Simulación

Yanink Neried Caro Vega y Lizbeth Naranjo Albarrán

Proyecto I: Métodos Estadísticos para la Investigación en VIH/SIDA

Se presentan varios ejemplos de simulación:

- Simular datos.
- Obtener distribución de probabilidad usando información subjetiva
- Simular una respuesta de un modelo de regresión lineal.
- Simular una respuesta de un modelo de regresión logística.
- Simular una respuesta de un modelo de supervivencia o sobrevida.
- CCASAnet

Scripts y bases de datos simulados de análisis y estudios publicados en CCASAnet:

http://biostat.mc.vanderbilt.edu/wiki/Main/ArchivedAnalyses

http://ccasanet.vanderbilt.edu/

Cuando se tienen diferentes fuentes o valor puntuales que provienen de distintos estudios, se puden considerar rangos para los parámetros, y calcular una distribución de manera subjetiva.

Simular datos simples

La simualción de los datos dependen de las características y del tipo de datos que se necesiten simular.

Es común que cuando el interés es generar variables relacionadas a ciertas características se elijan las distribuciones más conocidas.

- Si los datos son binarios, $X \in \{0, 1\}$, entonces usar $X \sim Bernoulli(\theta)$.

– Género (hombre, mujer), donde la mitad de la población son hombres, $X \sim Bernoulli(0.5)$.

– Enfermedad (enfermo, sano), donde el 20% de la población está enferma, $X \sim Bernoulli(0.2)$.

- Si los datos están definidos en $X \in \mathbb{R}$, frecuentemente se usa la v.a. $X \sim Normal(\mu, \sigma^2)$ o $X \sim Uniforme(a, b)$.

– Niveles de CD4 con media 300 y desviación estándar 20, $X \sim Normal(\mu = 300, \sigma = 20)$.

– Edades definidas en un rango de 20 a 60, $X \sim Uniforme(20, 60)$.

- Si los datos se refieren a probabilidades $\theta \in [0, 1]$ entonces usar $\theta \in Beta(a, b)$.

– La probabilidad de enfermarse se distribuye como $\theta \sim Beta(5, 15)$.

- Si los datos se refieren a conteos $X \in \{0, 1, 2, \ldots\}$, entonces usar $X \sim Poisson(\lambda)$.

– Se espera que lleguen 5 pacientes a un hospital en el intervalo de un minuto, $X \sim Poisson(5)$.

Obtener distribución de probabilidad usando información subjetiva

Si se tiene información acerca del comportamiento de unos datos, ésta se puede traducir en una distribución de probabilidad.

Este procedimiento en utilizado en estadística Bayesiana para construir distribuciones iniciales informativas (*informative prior distribution*), y se le conoce como **elicitar** (*elicite*).

Ejemplo Normal con Percentiles y Cuantiles

Se tiene la información de que la temperatura diaria de la ciudad de México en octubre está resumida en los siguientes cuantiles:

Percentile	Cuantile
0.50	25 ºC
0.90	30 ºC

Si suponemos que los datos siguen una distribución $Normal(\mu, \sigma^2)$, entonces es necesario calcular μ y σ.

- A partir del cuantil 0.50, como la distribución normal es simétrica, la mediana es igual a la media, por tanto, $\mu = 25$.
- Por propiedades de la distribución normal, si $X \sim Normal(\mu, \sigma^2)$, entonces $Z = \frac{X-\mu}{\sigma} \sim Normal(0, 1)$. Se sabe que $P[X \leq 30] = 0.90$, entonces

$$P\left(Z \leq \frac{30-\mu}{\sigma}\right) = 0.90$$

El cuantil 0.90 de una normal estandar es 1.281552, entonces $\frac{30-\mu}{\sigma} = 1.281552$, y como $\mu = 25$,

```
(30-25)/qnorm(0.90)
```

```
## [1] 3.901521
```

Entonces despejando $\sigma = \frac{30-25}{1.281552} = 3.901521$.

Ejemplo Beta con Media y Percentiles

Se tiene información acerca de la prevalencia de una enfermedad, θ que es la proporción de sujetos infectados. Se sabe que la tasa de infección se encuentra entre 0.05 y 0.20 con un 95% de probabilidad, y que la prevalencia promedio es de 0.10.

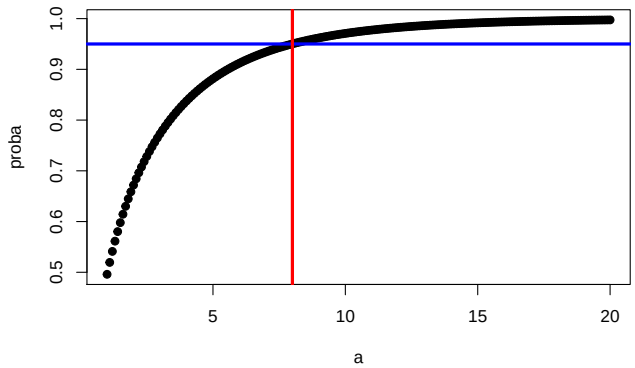
Si nosotros queremos traducir (*elicitar*) esta información en una distribución de tipo $\theta \sim Beta(a, b)$. ¿Cuáles son los valores de a y b? Tenemos que resolver el sistema de ecuaciones:

$$\begin{aligned} \mathbb{E}[\theta] = \frac{a}{a+b} &= 0.10 \\ \mathbb{P}[0.05 < \theta < 0.20] &= 0.95 \end{aligned}$$

¿Cómo se calcula? Una simple solución es la siguiente:

```
a = seq(1,20,by=0.1)    ### Conjuntos de posibles valores para a.
b = a*(1-0.10)/0.10    ### A partir de la media se despeja b.
proba = c()    ### Calcular la probabilidad acumulada P[0.05<\theta<0.20]
for(i in 1:length(a)){
  proba[i] = pbeta(0.20,a[i],b[i]) - pbeta(0.05,a[i],b[i]) }
```

Lo cual resulta aproximadamente en $a = 8$ y $b = 72$.



Regresión Lineal

- Simular una respuesta de un modelo de regresión lineal.

Se tiene la variable de interés Y (variable respuesta o dependiente), $Y \in \mathbb{R}$, y se relaciona con un conjunto de covariables $x_1, \ldots, x_k$, a través de un modelo de regresión lineal:

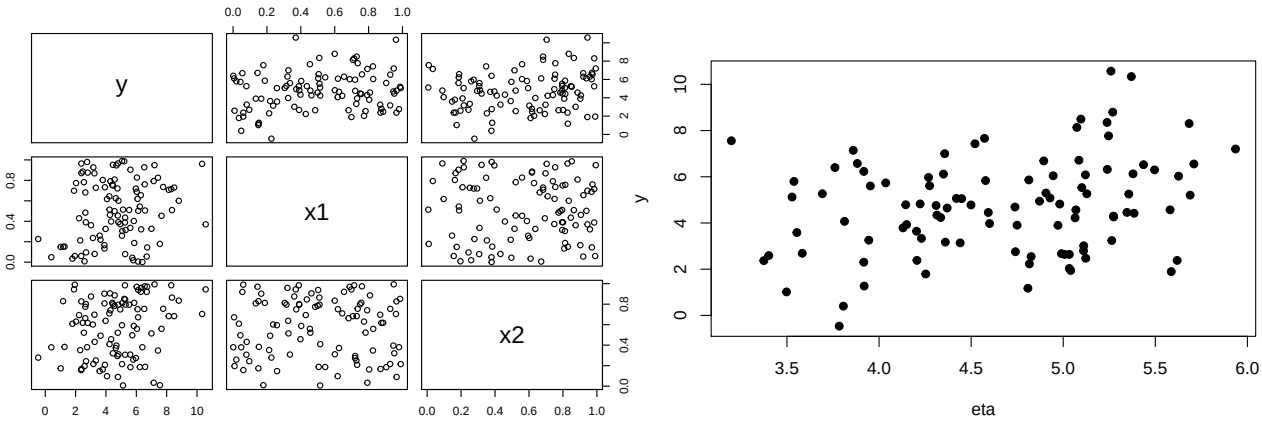
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon, \quad \varepsilon \sim Normal(0, \sigma^2)$$

```
n = 100   ### numero de sujetos
betas = as.vector(c(1,2,3))   ### coeficientes de regresion
sigma2 = 4   ### varianza
set.seed(12345)
x1 = runif(n)   ### covariable x1
x2 = runif(n)   ### covariable x2
epsilon = rnorm(n,0,sqrt(sigma2))   ### simular error aleatorio
eta = betas[3] + betas[1]*x1 + betas[2]*x2   ### predictor lineal
y = eta + epsilon   ### respuesta aleatoria
head(cbind(y,x1,x2,eta,epsilon),4)
```

```
##             y        x1        x2      eta    epsilon
## [1,] 4.757686 0.7209039 0.2944654 4.309835  0.4478508
## [2,] 2.797834 0.8757732 0.6172537 5.110281 -2.3124467
## [3,] 6.554368 0.7609823 0.9742741 5.709531  0.8448371
## [4,] 2.473038 0.8861246 0.6182120 5.122549 -2.6495105
```

```
pairs(cbind(y,x1,x2))
plot(eta,y, pch=19)
modelo <- lm(y ~ x1+x2)   ### estimar parametros
summary(modelo)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.391 -1.165 -0.026  1.244  5.425
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2049     0.5745   5.579 2.19e-07 ***
## x1            1.2762     0.6809   1.874   0.0639 .
## x2            1.5547     0.7159   2.172   0.0323 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.037 on 97 degrees of freedom
## Multiple R-squared:  0.07822,    Adjusted R-squared:  0.05922
## F-statistic: 4.116 on 2 and 97 DF,  p-value: 0.01925
```



Regresión Logística

- Simular una respuesta de un modelo de regresión logística.

Se tiene la variable de interés Y binaria (variable respuesta o dependiente), $Y \in \{0, 1\}$, y se relaciona con un conjunto de covariables $x_1, \ldots, x_k$, a través de un modelo de regresión logística:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

o equivalentemente

$$p = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right)}$$

que es igual a la función de distribución acumulada (fda) de una distribución logística.

Si $X \sim Logistica(0, 1)$ su fda es: $F_X(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)}$.

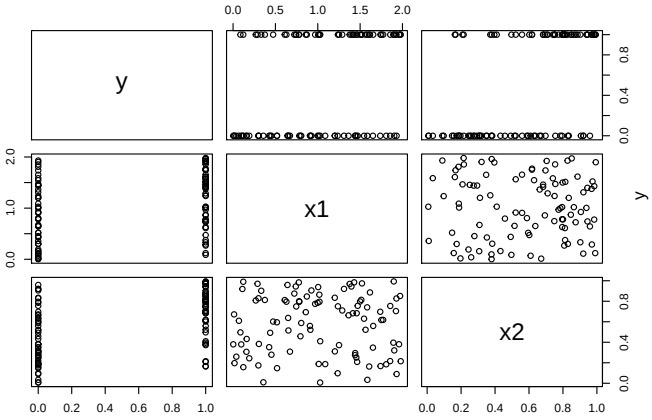
Regresión logística usando coeficientes para la probabilidad de éxito

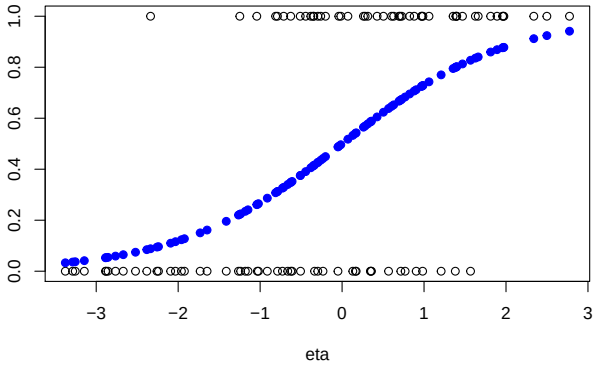
```
n = 100   ### numero de sujetos
K = 3   ### numero parametros de regresion
betas = as.vector(c(2,3,-4))   ### coeficientes de regresion
set.seed(12345)
x1 = runif(n,0,2)   ### covariable x1
x2 = runif(n)   ### covariable x2
eta = betas[3] + betas[1]*x1 + betas[2]*x2   ### combinacion lineal
p = plogis(eta)   ### se calcula p=P[Y=1]
u = runif(n)
y = ifelse(p>u,1,0)   ### simular la variable respuesta Y de manera aleatoria
y = rep(NA,n)
for(i in 1:n){
  y[i] = rbinom(1,size=1,prob=p[i])   ### Bernoulli
}
head(cbind(y,x1,x2,eta,p,u),4)
```

```
##      y       x1        x2        eta         p         u
## [1,] 0 1.441808 0.2944654 -0.2329882 0.4420150 0.5885923
## [2,] 1 1.751546 0.6172537  1.3548538 0.7949220 0.8925918
## [3,] 1 1.521965 0.9742741  1.9667517 0.8772618 0.1237949
## [4,] 1 1.772249 0.6182120  1.3991344 0.8020465 0.5133090
```

```
pairs(cbind(y,x1,x2))
plot(eta,y)
points(eta,p,col="blue",pch=19)
modelo <- glm(y ~ x1+x2, family=binomial(link="logit"))   ### estimar parametros
summary(modelo)
```

```
##
## Call:
## glm(formula = y ~ x1 + x2, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0774  -0.7861  -0.2138   0.8035   2.0635
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.5652     0.9841  -4.639 3.50e-06 ***
## x1            1.6086     0.4804   3.348 0.000813 ***
## x2            4.8806     1.0471   4.661 3.14e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.589  on 99  degrees of freedom
## Residual deviance:  97.873  on 97  degrees of freedom
## AIC: 103.87
##
## Number of Fisher Scoring iterations: 5
```





Odds ratio

Algunas veces los resultados los presentan como odds, entonces es cuestión de entender qué significa este valor.

Suponga que se estudia la prevalencia de una enfermedad, es decir, la probabilidad de que un paciente se enferme, $p = P[Y = 1]$. Los odds (*momios*) se definen como:

$$odds = \frac{p}{1-p}$$

Esto es la razón de enfermos sobre sanos. Por ejemplo, si $p = 0.2$, entonces

$$odds = \frac{0.2}{1-0.2} = \frac{0.2}{0.8} = \frac{2}{8} = \frac{1}{4}$$

indicando que por cada 1 enfermo existen 4 sanos.

Si se tienen los odds de dos poblaciones, por ejemplo, de mujeres $odds_1$ y hombres $odds_2$, entonces el *odds ratio* es:

$$\frac{odds_1}{odds_2} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

donde en este caso $p_1 = P[Y_{mujer} = 1]$ y $p_2 = P[Y_{hombre} = 1]$.

Por ejemplo, si $p_1 = 0.2$, y $p_2 = 0.25$ entonces

$$oddsratio = \frac{odds_1}{odds_2} = \frac{\frac{0.2}{0.8}}{\frac{0.25}{0.75}} = \frac{\frac{1}{4}}{\frac{1}{3}} = \frac{3}{4}$$

indicando que por cada 3 enfermo mujer existen 4 enfermos hombres.

Modelos paramétricos de supervivencia

- Simular una respuesta de un modelo de supervicencia/sobrevida.
- Genera un modelo de supervivencia paramétrico, usando una función de supervivencia del tipo:

$$S(t) = P[T > t]$$

con T v.a. Weibull, exponential, gamma, log-logistic, o log-normal.

Covariables

```
library(survival)
library(muhaz)
N = 1000   ### numero de sujetos
K = 2   ### numero parametros de regresion
betas = as.vector(c(3,1))   ### coeficientes de regresion
x1 = runif(N)   ### covariable x1
x2 = sample(x=c(0,1), size=N, prob=c(0.5,0.5), replace=TRUE)   ### covariable x2 binaria
XBeta = betas[1]*x1 + betas[2]*x2   ### combinacion lineal
```

Simular censuras

```
censuraProporcion = 0.2   ### proporcion de datos censurados
censura = sample(x=c(0,1), size=N, prob = c(censuraProporcion, 1-censuraProporcion), replace=TRUE)
head(cbind(x1,x2,censura),4)
```

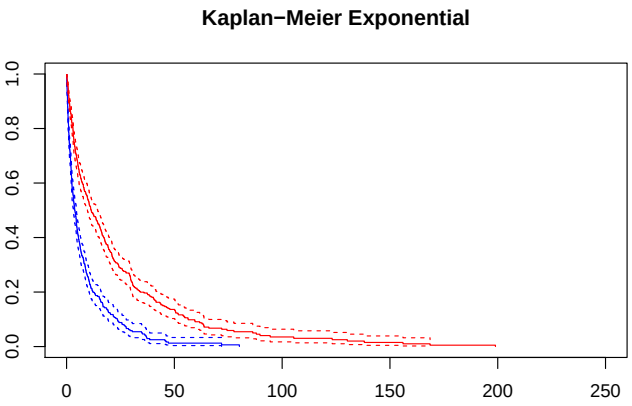
```
##              x1 x2 censura
## [1,] 0.07548045  0       1
## [2,] 0.47438424  0       0
## [3,] 0.26458955  0       0
## [4,] 0.23074607  0       1
```

Simular tiempos de supervivencia de una familia parametrica Exponencial

```
tiempoExp = rexp(n=N, exp(-XBeta))   ### exponential
ajusteExp <- survfit(Surv(tiempoExp,censura)~x2, type = "kaplan-meier",
conf.type="log-log", conf.int=0.95)
plot(ajusteExp,conf.int=T,main="Kaplan-Meier Exponential",
     col=c("blue","blue","blue","red","red","red"),xlim=c(0,250))
modExp <- survreg(Surv(tiempoExp, censura) ~ 0+x1+x2, dist="exponential")
summary(modExp)
```

```
##
## Call:
## survreg(formula = Surv(tiempoExp, censura) ~ 0 + x1 + x2, dist = "exponential")
##     Value Std. Error    z      p
## x1 3.1964     0.0759 42.1 <2e-16
## x2 1.1900     0.0614 19.4 <2e-16
##
## Scale fixed at 1
##
```

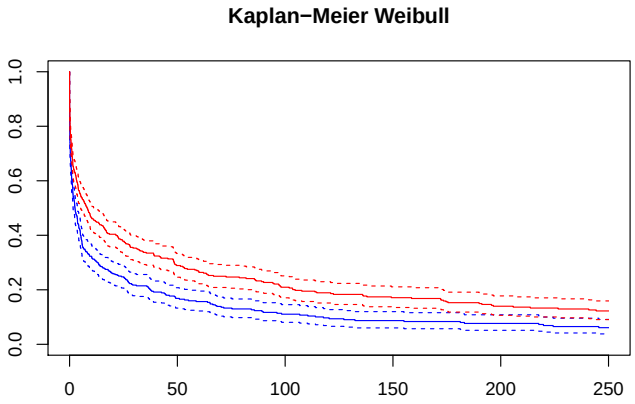
```
## Exponential distribution
## Loglik(model)= -2602.2   Loglik(intercept only)= -2948.4
##  Chisq= 692.35 on 1 degrees of freedom, p= 1.4e-152
## Number of Newton-Raphson Iterations: 5
## n= 1000
```



Simular tiempos de supervivencia de una familia parametrica Weibull

```
sigma = 3   ### parametro para weibull
tiempoWeibull = rweibull(n=N, shape=1/sigma, scale=exp(XBeta))   ### Weibull
ajusteWeibull <- survfit(Surv(tiempoWeibull,censura)~x2, type = "kaplan-meier",
conf.type="log-log", conf.int=0.95)
plot(ajusteWeibull,conf.int=T,main="Kaplan-Meier Weibull",
     col=c("blue","blue","blue","red","red","red"),xlim=c(0,250))
modWeibull <- survreg(Surv(tiempoWeibull, censura) ~ 0+x1+x2, dist="weibull")
summary(modWeibull)
```

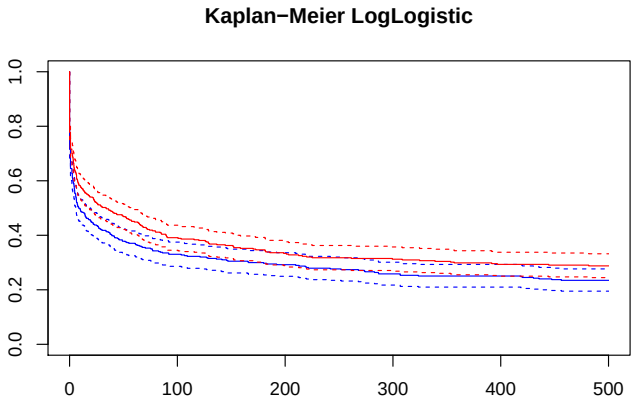
```
##
## Call:
## survreg(formula = Surv(tiempoWeibull, censura) ~ 0 + x1 + x2,
##     dist = "weibull")
##             Value Std. Error     z       p
## x1          3.636      0.224 16.21 < 2e-16
## x2          1.308      0.182  7.17 7.4e-13
## Log(scale) 1.076      0.027 39.80 < 2e-16
##
## Scale= 2.93
##
## Weibull distribution
## Loglik(model)= -2507   Loglik(intercept only)= -2552.3
##  Chisq= 90.66 on 1 degrees of freedom, p= 1.7e-21
## Number of Newton-Raphson Iterations: 5
## n= 1000
```



Simular tiempos de supervivencia de una familia parametrica LogLogistic

```
scale = 3   ### parametro para loglogistic
gamma = (1/scale)   ### parametro para loglogistic
U = runif(n=N)
tiempoLogLogistic = ( (U/(1 - U))^(1/gamma) ) * exp(XBeta)   ### loglogistic
ajusteLogLogistic <- survfit(Surv(tiempoLogLogistic,censura)~x2, type = "kaplan-meier",
conf.type="log-log", conf.int=0.95)
plot(ajusteLogLogistic,conf.int=T,main="Kaplan-Meier LogLogistic",
     col=c("blue","blue","blue","red","red","red"),xlim=c(0,500))
modLogLogistic <- survreg(Surv(tiempoLogLogistic, censura) ~ 0+x1+x2, dist="loglogistic")
summary(modLogLogistic)
```

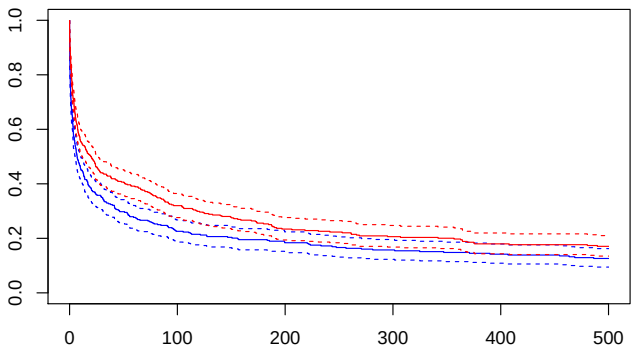
```
##
## Call:
## survreg(formula = Surv(tiempoLogLogistic, censura) ~ 0 + x1 +
##     x2, dist = "loglogistic")
##              Value Std. Error     z       p
## x1          3.6977     0.4041  9.15 < 2e-16
## x2          1.6071     0.3277  4.90 9.4e-07
## Log(scale) 1.1711     0.0292 40.06 < 2e-16
##
## Scale= 3.23
##
## Log logistic distribution
## Loglik(model)= -4175.8   Loglik(intercept only)= -4179.7
##  Chisq= 7.72 on 1 degrees of freedom, p= 0.0055
## Number of Newton-Raphson Iterations: 3
## n= 1000
```



Simular tiempos de supervivencia de una familia parametrica LogNormal

```
scale = 3 ### parametro para lognormal
tiempoLogNormal = rlnorm(n = N, meanlog = XBeta, sdlog = scale)   ### lognormal
ajusteLogNormal <- survfit(Surv(tiempoLogNormal,censura)~x2, type = "kaplan-meier",
conf.type="log-log", conf.int=0.95)
plot(ajusteLogNormal,conf.int=T,main="Kaplan-Meier LogNormal",
     col=c("blue","blue","blue","red","red","red"),xlim=c(0,500))
modLogNormal <- survreg(Surv(tiempoLogNormal, censura) ~ 0+x1+x2, dist="lognormal")
summary(modLogNormal)
```

```
##
## Call:
## survreg(formula = Surv(tiempoLogNormal, censura) ~ 0 + x1 + x2,
##     dist = "lognormal")
##             Value Std. Error     z       p
## x1          3.853      0.228 16.93 < 2e-16
## x2          1.163      0.186  6.24 4.3e-10
## Log(scale) 1.147      0.025 45.80 < 2e-16
##
## Scale= 3.15
##
## Log Normal distribution
## Loglik(model)= -3789.5   Loglik(intercept only)= -3831.9
##  Chisq= 84.75 on 1 degrees of freedom, p= 3.4e-20
## Number of Newton-Raphson Iterations: 3
## n= 1000
```



Usando base de datos CCASAnet

Scripts y bases de datos simulados de análisis y estudios publicados en CCASAnet:

http://biostat.mc.vanderbilt.edu/wiki/Main/ArchivedAnalyses

```
art <- read.csv(paste0(dir,"art_sim.csv"))
basic <- read.csv(paste0(dir,"basic_sim.csv"))
follow <- read.csv(paste0(dir,"follow_sim.csv"))
lab_cd4 <- read.csv(paste0(dir,"lab_cd4_sim.csv"))
lab_rna <- read.csv(paste0(dir,"lab_rna_sim.csv"))
visit <- read.csv(paste0(dir,"visit_sim.csv"))

dim(art)
```

```
## [1] 40516    18
```

```
dim(basic)
```

```
## [1] 25435    19
```

```
dim(follow)
```

```
## [1] 25435     8
```

```
dim(lab_cd4)
```

```
## [1] 233325      5
```

```
dim(lab_rna)
```

```
## [1] 171040      4
```

```
dim(visit)
```

```
## [1] 25435     5
```

```
head(art,3)
```

```
##   patient      site      art_id     art_sd pi nnrti1 nnrti2 nnrti nrti t20 ccr5
## 1    ar.1 argentina 3TC,AZT,NVP 2007-05-16  0      1      0     1    2   0    0
## 2    ar.1 argentina 3TC,AZT,EFV 2007-05-30  0      1      0     1    2   0    0
## 3    ar.1 argentina 3TC,ABC,AZT 2007-08-03  0      0      0     0    3   0    0
##   ii1 ii2 rtv_drug numdrugs art_class     art_ed               art_rs
## 1   0   0       No        3     HAART 2007-05-28 Toxicity Dermatologic
## 2   0   0       No        3     HAART 2007-07-04              Unknown
## 3   0   0       No        3     HAART
```

```r
head(basic,3)
```

```
##   patient      site baseline_d baseline_d_num male      age    birth_d
## 1    ar.1 argentina 2007-04-13          754.5    1 34.16329 1973-02-12
## 2    ar.2 argentina 2010-07-06         1934.5    0 45.89359 1964-08-13
## 3    ar.3 argentina 2011-03-28         2199.5    1 47.23421 1964-01-02
##   hivdiagnosis_d_num hivdiagnosis_d aids_y aids.miss                mode
## 1           3030.601     2007-04-13      9         0   Homosexual contact
## 2          -1270.448     1999-01-07      0         0  Injecting drug user
## 3          -1052.261     1999-08-13      0         0 Heterosexual contact
##   clinicaltrial_y mode_oth recart_y aids_d birth_d_a aids_cl_y aids_cl_d
## 1               0       NA        0               D         9
## 2               0       NA        0               D         0
## 3               0       NA        9               D         0
```

```r
head(follow,3)
```

```
##   patient      site  l_alive_d death_y death_d death_d_a drop_rs drop_rs_oth
## 1    ar.1 argentina 2013-02-04       0                D
## 2   ar.10 argentina 2013-02-13       0                D
## 3  ar.100 argentina 2013-07-12       0                D
```

```r
head(lab_cd4,3)
```

```
##   patient      site      cd4_d cd4_v time
## 1    ar.1 argentina 2007-01-19   405    0
## 2    ar.1 argentina 2008-05-07   490  474
## 3    ar.1 argentina 2009-08-26   238  950
```

```r
head(lab_rna,3)
```

```
##   patient      site      rna_d rna_v
## 1    ar.1 argentina 2007-01-19 74724
## 2    ar.1 argentina 2013-02-04   -40
## 3    ar.1 argentina 2011-11-03   399
```

```r
head(visit,3)
```

```
##   patient      site    visit_d cdcstage whostage
## 1    ar.1 argentina 2007-04-13       NA       NA
## 2    ar.2 argentina 2010-07-06       NA       NA
## 3    ar.3 argentina 2011-03-28       NA       NA
```

Cambiar formato de las fechas.

```r
class(basic$birth_d)
```

```
## [1] "factor"
```

```r
basic$birth_d <- as.Date(basic$birth_d,"%Y-%m-%d")
class(basic$birth_d)
```

```
## [1] "Date"
```

```r
basic$baseline_d <- as.Date(basic$baseline_d,"%Y-%m-%d")
follow$l_alive_d <- as.Date(follow$l_alive_d,"%Y-%m-%d")
visit$visit_d <- as.Date(visit$visit_d,"%Y-%m-%d")
lab_cd4$cd4_d <- as.Date(lab_cd4$cd4_d,"%Y-%m-%d")
```

Frequency of non-communicable diseases in people 50 years of age and older receiving HIV care in Latin America

** Analyses were performed according to age at enrollment into HIV care ($<$ 50yo and $\geq$ 50yo) (years old). Most common NCDs (non-communicable diseases) at the last visit in each age-group at enrollment were dyslipidemia (36% in $<$ 50yo and 28% in $\geq$ 50yo), hypertension (17% and 18%), psychiatric disorders (15% and 10%), and diabetes (11% and 12%). **

Unir bases de datos.

```
datos_bf <- merge(basic,follow, by=c("patient","site"),all=TRUE)
datos_bfv <- merge(datos_bf,visit, by=c("patient","site"),all=TRUE)
datos_bfv4 <- merge(datos_bfv,lab_cd4, by=c("patient","site"), all.x=FALSE,all.y=FALSE,nu.dups=TRUE)

### Edad al momento de la muestra de CD4
datos_bfv4$edad <- as.numeric(difftime(datos_bfv4$cd4_d,datos_bfv4$birth_d, units="auto"))/365.25

patient50agemayor = unique(datos_bfv4$patient[datos_bfv4$edad>=50 & datos_bfv4$age>=50])
patient50agemenor = unique(datos_bfv4$patient[datos_bfv4$edad>=50 & datos_bfv4$age<50])
(N50mayor = n_distinct(patient50agemayor))
```

```
## [1] 2255
```

```
(N50menor = n_distinct(patient50agemenor))
```

```
## [1] 2737
```

Seleccionar muestra de los que cumplen las características.

```
diabetes50mayor = sample(c(1,0),size=N50mayor,replace=TRUE,prob=c(12,88))
table(diabetes50mayor)/N50mayor
```

```
## diabetes50mayor
##         0         1
## 0.8851441 0.1148559
```

```
diabetes50menor = sample(c(1,0),size=N50menor,replace=TRUE,prob=c(11,89))
table(diabetes50menor)/N50menor
```

```
## diabetes50menor
##         0         1
## 0.8801608 0.1198392
```

```
diseases = data.frame(patient=basic$patient, site=basic$site)
diseases$diabetes_y = 0
diseases$diabetes_d = as.Date(NA)
```

Simular la fecha en la que se diagnostica, tomando como base las fechas de CD4.

```
for(i in 1:N50mayor){
  if(diabetes50mayor[i]==1){
    id = patient50agemayor[i]
    datos_id = subset(datos_bfv4,patient==id)
    id.num = which(diseases$patient==as.character(id))
    diseases$diabetes_y[id.num] = 1
    diseases$diabetes_d[id.num] = sample(datos_id$cd4_d,size=1)
  }
}
for(i in 1:N50menor){
  if(diabetes50menor[i]==1){
    id = patient50agemenor[i]
    datos_id = subset(datos_bfv4,patient==id & datos_bfv4$edad>=50)
    id.num = which(diseases$patient==as.character(id))
    diseases$diabetes_y[id.num] = 1
    diseases$diabetes_d[id.num] = sample(datos_id$cd4_d,size=1)
  }
}
table(diseases$diabetes_y)/(N50mayor+N50menor)
```

```
##
##         0         1
## 4.9775641 0.1175881
```

```
head(diseases)
```

```
##   patient      site diabetes_y diabetes_d
## 1    ar.1 argentina          0       <NA>
## 2    ar.2 argentina          0       <NA>
## 3    ar.3 argentina          0       <NA>
## 4    ar.4 argentina          0       <NA>
## 5    ar.5 argentina          0       <NA>
## 6    ar.6 argentina          0       <NA>
```

Increased Mortality After Tuberculosis Treatment Completion in Persons Living With Human Immunodeficiency Virus in Latin America

We assessed the association between cured tuberculosis (TB) and mortality among persons living with human immunodeficiency virus (HIV) in Latin America. We compared survival among persons with and without TB at enrollment in HIV care, starting 9 months after clinic enrollment. In multivariable analysis, TB was associated with higher long-term mortality (hazard ratio, 1.57; 95% confidence interval, 1.25–1.99).

```
datos_bfvd <- merge(datos_bfv,diseases, by=c("patient","site"), all.x=FALSE,all.y=FALSE,nu.dups=TRUE)
N = dim(datos_bfvd)[1]
datos_bfvd$tuberculosis_y = sample(c(1,0),size=N,replace=TRUE,prob=c(6.8,93.2))
```

```
attach(datos_bfvd)

beta_HR_TB = log(1.57)
beta_HR_CD4 = log(1.57)
beta_HR_age = log(1.56)
beta_HR_edu = log(1.24)

CD4count = rnorm(N,mean=227,sd=45) ### median baseline CD4 count was 227 cells/mm3 (IQR,90-386)
lowerCD4count = ifelse(CD4count<227,1,0)
educa = rbinom(N,size=1,prob=0.15) ### patients (15%) had no education or primary school only
ageolder = ifelse(age>35,1,0)

Xbeta = tuberculosis_y*beta_HR_TB + lowerCD4count*beta_HR_CD4 +
  educa*beta_HR_edu + ageolder*beta_HR_age

sigma = 3   ### parametro para weibull
tiempo = rweibull(n=N, shape=1/sigma, scale=exp(XBeta))

muerte = rep(0,N)
muerte[tuberculosis_y==1] = sample(c(1,0),size=sum(tuberculosis_y==1),replace=TRUE,
                                   prob=c(10.2,89.8))
muerte[tuberculosis_y==0] = sample(c(1,0),size=N-sum(tuberculosis_y==1),replace=TRUE,
                                   prob=c(5.6,94.4))

mod_cox <- coxph(Surv(tiempo, muerte) ~ tuberculosis_y + lowerCD4count + educa + ageolder )
summary(mod_cox)
```

```
## Call:
## coxph(formula = Surv(tiempo, muerte) ~ tuberculosis_y + lowerCD4count +
##     educa + ageolder)
##
##   n= 25435, number of events= 1587
##
##                     coef exp(coef)  se(coef)      z Pr(>|z|)
## tuberculosis_y  0.559050  1.749010  0.079586  7.025 2.15e-12 ***
## lowerCD4count   0.004293  1.004302  0.050232  0.085   0.9319
## educa          -0.011898  0.988172  0.070436 -0.169   0.8659
## ageolder        0.090066  1.094246  0.051131  1.761   0.0782 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## tuberculosis_y     1.7490     0.5718    1.4964     2.044
## lowerCD4count      1.0043     0.9957    0.9101     1.108
## educa              0.9882     1.0120    0.8608     1.134
## ageolder           1.0942     0.9139    0.9899     1.210
##
## Concordance= 0.532  (se = 0.008 )
## Likelihood ratio test= 45.92  on 4 df,   p=3e-09
## Wald test            = 52.51  on 4 df,   p=1e-10
## Score (logrank) test = 53.81  on 4 df,   p=6e-11
```

```r
km <- survfit(Surv(tiempo, muerte) ~ tuberculosis_y, type = "kaplan-meier",conf.type="log-log")
plot(km, conf.int=T,xlab="time", ylab="Survival", lty=c(1,4),
     lwd=2, xlim=c(0,300), col=c("blue","blue","blue","red","red","red"))
legend("topright", c("No Baseline TB","Baseline TB"), lty=c(1,4),col=c("blue","red"))

plot(km, conf.int=T,xlab="time", fun="event", ylab="Probability of Death", lty=c(1,4),
     lwd=2, xlim=c(0,300), col=c("blue","blue","blue","red","red","red"))
legend("topright", c("No Baseline TB","Baseline TB"), lty=c(1,4),col=c("blue","red"))
```