Regresion Logistica

Cargar los datos

Cargamos los datos que se usaron durante el análisis descriptivo de los datos y mostramos las variables a usar.

```
datos_it0 <- read.csv("../data/bd_trabajo.csv")
names(datos_it0)
```

```
##  [1] "id"
##  [2] "diagnostico"
##  [3] "d1"
##  [4] "d2"
##  [5] "d3"
##  [6] "d4"
##  [7] "d5"
##  [8] "healthCode1"
##  [9] "punto_medicacion"
## [10] "F0semitoneFrom27.5Hz_sma3nz_amean"
## [11] "F0semitoneFrom27.5Hz_sma3nz_stddevNorm"
## [12] "F0semitoneFrom27.5Hz_sma3nz_percentile20.0"
## [13] "F0semitoneFrom27.5Hz_sma3nz_percentile50.0"
## [14] "F0semitoneFrom27.5Hz_sma3nz_percentile80.0"
## [15] "F0semitoneFrom27.5Hz_sma3nz_pctlrange0.2"
## [16] "F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope"
## [17] "F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope"
## [18] "F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope"
## [19] "F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope"
## [20] "jitterLocal_sma3nz_amean"
## [21] "jitterLocal_sma3nz_stddevNorm"
## [22] "shimmerLocaldB_sma3nz_amean"
## [23] "shimmerLocaldB_sma3nz_stddevNorm"
## [24] "edad"
## [25] "cuidado"
## [26] "estimulacion_cerebral"
## [27] "anio_diagnostico"
## [28] "educacion"
## [29] "trabajo"
## [30] "genero"
## [31] "estado_civil"
## [32] "anio_medicacion"
## [33] "anio_enfermedad"
## [34] "facilidad_celular"
## [35] "fumo"
## [36] "cirugias"
```

```
## [37] "anios_fumo"
## [38] "healthCode"
```

No vamos a mostrar las medidas resumenes o los gráficos de todas las variables ya que lo hicimos durante el análisis descriptivo.

Regresion Logística

En esta regresión vamos a utilizar como variable respuesta a **diagnostico**, usando las variables **d1**, **d2**, **d3**, **d4**, **d5** y **edad**. Para trabajar vamos a quedarnos únicamente con estas variables en el dataset.

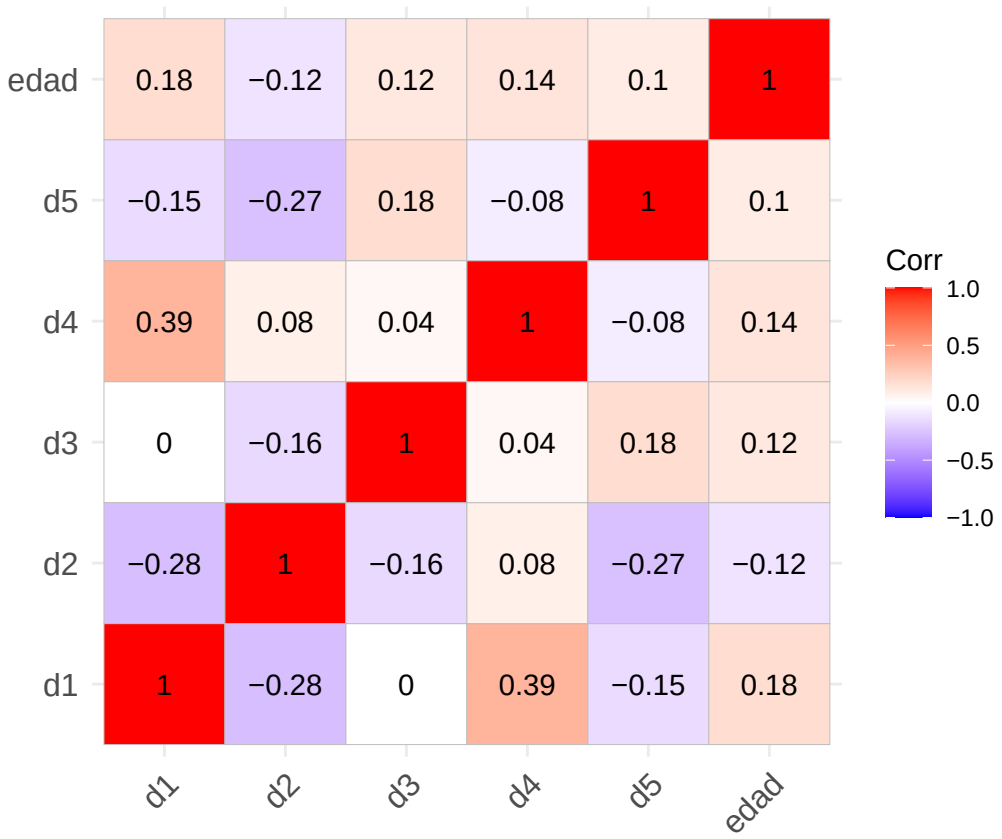
```
datos_it1 <- datos_it0 %>%
  select(diagnostico, d1, d2, d3, d4, d5, edad)
names(datos_it1)
```

```
## [1] "diagnostico" "d1"          "d2"          "d3"          "d4"
## [6] "d5"          "edad"
```

¿Colinealidad/Multicolinealidad?

Empezamos realizando la matriz de correlación de las varaibles para descartar la colinealidad

```
datos_it1 %>%
  select(-diagnostico) %>%
  cor() %>%
  ggcorrplot(lab = TRUE)
```



No se observa ningun coeficiente de correlación mayor/menor a .5/-.5. Ahora, buscamos multicolinealidad usando el VIF. Para esto creamos una regresión lineal múltiple para verficiar si existe multicolinealidad entre todas las variables predictoras.

```
vif(lm(d1 ~ edad + d2 + d3 + d4 + d5, data = datos_it1))
```

```
##     edad       d2       d3       d4       d5
## 1.051162 1.109203 1.060432 1.036321 1.111096
```

Se descarta multicolinealidad porque ningún VIF supera el valor de 10.

Empezando con los modelos

Comenzamos realizando el modelo 0, donde se incluyen todas las variables posibles

```
modelo_it0 <- glm(diagnostico ~ ., data = datos_it1, family = "binomial")
summary(modelo_it0)
##
## Call:
## glm(formula = diagnostico ~ ., family = "binomial", data = datos_it1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5698  -0.7436  -0.4146   0.8531   2.3973
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.513620   0.437964 -19.439  < 2e-16 ***
## d1           0.087467   0.012478   7.009 2.39e-12 ***
## d2          -0.003792   0.002220  -1.708   0.0876 .
## d3           0.220782   0.398317   0.554   0.5794
## d4           0.111645   2.591415   0.043   0.9656
## d5           0.757470   2.769715   0.273   0.7845
## edad         0.103631   0.004972  20.843  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3023.1  on 2217  degrees of freedom
## Residual deviance: 2261.8  on 2211  degrees of freedom
## AIC: 2275.8
##
## Number of Fisher Scoring iterations: 4

r2 <- calcular_pseudo_R2(modelo_it0)
r2
## pseudo_R2   p_valor
## 0.2518205 0.0000000
```

Observamos que las variables **d1**, **d2** y **edad** resultan significativas, pero **d3**, **d4** y **d5** no.

Además, armaremos una tabla donde se especificara el AIC, R2 y alguna de observación de cada modelo armado, a manera comparación.

```
tabla_resumen_modelos <- data.frame(Nombre = "modelo_it0",
                                    AIC = modelo_it0$aic,
                                    R2 = r2[1],
                                    Observacion = "Todas las variables",
                                    row.names = NULL)
rm(r2)
```

Usamos la función step() para seleccionar aquellas variables más importantes.

```
step(modelo_it0, direction = "both")
```

```
## Start:  AIC=2275.82
## diagnostico ~ d1 + d2 + d3 + d4 + d5 + edad
##
##        Df Deviance    AIC
## - d4    1   2261.8 2273.8
## - d5    1   2261.9 2273.9
## - d3    1   2262.1 2274.1
## <none>      2261.8 2275.8
## - d2    1   2264.8 2276.8
## - d1    1   2312.7 2324.7
## - edad  1   2844.0 2856.0
```

```
##
## Step:  AIC=2273.82
## diagnostico ~ d1 + d2 + d3 + d5 + edad
##
##        Df Deviance    AIC
## - d5    1   2261.9 2271.9
## - d3    1   2262.1 2272.1
## <none>      2261.8 2273.8
## - d2    1   2264.9 2274.9
## + d4    1   2261.8 2275.8
## - d1    1   2324.0 2334.0
## - edad  1   2847.3 2857.3
##
## Step:  AIC=2271.89
## diagnostico ~ d1 + d2 + d3 + edad
##
##        Df Deviance    AIC
## - d3    1   2262.2 2270.2
## <none>      2261.9 2271.9
## - d2    1   2265.6 2273.6
## + d5    1   2261.8 2273.8
## + d4    1   2261.9 2273.9
## - d1    1   2327.4 2335.4
## - edad  1   2853.4 2861.4
##
## Step:  AIC=2270.25
## diagnostico ~ d1 + d2 + edad
##
##        Df Deviance    AIC
## <none>      2262.2 2270.2
## + d3    1   2261.9 2271.9
## + d5    1   2262.1 2272.1
## + d4    1   2262.2 2272.2
## - d2    1   2266.5 2272.5
## - d1    1   2327.4 2333.4
## - edad  1   2863.8 2869.8


##
## Call:  glm(formula = diagnostico ~ d1 + d2 + edad, family = "binomial",
##     data = datos_it1)
##
## Coefficients:
## (Intercept)           d1           d2         edad
##   -8.434462     0.086348    -0.004163     0.104027
##
## Degrees of Freedom: 2217 Total (i.e. Null);  2214 Residual
## Null Deviance:       3023
## Residual Deviance: 2262  AIC: 2270
```

Aqui tenemos dos opciones: la primera es mantener las varaibles **d3**, **d4** y **d5** porque nos interesean manetenerlas en el modelo. El segundo es exlcuirlas. Nosotros decidimos tomar la segunda deción. A partir del resultado creamos el primer modelo

```r
modelo_it1 <- glm(formula = diagnostico ~ d1 + d2 + edad, family = "binomial",
    data = datos_it1)
summary(modelo_it1)
##
## Call:
## glm(formula = diagnostico ~ d1 + d2 + edad, family = "binomial",
##     data = datos_it1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5670  -0.7448  -0.4159   0.8455   2.4236
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.434462   0.414288 -20.359  < 2e-16 ***
## d1           0.086348   0.010898   7.924 2.31e-15 ***
## d2          -0.004163   0.002031  -2.050   0.0404 *
## edad         0.104027   0.004927  21.113  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3023.1  on 2217  degrees of freedom
## Residual deviance: 2262.3  on 2214  degrees of freedom
## AIC: 2270.3
##
## Number of Fisher Scoring iterations: 4

r2 <- calcular_pseudo_R2(modelo_it1)
r2
## pseudo_R2   p_valor
## 0.2516767 0.0000000
```

Todas las variables resultan significativas. El AIC es de 2270.25, que es levemente mejor al modelo 0. De aquí notamos que:

- Por cada incremento de una unidad en **d1**, los odds de tener un diagnóstico positivo se incrementan en 9.02%, manteniendo el resto de variables constantes.
- Por cada incremento de una unidad de **d2**, los odss de tener un diagnóstico positivo se reducen en 0.42%, manteniendo el resto de variables constantes.
- Por cada incremento de un año en la **edad**, los odds de tener un diagnóstico positivo se incrementan en 10.96%, manteniendo el resto de variables constantes.

```r
tabla_resumen_modelos <- tabla_resumen_modelos %>%
  add_row(Nombre = "modeo_it1", AIC = modelo_it1$aic, R2 = r2[1], Observacion = "Primer modelo con sele
rm(r2)
```

Puntos Influyentes

Empezamos buscando los puntos influyentes con la distancia_cook

```
puntos_influyentes <- data.frame(d_cook = cooks.distance(modelo_it1)) %>% arrange(-d_cook)
head(puntos_influyentes, n = 15)
```

```
##           d_cook
## 368  0.027452786
## 125  0.009969629
## 2    0.009638075
## 331  0.008439323
## 4    0.006828532
## 89   0.006617408
## 620  0.006537773
## 44   0.006285796
## 76   0.006222473
## 45   0.006130574
## 1628 0.006084733
## 36   0.005943072
## 628  0.005482373
## 23   0.005304204
## 186  0.005144823
```

En general, las distancias son menores a uno por lo que no se consideran influyentes. Sin embargo, analizsamos la observación nro 368 por se la que presta una distancia de cook mucho mayor al resto.

```
datos_it1[368,]
```

```
##     diagnostico       d1        d2       d3       d4      d5 edad
## 368           0 25.10758 -196.4837 0.330796 0.107532 0.16903   68
```

Esta observación presenta un valor de d2 muy bajo.

Vamos a explorar los modelos si excluimos dicha observación.

```
datos_it2 <- datos_it1[-368,]
```

Volvemos a realizar los modelos.

```
modelo_it2 <- glm(diagnostico ~ ., data = datos_it2, family = "binomial")
summary(modelo_it2)
##
## Call:
## glm(formula = diagnostico ~ ., family = "binomial", data = datos_it2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5724  -0.7434  -0.4130   0.8531   2.4055
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.530172   0.438567 -19.450  < 2e-16 ***
## d1           0.086401   0.012486   6.920 4.52e-12 ***
## d2          -0.004375   0.002241  -1.953   0.0509 .
## d3           0.193723   0.398726   0.486   0.6271
```

```
## d4           0.423413   2.596834   0.163   0.8705
## d5           1.426176   2.789018   0.511   0.6091
## edad         0.103769   0.004977  20.850  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3022.0  on 2216  degrees of freedom
## Residual deviance: 2257.9  on 2210  degrees of freedom
## AIC: 2271.9
##
## Number of Fisher Scoring iterations: 4

r2 <- calcular_pseudo_R2(modelo_it2)
r2
## pseudo_R2   p_valor
## 0.2528374 0.0000000
```

Para este modelo resultan **d1**, **d2** y **edad** significativas.

```
tabla_resumen_modelos <- tabla_resumen_modelos %>%
  add_row(Nombre = "modelo_it2", AIC = modelo_it2$aic, R2 = r2[1], Observacion = "Modelo todas variables
rm(r2)
```

Ahora seleccionamos variables con stepwise

```
step(modelo_it2, direction = "both")
```

```
## Start:  AIC=2271.92
## diagnostico ~ d1 + d2 + d3 + d4 + d5 + edad
##
##        Df Deviance    AIC
## - d4    1   2257.9 2269.9
## - d3    1   2258.2 2270.2
## - d5    1   2258.2 2270.2
## <none>      2257.9 2271.9
## - d2    1   2261.8 2273.8
## - d1    1   2307.5 2319.5
## - edad  1   2840.8 2852.8
##
## Step:  AIC=2269.94
## diagnostico ~ d1 + d2 + d3 + d5 + edad
##
##        Df Deviance    AIC
## - d3    1   2258.2 2268.2
## - d5    1   2258.2 2268.2
## <none>      2257.9 2269.9
## - d2    1   2261.9 2271.9
## + d4    1   2257.9 2271.9
## - d1    1   2319.4 2329.4
## - edad  1   2844.4 2854.4
```

```
##
## Step:  AIC=2268.19
## diagnostico ~ d1 + d2 + d5 + edad
##
##        Df Deviance    AIC
## - d5    1   2258.5 2266.5
## <none>      2258.2 2268.2
## + d3    1   2257.9 2269.9
## + d4    1   2258.2 2270.2
## - d2    1   2262.4 2270.4
## - d1    1   2319.5 2327.5
## - edad  1   2852.3 2860.3
##
## Step:  AIC=2266.53
## diagnostico ~ d1 + d2 + edad
##
##        Df Deviance    AIC
## <none>      2258.5 2266.5
## + d5    1   2258.2 2268.2
## + d3    1   2258.2 2268.2
## + d4    1   2258.5 2268.5
## - d2    1   2264.1 2270.1
## - d1    1   2321.8 2327.8
## - edad  1   2861.7 2867.7
```

```
##
## Call:  glm(formula = diagnostico ~ d1 + d2 + edad, family = "binomial",
##     data = datos_it2)
##
## Coefficients:
## (Intercept)           d1           d2         edad
##   -8.424211     0.085258    -0.004821     0.104276
##
## Degrees of Freedom: 2216 Total (i.e. Null);  2213 Residual
## Null Deviance:       3022
## Residual Deviance: 2259  AIC: 2267
```

Creamos el modelo con los resultados

```
modelo_it3 <- glm(formula = diagnostico ~ d1 + d2 + edad, family = "binomial", data = datos_it2)
summary(modelo_it3)
```

```
##
## Call:
## glm(formula = diagnostico ~ d1 + d2 + edad, family = "binomial",
##     data = datos_it2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5674  -0.7465  -0.4144   0.8437   2.4253
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -8.424211   0.414310 -20.333  < 2e-16 ***
## d1           0.085258   0.010908   7.816 5.44e-15 ***
## d2          -0.004821   0.002058  -2.343   0.0191 *
## edad         0.104276   0.004935  21.128  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3022.0  on 2216  degrees of freedom
## Residual deviance: 2258.5  on 2213  degrees of freedom
## AIC: 2266.5
##
## Number of Fisher Scoring iterations: 4
```

```
r2 <- calcular_pseudo_R2(modelo_it3)
r2
```

```
## pseudo_R2   p_valor
## 0.2526338 0.0000000
```

Se obtiene que todas las variables son significativas. Se observa un AIC de 2266.53, que es el mejor obtenido hasta ahora.

- Por cada incremento de una unidad en **d1**, los odds de tener un diagnóstico postivo incrementan en 8.9%, manteniendo el resto de variables constantes.
- Por cada incremento de una unidad en **d2**, los odds de tener un diagnóstico positivo se reducen en 0.48%, manteniendo el resto de variables constantes.
- Por cada incremento de un año en la **edad**, los odds de tener un diagnóstico positivo se incrementan en 10.99% manteniendo el resto de variables constantes.

```
tabla_resumen_modelos <- tabla_resumen_modelos %>%
  add_row(Nombre = "modelo_it3", AIC = modelo_it3$aic, R2 = r2[1], Observacion = "Se selecciono variable

rm(r2)
```

Volvemos a buscar puntos influyentes

```
puntos_influyentes <- data.frame(d_cook = cooks.distance(modelo_it3)) %>% arrange(-d_cook)
head(puntos_influyentes, n = 15)
```

```
##           d_cook
## 2    0.010727185
## 125  0.009933780
## 331  0.008311292
## 44   0.007338478
## 89   0.006713961
## 620  0.006543840
## 76   0.006418496
## 45   0.006395895
## 1628 0.006169848
## 36   0.006151145
```

```
## 4    0.006104185
## 628  0.005468280
## 23   0.005425973
## 186  0.005054355
## 2038 0.004970606
```

Observamos que ninguna d_cook es mayor a 1 y tampoco que haya algunos puntos más alejados que otros. Por esto, no eleminamos ninguna observación.

Como no se observaron cambios significativos al excluir la observación 368, vamos a seguir el análisis con los datos completos (datos_it1 y modelo_it1).

```
rm(datos_it2)
```