Tarea 8

Dalia Camacho

8-Simulación de modelos de regresión

Los datos beauty consisten en evaluaciones de estudiantes a profesores, los estudiantes calificaron belleza y calidad de enseñanza para distintos cursos en la Universidad de Texas. Las evaluaciones de curso se realizaron al final del semestre y tiempo después 6 estudiantes que no llevaron el curso realizaron los juicios de belleza.

Ajustamos el siguiente modelo de regresión lineal usando las variables *edad* (age), *belleza* (btystdave), *sexo* (female) e *inglés no es primera lengua* (nonenglish) para predecir las evaluaciones del curso (courseevaluation).
.

```
library(MASS)
suppressMessages(library(tidyverse))
```

```
beauty    <- readr::read_csv("https://raw.githubusercontent.com/tereom/est-computacional-2018/master/da
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   btystdave = col_double(),
##   btystdf2u = col_double(),
##   btystdfl = col_double(),
##   btystdfu = col_double(),
##   btystdm2u = col_double(),
##   btystdml = col_double(),
##   btystdmu = col_double(),
##   courseevaluation = col_double(),
##   percentevaluating = col_double(),
##   profevaluation = col_double(),
##   btystdvariance = col_double(),
##   btystdavepos = col_double(),
##   btystdaveneg = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
fit_score <- lm(courseevaluation ~ age + btystdave + female + nonenglish,
                data = beauty)
```

1. La instructora A es una mujer de 50 años, el inglés es su primera lengua y tiene un puntaje de belleza de -1. El instructor B es un hombre de 60 años, su primera lengua es el inglés y tiene un puntaje de belleza de -0.5. Simula 1000 generaciones de la evaluación del curso de estos dos instructores. En tus simulaciones debes incorporar la incertidumbre en los parámetros y en la predicción.

Para hacer las simulaciones necesitarás la distribución del vector de coeficientes β, este es normal con media:

```
coef(fit_score)
```

```
##  (Intercept)          age    btystdave       female   nonenglish
##  4.244464824 -0.002585912  0.141031893 -0.210304324 -0.332233708
```

```
mean_coefs <- coef(fit_score)
```

y matriz de varianzas y covarianzas $\sigma^2 V$, donde V es:

```r
summary(fit_score)$cov.unscaled
```

```
##               (Intercept)           age     btystdave        female
## (Intercept)  0.070758980 -1.331151e-03 -3.787757e-03 -1.049379e-02
## age         -0.001331151  2.653270e-05  8.781697e-05  1.324028e-04
## btystdave   -0.003787757  8.781697e-05  3.826989e-03 -2.709254e-04
## female      -0.010493789  1.324028e-04 -2.709254e-04  9.662597e-03
## nonenglish  -0.002199634 -1.791673e-06 -1.206447e-04 -5.576679e-05
##                nonenglish
## (Intercept) -2.199634e-03
## age         -1.791673e-06
## btystdave   -1.206447e-04
## female      -5.576679e-05
## nonenglish   3.801753e-02
```

```r
Cov_params <- summary(fit_score)$cov.unscaled
```

y σ se calcula como $\sigma = \hat{\sigma}\sqrt{(df)/X}$, donde X es una generación de una distribución χ^2 con df (458) grados de libertad $\hat{\sigma}$ es:

```r
summary(fit_score)$sigma
```

```
## [1] 0.5320521
```

```r
Sigma <- summary(fit_score)$sigma
```

y *df* (los grados de libertad) se obtienen:

```r
summary(fit_score)$df[2]
```

```
## [1] 458
```

```r
DF <- summary(fit_score)$df[2]
```

Realizamos la simulaciones para los profesores A y B

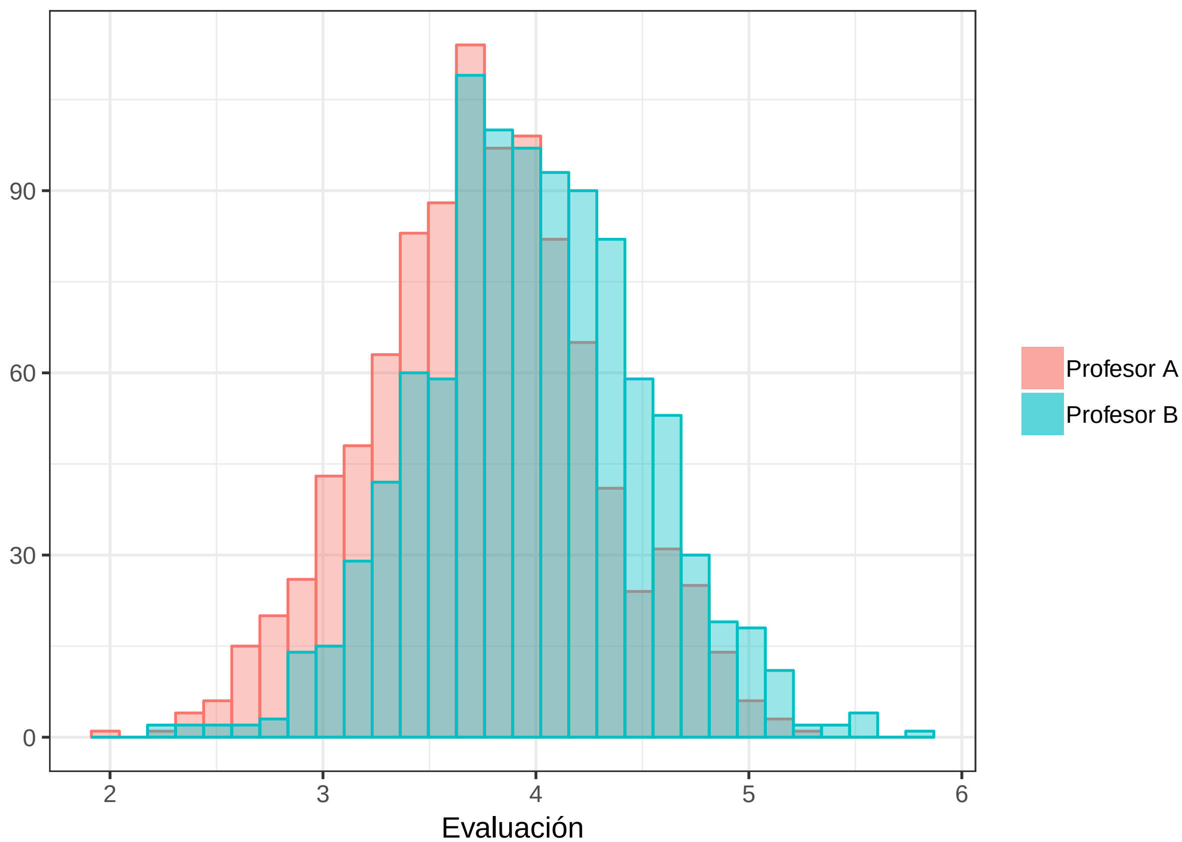
```r
# Fihjamos semilla y número de simulaciones
set.seed(44578)
Nsims     <- 1000

# Definimos las características de los profesores
caract_A  <- c(1,age = 50, btystdave = -1.0, female = 1, nonenglish = 0)
caract_B  <- c(1,age = 60, btystdave = -0.5, female = 0, nonenglish = 0)

# Definimos la simulación de la evaluación de los profesores
sims <- function(){
  Sigma2    <- (Sigma * sqrt((DF) / rchisq(1, DF)))^2
  simParams <- mvrnorm(1, mean_coefs, Sigma2*Cov_params)

  predmuA  <- simParams%*%caract_A
  predsimA <- rnorm(1,predmuA, sqrt(Sigma2))

  predmuB  <- simParams%*%caract_B
  predsimB <- rnorm(1,predmuB, sqrt(Sigma2))
  return(c(predsimA, predsimB))
}

# Simulamos las evaluaciones
```

```r
Sims <- rerun(Nsims, sims())
SimsA <- unlist(Sims)[seq(1,2000,by=2)]
SimsB <- unlist(Sims)[seq(2,2000,by=2)]
```

Una vez que obtengas una simulación del vector β generas simulaciones para los profesores usando el modelo de regresión lineal y las simulaciones de los parámetros.

- Realiza un histograma de la diferencia entre la evaluación del curso para A y B.

```r
ggplot()+theme_bw()+
  geom_histogram( aes(SimsA, col="Profesor A", fill="Profesor A"), alpha=0.4, bins = 30)+
  geom_histogram( aes(SimsB, col="Profesor B", fill="Profesor B"), alpha=0.4, bins = 30)+
  guides(colour=FALSE)+ theme(legend.title = element_blank())+
  xlab("Evaluación")+ylab("")
```



- ¿Cuál es la probabilidad de que A obtenga una calificación mayor?

```r
AmayorB <- length(which(SimsA>SimsB))
AmayorB/Nsims
```

```
## [1] 0.373
```

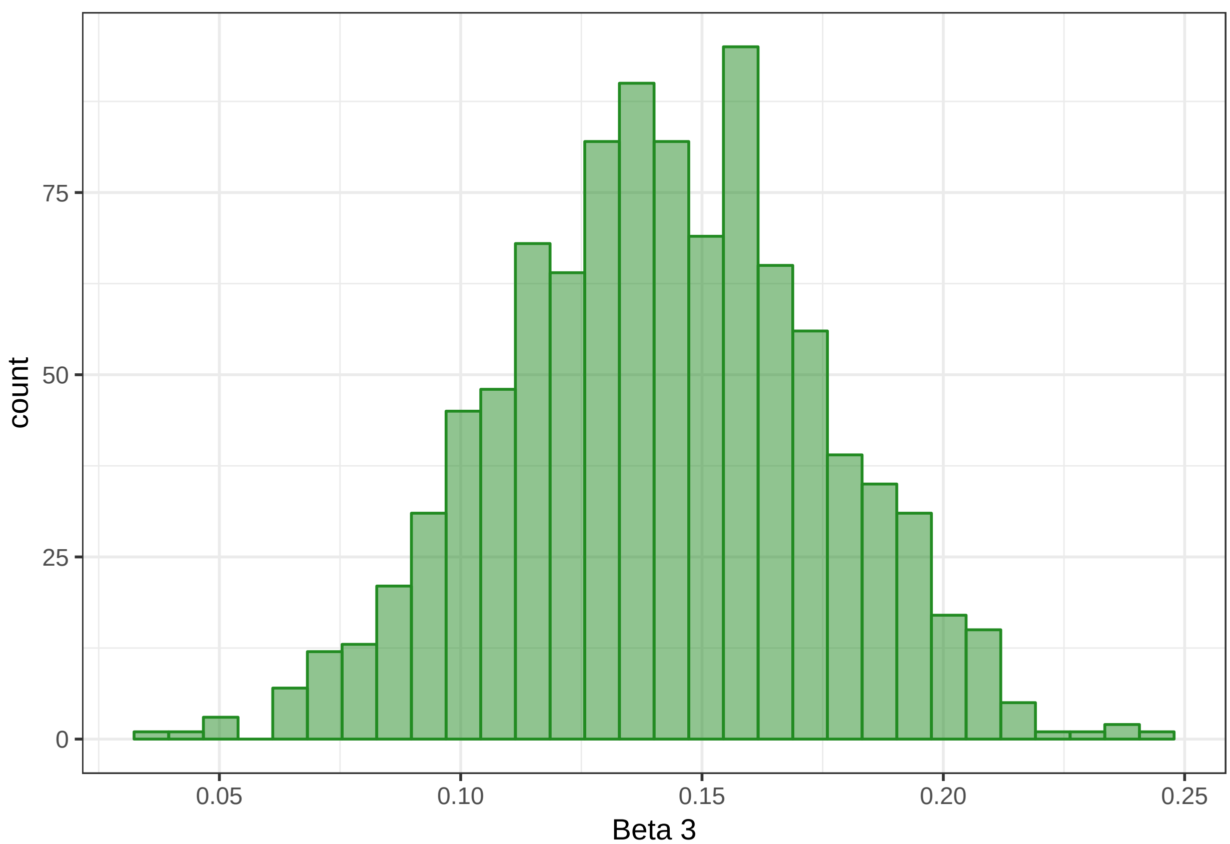
La probabilida de que A obtenga una calificación mayor a B es 0.373.

2. En el inciso anterior obtienes simulaciones de la distribución conjunta $p(\tilde{y}, \beta, \sigma^2)$ donde β es el vector de coeficientes de la regresión lineal. Para este ejercicio nos vamos a enfocar en el coeficiente de belleza (β_3), realiza 6000 simulaciones del modelo (como en el inciso anterior) y guarda las realizaciones de β_3.

- Genera un histograma con las simulaciones de β_3.

```
sims2 <- function(){
  Sigma2    <- (Sigma * sqrt((DF) / rchisq(1, DF)))^2
  simParams <- mvrnorm(1, mean_coefs, Sigma2*Cov_params)
  beta3     <- simParams[3]
  predmuA   <- simParams%*%caract_A
  predsimA <- rnorm(1,predmuA, sqrt(Sigma2))

  predmuB   <- simParams%*%caract_B
  predsimB <- rnorm(1,predmuB, sqrt(Sigma2))
  return(c(predsimA, predsimB, beta3))
}

# Simulamos las evaluaciones
Sims2  <- rerun(Nsims, sims2())
SimsA  <- unlist(Sims2)[seq(1,3000,by=3)]
SimsB  <- unlist(Sims2)[seq(2,3000,by=3)]
Simsb3 <- unlist(Sims2)[seq(3,3000,by=3)]
```

```
ggplot()+theme_bw()+
  geom_histogram(aes(Simsb3), col="forestgreen", fill="forestgreen", alpha=0.5, bins = 30)+
  xlab("Beta 3")
```



- Calcula la media y desviación estándar de las simulaciones y comparalas con la estimación y desviación estándar del coeficiente obtenidas usando summary.

```
summary(fit_score)
```

```
##
```

```
## Call:
## lm(formula = courseevaluation ~ age + btystdave + female + nonenglish,
##     data = beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.87539 -0.35399  0.04531  0.38321  1.02355
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.244465   0.141529  29.990  < 2e-16 ***
## age         -0.002586   0.002741  -0.944  0.34589
## btystdave    0.141032   0.032914   4.285 2.23e-05 ***
## female      -0.210304   0.052300  -4.021 6.77e-05 ***
## nonenglish  -0.332234   0.103740  -3.203  0.00146 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5321 on 458 degrees of freedom
## Multiple R-squared:  0.0885, Adjusted R-squared:  0.08054
## F-statistic: 11.12 on 4 and 458 DF,  p-value: 1.294e-08
```

```r
mean(Simsb3)
```

```
## [1] 0.1414527
```

```r
sd(Simsb3)
```

```
## [1] 0.03284326
```

Tanto el promedio como el error estándar de β_3 son iguales a los de la estimación de los coeficientes en las primeras tres cifras significativas.