ANOVAFACTORIAL

Juan C. Correa

3/15/2021

El análisis de varianza factorial es una extensión del análisis de varianza que estudiamos previamente. Al igual que en el análisis de varianza unifactorial, en el análisis de varianza factorial, se analiza una variable dependiente, pero ahora se incluyen más variables independientes. Acá vamos a analizar la influencia conjunta de dos variables (Entrenamiento y Gerente) sobre una variable dependiente (TCH), de la base de datos que encontramos a continuación

```
setwd("~/Documents/GitHub/Pantaleon")
library(readxl)
AOV <- read_excel("PantaleonDatos2.xls")
```

```
## New names:
## * TOTAL -> TOTAL...25
## * CONTAR -> CONTAR...26
## * edad -> edad...35
## * edad -> edad...41
## * TOTAL -> TOTAL...90
## * ...
```

Dado el objetivo hacia el cual nos dirigimos, resulta fundamental entender cómo cambia la sintaxis que debemos correr para obtener los resultados.

```
resultado_TCH <- aov(AOV$TCH ~ AOV$Entrenamiento + AOV$Gerente)
summary(resultado_TCH)
```

```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## AOV$Entrenamiento     2  34302   17151  41.830 <2e-16 ***
## AOV$Gerente           1   2262    2262   5.516  0.019 *
## Residuals          1526 625677     410
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obsérvese que ahora, tenemos como resultado una tabla dentro de la cual se indica el valor del estadístico F para cada variable independiente. Así por ejemplo, la variable Entrenamiento muestra una influencia estadísticamente significativa sobre TCH ($F = 41.830$, $df = 2$, $p < 2e\text{-}16$). Asimismo, la variable Gerente muestra una influencia estadísticamente significativa sobre TCH ($F = 5.516$, $df = 1$, $p = 0.019$). La sinxtaxis anterior, sin embargo, solo modela la acción separada de cada variable independiente, sin considerar la interación entre ellas. Para considerar la interacción entre estas variables debemos hacer una pequeña modificación a la sintaxis anterior, reemplazando el simbolo + por el simbolo * dentro de los argumentos de la función aov. Obsérvese que en los resultados, ahora se obtendrá una línea adicional donde aparece AOV$Entrenamiento:AOV$Gerente para indicar la interacción entre esas variables. Obsérvese que la estructura de la tabla del anova factorial con el cálculo de interacción entre variables es similar a la tabla del anova factorial sin interacción y del anova unifactorial.

```
resultado_TCH <- aov(AOV$TCH ~ AOV$Entrenamiento * AOV$Gerente)
summary(resultado_TCH)
```

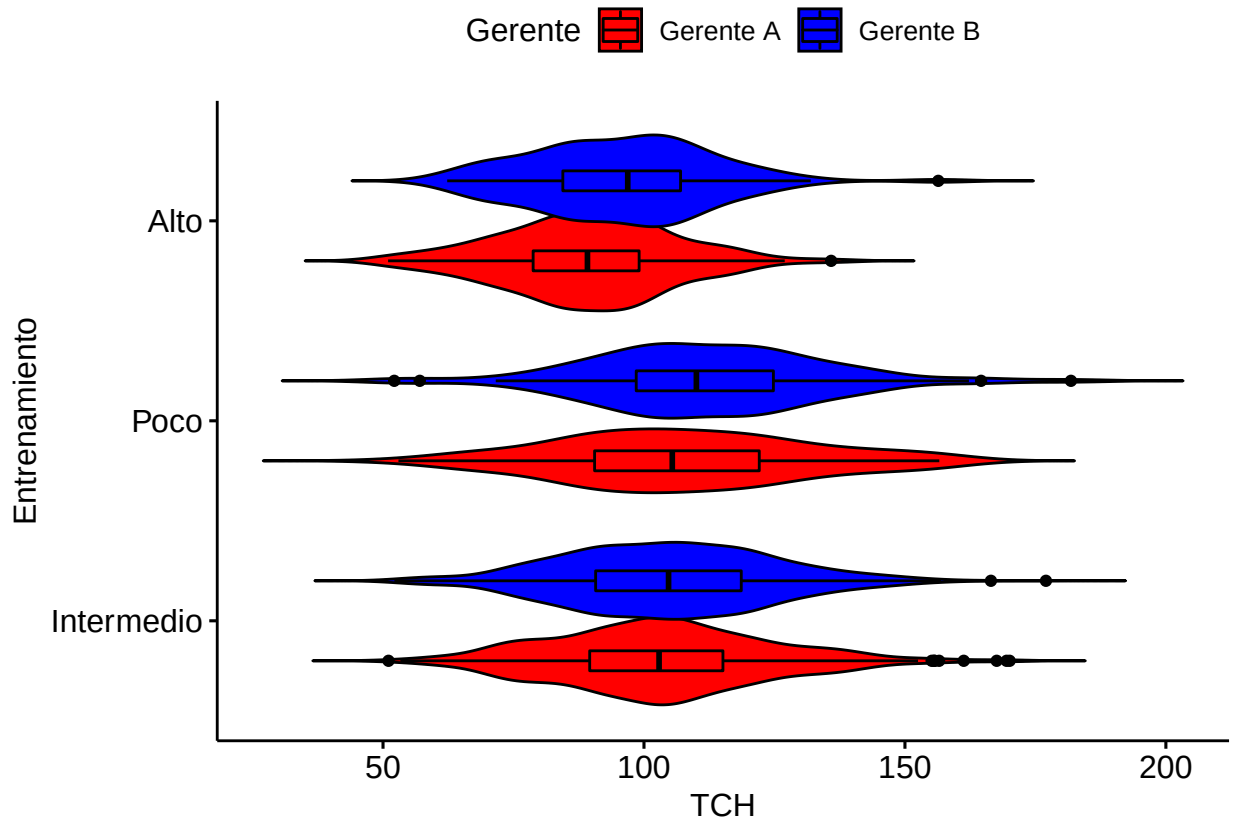
```
##                                  Df Sum Sq Mean Sq F value Pr(>F)
## AOV$Entrenamiento                 2  34302   17151  41.863 <2e-16 ***
## AOV$Gerente                       1   2262    2262   5.520 0.0189 *
## AOV$Entrenamiento:AOV$Gerente     2   1303     652   1.591 0.2041
## Residuals                      1524 624373     410
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ahora, en orden de importancia, según su valor F, Entrenamiento es el factor con mayor efecto sobre TCH, seguido por Gerente, mientras que la interacción entre Gerente y Entrenamiento no resulta significativa.

```
library("ggpubr")
```

```
## Loading required package: ggplot2
```

```
ggviolin(AOV, x = "Entrenamiento", y = "TCH", merge = TRUE, fill = "Gerente",
         orientation="horiz",
         palette = c("red", "blue"),
         add = "boxplot", color = "black")
```




```
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
group_by(AOV, Entrenamiento, Gerente) %>%
  summarise(
    count = n(),
    mean = mean(TCH, na.rm = TRUE),
    sd = sd(TCH, na.rm = TRUE)
  )
```

```
## `summarise()` has grouped output by 'Entrenamiento'. You can override using the `.groups` argument.
```

```
## # A tibble: 6 x 5
## # Groups:   Entrenamiento [3]
##   Entrenamiento Gerente   count  mean    sd
##   <chr>         <chr>     <int> <dbl> <dbl>
## 1 Alto          Gerente A   113  89.1  16.4
## 2 Alto          Gerente B    96  95.5  17.1
## 3 Intermedio    Gerente A   563 103.   20.7
## 4 Intermedio    Gerente B   578 105.   20.1
## 5 Poco          Gerente A    89 108.   23.9
## 6 Poco          Gerente B    91 112.   22.0
```

```
model.tables(resultado_TCH, type="means", se = TRUE)
```

```
## Design is unbalanced - use se.contrast() for se's
```

```
## Tables of means
## Grand mean
##
## 102.9087
##
##  AOV$Entrenamiento
##     Alto Intermedio  Poco
##       92      103.8 109.8
## rep  209     1141.0 180.0
##
##  AOV$Gerente
##     Gerente A Gerente B
##         101.7     104.1
## rep     765.0     765.0
##
##  AOV$Entrenamiento:AOV$Gerente
##                   AOV$Gerente
## AOV$Entrenamiento Gerente A Gerente B
##        Alto         89.1      95.5
##        rep         113.0      96.0
##        Intermedio 103.1     104.5
##        rep         563.0     578.0
##        Poco        107.5     112.0
##        rep          89.0      91.0
```

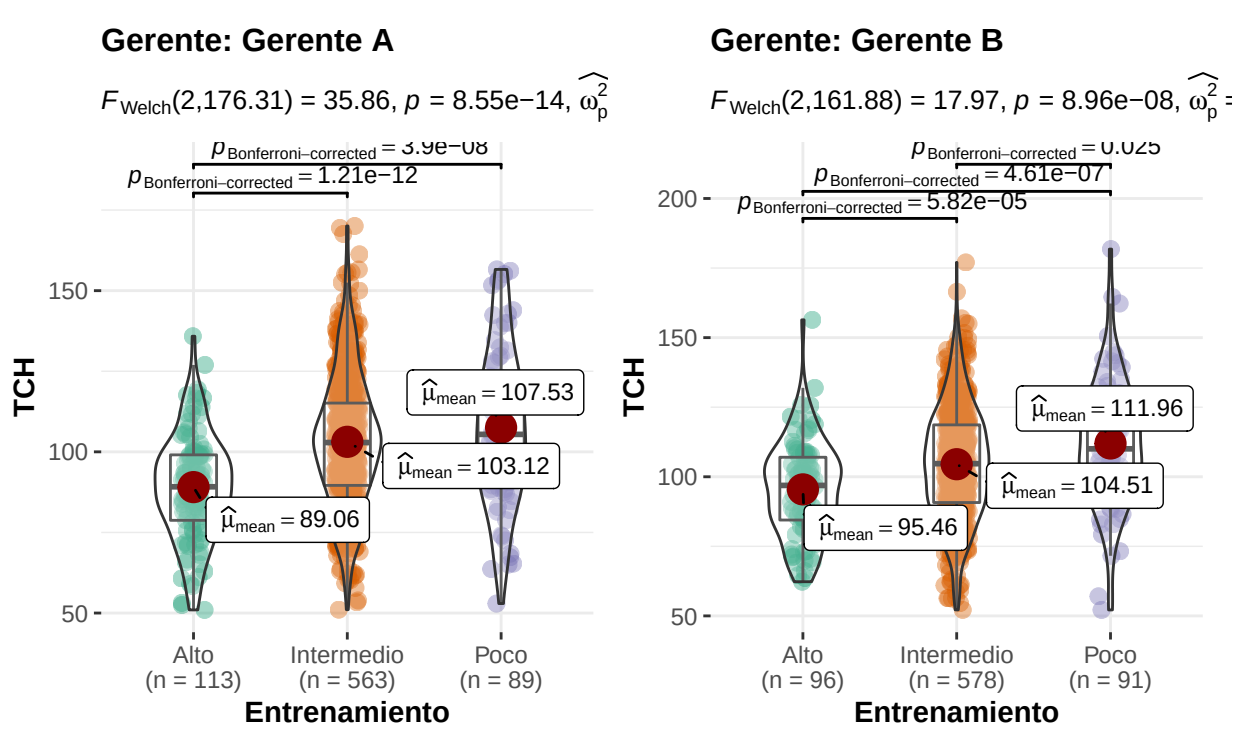
```
library(ggstatsplot)
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
```

```
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
## In case you would like cite this package, cite it as:
##      Patil, I. (2018). ggstatsplot: "ggplot2" Based Plots with Statistical Details. CRAN.
##      Retrieved from https://cran.r-project.org/web/packages/ggstatsplot/index.html
```

```
grouped_ggbetweenstats(data =AOV,
                       x = Entrenamiento, y = TCH,
                       grouping.var = Gerente,
                       p.adjust.method = "bonferroni")
```



20.33, $\widehat{R^2}^{\text{posterior}}_{\text{Bayesian}}$ = 0.06, $\text{CI}^{\text{HDI}}_{95\%}$ [0.03, 0.09], $r^{\text{JZS}}_{\text{Cauchy}}$ (B = 0.71 −10.81, $\widehat{R^2}^{\text{posterior}}_{\text{Bayesian}}$ = 0.04, $\text{CI}^{\text{HDI}}_{95\%}$ [0.02, 0.07], $r^{\text{JZS}}_{\text{Cauchy}}$ = 0.71

mes–Howell test; Comparisons shown: **only significant** **mes–Howell test**; Comparisons shown: **only significant**

```
TukeyHSD(resultado_TCH)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = AOV$TCH ~ AOV$Entrenamiento * AOV$Gerente)
##
## $`AOV$Entrenamiento`
##                       diff       lwr       upr     p adj
## Intermedio-Alto 11.822393  8.249598 15.395188 0.0000000
## Poco-Alto       17.771771 12.943156 22.600386 0.0000000
## Poco-Intermedio  5.949378  2.141093  9.757664 0.0007477
##
## $`AOV$Gerente`
```

```
##                         diff       lwr      upr     p adj
## Gerente B-Gerente A 2.430277 0.4002254 4.460329 0.0189898
##
## $`AOV$Entrenamiento:AOV$Gerente`
##                                                 diff         lwr       upr
## Intermedio:Gerente A-Alto:Gerente A        14.055756   8.1023782 20.009133
## Poco:Gerente A-Alto:Gerente A              18.473834  10.2887159 26.658953
## Alto:Gerente B-Alto:Gerente A               6.401615  -1.6148294 14.418060
## Intermedio:Gerente B-Alto:Gerente A        15.451589   9.5111390 21.392040
## Poco:Gerente B-Alto:Gerente A              22.901423  14.7667766 31.036069
## Poco:Gerente A-Intermedio:Gerente A         4.418078  -2.1699900 11.006147
## Alto:Gerente B-Intermedio:Gerente A        -7.654140 -14.0314338 -1.276847
## Intermedio:Gerente B-Intermedio:Gerente A   1.395834  -2.0240254  4.815693
## Poco:Gerente B-Intermedio:Gerente A         8.845667   2.3204120 15.370922
## Alto:Gerente B-Poco:Gerente A             -12.072219 -20.5706497 -3.573788
## Intermedio:Gerente B-Poco:Gerente A        -3.022245  -9.5986337  3.554144
## Poco:Gerente B-Poco:Gerente A               4.427589  -4.1824293 13.037606
## Intermedio:Gerente B-Alto:Gerente B         9.049974   2.6847469 15.415201
## Poco:Gerente B-Alto:Gerente B              16.499807   8.0499770 24.949638
## Poco:Gerente B-Intermedio:Gerente B         7.449833   0.9363704 13.963296
##                                                p adj
## Intermedio:Gerente A-Alto:Gerente A       0.0000000
## Poco:Gerente A-Alto:Gerente A             0.0000000
## Alto:Gerente B-Alto:Gerente A             0.2034279
## Intermedio:Gerente B-Alto:Gerente A       0.0000000
## Poco:Gerente B-Alto:Gerente A             0.0000000
## Poco:Gerente A-Intermedio:Gerente A       0.3941369
## Alto:Gerente B-Intermedio:Gerente A       0.0082861
## Intermedio:Gerente B-Intermedio:Gerente A 0.8536228
## Poco:Gerente B-Intermedio:Gerente A       0.0015941
## Alto:Gerente B-Poco:Gerente A             0.0007526
## Intermedio:Gerente B-Poco:Gerente A       0.7788959
## Poco:Gerente B-Poco:Gerente A             0.6852852
## Intermedio:Gerente B-Alto:Gerente B       0.0007414
## Poco:Gerente B-Alto:Gerente B             0.0000004
## Poco:Gerente B-Intermedio:Gerente B       0.0142936
```

Un aspecto que no se ha considerado acá es el relativo al número de observaciones que hay por cada combinación de Gerente y Entrenamiento. Dado que en este caso, el número es desigual, estamos en presencia de un diseño no balanceado y esto nos lleva a introducir una pequeña modificación en el cálculo del análisis de varianza al especificar que se trata de un diseño de análisis de varianza tipo III

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
Anova(resultado_TCH, type = "III")
```

```
## Anova Table (Type III tests)
##
```

```
## Response: AOV$TCH
##                                Sum Sq   Df   F value    Pr(>F)
## (Intercept)                    896298    1 2187.7267 < 2.2e-16 ***
## AOV$Entrenamiento               22195    2   27.0873 2.757e-12 ***
## AOV$Gerente                      2127    1    5.1919   0.02283 *
## AOV$Entrenamiento:AOV$Gerente    1303    2    1.5906   0.20415
## Residuals                      624373 1524
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obsérvese que ahora el valor de la F para entrenamiento ha cambiado considerablemente, solo en función del tipo de cálculo que se ha desarrollado para este ejercicio. Una explicación sobre los tipos de cálculos existentes para los análisis de varianza factorial puede encontrarse [aquí]: ([https://towardsdatascience.com/anovas-three-types-of-estimating-sums-of-squares-don-t-make-the-wrong-choice-91107c77a27a])

Finalmente, y al igual que lo hicimos con el caso del análisis de varianza unidireccional, debemos volver a chequear el supuesto de homoscedasticidad, aplicando las herramientas de cálculos ofrecidas por la prueba de Breusch-Pagan y de varianza no constante.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(resultado_TCH)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  resultado_TCH
## BP = 18.32, df = 5, p-value = 0.002571
```

```
library(car)
ncvTest(lm(AOV$TCH ~ AOV$Entrenamiento*AOV$Gerente))
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 14.62692, Df = 1, p = 0.00013103
```

Con estos resultados, volvemos a comprobar que el modelo no se ajusta y los cálculos obtenidos están sesgados, de modo pues que hace falta aplicar alguna otra técnica. Las variantes de regresión múltiple, especialmente regresión ordinal, regresión kernel, regresión quantil. Estudiaremos algunas de esas técnicas más adelante. Por lo pronto, podríamos calcular un análisis de varianza robusto.

```
library(WRS2)
t1way(AOV$TCH ~ AOV$Entrenamiento*AOV$Gerente)
```

```
## Call:
## t1way(formula = AOV$TCH ~ AOV$Entrenamiento * AOV$Gerente)
##
## Test statistic: F = 44.5781
## Degrees of freedom 1: 2
## Degrees of freedom 2: 204.91
```

```
## p-value: 0
##
## Explanatory measure of effect size: 0.43
## Bootstrap CI: [0.32; 0.5]
```