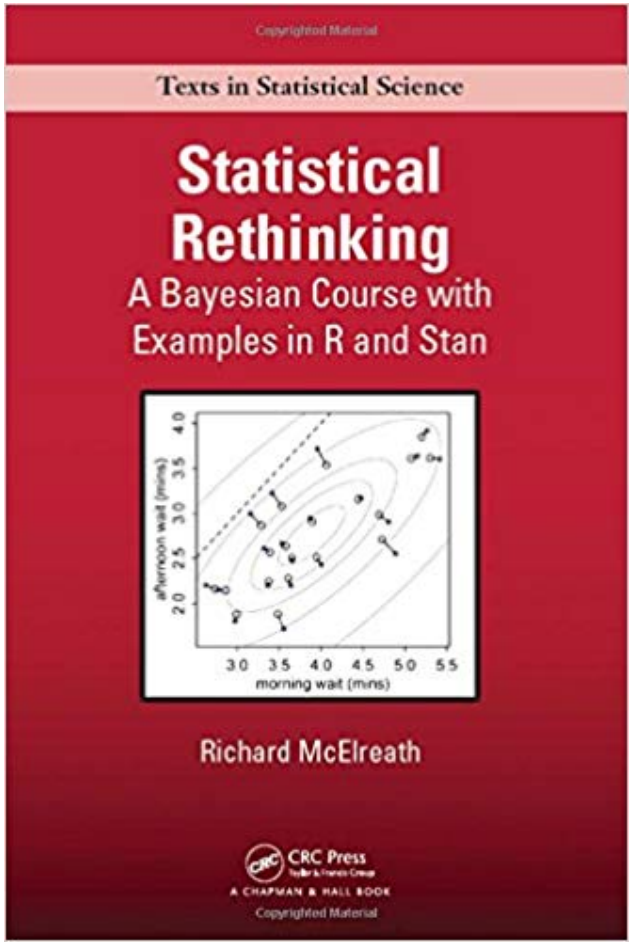
Computar para

- ilustrar
- aplicar
- modificar
- simular

... entender

R en la clase teorica!


"We don't have to teach data science,
it's just a fancy word for statistics"
"Why would we teach programming?
This is a statistics course"
UP
iversidad
Palermo


Texts in Statistical Science
Statistical Rethinking
A Bayesian Course with Examples in R and Stan
afternoon wait (mins)
morning wait (mins)
Richard McElreath
CRC Press
A CHAPMAN & HALL BOOK

Home
Material
Tutoriales
Trabajos Prácticos
Profesores
BIG DATA Y MINERÍA DE DATOS
UDESA
Perspectivas, ideas y herramientas para economistas
Walter Sosa Escudero
Asistente: Florencia Hnilo
PROGRAMA

Teorema Central del Limite

Teorema Central del Limite:

- ‘Todo es normal’
- ‘Para $n > 30$ la normal aproxima bien cualquier distribucion‘.
- ‘Por el TCL, $n > 30$ ya es una muestra grande‘.

Quora

Search for questions, people, and topics

What is the importance of the central limit theorem?

9 Answers

[illegible], K-12 math teacher

Updated Nov 2, 2016 · Author has **99** answers and **245.7k** answer views

CLT is important because under certain condition, you can approximate some distribution with Normal distribution **although the distribution is not Normally distributed**.

[redacted] · 19 ago.

En respuesta a @wsosaescudero

Last course some student told me that her stats professor set the threshold at 50 obs. Mmmmmmm

Traducir Tweet

1

[redacted] · 19 ago.

En respuesta a @wsosaescudero

Me la paso haciendo pruebas no parametricas. Jamas tuve datos q se ajustaran a una normal 😩😩😩😩

1

Teorema Central del Limite

Lindeberg-Levy: $\{z_n\}$ una sucesion de variables aleatorias i.i.d. con $E(z_i) = \mu$ y $V(z_i) = \sigma^2$ (ambas finitas). Entonces

$$\frac{\bar{z}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

En la practica

Si para $n \to \infty$

$$\frac{\bar{z}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1),$$

entonces, para n 'suficientemente grande':

$$\frac{\bar{z}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

de modo que:

$$\bar{z}_n \sim N(\mu, \sigma^2/n)$$

TCL como aproximacion, errada pero util

Entonces:

- TCL no aplica a z_n sino a $\bar{z}_n$.
- Exacto para $(\bar{z}_n - \mu)/(\sigma/\sqrt{n})$, $n \to \infty$
- Aproximado para $\bar{z}_n$, n 'grande'.

Demostracion?

THEOREM B *Central Limit Theorem*

Let $X_1, X_2, \ldots$ be a sequence of independent random variables having mean 0 and variance σ^2 and the common distribution function F and moment-generating function M defined in a neighborhood of zero. Let

$$S_n = \sum_{i=1}^{n} X_i$$

Then

$$\lim_{n\to\infty} P\left(\frac{S_n}{\sigma\sqrt{n}} \le x\right) = \Phi(x), \quad -\infty < x < \infty$$

of Let $Z_n = S_n/(\sigma\sqrt{n})$. We will show that the mgf of Z_n tends to the mgf of the standard normal distribution. Since S_n is a sum of independent random variables,

$$M_{S_n}(t) = [M(t)]^n$$

and

$$M_{Z_n}(t) = \left[M\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$

$M(s)$ has a Taylor series expansion about zero:

$$M(s) = M(0) + sM'(0) + \tfrac{1}{2}s^2M''(0) + \varepsilon_s$$

where $\varepsilon_s/s^2 \to 0$ as $s \to 0$. Since $E(X) = 0$, $M'(0) = 0$, and $M''(0) = \sigma^2$. As $n \to \infty$, $t/(\sigma\sqrt{n}) \to 0$, and

$$M\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 + \frac{1}{2}\sigma^2\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + \varepsilon_n$$

where $\varepsilon_n/(t^2/(n\sigma^2)) \to 0$ as $n \to \infty$. We thus have

$$M_{Z_n}(t) = \left(1 + \frac{t^2}{2n} + \varepsilon_n\right)^n$$

It can be shown that if $a_n \to a$, then

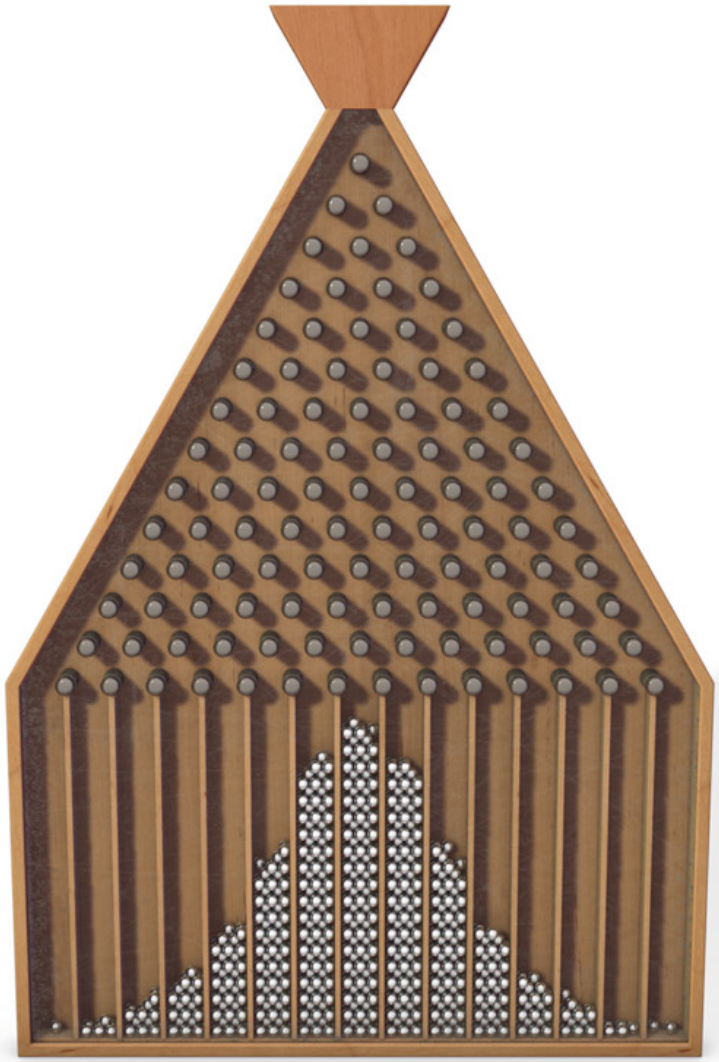
$$\lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = e^a$$

From this result, it follows that

$$M_{Z_n}(t) \to e^{t^2/2} \quad \text{as } n \to \infty$$

Rice, J., 1995, *Mathematical Statistics and Data Analysis*, Duxbury, New York

Quincunx?



Anecdotas?

La Plata, Buenos Aires, circa 1998....

Computar!

$\{z_n\}$ una sucesion de variables aleatorias i.i.d. con $E(z_i) = \mu$ y $V(z_i) = \sigma^2$ (ambas finitas).

```
#Una sucesion
N<-1000

#Cual? Cualquiera.Uniforme
a<-0; b<-1

z<-1:N
for(i in 1:N){
  z[i]<-runif(1,min=a, max=b)
  }

#x<-runif(N,min=a, max=b)
```

$\{z_n\}$ una sucesion de variables aleatorias i.i.d. con $E(z_i) = \mu$ y $V(z_i) = \sigma^2$ (ambas finitas). Entonces

$$\frac{\bar{z}_n - \mu}{\sigma/\sqrt{n}}$$

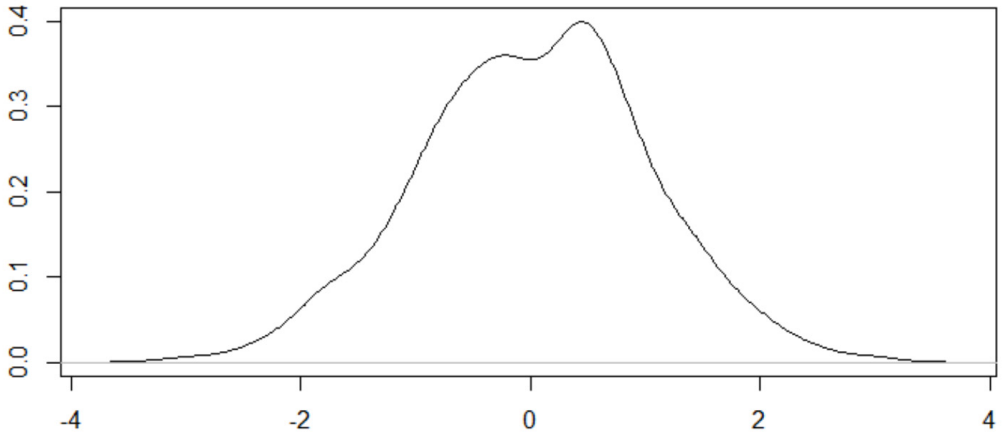
```
ez<-(a+b)/2
vz<-(1/12)*(b-a)^2

zn<-(mean(z)-ez)/(sqrt(vz)/sqrt(N))
```

$\{z_n\}$ una sucesion de variables aleatorias i.i.d. con $E(z_i) = \mu$ y $V(z_i) = \sigma^2$ (ambas finitas). Entonces

$$\frac{\bar{z}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d}$$

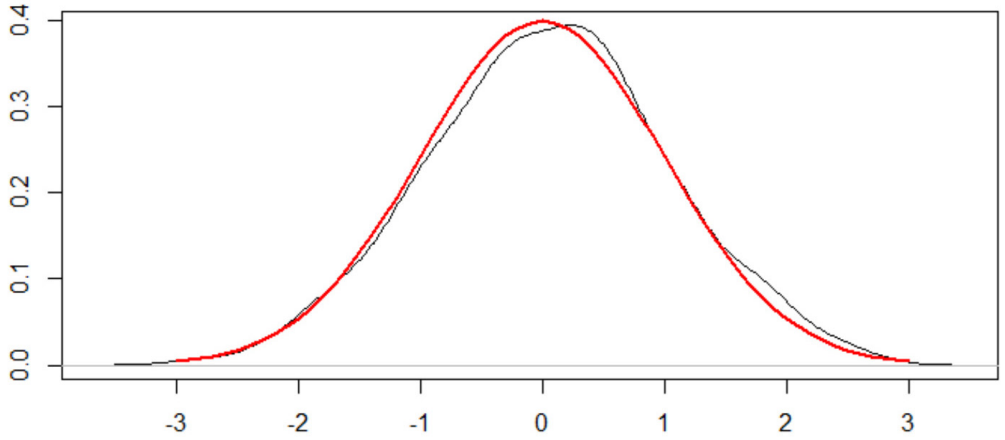
```
R<-1000 #Ojo: Replicas!
zetas<-1:R
for(j in 1:R){
  z<-runif(N,min=a,max=b)
  zetas[j]<-(mean(z)-ez)/(sqrt(vz)/sqrt(N))
  }
plot(density(zetas))
```



$\{z_n\}$ una sucesion de variables aleatorias i.i.d. con $E(z_i) = \mu$ y $V(z_i) = \sigma^2$ (ambas finitas). Entonces

$$\frac{\bar{z}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

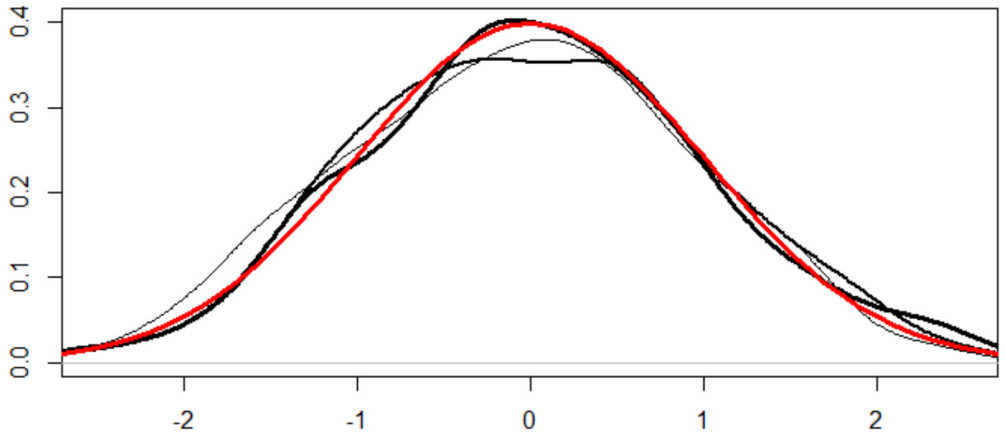
```
s<-seq(-3,3,by=0.1)
lines(s,dnorm(s, mean=0, sd=1), col=2, lwd=2)
```



$\{z_n\}$ una sucesion de variables aleatorias i.i.d. con $E(z_i) = \mu$ y $V(z_i) = \sigma^2$ (ambas finitas). Entonces

$$\frac{\bar{z}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

```
ENE<-c(3,10,30)
for(h in 1:length(ENE)){
  R<-1000
  zetas<-1:R
  for(j in 1:R){
    z<-runif(ENE[h],min=a,max=b)
    zetas[j]<-(mean(z)-ez)/(sqrt(vz)/sqrt(ENE[h]))
    }
  if(h==1) plot(density(zetas))
  lines(density(zetas), lwd=h)
}
lines(s,dnorm(s, mean=0, sd=1), col=2, lwd=3)
```



Codigo R para:

- Entender el teorema (ida y vuelta)
- Simular
- Paso previo (Shiny, ggplot2, Aprendiendo a Computar)

Insesgadez

$$Y = X\beta + u, \qquad \hat{\beta} = (X'X)^{-1}X'Y$$

1. $\rho(X) = K$
2. $E(u|X) = 0$

$\hat{\beta}$ es insesgado: $E(\hat{\beta}) = \beta$

¿Insesgado?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.90831    0.39522   2.298   0.0237 *
x            1.03769    0.06816  15.224   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.733 on 98 degrees of freedom
Multiple R-squared:  0.7028,    Adjusted R-squared:  0.6998
F-statistic: 231.8 on 1 and 98 DF,  p-value: < 2.2e-16
```

Demostracion

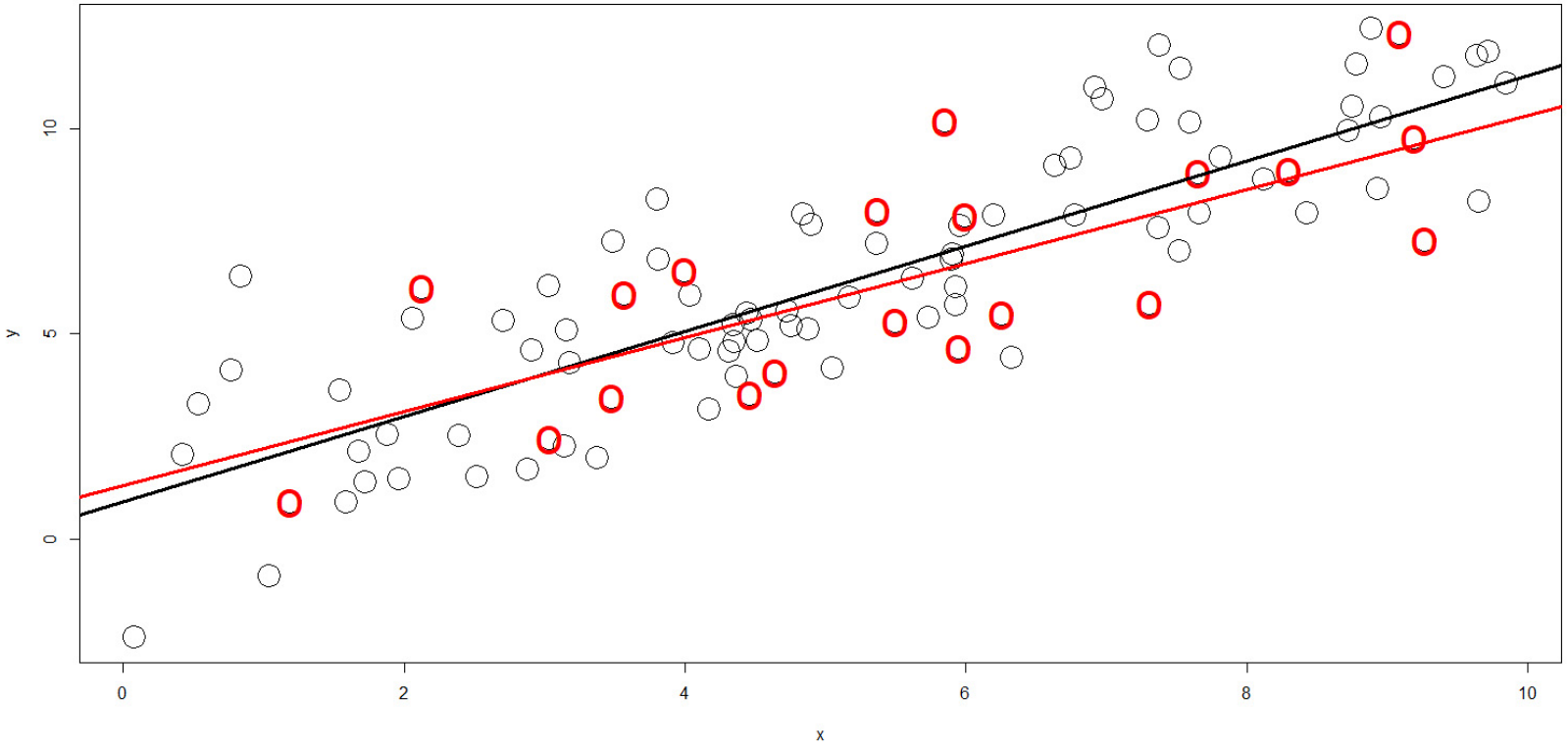
Prueba de la insesgadez

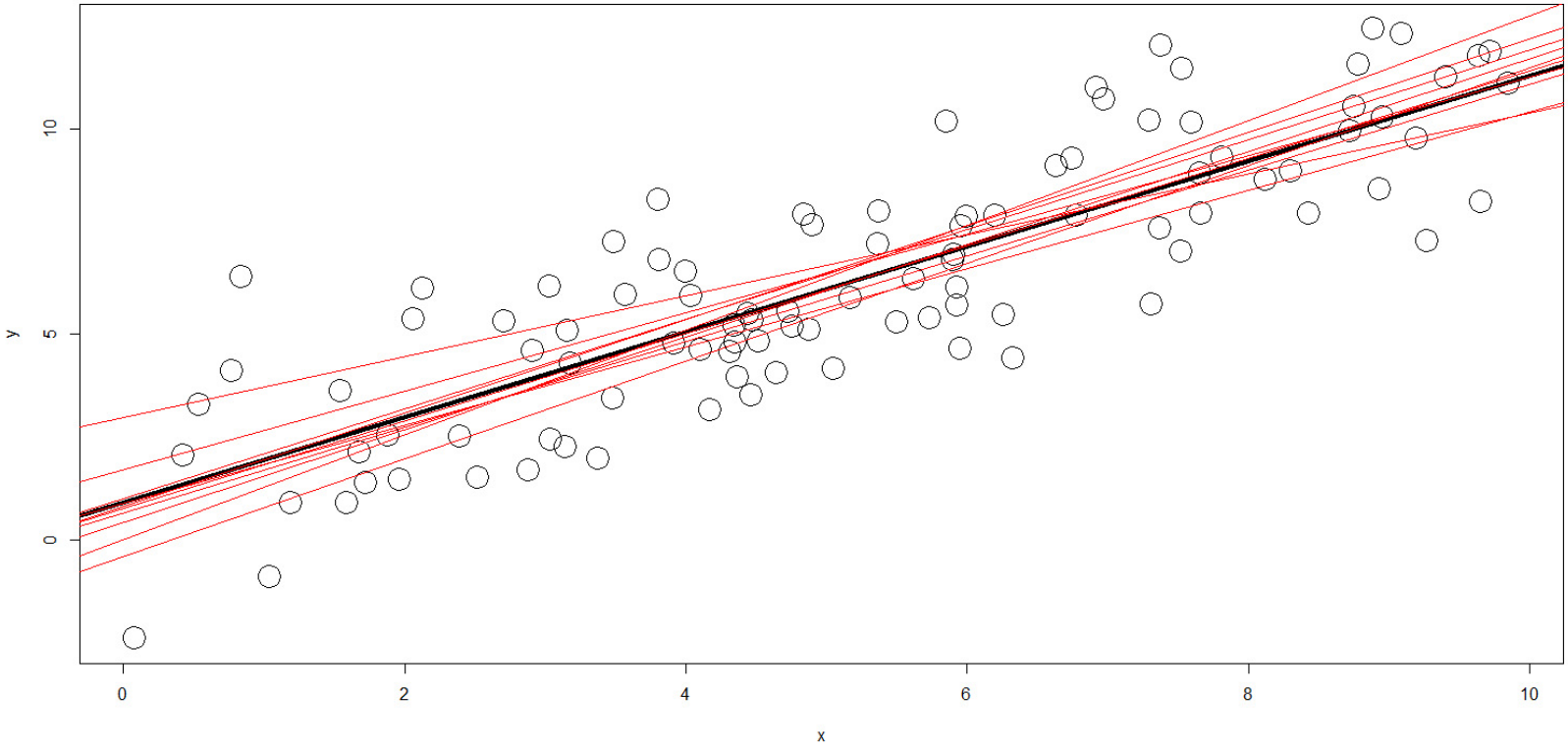
$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'Y \\
&= (X'X)^{-1}X'(X\beta + u) \\
&= \beta + (X'X)^{-1}X'u \\
E(\hat{\beta}|X) &= \beta + E\left[(X'X)^{-1}X'u|X\right] \\
&= \beta + (X'X)^{-1}X'E\left[u|X\right] \\
&= \beta
\end{aligned}$$

Usando la ley de esperanzas iteradas

$$E(\hat{\beta}) = E(E(\hat{\beta}|X)) = E(\beta) = \beta$$

Insesgadez: propiedad del rifle (estimador) y no del disparo (estimacion)


y
x
0
5
10
0
2
4
6
8
10


y
0
5
10
0
2
4
6
8
10
x

```r
x<-runif(100)*10
y<- 1+x+rnorm(100)*2

plot(x,y,cex=3)
abline(lm(y~x), lwd=4)

REP<-10
coef<-matrix(0, REP,2)
for(i in 1:REP){
  ese<-sample(100,20)
  modelo<-lm(lm(y[ese]~x[ese]))
  coef[i,]<-modelo$coefficients
  abline(lm(y[ese]~x[ese]), col=2)
  }
```

Examen

- En que parte del codigo se hacen los supuestos?
- Cuanto tiene que valer `REP` para que se vea el resultado?
- Y N?
- Como habria que cambiar el codigo para que $E(u|X) \neq 0$?
- Que pasa si cambiamos `runif`?

En sintesis

- Codigo como parte del razonamiento.
- No mera aplicacion ('programar el simulador').
- Dimensiones de formalidad: matematica, comunicacional, visual, expositiva.

4.4.2. Fitting the model. The code needed to fit this model via quadratic approximation is a straightforward modification of the kind of code you've already seen. All we have to do is incorporate our new model for the mean into the model specification inside `map` and be sure to add our new parameters to the `start` list. Let's repeat the model definition, now with the corresponding R code on the right-hand side:

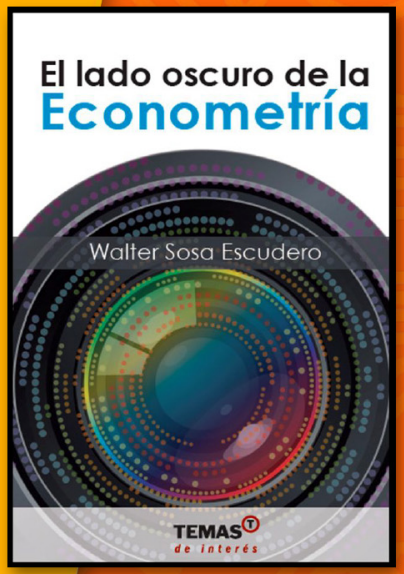
$h_i \sim \text{Normal}(\mu_i, \sigma)$	`height ~ dnorm(mu,sigma)`
$\mu_i = \alpha + \beta x_i$	`mu <- a + b*weight`
$\alpha \sim \text{Normal}(178, 100)$	`a ~ dnorm(156,100)`
$\beta \sim \text{Normal}(0, 10)$	`b ~ dnorm(0,10)`
$\sigma \sim \text{Uniform}(0, 50)$	`sigma ~ dunif(0,50)`

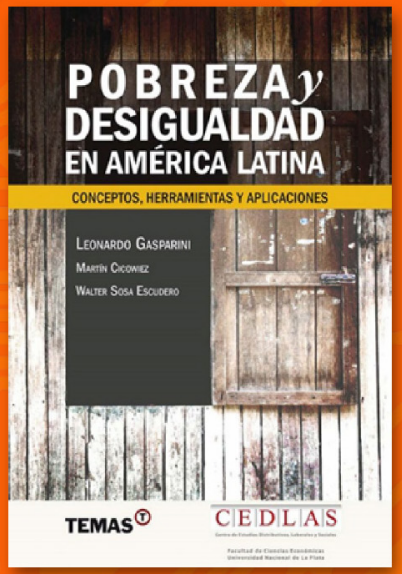
Notice that the linear model, in the R code on the right-hand side, uses the R assignment

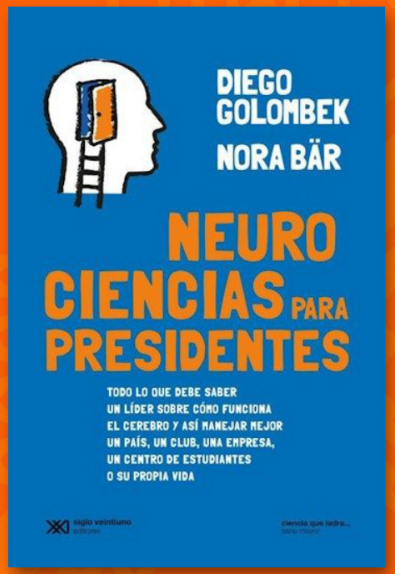
McElreath, R., 2016, Statistical Rethinking, CRC Press.

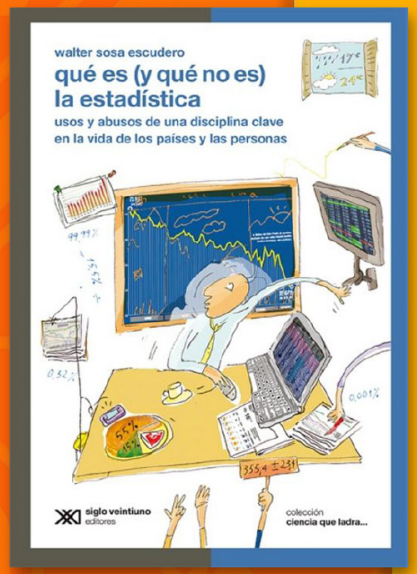
Enseñar estadistica como si los ultimos 20 años hubiesen ocurrido

- Revolucion pedagogica. Rol integral de la programacion. Go R!
- Educar para *pensar* en terminos de datos y metodos. No simplemente aplicar.


El lado oscuro de la
Econometría
Walter Sosa Escudero
TEMAS
de interés


POBREZA y
DESIGUALDAD
EN AMÉRICA LATINA
CONCEPTOS, HERRAMIENTAS Y APLICACIONES
Leonardo Gasparini
Martín Cicowiez
Walter Sosa Escudero
TEMAS
CEDLAS


DIEGO
GOLOMBEK
NORA BÄR
NEURO
CIENCIAS PARA
PRESIDENTES
TODO LO QUE DEBE SABER
UN LÍDER SOBRE CÓMO FUNCIONA
EL CEREBRO Y ASÍ MANEJAR MEJOR
UN PAÍS, UN CLUB, UNA EMPRESA,
UN CENTRO DE ESTUDIANTES
O SU PROPIA VIDA
siglo veintiuno


walter sosa escudero
qué es (y qué no es)
la estadística
usos y abusos de una disciplina clave
en la vida de los países y las personas
siglo veintiuno
editores
colección
ciencia que ladra...

Walter Sosa Escudero

Profesor Plenario, Director

Departamento de Economía

Universidad de San Andrés

Investigador Principal

CONICET

wsosa@udesa.edu.ar

@wsosaescudero

waltersosa.weebly.com

PhD en Economía (University of Illinois), Lic. en Economía (UBA). Especialista en econometría y estadística aplicada a cuestiones sociales. Director del Departamento de Economía de la Universidad de San Andrés. Investigador principal del CONICET. Autor de numerosas publicaciones académicas, incluyendo artículos en Journal of Econometrics o Econometric Theory, entre otros. Orador en TEDXRioDeLaPlata y autor de Que es (y que no es) la Estadística. Escribe frecuentemente para los principales diarios argentinos. Ha recibido los premios Konex, Fulvio Pagani/ARCOR, Bernardo Houssay, EGSO Award (como el mejor profesor del PhD en Economía de la Universidad de Illinois) y elegido por la Fundación BGH como uno de los 100 argentinos innovadores. Ex presidente de la Asociación Argentina de Economía Política. Miembro de número de la Academia Nacional de Ciencias Económicas.