Limpieza, transformacion, EDA y visualizacion de datos

Queremos responder la siguiente pregunta ¿Existe una variación importante entre el número de semanas de gestación obtenido a partir de la última menstruación y el determinado por ultrasonido. ¿Esta variación se relaciona con la edad o el nivel educativo?

Para esto lo que hacemos primero es seleccionar las variables importantes, estas son las que nos servirán para responder las preguntas, además a las variables categóricas las convertidos en factores, además contaremos la cantidad de datos faltantes

sapply(datos1, function(x) mean(is.na(x)) * 100)

```
sapply(datos1, function(x) mean(is.na(x)) * 100)
```

```
##                      edad             estado_civil                nivel_edu
##              0.000000000              1.307899090              1.853287841
##                ocupacion                 religion edad_primera_menstruacion
##             20.393920596              3.884925558              1.987696443
##   edad_inicio_vida_sexual         semanas_embarazo          numero_embarazos
##              2.375413565              7.785359801              2.202233251
##           numero_abortos            numero_partos           numero_cesareas
##              5.942411084              4.903329198              5.526261373
##              numero_iles       recibio_consejeria       uso_anticonceptivo
##              3.701406121             20.499896609             11.543631100
##                  entidad              se_complica        procedimiento_ile
##              0.005169562             10.119416873              0.000000000
##     semanas_gestacion_usg
##              0.000000000
```

Posteriormente realizamos una imputación usando MICE, el metodo pmm y con esto finalmente tendremos nuestra base de datos más completa, solo elegiendo imputar cuando no falten más del 30% de los datos a la variable

```
sapply(imputacion, function(x) mean(is.na(x)) * 100)
```

```
##                      edad             estado_civil                nivel_edu
##                   0.00000                  0.00000                  0.00000
##                ocupacion                 religion edad_primera_menstruacion
##                   0.00000                  0.00000                  0.00000
##   edad_inicio_vida_sexual         semanas_embarazo          numero_embarazos
##                   0.00000                  0.00000                  0.00000
##           numero_abortos            numero_partos           numero_cesareas
##                   0.00000                  0.00000                  0.00000
##              numero_iles       recibio_consejeria       uso_anticonceptivo
##                   0.00000                  0.00000                  0.00000
##                  entidad              se_complica        procedimiento_ile
##                   0.00000                  0.00000                  0.00000
##     semanas_gestacion_usg             numero_hijos       anticonceptivo_post
##                   0.00000                 34.70844                 61.90033
```

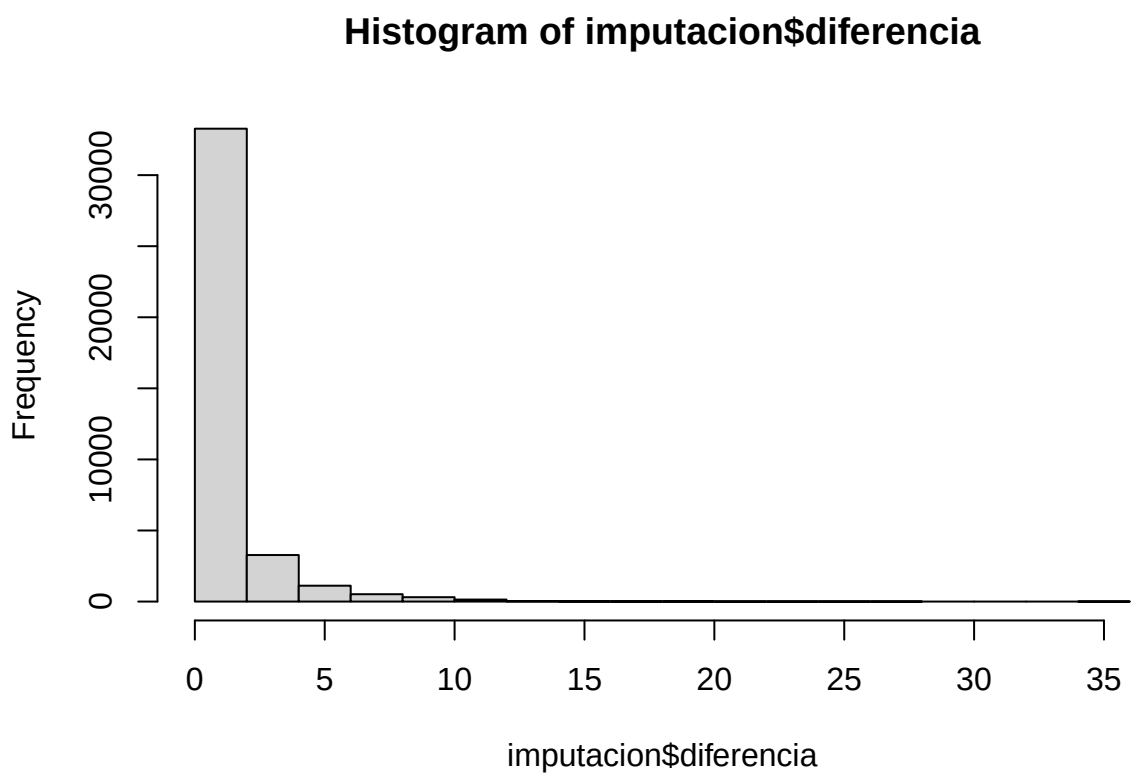
Después de haber realizado la imputación, creamos una nueva columna de la diferencia absoluta de las semanas de embarazo por ultima menstruacion y de las de embarazo por ultrasonido, encontrando que el máximo de esta diferencia es de 35 semanas, y decidimos hacer una prueba para ver si al cuantil 70 habia suficiente diferencia, lo cual no sucede y por eso podemos concluir que no hay diferencia significativa.

```
wilcox.test(imputacion$diferencia, mu = q_05, alternative = "less")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  imputacion$diferencia
## V = 184340133, p-value = 1
## alternative hypothesis: true location is less than 1
```

Esto se apoya en que se estima a nivel mundial que una cantidad estimada del 14 % al 25 % de las mujeres en edad de procrear tiene irregularidades menstruales https://espanol.nichd.nih.gov/salud/temas/menstruation/informacion/mujeres y en general, solo estas mujeres muy irregulares presentan diferencias significativas en la diferencia de sus semanas, por lo que además viendo el histograma es lógico pensar que realmente las diferencias no son significativas considerando a 2 semanas como el valor de corte, y son muy pocas las mujeres que tienen más de dos semanas para considerar que es una cantidad significativa.

```
hist(imputacion$diferencia)
```



Y tambien en el histograma se puede ver que muy pocas mujeres llegan a presentar una diferencia de hasta más de un mes, por lo que en general no es una diferencia significativa

```r
summary(modelo)
```

```
##
## Call:
## lm(formula = diferencia ~ edad + nivel_edu, data = imputacion)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.458 -1.234 -0.338  0.635 33.766
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.122752   0.047276  23.749   <2e-16 ***
## edad        -0.003437   0.001529  -2.248   0.0246 *
## nivel_edu    0.065676   0.006897   9.522   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.866 on 38685 degrees of freedom
## Multiple R-squared:  0.002502,   Adjusted R-squared:  0.00245
## F-statistic: 48.51 on 2 and 38685 DF,  p-value: < 2.2e-16
```

```r
cor(imputacion$diferencia,imputacion$nivel_edu)
```

```
## [1] 0.0486957
```

```r
cor(imputacion$diferencia,imputacion$edad)
```

```
## [1] -0.01279947
```

```r
print(test1)
```

```
##
##  Spearman's rank correlation rho
##
## data:  imputacion$diferencia and imputacion$nivel_edu
## S = 9.1016e+12, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.05693824
```

```r
print(test2)
```

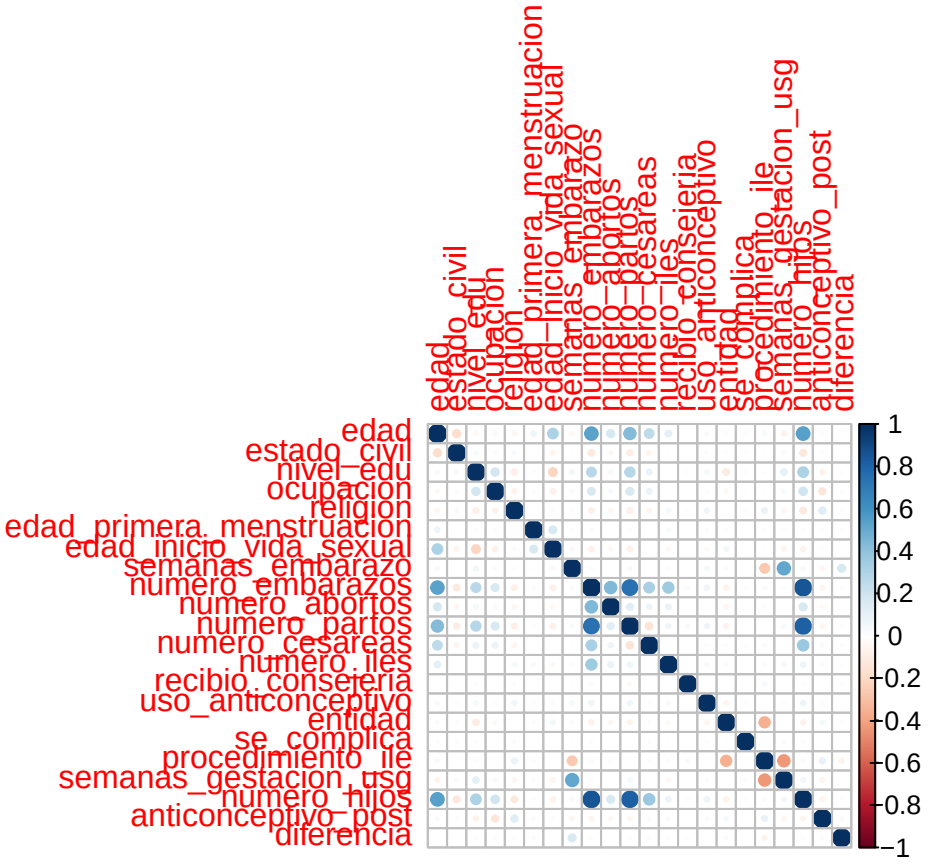
```
##
##  Spearman's rank correlation rho
##
## data:  imputacion$diferencia and imputacion$edad
## S = 9.7973e+12, p-value = 0.002898
## alternative hypothesis: true rho is not equal to 0
```

```
## sample estimates:
##         rho
## -0.01514206
```

Después para ver si había relación de la diferencia con la edad o el nivel educativo, se calculo la correlacion y se hizo un modelo de regresión, encontrando que sí hay correlación, pero esta es muy débil

Por lo que podemos concluir que, al calcular las correlaciones y hacer una prueba para ver si la correlacion es diferente de cero, encontramos que lo es, sin embargo el valor de correlacion obtenido tanto aqui como al tratar de ajustar un modelo de regresion es muy pequeño, por lo que podemos concluir que la correlación no es lo suficiente significativa para decir que se relaciona con alguna de las dos variables.

Ahora buscamos responder ¿Cómo caracterizaría a la población que más utiliza la ile?, y para esto decidimos calcular primero las correlaciones entre las variables:




```
##           edad estado_civil   nivel_edu  ocupacion religion
## [1,] 0.1268556  0.001362224 0.007603582 0.02989215 0.032779
##      edad_primera_menstruacion edad_inicio_vida_sexual semanas_embarazo
## [1,]               0.004674742              0.04426313       0.01743002
##      numero_embarazos numero_abortos numero_partos numero_cesareas numero_iles
## [1,]         0.383588     0.01684872    0.05151948      0.02399504           1
##      recibio_consejeria uso_anticonceptivo    entidad se_complica
## [1,]        0.001129692         0.07910471 0.03101418  0.02982005
##      procedimiento_ile semanas_gestacion_usg numero_hijos anticonceptivo_post
## [1,]        0.05104712            0.06010056           NA                  NA
##       diferencia
## [1,] 0.001410426
```

Además como observación, hay un caso de 9 iles, dos casos de 6 iles, 4 casos de 5 iles, siendo en general datos muy atipicos en los que hay tantos iles, y también podemos ver que el numero de embarazos y la edad son las variables que mejor nos permitirian caracterizar a la poblacion, al ser estas las más correlacionadas.

```
##
## Call:
## lm(formula = numero_iles ~ edad + numero_embarazos, data = imputacion)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5222 -0.2337 -0.0692  0.0054  8.5441
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.0616691  0.0101746   6.061 1.36e-09 ***
## edad             -0.0093325  0.0004482 -20.824  < 2e-16 ***
## numero_embarazos  0.1662072  0.0020704  80.277  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4594 on 38685 degrees of freedom
## Multiple R-squared:  0.1566, Adjusted R-squared:  0.1566
## F-statistic:  3591 on 2 and 38685 DF,  p-value: < 2.2e-16
```

Entonces después de hacer un modelo lineal multiple con solo esas dos variables, encontramos que realmente su ajuste es bastante pobre, al tener una R^2 solo de 0.15, bastante insuficiente, por lo que tomando el enfoque de stepwise de hacer un modelo con todas las variables y solo quedarnos con las mas importantes llegamos a esto:

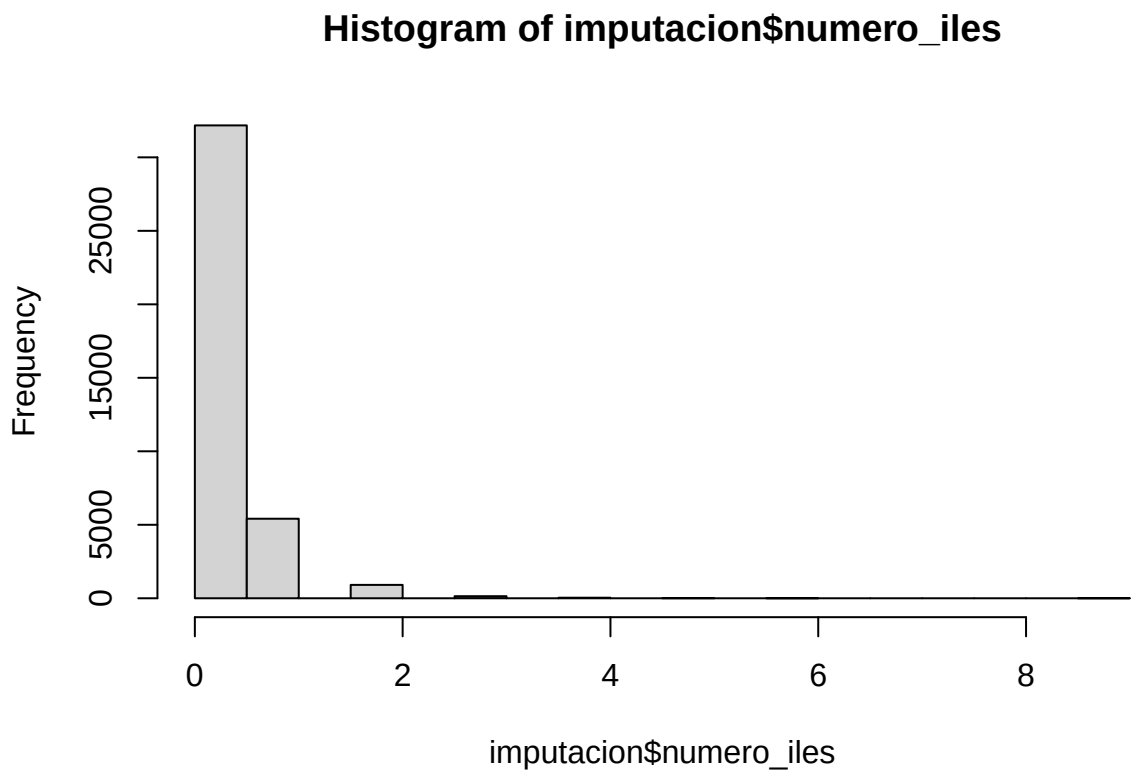
```
# Resumen del modelo final
summary(modelo_step)
```

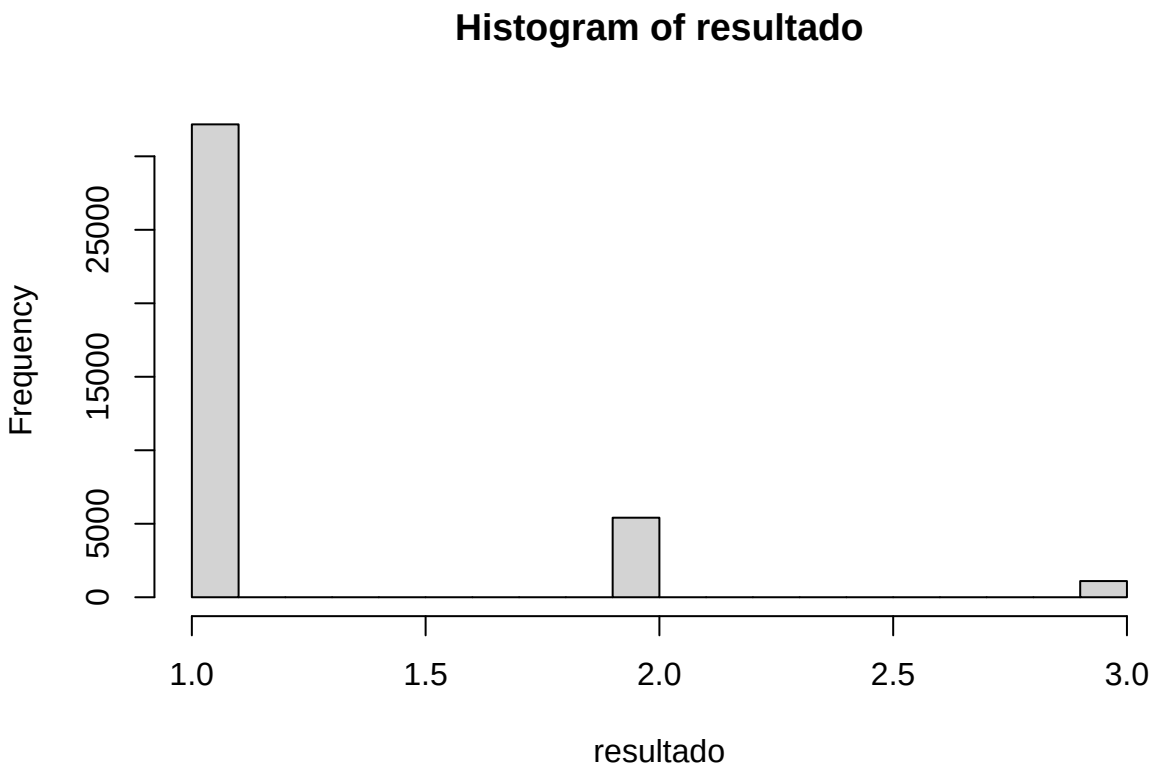
```
##
## Call:
## lm(formula = numero_iles ~ edad + estado_civil + religion + edad_primera_menstruacion +
##     edad_inicio_vida_sexual + semanas_embarazo + numero_embarazos +
##     numero_abortos + numero_partos + numero_cesareas + uso_anticonceptivo +
##     entidad + numero_hijos + anticonceptivo_post + diferencia,
##     data = imputacion)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7732  -0.0843  -0.0603  -0.0356   4.8977
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -0.6055496  0.0380384 -15.919  < 2e-16 ***
## edad                       0.0019110  0.0007075   2.701  0.00693 **
## estado_civil               0.0052846  0.0030877   1.712  0.08702 .
## religion                   0.0028281  0.0017605   1.606  0.10821
## edad_primera_menstruacion  0.0051881  0.0018389   2.821  0.00479 **
## edad_inicio_vida_sexual   -0.0034904  0.0015192  -2.298  0.02161 *
## semanas_embarazo          -0.0036931  0.0012940  -2.854  0.00433 **
```

```
## numero_embarazos           0.5904439  0.0064160  92.027  < 2e-16 ***
## numero_abortos            -0.4625022  0.0092671 -49.908  < 2e-16 ***
## numero_partos             -0.4792729  0.0117028 -40.954  < 2e-16 ***
## numero_cesareas           -0.4785293  0.0126537 -37.817  < 2e-16 ***
## uso_anticonceptivo         0.0016521  0.0007621   2.168  0.03019 *
## entidad                   -0.0016157  0.0008703  -1.856  0.06343 .
## numero_hijos              -0.1052024  0.0108966  -9.655  < 2e-16 ***
## anticonceptivo_post        0.0021391  0.0009614   2.225  0.02611 *
## diferencia                 0.0033810  0.0016661   2.029  0.04246 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3063 on 8907 degrees of freedom
##   (29765 observations deleted due to missingness)
## Multiple R-squared:  0.5027, Adjusted R-squared:  0.5018
## F-statistic: 600.2 on 15 and 8907 DF,  p-value: < 2.2e-16
```

Un modelo que en vez de usar 2 variables como el anterior, termina usando 15 variables de las 21 que habia en total, con esto terminamos con un modelo con mejor ajuste, pero dificil de interpretar debido a tantas variables, ademas de que debido a lo discretas que son estas variables, buscar realizar un PCA no seria adecuado, ademas crear indices entre estas variables tampoco parece algo tan viable, por lo que usando un ultimo enfoque llegamos a otra alternativa de modelo.

Para este ultimo enfoque se decidió asignar 1 a las personas con 0 iles, un 2 a las personas con una ile y 3 a las personas con más de una Ile, viendo la comparativa de la distribución antes de esta asignación contra la nueva después de la asignación:





Aquí podemos ver que igual la mayor cantidad de la población no ha sufrido ILE, una pequeña cantidad ha usado al menos una vez, y una infima cantidad ha usado más de 2, entonces al ver las similitudes de esto con una variable de conteo al solo tener numeros enteros y positivos buscamos ajustar diversos modelos, algunos ajustaron mejor, aunque siempre teniendo el problema de lo complicado que se volvio encontrar una transformación adecuada para esta variable que se busca explicar.

```
summary(modelo_poisson)
```

```
##
## Call:
## glm(formula = resultado ~ edad + estado_civil + nivel_edu + ocupacion +
##     religion + edad_primera_menstruacion + edad_inicio_vida_sexual +
##     semanas_embarazo + numero_embarazos + numero_abortos + numero_partos +
##     numero_cesareas + recibio_consejeria + uso_anticonceptivo +
##     entidad + se_complica + procedimiento_ile + semanas_gestacion_usg +
##     numero_hijos + anticonceptivo_post + diferencia, family = poisson,
##     data = imputacion)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.3243  -0.1290  -0.0906  -0.0517   2.3918
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -0.238542   0.382295  -0.624 0.532644
## edad                        0.006059   0.002168   2.794 0.005201 **
```

```
## estado_civil               0.005098   0.009361   0.545 0.586028
## nivel_edu                 -0.004722   0.007829  -0.603 0.546389
## ocupacion                  0.005602   0.008540   0.656 0.511805
## religion                   0.006819   0.005351   1.274 0.202530
## edad_primera_menstruacion  0.008409   0.005593   1.503 0.132747
## edad_inicio_vida_sexual   -0.008229   0.004657  -1.767 0.077186 .
## semanas_embarazo          -0.001507   0.004595  -0.328 0.743002
## numero_embarazos           0.186498   0.008091  23.050  < 2e-16 ***
## numero_abortos            -0.129554   0.022687  -5.710 1.13e-08 ***
## numero_partos             -0.107716   0.031102  -3.463 0.000534 ***
## numero_cesareas           -0.117688   0.034685  -3.393 0.000691 ***
## recibio_consejeria         0.001805   0.354361   0.005 0.995936
## uso_anticonceptivo         0.001418   0.002338   0.607 0.544146
## entidad                   -0.001927   0.002931  -0.657 0.510879
## se_complica               -0.073394   0.243042  -0.302 0.762668
## procedimiento_ile         -0.004410   0.030569  -0.144 0.885295
## semanas_gestacion_usg     -0.001185   0.006006  -0.197 0.843635
## numero_hijos              -0.082267   0.033606  -2.448 0.014366 *
## anticonceptivo_post        0.002777   0.002962   0.937 0.348579
## diferencia                 0.002882   0.005115   0.563 0.573189
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1013.2  on 8922  degrees of freedom
## Residual deviance:  676.7  on 8901  degrees of freedom
##   (29765 observations deleted due to missingness)
## AIC: 19344
##
## Number of Fisher Scoring iterations: 9
```

```
summary(modelo_seleccionado)
```

```
##
## Call:
## glm(formula = resultado ~ edad + nivel_edu + ocupacion + edad_primera_menstruacion +
##     edad_inicio_vida_sexual + semanas_embarazo + numero_embarazos +
##     numero_abortos + numero_partos + numero_cesareas + uso_anticonceptivo +
##     entidad + se_complica + semanas_gestacion_usg + numero_hijos +
##     anticonceptivo_post + diferencia, family = Gamma(link = "log"),
##     data = imputacion)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6802  -0.0625  -0.0453  -0.0274   3.5438
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -0.3489214  0.0306388 -11.388  < 2e-16 ***
## edad                       0.0008043  0.0005711   1.408 0.159033
## nivel_edu                 -0.0032511  0.0020197  -1.610 0.107494
## ocupacion                  0.0036370  0.0022150   1.642 0.100629
## edad_primera_menstruacion  0.0024800  0.0014477   1.713 0.086746 .
```

```
## edad_inicio_vida_sexual   -0.0025245  0.0012010  -2.102 0.035581 *
## semanas_embarazo          -0.0057346  0.0011880  -4.827 1.41e-06 ***
## numero_embarazos           0.3831119  0.0050508  75.851  < 2e-16 ***
## numero_abortos            -0.2444721  0.0072902 -33.534  < 2e-16 ***
## numero_partos             -0.3509329  0.0092363 -37.995  < 2e-16 ***
## numero_cesareas           -0.3478999  0.0099802 -34.859  < 2e-16 ***
## uso_anticonceptivo         0.0015049  0.0005990   2.512 0.012015 *
## entidad                   -0.0015180  0.0006874  -2.208 0.027248 *
## se_complica               -0.0719159  0.0603819  -1.191 0.233678
## semanas_gestacion_usg      0.0040031  0.0014215   2.816 0.004871 **
## numero_hijos              -0.0305524  0.0085985  -3.553 0.000383 ***
## anticonceptivo_post        0.0026704  0.0007564   3.530 0.000417 ***
## diferencia                 0.0040893  0.0013195   3.099 0.001947 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.05811499)
##
##     Null deviance: 731.80  on 8922  degrees of freedom
## Residual deviance: 318.07  on 8905  degrees of freedom
##   (29765 observations deleted due to missingness)
## AIC: -2589.3
##
## Number of Fisher Scoring iterations: 7
```

Por lo que, despues de evaluar diversos modelos de conteo, elegimos un modelo Gamma con el metodo stepwise al tener el menor AIC de todos, quedandonos con las mismas 15 variables del modelo lineal multiple. Por lo tanto hemos obtenido que sí es posible caracterizar a la variable numero de iles, todo dependiendo del enfoque que queremos buscar, y también el número de variables dependerá si queremos un mejor ajuste a cambio de una interpretación más compleja o si buscamos una mejor interpretación a cambio de un peor ajuste.