Metodología Cuantitativa Avanzada I

Comparación de modelos GLM y Correlaciones

Carrasco, D., PhD

Centro de Medición MIDE UC

PSI4035

Santiago, Marzo 23 de 2022

Centro UC
Medición - MIDE

Taller

Comparación de modelos

Evalación global de los modelos

Comparación de modelos

Vik (2014, p50) emplea la siguiente tabla para evaluar al modelo ajusstado

- El estadístico F es un "ratio"
- Es la relación entre los MS (i.e., *mean square*) de cada modelo ajustado.
- Cada uno de estos MS proviene de la suma de cuadrados.

- Es la division de los SS por los grado de libertad (*df*).

- El modelo compacto o nulo, posee una suma de cuadrados (SS) de 80
    - esta es la variabilidad de las observaciones con respecto a la media.

- El modelo aumentado o modelo con un predictor, posee una suma de cuadrados de 15.617
    - esta es la variabilidad que nos queda, luego de restar los valores esperados sobre los valores observados (ver Tabla 3.3)

- El error explicado, es 64.383, porque es lo que logro reducirse en error

Tabla 4.7

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

- Mientras más grande sea F, esto significa que hay más varianza explicada, que varianza por explicar.
- La distribución muestral F (i.e., la forma que sigue la distribución) es asimétrica positiva.
    - Esto quiere decir que, a mayores valores de F, uno espera que las chances de observar valores F de gran tamaño sea muy pequeña. Si nuestro valor F observado posee chances menores a %5 (por ejemplo), es convencional afirmar que nuestros resultados estan por sobre el azar.
    - En otras palabras, que nuestra distribucion de F, en muy pocas ocasiones genera nuestros datos observados. Y por tanto, podemos rechazar la hipotesis nula (ver Huck, 2012).
    - Con lo anterior, planteamos que nuestro modelo ajusta a los datos con tal o cual R2.

Taller

Comparación de modelos

Reajustemos los modelos

Abrir y preparar datos

```
#-------------------------------------------------------------------
# datos
#-------------------------------------------------------------------

#-----------------------------------------------------------
# tabla 3.2
#-----------------------------------------------------------

data_table_3_2 <- read.table(
text="
person    y    x    x_q   xy   z
1         2    8     64   16   1
2         3    9     81   27   2
3         3    9     81   27   1
4         4   10    100   40   2
5         7    6     36   42   1
6         5    7     49   35   2
7         5    4     16   20   1
8         7    5     25   35   2
9         8    3      9   24   1
10        9    1      1    9   2
11        9    2      4   18   1
12       10    2      4   20   2

",
header=TRUE, stringsAsFactors = FALSE)

# Nota: agregamos a la variable z,
#       para ilustrar como se ve un
#       modelo que no explica a y.

#-----------------------------------------------------------
# preparar datos
#-----------------------------------------------------------

data_model <- data_table_3_2 %>%
              mutate(x_g = mean(x, na.rm = TRUE)) %>%
              mutate(x_cgm = x - x_g) %>%
              dplyr::select(y, x, x_cgm, z)
```

Ajustar modelos

```
#-------------------------------------------------------------------
# datos
#-------------------------------------------------------------------

#-------------------------------------
# formulas
#-------------------------------------

f00 <- as.formula(y ~ + 1)
f01 <- as.formula(y ~ + 1 + x)
f02 <- as.formula(y ~ + 1 + x_cgm)
f03 <- as.formula(y ~ + 1 + z)

#-------------------------------------
# ajustar modelos
#-------------------------------------

m00 <- lm(f00, data = data_model)
m01 <- lm(f01, data = data_model)
m02 <- lm(f02, data = data_model)
m03 <- lm(f03, data = data_model)

#-------------------------------------
# comparar modelos de forma sintética
#-------------------------------------

texreg::screenreg(
    list(m00, m01, m02, m03),
    star.symbol = "*",
    center = TRUE,
    doctype = FALSE,
    dcolumn = TRUE,
    booktabs = TRUE,
    single.row = FALSE
    )
```

Ajustar modelos

```
#-------------------------------------------------------------
# datos
#-------------------------------------------------------------

#--------------------------------------
# formulas
#--------------------------------------

f00 <- as.formula(y ~ + 1)
f01 <- as.formula(y ~ + 1 + x)
f02 <- as.formula(y ~ + 1 + x_cgm)
f03 <- as.formula(y ~ + 1 + z)

#--------------------------------------
# ajustar modelos
#--------------------------------------

m00 <- lm(f00, data = data_model)
m01 <- lm(f01, data = data_model)
m02 <- lm(f02, data = data_model)
m03 <- lm(f03, data = data_model)

#--------------------------------------
# comparar modelos de forma sintética
#--------------------------------------

texreg::screenreg(
    list(m00, m01, m02, m03),
    star.symbol = "*",
    center = TRUE,
    doctype = FALSE,
    dcolumn = TRUE,
    booktabs = TRUE,
    single.row = FALSE
    )
```

Resultados de los modelos

```
=================================================
             Model 1    Model 2    Model 3    Model 4
-------------------------------------------------
(Intercept)   6.00 ***  10.27 ***   6.00 ***   5.00
              (0.78)     (0.76)     (0.36)     (2.56)
x                        -0.78 ***
                         (0.12)
x_cgm                               -0.78 ***
                                    (0.12)
z                                                0.67
                                                 (1.62)
-------------------------------------------------
R^2            0.00       0.80       0.80       0.02
Adj. R^2       0.00       0.79       0.79      -0.08
Num. obs.     12         12         12         12
=================================================
 *** p < 0.001; ** p < 0.01; * p < 0.05
```

El modelo 3 de la tabla anterior, llamado `m02` en nuestro código, es nuestro modelo de interes. Este fue ajustado con la formula `y ~ + 1 + x_cgm`. En la siguiente lámina vamos a aplicar la prueba de ANOVA o prueba F, para realizar una evaluación global.

Evaluación Global del Modelo

```
#-----------------------------------------------------------
# tabla 4.7 evaluación global
#-----------------------------------------------------------

# tabla F de modelo aumentado
anova(m02)

# R2 del modelo
summary(m02)$r.squared
```

```
# tabla F de modelo aumentado
> anova(m02)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value     Pr(>F)
x_cgm      1 64.383  64.383  41.227 0.00007627 ***
Residuals 10 15.617   1.562
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # R2 del modelo
> summary(m02)$r.squared
[1] 0.8047897
```

Tabla 4.7

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

Centro UC
Medición - MIDE

Comparación entre Modelos

Tabla 4.7

```
#-----------------------------------------------------------
# tabla 4.7 comparación de modelos (Vik, 2014, p50)
#-----------------------------------------------------------

# tabla F de la comparación de modelos

anova(m00, m02)

# R2 del modelo
summary(m02)$r.squared
```

```
# tabla F de modelo aumentado
>anova(m00,  m02)
Analysis of Variance Table

Model 1: y ~ +1
Model 2: y ~ +1 + x_cgm
  Res.Df    RSS Df Sum of Sq      F     Pr(>F)
1     11 80.000
2     10 15.617  1    64.383 41.227 0.00007627 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # R2 del modelo
> summary(m02)$r.squared
[1] 0.8047897
```

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

Taller

Evaluación Global

Donde se encuentra cada componente en el output de R

Evaluación Global

```
#-------------------------------------------------------
# tabla 4.7 evaluación global
#-------------------------------------------------------

# tabla F de modelo aumentado
anova(m02)

# R2 del modelo
summary(m02)$r.squared
```

```
# tabla F de modelo aumentado
> anova(m02)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value     Pr(>F)
x_cgm      1 64.383  64.383  41.227 0.00007627 ***
Residuals 10 15.617   1.562
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # R2 del modelo
> summary(m02)$r.squared
[1] 0.8047897
```

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

Evaluación Global

```
#-------------------------------------------------------
# tabla 4.7 evaluación global
#-------------------------------------------------------

# tabla F de modelo aumentado
anova(m02)

# R2 del modelo
summary(m02)$r.squared
```

```
# tabla F de modelo aumentado
> anova(m02)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value     Pr(>F)
x_cgm      1 64.383  64.383  41.227 0.00007627 ***
Residuals 10 15.617   1.562
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # R2 del modelo
> summary(m02)$r.squared
[1] 0.8047897
```

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

Evaluación Global

```
#-------------------------------------------------------
# tabla 4.7 evaluación global
#-------------------------------------------------------

# tabla F de modelo aumentado
anova(m02)

# R2 del modelo
summary(m02)$r.squared
```

```
# tabla F de modelo aumentado
> anova(m02)
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value     Pr(>F)
x_cgm       1 64.383  64.383  41.227 0.00007627 ***
Residuals  10 15.617   1.562
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # R2 del modelo
> summary(m02)$r.squared
[1] 0.8047897
```

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

Evaluación Global

```
#-------------------------------------------------------
# tabla 4.7 evaluación global
#-------------------------------------------------------

# tabla F de modelo aumentado
anova(m02)

# R2 del modelo
summary(m02)$r.squared
```

```
# tabla F de modelo aumentado
> anova(m02)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value     Pr(>F)
x_cgm      1 64.383  64.383  41.227 0.00007627 ***
Residuals 10 15.617   1.562
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # R2 del modelo
> summary(m02)$r.squared
[1] 0.8047897
```

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

Evaluación Global

```
#-------------------------------------------------------
# tabla 4.7 evaluación global
#-------------------------------------------------------

# tabla F de modelo aumentado
anova(m02)

# R2 del modelo
summary(m02)$r.squared
```

```
# tabla F de modelo aumentado
> anova(m02)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value     Pr(>F)
x_cgm      1 64.383  64.383  41.227 0.00007627 ***
Residuals 10 15.617   1.562
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # R2 del modelo
> summary(m02)$r.squared
[1] 0.8047897
```

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

Evaluación Global

```
#-------------------------------------------------------
# tabla 4.7 evaluación global
#-------------------------------------------------------

# tabla F de modelo aumentado
anova(m02)

# R2 del modelo
summary(m02)$r.squared
```

```
# tabla F de modelo aumentado
> anova(m02)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value     Pr(>F)
x_cgm      1 64.383  64.383  41.227 0.00007627 ***
Residuals 10 15.617   1.562
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # R2 del modelo
> summary(m02)$r.squared
[1] 0.8047897
```

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

Comparación de modelos

```
#-----------------------------------------------------------
# tabla 4.7 comparación de modelos (Vik, 2014, p50)
#-----------------------------------------------------------

# tabla F de la comparación de modelos

anova(m00, m02)

# R2 del modelo
summary(m02)$r.squared
```

```
# tabla F de modelo aumentado
>anova(m00,  m02)
Analysis of Variance Table

Model 1: y ~ +1
Model 2: y ~ +1 + x_cgm
  Res.Df    RSS Df Sum of Sq      F     Pr(>F)
1     11 80.000
2     10 15.617  1    64.383 41.227 0.00007627 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # R2 del modelo
> summary(m02)$r.squared
[1] 0.8047897
```

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

Taller

Comparación de Modelos

Donde se encuentra cada componente en el output de R

Comparación de modelos

```
#-----------------------------------------------------------
# tabla 4.7 comparación de modelos (Vik, 2014, p50)
#-----------------------------------------------------------

# tabla F de la comparación de modelos

anova(m00, m02)

# R2 del modelo
summary(m02)$r.squared
```

```
# tabla F de modelo aumentado
>anova(m00,  m02)
Analysis of Variance Table

Model 1: y ~ +1
Model 2: y ~ +1 + x_cgm
  Res.Df    RSS Df Sum of Sq      F     Pr(>F)
1     11 80.000
2     10 15.617  1    64.383 41.227 0.00007627 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # R2 del modelo
> summary(m02)$r.squared
[1] 0.8047897
```

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

Comparación de modelos

```
#-----------------------------------------------------------
# tabla 4.7 comparación de modelos (Vik, 2014, p50)
#-----------------------------------------------------------

# tabla F de la comparación de modelos

anova(m00, m02)

# R2 del modelo
summary(m02)$r.squared
```

```
# tabla F de modelo aumentado
>anova(m00,  m02)
Analysis of Variance Table

Model 1: y ~ +1
Model 2: y ~ +1 + x_cgm
  Res.Df    RSS Df Sum of Sq      F     Pr(>F)
1     11 80.000
2     10 15.617  1    64.383 41.227 0.00007627 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # R2 del modelo
> summary(m02)$r.squared
[1] 0.8047897
```

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

Comparación de modelos

```
#-----------------------------------------------------------
# tabla 4.7 comparación de modelos (Vik, 2014, p50)
#-----------------------------------------------------------

# tabla F de la comparación de modelos

anova(m00, m02)

# R2 del modelo
summary(m02)$r.squared
```

```
# tabla F de modelo aumentado
>anova(m00,  m02)
Analysis of Variance Table

Model 1: y ~ +1
Model 2: y ~ +1 + x_cgm
  Res.Df    RSS Df Sum of Sq      F     Pr(>F)
1     11 80.000
2     10 15.617  1    64.383 41.227 0.00007627 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # R2 del modelo
> summary(m02)$r.squared
[1] 0.8047897
```

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

Comparación de modelos

```
#-----------------------------------------------------------
# tabla 4.7 comparación de modelos (Vik, 2014, p50)
#-----------------------------------------------------------

# tabla F de la comparación de modelos

anova(m00, m02)

# R2 del modelo
summary(m02)$r.squared
```

```
# tabla F de modelo aumentado
>anova(m00,  m02)
Analysis of Variance Table

Model 1: y ~ +1
Model 2: y ~ +1 + x_cgm
  Res.Df    RSS Df Sum of Sq      F     Pr(>F)
1     11 80.000
2     10 15.617  1    64.383 41.227 0.00007627 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # R2 del modelo
> summary(m02)$r.squared
[1] 0.8047897
```

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

Comparación de modelos

```
#-----------------------------------------------------------
# tabla 4.7 comparación de modelos (Vik, 2014, p50)
#-----------------------------------------------------------

# tabla F de la comparación de modelos

anova(m00, m02)

# R2 del modelo
summary(m02)$r.squared
```

```
# tabla F de modelo aumentado
>anova(m00,  m02)
Analysis of Variance Table

Model 1: y ~ +1
Model 2: y ~ +1 + x_cgm
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     11 80.000
2     10 15.617  1    64.383 41.227 0.00007627 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # R2 del modelo
> summary(m02)$r.squared
[1] 0.8047897
```

Table 4.7 Summary Table

Source	*SS*	*df*	*MS*	*F*	R^2
Model comparison	64.383	1	64.383	41.218	.805
Residual	15.617	10	1.562		
Total	80	11			

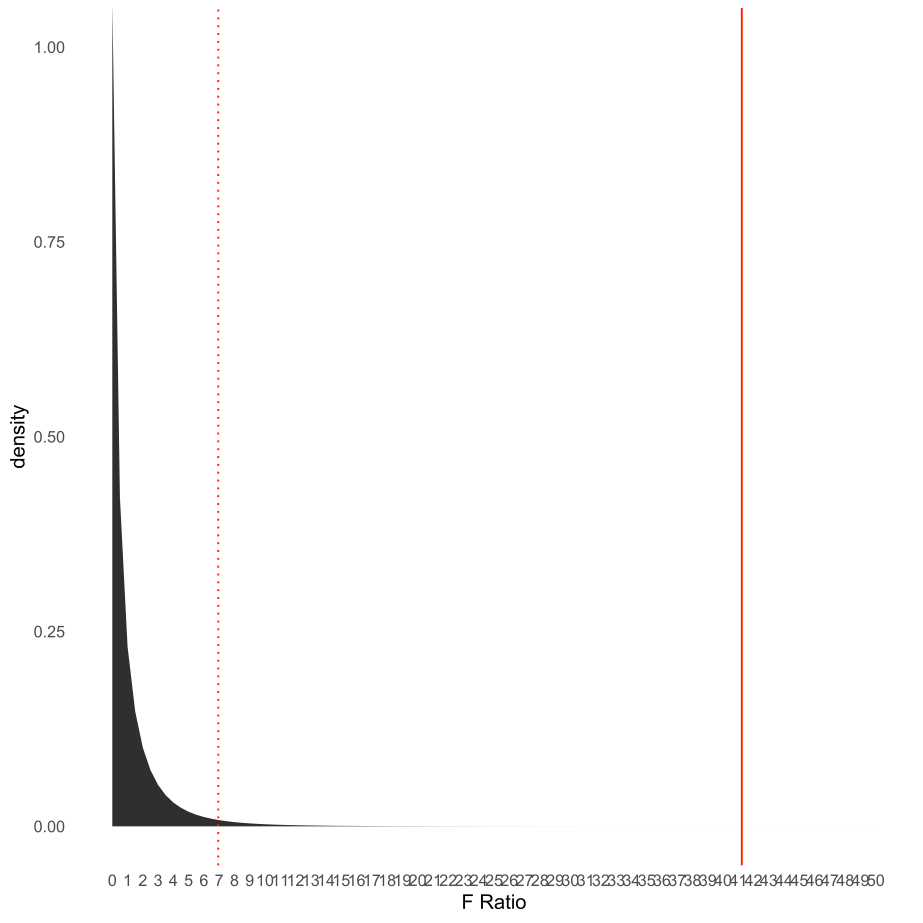
NOTE: SSR is the reduction in error (SST – SSE_A), SST is the error of Model C, and SSE_A is the error of Model A.

Taller

¿Qué tan típico es nuestro modelo?

Cuáles son las chances de los resultados que estamos observando

```r
#-------------------------------------------------------------------
# p value de la comparación modelos (m00, m02)
#-------------------------------------------------------------------

# opciones de consola
options(scipen = 999)
options(digits = 7)

# valor p de la comparación de modelos
anova(m00, m02) %>%
broom::tidy() %>%
knitr::kable(., digits = 7)

# ------------------------------------------------
# f value
# ------------------------------------------------

f_value <- anova(m00, m02) %>%
           broom::tidy() %>%
           mutate(model = c('compact', 'augmented')) %>%
           dplyr::filter(model == 'augmented') %>%
           dplyr::select(statistic) %>%
           pull() %>%
           as.numeric()

# ------------------------------------------------
# p value
# ------------------------------------------------

df_1 <- 1  # cantidad de parámetros fijos del modelo
df_2 <- 10 # grados de libertad restantes (n_total - df_1 - 1)

pf(f_value, df1 = df_1, df2 = df_2, lower.tail = FALSE)

# ------------------------------------------------
# f critic
# ------------------------------------------------

f_critic <- qf(.975, df1 = df_1, df2 = df_2)

# ------------------------------------------------
# visualization
# ------------------------------------------------

library(ggplot2)
f_m02 <-  ggplot(data.frame(x = c(0, 50)), aes(x)) +
  stat_function(fun = df, args = list(df1 = df_1, df2 = df_2), geom
  geom_vline(xintercept = f_value, color = 'red') +
  geom_vline(xintercept = f_critic, color = 'red', linetype = 'dotte
  scale_x_continuous(breaks=seq(0, 50, 1)) +
  # xlim(0,10) +
  ylim(0,1) +
  labs(
    x = 'F Ratio',
    y = 'density') +
  theme_minimal() +
  theme(
  panel.background = element_blank(),
  panel.grid.minor = element_blank(),
  panel.grid.major = element_blank(),
  axis.text.x=element_text(size=1)
  )

# visualize plot
f_m02
```



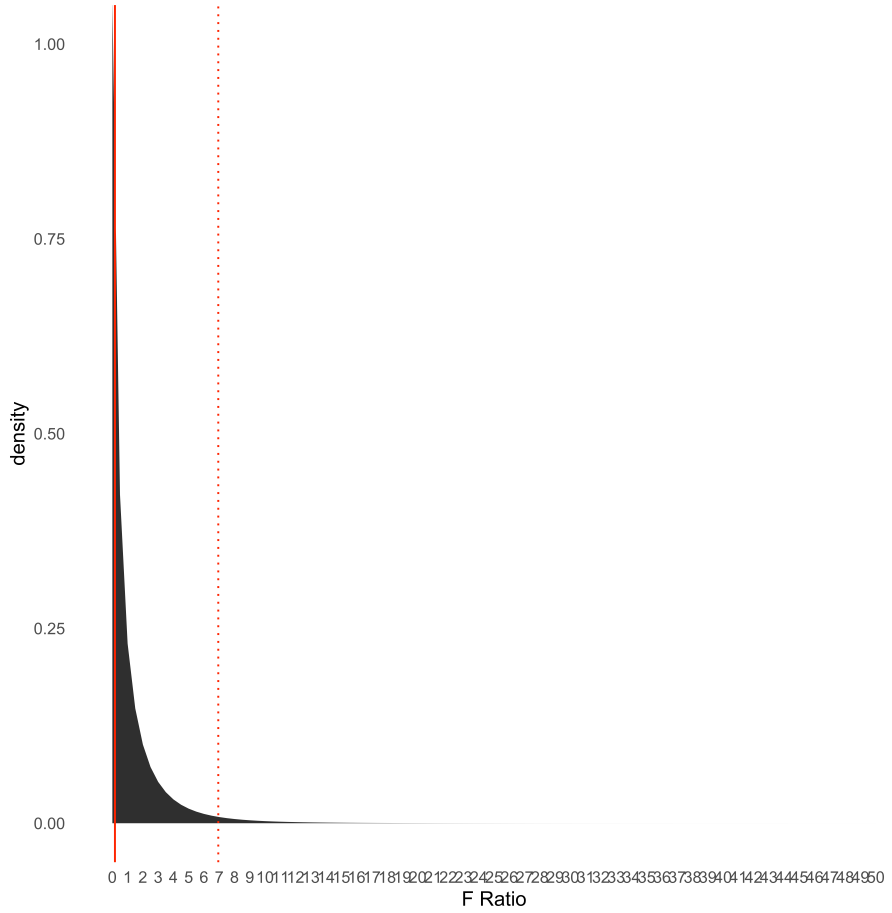
Nota: distrubición de estadístico F. Línea roja es nuestro F observado.

Taller

Modelo no predictivo

Cuáles son las chances de un modelo que no ajusta

```r
#-------------------------------------------------------------------
# p value de la comparación modelos (m00, m03)
#-------------------------------------------------------------------

# opciones de consola
options(scipen = 999)
options(digits = 7)

# valor p de la comparación de modelos
anova(m00, m03) %>%
broom::tidy() %>%
knitr::kable(., digits = 7)

# ------------------------------------------------
# f value
# ------------------------------------------------

f_value_null <- anova(m00, m03) %>%
                broom::tidy() %>%
                mutate(model = c('compact', 'augmented')) %>%
                dplyr::filter(model == 'augmented') %>%
                dplyr::select(statistic) %>%
                pull() %>%
                as.numeric()

# ------------------------------------------------
# p value
# ------------------------------------------------

df_1 <- 1  # cantidad de parámetros fijos del modelo
df_2 <- 10 # grados de libertad restantes (n_total - df_1 - 1)

pf(f_value_null, df1 = df_1, df2 = df_2, lower.tail = FALSE) %>%
r4sda::decimal(., 7)


# ------------------------------------------------
# f critic
# ------------------------------------------------

f_critic <- qf(.975, df1 = df_1, df2 = df_2)

# ------------------------------------------------
# visualization
# ------------------------------------------------

library(ggplot2)
f_m03 <- ggplot(data.frame(x = c(0, 50)), aes(x)) +
  stat_function(fun = df, args = list(df1 = df_1, df2 = df_2), geom
  geom_vline(xintercept = f_value_null, color = 'red') +
  geom_vline(xintercept = f_critic, color = 'red', linetype = 'dotte
  scale_x_continuous(breaks=seq(0, 50, 1)) +
  # xlim(0,10) +
  ylim(0,1) +
  labs(
    x = 'F Ratio',
    y = 'density') +
  theme_minimal() +
  theme(
  panel.background = element_blank(),
  panel.grid.minor = element_blank(),
  panel.grid.major = element_blank(),
  axis.text.x=element_text(size=1)
  )


# show plot
f_m03
```



Nota: distrubición de estadístico F. Línea roja es nuestro F del modelo m03.

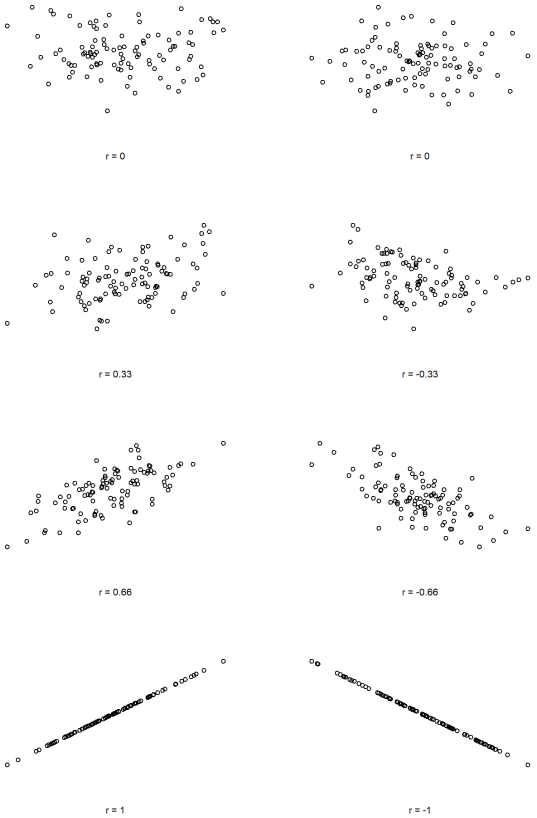
Modelo de interés `m02`

Taller

Correlaciones

Cuáles son las chances de un modelo que no ajusta

Correlaciones esperadas



Referencias: Navarro (2019)

]

Correlaciones no esperadas

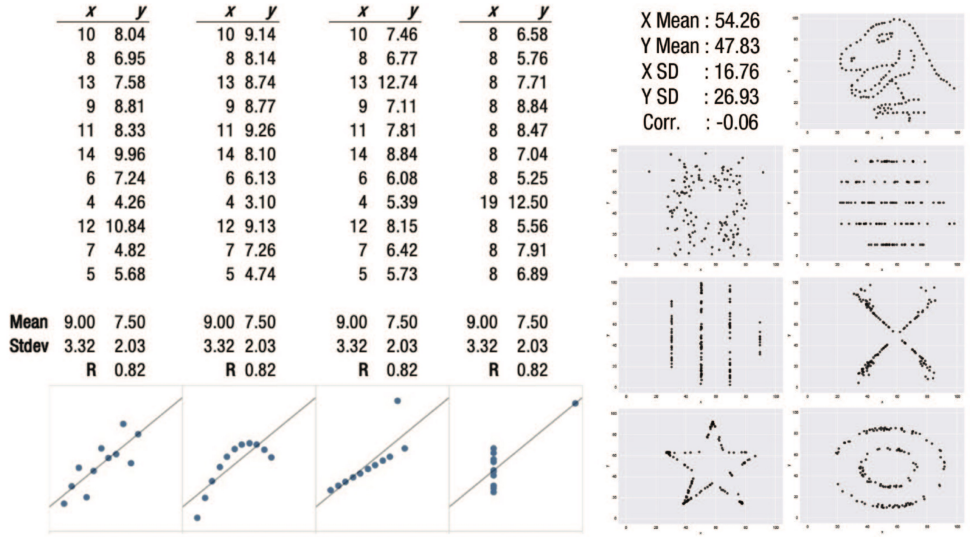


Fig. 1. Examples of how visualizations can let viewers see beyond summary statistics. At left, four sets of 11 numbers have identical statistics but dramatically different patterns, as revealed by the scatterplots below each column. At right is a more extreme example of nine dramatically different scatterplots (including one that looks suspiciously like a dinosaur) depicting data with identical statistics, down to the second decimal place. The graphs on the right are adapted with permission of the Association for Computing Machinery, from "Same Stats, Different Graphs: Generating Datasets With Varied Appearance and Identical Statistics Through Simulated Annealing," by J. Matejka and G. Fitzmaurice, *CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (https://doi.org/10.1145/3025453.3025912). Copyright 2017 Association for Computing Machinery.

Referencias: Franconeri et al. (2021)

]

Correlaciones

```
#-------------------------------------------------------------------------------
# datos
#-------------------------------------------------------------------------------

#-------------------------------------------------------------
# tabla 3.2
#-------------------------------------------------------------

data_table_3_2 <- read.table(
text="
person   y    x    x_q   xy   z
1        2    8     64   16   1
2        3    9     81   27   2
3        3    9     81   27   1
4        4   10    100   40   2
5        7    6     36   42   1
6        5    7     49   35   2
7        5    4     16   20   1
8        7    5     25   35   2
9        8    3      9   24   1
10       9    1      1    9   2
11       9    2      4   18   1
12      10    2      4   20   2

",
header=TRUE, stringsAsFactors = FALSE)

# Nota: agregamos a la variable z,
#       para ilustrar como se ve un
#       modelo que no explica a y.

#-------------------------------------------------------------
# preparar datos
#-------------------------------------------------------------

library(dplyr)
data_model <- data_table_3_2 %>%
              mutate(x_g = mean(x, na.rm = TRUE)) %>%
              mutate(x_cgm = x - x_g) %>%
              dplyr::select(y, x, x_cgm, z)


#-------------------------------------------------------------
# correlación base::cor
#-------------------------------------------------------------

cor(data_model$y, data_model$x)

#-------------------------------------------------------------
# correlación base::cor.test
#-------------------------------------------------------------

cor.test(data_model$y, data_model$x)
```

Output

```
> #-------------------------------------------------------------
> # tabla 3.2
> #-------------------------------------------------------------
>
> data_table_3_2 <- read.table(
+ text="
+ person   y    x    x_q   xy   z
+ 1        2    8     64   16   1
+ 2        3    9     81   27   2
+ 3        3    9     81   27   1
+ 4        4   10    100   40   2
+ 5        7    6     36   42   1
+ 6        5    7     49   35   2
+ 7        5    4     16   20   1
+ 8        7    5     25   35   2
+ 9        8    3      9   24   1
+ 10       9    1      1    9   2
+ 11       9    2      4   18   1
+ 12      10    2      4   20   2
+
+ ",
+ header=TRUE, stringsAsFactors = FALSE)
>
> # Nota: agregamos a la variable z,
> #       para ilustrar como se ve un
> #       modelo que no explica a y.
>
> #-------------------------------------------------------------
> # preparar datos
> #-------------------------------------------------------------
>
> library(dplyr)
> data_model <- data_table_3_2 %>%
+               mutate(x_g = mean(x, na.rm = TRUE)) %>%
+               mutate(x_cgm = x - x_g) %>%
+               dplyr::select(y, x, x_cgm, z)
>
>
> #-------------------------------------------------------------
> # correlación base::cor
> #-------------------------------------------------------------
>
> cor(data_model$y, data_model$x)
[1] -0.8971007
>
> #-------------------------------------------------------------
> # correlación base::cor.test
> #-------------------------------------------------------------
>
> cor.test(data_model$y, data_model$x)

  Pearson's product-moment correlation

data:  data_model$y and data_model$x
t = -6.4208, df = 10, p-value = 7.627e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9710564 -0.6661805
sample estimates:
       cor
-0.8971007
```

Diagonal inferior de correlaciones

```
#-------------------------------------------------------------------------------
# datos
#-------------------------------------------------------------------------------

#-----------------------------------------------------------
# tabla 3.2
#-----------------------------------------------------------

data_table_3_2 <- read.table(
text="
person   y    x    x_q   xy   z
1        2    8     64   16   1
2        3    9     81   27   2
3        3    9     81   27   1
4        4   10    100   40   2
5        7    6     36   42   1
6        5    7     49   35   2
7        5    4     16   20   1
8        7    5     25   35   2
9        8    3      9   24   1
10       9    1      1    9   2
11       9    2      4   18   1
12      10    2      4   20   2

",
header=TRUE, stringsAsFactors = FALSE)

# Nota: agregamos a la variable z,
#       para ilustrar como se ve un
#       modelo que no explica a y.

#-----------------------------------------------------------
# preparar datos
#-----------------------------------------------------------

library(dplyr)
data_model <- data_table_3_2 %>%
              mutate(x_g = mean(x, na.rm = TRUE)) %>%
              mutate(x_cgm = x - x_g) %>%
              dplyr::select(y, x, x_cgm, z)

#-----------------------------------------------------------
# con corrr::correlate
#-----------------------------------------------------------

data_model %>%
corrr::correlate() %>%
corrr::shave()  %>%
corrr::fashion()
```

Output

```
> #-------------------------------------------------------------------------------
> # datos
> #-------------------------------------------------------------------------------
>
> #-----------------------------------------------------------
> # tabla 3.2
> #-----------------------------------------------------------
>
> data_table_3_2 <- read.table(
+ text="
+ person   y    x    x_q   xy   z
+ 1        2    8     64   16   1
+ 2        3    9     81   27   2
+ 3        3    9     81   27   1
+ 4        4   10    100   40   2
+ 5        7    6     36   42   1
+ 6        5    7     49   35   2
+ 7        5    4     16   20   1
+ 8        7    5     25   35   2
+ 9        8    3      9   24   1
+ 10       9    1      1    9   2
+ 11       9    2      4   18   1
+ 12      10    2      4   20   2
+
+ ",
+ header=TRUE, stringsAsFactors = FALSE)
>
> # Nota: agregamos a la variable z,
> #       para ilustrar como se ve un
> #       modelo que no explica a y.
>
> #-----------------------------------------------------------
> # preparar datos
> #-----------------------------------------------------------
>
> library(dplyr)
> data_model <- data_table_3_2 %>%
+               mutate(x_g = mean(x, na.rm = TRUE)) %>%
+               mutate(x_cgm = x - x_g) %>%
+               dplyr::select(y, x, x_cgm, z)
>
> #-----------------------------------------------------------
> # con corrr::correlate
> #-----------------------------------------------------------
>
> data_model %>%
+ corrr::correlate() %>%
+ corrr::shave()  %>%
+ corrr::fashion()

Correlation method: 'pearson'
Missing treated using: 'pairwise.complete.obs'

   term    y    x x_cgm z
1     y
2     x -.90
3 x_cgm -.90 1.00
4     z  .13  .06   .06
```

Muchas gracias!

Carrasco, D., PhD
Centro de Medición MIDE UC,
Pontificia Universidad Católica de Chile
https://dacarras.github.io/

Referencias

Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The Science of Visual Data Communication: What Works. Psychological Science in the Public Interest, 22(3), 110–161. https://doi.org/10.1177/15291006211051956

Huck, S. W. (2012). Bivariate, Multiple, and Logistic Regression. In Reading Statistics and Research (6th ed., pp. 367–403). Pearson Education.

Navarro, D. (2013). Learning statistics with R: A tutorial for psychology students and other beginners.

Vik, P. (2014). Regression, ANOVA, and the general linear model: A statistics primer. Sage.