M5_AG1_MODELO LINEAL GENERALIZADOS. INFERENCIA

Jose Godoy & Bryan Casanova

2022-10-17

Lectura de datos

```
data <- read.csv("https://www-eio.upc.edu/~pau/cms/rdata/csv/COUNT/affairs.csv")

#Remove index
data <- data[, -1]

library(pander)
pander(summary(data))
```

Table 1: Table continues below

naffairs	kids	vryunhap	unhap
Min. : 0.000	Min. :0.0000	Min. :0.00000	Min. :0.0000
1st Qu.: 0.000	1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:0.0000
Median : 0.000	Median :1.0000	Median :0.00000	Median :0.0000
Mean : 1.456	Mean :0.7155	Mean :0.02662	Mean :0.1098
3rd Qu.: 0.000	3rd Qu.:1.0000	3rd Qu.:0.00000	3rd Qu.:0.0000
Max. :12.000	Max. :1.0000	Max. :1.00000	Max. :1.0000

Table 2: Table continues below

avgmarr	hapavg	vryhap	antirel
Min. :0.0000	Min. :0.0000	Min. :0.000	Min. :0.00000
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.000	1st Qu.:0.00000
Median :0.0000	Median :0.0000	Median :0.000	Median :0.00000
Mean :0.1547	Mean :0.3228	Mean :0.386	Mean :0.07987
3rd Qu.:0.0000	3rd Qu.:1.0000	3rd Qu.:1.000	3rd Qu.:0.00000
Max. :1.0000	Max. :1.0000	Max. :1.000	Max. :1.00000

Table 3: Table continues below

notrel	slghtrel	smerel	vryrel
Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.0000

notrel	slghtrel	smerel	vryrel
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000
Median :0.0000	Median :0.0000	Median :0.0000	Median :0.0000
Mean :0.2729	Mean :0.2146	Mean :0.3161	Mean :0.1165
3rd Qu.:1.0000	3rd Qu.:0.0000	3rd Qu.:1.0000	3rd Qu.:0.0000
Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :1.0000

Table 4: Table continues below

yrsmarr1	yrsmarr2	yrsmarr3	yrsmarr4
Min. :0.00000	Min. :0.0000	Min. :0.0000	Min. :0.0000
1st Qu.:0.00000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000
Median :0.00000	Median :0.0000	Median :0.0000	Median :0.0000
Mean :0.08652	Mean :0.1464	Mean :0.1747	Mean :0.1364
3rd Qu.:0.00000	3rd Qu.:0.0000	3rd Qu.:0.0000	3rd Qu.:0.0000
Max. :1.00000	Max. :1.0000	Max. :1.0000	Max. :1.0000

yrsmarr5	yrsmarr6
Min. :0.0000	Min. :0.0000
1st Qu.:0.0000	1st Qu.:0.0000
Median :0.0000	Median :0.0000
Mean :0.1165	Mean :0.3394
3rd Qu.:0.0000	3rd Qu.:1.0000
Max. :1.0000	Max. :1.0000

Ejercicio 1

Para poder identificar cuales son las variables que tienen mayor impacto en la incredibilidad se desarrolla una regresión lineal, donde el estimador de cada variable ayudará a poder el impacto en el modelo

Antes de hacer un modelo, verificaremos utilizando `Stepwise` para poder ver cuales con las variables a considerar

```
data_binary <- data
data_binary$naffairs <- ifelse(data$naffairs>0, 1, 0)
step_wise <- step(glm(formula = 'naffairs ~ .', data = data_binary, family = binomial), trace = FALSE)
pander(summary(step_wise))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.986	0.2405	-8.258	1.477e-16
vryunhap	1.532	0.551	2.78	0.005429
unhap	1.655	0.3213	5.15	2.601e-07
avgmarr	0.8122	0.3071	2.645	0.008171
hapavg	0.5141	0.2604	1.975	0.04831
antirel	1.43	0.3534	4.047	5.198e-05
notrel	0.4832	0.258	1.873	0.06111
slghtrel	0.867	0.26	3.335	0.000853
yrsmarr1	-1.158	0.4945	-2.342	0.0192

	Estimate	Std. Error	z value	Pr(>|z|)
yrsmarr2	-0.8518	0.3482	-2.446	0.01444

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	675.4 on 600 degrees of freedom
Residual deviance:	604.0 on 591 degrees of freedom

```
formula <- step_wise$formula
model_glm <- glm(formula, data = data_binary, family = binomial)
pander(model_glm)
```

Table 8: Fitting generalized (binomial/logit) linear model: formula

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.986	0.2405	-8.258	1.477e-16
vryunhap	1.532	0.551	2.78	0.005429
unhap	1.655	0.3213	5.15	2.601e-07
avgmarr	0.8122	0.3071	2.645	0.008171
hapavg	0.5141	0.2604	1.975	0.04831
antirel	1.43	0.3534	4.047	5.198e-05
notrel	0.4832	0.258	1.873	0.06111
slghtrel	0.867	0.26	3.335	0.000853
yrsmarr1	-1.158	0.4945	-2.342	0.0192
yrsmarr2	-0.8518	0.3482	-2.446	0.01444

Como era de esperarse, los matrimonios que son infelices o muy infelices son propensos a caer en una inefabilidad. También hay un fuerte impacto en las personas que son anti-religiosas.

Ejercicio 2

```
step_wise_2 <- step(glm(formula = 'naffairs ~ .', data = data), trace = FALSE)
model_glm_2 <- glm(step_wise_2$formula, data = data)
pander(model_glm_2)
```

Table 9: Fitting generalized (gaussian/identity) linear model: step_wise_2$formula

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.053	0.2277	4.624	4.617e-06
vryunhap	2.312	0.7845	2.947	0.003335
unhap	2.599	0.4089	6.357	4.127e-10
antirel	2.076	0.4884	4.251	2.478e-05
notrel	0.9203	0.3136	2.934	0.003471
slghtrel	1.197	0.3346	3.577	0.0003756
yrsmarr1	-1.31	0.4709	-2.781	0.005589

	Estimate	Std. Error	t value	Pr(>|t|)
yrsmarr2	-1.637	0.3847	-4.254	2.436e-05
yrsmarr3	-1.03	0.361	-2.853	0.004484
yrsmarr4	-0.6215	0.3877	-1.603	0.1095

cambiar Al hacer esa modificación hay un cambio en los estimadores de las variables. Pero al mismo tiempo, las variables que tienen mayor impacto siguen siendo las mismas.

Ejercicio 3

La variable `yrsmarr` es una descomposición de una variable general. Al ser una descomposiciión generada por intervalos, esta variable que era continua se está comportando como una variable categorica, por lo que lo recomendado sería trabajar con una unica columna que indique la cantidad de años casados. Complementario a lo anterior, al tener 6 columnas para esta variable (siendo binarias y excluyentes entre si) al sumar estas columnas se tiene una combinación lineal del intercepto y por lo tanto el modelo no se puede calcular.

Ejercicio 4

```
new_data <- data[, c("naffairs" , "kids", "antirel",  "notrel" ,   "slghtrel", "smerel" ,   "vryrel" ,
formula <- as.formula("naffairs ~ .")
model_glm_ex4 <- glm(formula, data = new_data)
pander(model_glm_ex4)
```

Table 10: Fitting generalized (gaussian/identity) linear model: formula

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.672	0.5484	3.049	0.002398
kids	-0.2297	0.3782	-0.6074	0.5438
antirel	2.132	0.6074	3.51	0.0004824
notrel	1.212	0.4646	2.609	0.0093
slghtrel	1.416	0.4794	2.953	0.003268
smerel	0.05823	0.4496	0.1295	0.897
yrsmarr1	-2.006	0.5804	-3.456	0.0005885
yrsmarr2	-2.148	0.4936	-4.352	1.592e-05
yrsmarr3	-1.492	0.4072	-3.663	0.0002719
yrsmarr4	-0.6314	0.4222	-1.495	0.1353
yrsmarr5	-0.3472	0.4434	-0.7831	0.4339

```
data_frame_predict <- data.frame("kids"=0,   "antirel" = 0,  "notrel" = 1,   "slghtrel"= 0, "smerel" = 0
predict(model_glm_ex4, data_frame_predict)
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
##        1
## 2.537195
```

Lo más probable es que dado el perfil del enunciado, asumiendo que tiene más de 10 años de casado y menos de 15, es que ocurra una infidelidad

Pregunta 5

```
confint(model_glm_ex4, type="response")
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %     97.5 %
## (Intercept)  0.5973181  2.7469758
## kids        -0.9710284  0.5115598
## antirel      0.9414071  3.3222825
## notrel       0.3017244  2.1228624
## slghtrel     0.4762238  2.3553667
## smerel      -0.8230052  0.9394616
## vryrel              NA         NA
## yrsmarr1    -3.1434566 -0.8681464
## yrsmarr2    -3.1154437 -1.1805624
## yrsmarr3    -2.2896596 -0.6934113
## yrsmarr4    -1.4589790  0.1961529
## yrsmarr5    -1.2163185  0.5218283
## yrsmarr6            NA         NA
```

Pregunta 6

hist y qqlot

Pregunta 7

```
formula_interaccion <- c("naffairs ~ kids + vryunhap + unhap + avgmarr + hapavg + vryhap + antirel + no
            slghtrel + smerel + vryrel + yrsmarr1*kids + yrsmarr2*kids + yrsmarr3*kids +
            yrsmarr4*kids + yrsmarr5*kids + yrsmarr6*kids")
step_wise_7 <- step(glm(formula = formula_interaccion, data = data), trace = FALSE)
model_glm_7 <- glm(step_wise_7$formula, data = data)
summary(model_glm_7)
```

```
##
## Call:
## glm(formula = step_wise_7$formula, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.1330  -1.6251  -0.8465   0.3230  11.0099
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.0501     0.4748   4.318 1.85e-05 ***
## kids           -1.0464     0.4447  -2.353 0.018953 *
## vryunhap        2.2256     0.7808   2.850 0.004520 **
## unhap           2.6374     0.4077   6.470 2.07e-10 ***
## antirel         2.0734     0.4856   4.270 2.28e-05 ***
```

```
## notrel          0.9017     0.3138   2.874 0.004204 **
## slghtrel        1.1695     0.3333   3.509 0.000485 ***
## yrsmarr1       -2.6804     0.6566  -4.083 5.07e-05 ***
## yrsmarr2       -2.7549     0.6013  -4.581 5.64e-06 ***
## yrsmarr3       -1.3267     0.3805  -3.486 0.000526 ***
## yrsmarr4       -0.7075     0.3873  -1.827 0.068217 .
## kids:yrsmarr1   3.2707     1.2055   2.713 0.006859 **
## kids:yrsmarr2   1.5718     0.8752   1.796 0.073016 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 9.289478)
##
##     Null deviance: 6529.1  on 600  degrees of freedom
## Residual deviance: 5462.2  on 588  degrees of freedom
## AIC: 3060
##
## Number of Fisher Scoring iterations: 2
```

Vemos que el modelo considero algunas de las interacciones entre la tenencia de hijos y los años de matrimonio
Falta comparar AIC

Pregunta 8

Dado que en el modelo anterior solo quedaron algunas interacciones entre la tenencia de hijos y años de matrimonio, crearemos un nuevo modelo que considere solo estas interacciones

```
formula_interaccion_8 <- c("naffairs ~ yrsmarr1*kids + yrsmarr2*kids + yrsmarr3*kids +
           yrsmarr4*kids + yrsmarr5*kids + yrsmarr6*kids")
model_glm_7 <- glm(formula_interaccion_8, data = data)
summary(model_glm_7)
```

```
##
## Call:
## glm(formula = formula_interaccion_8, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.5000  -2.0051  -0.8696  -0.0465  11.2273
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.500      1.143   3.936 9.29e-05 ***
## yrsmarr1        -4.453      1.245  -3.576 0.000377 ***
## kids            -2.495      1.167  -2.139 0.032867 *
## yrsmarr2        -4.136      1.211  -3.416 0.000678 ***
## yrsmarr3        -3.000      1.264  -2.373 0.017953 *
## yrsmarr4        -2.591      1.503  -1.724 0.085213 .
## yrsmarr5        -1.786      1.674  -1.067 0.286471
## yrsmarr6            NA         NA      NA       NA
## yrsmarr1:kids    5.226      1.663   3.142 0.001761 **
## kids:yrsmarr2    2.904      1.412   2.056 0.040206 *
```

```
## kids:yrsmarr3    1.864      1.343   1.389 0.165488
## kids:yrsmarr4    2.149      1.568   1.371 0.171032
## kids:yrsmarr5    1.574      1.738   0.906 0.365445
## kids:yrsmarr6       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 10.45932)
##
##     Null deviance: 6529.1  on 600  degrees of freedom
## Residual deviance: 6160.5  on 589  degrees of freedom
## AIC: 3130.3
##
## Number of Fisher Scoring iterations: 2
```

En este caso, la beta de kids de manera independiente, es negativa para kids1, lo que quiere decir que, manteniendo el resto de las variables estable, para una pareja que tiene hijos el número de infidelidades en el último año disminuye en 2.5 veces. Al mismo tiempo, se muestra que al analizar las interacciones entre estas variables, vemos que para todos los años de matrimonio (salvo sobre 15 años en que no hay datos suficientes) hay un aumento en el número de infidelidades al tener hijo.

Pregunta 9

```
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
```

```
data %>%
  mutate(yrsmarr1 = ifelse(yrsmarr1>0, 1,0),
         yrsmarr2 = ifelse(yrsmarr2>0, 2,0),
         yrsmarr3 = ifelse(yrsmarr3>0, 5,0),
         yrsmarr4 = ifelse(yrsmarr4>0, 8,0),
         yrsmarr5 = ifelse(yrsmarr5>0, 12,0),
         yrsmarr6 = ifelse(yrsmarr6>0, 18,0),
         yrsmarr = yrsmarr1 + yrsmarr2 + yrsmarr3 + yrsmarr4 + yrsmarr5 + yrsmarr6,
         yrsmarrsquared = yrsmarr**2) %>%
  select(-yrsmarr1, -yrsmarr2, -yrsmarr3, -yrsmarr4, -yrsmarr5, - yrsmarr6) %>%
  group_by(yrsmarr, yrsmarrsquared) %>%
  summarize(Media = mean(naffairs))
```

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

```
## # A tibble: 6 x 3
## # Groups:   yrsmarr [6]
##   yrsmarr yrsmarrsquared Media
##     <dbl>          <dbl> <dbl>
## 1       1              1 0.519
## 2       2              4 0.466
## 3       5             25 1.09
## 4       8             64 1.61
## 5      12            144 1.89
## 6      18            324 2.10
```

Al considerar una cantidad de años de matrimonio dentro del rango para cada categoría y luego calcular el promedio de infidelidades para cada uno de ellos vemos que el único periodo en que disminuyen las infidelidades es desde un año hasta los cuatro años de matrimonio.