Andrés Felipe Paredes

John Sebastián Marín

# Taller en clase: Inferencia y comparación de modelos.

El equipo de data analytics propone 3 modelos para el problema planteado:

1. Shares=Todas las variables
2. Shares=Variables de la 1 a la 12
3. Shares=Variables de la 13 a la 30

Vamos a realizar el análisis comparativo de los modelos.

Primero, observemos que el modelo 2 y el modelo 3 están anidados en el modelo 1.

Por lo tanto, vamos a realizar la prueba F para comparar los modelos anidados.

Al realizar la prueba Anova entre el modelo 1 y el modelo 2 nos retorna un pvalor de casi 0, por lo que rechazamos la hipótesis nula de que el modelo restringido 2 es mejor que el modelo 1. Por lo tanto, el modelo 1 es mejor que el 2.

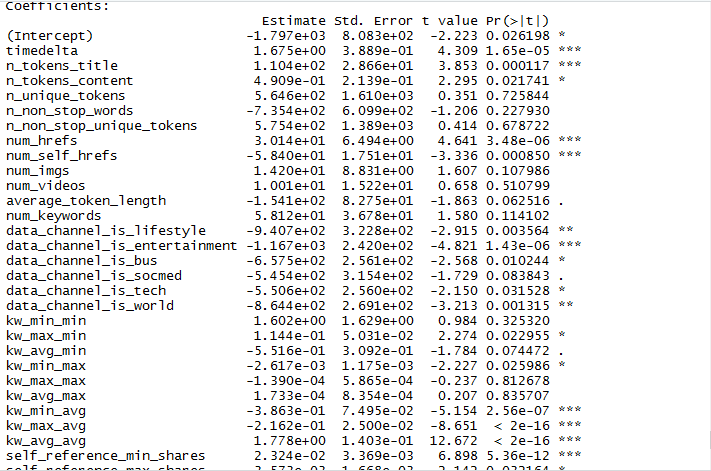
Al realizar la misma prueba entre el modelo 1 y el modelo 3 nos retorna un pvalor de casi 0, por lo que podemos concluir lo mismo que con el modelo 2. El modelo 1 es mejor que el modelo 3.

Sin embargo, después de analizar las variables individuales del modelo 1, observamos que hay varias variables que no afectan a la variable dependiente.

Estas son:

n\_unique\_tokens, n\_non\_stop\_words, n\_non\_stop\_unique\_tokens, num\_imgs, num\_videos, num\_keywords, kw\_min\_min, kw\_max\_max, kw\_avg\_max

Llegamos a esta conclusión revisando los valores p de cada variable.



Eliminando dichas variables podemos proponer un cuarto modelo sin estas variables.

Observemos que el modelo 4 está anidado en el modelo 1, por lo tanto usaremos F para comparar modelos anidados.

Utilizando Anova, la comparación nos retorna un pvalor de 0.4033. Con este pvalor no podemos rechazar la hipótesis nula de que el modelo 4 es mejor que el 1. Por lo tanto, podemos concluir que el modelo 4 es mejor que el modelo 1.

#Taller en clase: Inferencia y comparación de modelos.

data<-read.csv("DatosTallerInf.csv")

head(data,3)

data<-data[,c(-1,-2)]

head(data,3)

str(data)

head(data[1:12],3)

colnames(data[1:12])

fmla <- as.formula(paste("shares ~ ", paste(colnames(data[1:12]), collapse= "+")))

fmla2 <- as.formula(paste("shares~",paste(colnames(data[13:30]),collapse="+")))

res1<-lm(shares~.,data)

res2<-lm(fmla,data)

res3<-lm(fmla2,data)

anova(res2,res1)

#Con base en el pvalor podemos rechazar la hipótesis nula de que 2 es mejor que 1,

#por lo tanto 1 es mejor que 2

anova(res3,res1)

#Misma conclusión de arriba.

# Por lo tanto el modelo 1 aparentemente es el mejor de los 3 propuestos.

summary(res1)

res4<-lm(shares~.-n\_unique\_tokens-n\_non\_stop\_words-n\_non\_stop\_unique\_tokens

-num\_imgs-num\_videos-num\_keywords-kw\_min\_min-kw\_max\_max-kw\_avg\_max,data)

summary(res4)

anova(res4,res1)

#Con base en el pvalor 0.4033 no podemos rechazar la hipótesis nula,

#por lo tanto el modelo 4 es mejor.