Ejercicio 2

Selección de Variables

Considere la base de datos PimaIndiansDiabetes2 del paquete mlbench, sólo las observaciones con respuesta en todas las variables. Suponga que el objetivo del estudio es usar las ocho variables clínicas observadas en las pacientes para estudiar cuáles de éstas, adicionales o en lugar de la variable glucose, son los factores que ayudan a modelar mejor la probabilidad de presentar o no diabetes (var diabetes).

Solución:

Se nos solicitó realizar una regresión, para entender qué variables podrían ser las más influyentes en el diagnóstico de diabetes según algunas características clínicas y demográficas de la trribu Pima

Cargamos la base de datos y lo primero que observamos en los datos, es que tenemos múltiples datos sin información (*"NA"s*), por lo que, para el análisis se eliminarán dichas observaciones, como se indica anteriormente.

Tras haber realizado nuestro preprocesamiento de los datos, comenzamos realizando un análisis descriptivo del comportamiento de la variable diabetes, a través de varias de las características obtenidas, una posible primera idea es que un mayor nivel de glucosa plasmática, puede ser un indicativo de presentar diabetes, veamos si estamos en la correcto a través de la siguiente gráfica:

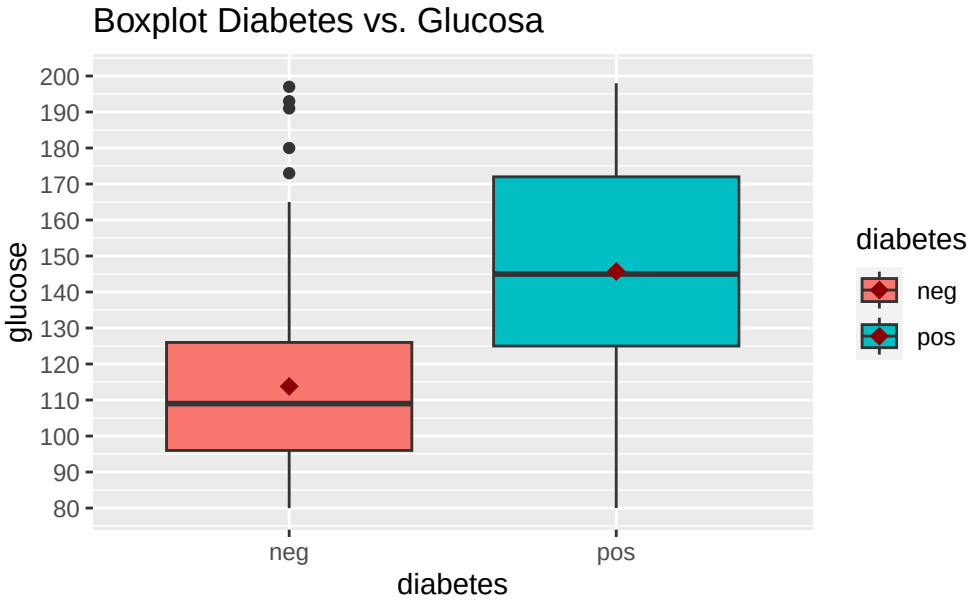

Figura 4: Diabetes vs Glucosa

Como podemos ver en la figura 4, en efecto, las mujeres que presentan diabetes diagnosticada, tienen un mayor nivel de glucosa plasmática, donde el promedio (que se representa como el rombo rojo) para las mujeres que presentan diabetes es, más o menos, de 145, un valor bastante alto a comparción de las mujeres que no presentan diabetes.

También, podemos realizar el análisis de ver qué relación hay entre el índice de masa corporal y la edad con la diabetes, lo anterior se visualizará a través del siguiente scatter plot:

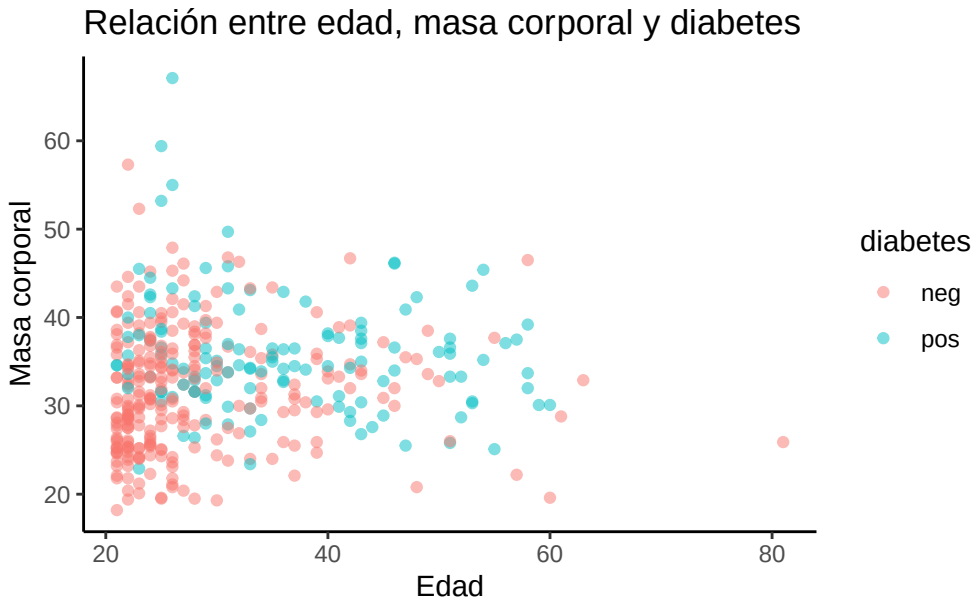

Figura 5: Relación entre edad, masa corporal y diabetes

Podemos observar en la figura 5, que este gráfico nos muestra que el índice de masa corporal tiene una relación más fuerte con la presencia de diabetes que la edad de las pacientes, ya que podemos ver como los puntos de las pacientes que presentan diabetes se encuentran desplazados más hacia arriba (mayor índice de masa corporal) a comparación de las pacientes que no tienen diabetes, sin embargo, sí se puede notar como los puntos de las mujeres que no presentan diabetes se encuentran mayormente concentrados entre las edades 20-30, y en el caso de las mujeres que sí presentan diabetes, podemos observar cómo se encuentran más dispersos entre los 20 y 60 años.

Hecho todo el análisis anterior, procedemos a realizar nuestra búsqueda del mejor conjunto de variables.

I. Considere un modelo para datos binarios con liga logit. Realice una selección de variables considerando sólo los efectos principales de las variables y usando: a) mejor subconjunto, b) un método stepwise y c) método lasso. En cada caso, presente el mejor modelo obtenido usando el criterio BIC.

Realizaremos los 3 métodos diferentes, donde mostraremos en la siguiente tabla, el BIC de el mejor modelo que se generó en cada caso.

Cuadro 4: BIC de cada modelo obtenido mediante los procesos indicados

Fórmula	Proceso	Familia	Liga	BIC
$diabetes \sim glucose + mass + pedigree + age$	Mejor subconjunto (Efectos principales)	Binomial	Logit	377.0913
$diabetes \sim glucose + mass + pedigree + age$	Step forward (Efectos principales)	Binomial	Logit	377.0913
$diabetes \sim glucose + mass + pedigree + age$	Optimización Lasso (Efectos principales)	Binomial	Logit	377.0913

(El proceso de cálculo de estos modelos y sus BIC se encuentran en los chunks *"bs1"*, *"bf1"* y *"Lasso 1"*)

Como podemos observar en el cuadro 4, todos los modelos obtenidos tienen el mismo BIC, lo cual no nos sirve para comparar dichos modelos, por lo que procedemos a la búsqueda de más modelos.

II. Considere un modelo para datos binarios con liga logit. Realice una selección de variables considerando en el modelo los efectos principales de las variables, así como su interacción y el cuadrado de las variables, sólo considerando: a) un método stepwise y b) método lasso. En cada caso, presente el mejor modelo obtenido usando el criterio BIC.

Realicemos los siguientes modelos solicitados, donde a continuación, mostraremos la siguiente tabla con el BIC de el mejor modelo que se generó en cada método.

Cuadro 5: BIC de cada modelo obtenido mediante los procesos indicados

Fórmula	Proceso	Familia	Liga	BIC
$diabetes \sim glucose + mass + pedigree + age + I(age^2)$	Step forward $(.^2 + I(variables)^2)$	Binomial	Logit	370.6345
$diabetes \sim glucose : mass + glucose : age + I(glucose^2)$	Optimización Lasso $(.^2 + I(variables)^2)$	Binomial	Logit	381.8993

(El proceso de cálculo de estos modelos y sus BIC se encuentran en los chunks *"bf2"* y *"Lasso 2"*)

Como podemos observar en la cuadro 5, el mejor modelo según el criterio BIC, es el que fue obtenido por medio del método step forward, donde obtuvimos un *BIC=370.63* y, que hasta este punto, ha sido el mejor modelo encontrado.

III. Considere posibles modificaciones a los incisos i) y ii) realizando lo siguiente. A) usar ligas probit o cloglog; B) usar el logaritmo como preprocesamiento a las variables. En cada caso, presente el mejor modelo obtenido usando el criterio BIC.
IV. Presente en una sola tabla los diferentes modelos obtenidos, así como el BIC de cada uno. Comente sobre los resultados, por ejemplo, qué variables aparecen en la mayoría de modelos, si parece necesario incluir interacciones o realizar un preprocesamiento a los datos y, considerando el mejor de todos, qué interpretación se puede dar a algunos de los coeficientes del modelo.

A continuación, en el siguiente cuadro 6, se muestran todos los mejores modelos encontrados para este ejercicio por medio de los métodos *"Mejor Subconjunto", "Step forward" y "Optimización Lasso"*, es decir, los mostrados previamente en el *inciso i)* y *ii)*, más los que se piden en el *inciso iii),* observemos que se realizaron dos modelos extra, debido a que, en la mayoría de los casos, los modelos obtenidos por medio del *método step forward* son los que tienen un menor BIC, por lo que se probaron otros dos últimos modelos con el preprocesamiento, en el que se había aplicado logaritmo a las variables númericas, solo cambiando de liga: probit y cloglog, en donde se consideró los efectos principales, sus interacciones y las variables al cuadrado, es decir, $.^2 + I(variables)^2$

Cuadro 6: BIC de cada modelo obtenido mediante los procesos indicados

Fórmula	Proceso	Familia	Liga	BIC
$diabetes \sim glucose + mass + pedigree + age$	Mejor subconjunto (Efectos principales)	Binomial	Logit	377.0913
$diabetes \sim glucose + mass + pedigree + age$	Step forward (Efectos principales)	Binomial	Logit	377.0913
$diabetes \sim glucose + mass + pedigree + age$	Optimización Lasso (Efectos principales)	Binomial	Logit	377.0913
$diabetes \sim glucose + mass + pedigree + age + I(age^2)$	Step forward $(.^2 + I(variables)^2)$	Binomial	Logit	370.6345
$diabetes \sim glucose : mass + glucose : age + I(glucose^2)$	Optimización Lasso $(.^2 + I(variables)^2)$	Binomial	Logit	381.8993
$diabetes \sim glucose + mass + age$	Mejor subconjunto (Efectos principales)	Binomial	Probit	378.0613
$diabetes \sim glucose + mass + age$	Step forward (Efectos principales)	Binomial	Probit	378.0613
$diabetes \sim glucose + mass + age$	Optimización Lasso (Efectos principales)	Binomial	Probit	378.0613
$diabetes \sim glucose + mass + age + I(age^2)$	Step forward $(.^2 + I(variables)^2)$	Binomial	Probit	371.2480
$diabetes \sim glucose : mass + glucose : age + I(glucose^2)$	Optimización Lasso $(.^2 + I(variables)^2)$	Binomial	Probit	382.0229
$diabetes \sim pregnant + glucose + mass$	Mejor subconjunto (Efectos principales)	Binomial	Cloglog	385.6527
$diabetes \sim glucose + triceps + age$	Step forward (Efectos principales)	Binomial	Cloglog	387.9489
$diabetes \sim glucose + mass + age$	Optimización Lasso (Efectos principales)	Binomial	Cloglog	386.6826
$diabetes \sim glucose + mass + I(pregnant^2)$	Step forward $(.^2 + I(variables)^2)$	Binomial	Cloglog	386.0716
$diabetes \sim glucose : mass + glucose : age + I(glucose^2)$	Optimización Lasso $(.^2 + I(variables)^2)$	Binomial	Cloglog	394.1181
$diabetes \sim glucose + mass + pedigree + age$	Mejor subconjunto (Efectos principales) - log(variables predictoras)	Binomial	Logit	370.5851
$diabetes \sim glucose + mass + pedigree + age$	Step forward (Efectos principales) - log(variables predictoras)	Binomial	Logit	370.5851
$diabetes \sim glucose + mass + pedigree + age$	Optimización Lasso (Efectos principales) - log(variables predictoras)	Binomial	Logit	370.5851
$diabetes \sim mass + pedigree + age + I(glucose^2) + I(age^2)$	Step forward $(.^2 + I(variables)^2)$ - log(variables predictoras)	Binomial	Logit	368.8146
$diabetes \sim glucose : mass + glucose : age + pedigree : age + I(glucose^2)$	Optimización Lasso $\log(.^2 + I(variables)^2)$ - log(variables predictoras)	Binomial	Logit	371.2793
$diabetes \sim mass + pedigree + age + I(glucose^2) + I(age^2)$	Step forward $(.^2 + I(variables)^2)$ - log(variables predictoras)	Binomial	Probit	370.0799
$diabetes \sim glucose + mass + age + I(age^2)$	Step forward $(.^2 + I(variables)^2)$ - log(variables predictoras)	Binomial	Cloglog	375.1478

(El proceso de cálculo de estos modelos y sus BIC se encuentran en los chunks *"probit"* ,*"cloglog"* y *"log"*)

Analizando el cuadro 6, podemos ver que en todos los modelos, las variables que más predominan son: glucose (*Concentración de glucosa plasmática*) , mass (*Índice de masa corporal*) y age (*Edad en años*), justo las variables que intuíamos en un principio que podían ser relevantes para determinar si una paciente podría tener diabetes o no.

Parece que haber considerado modelos con interacciones y términos cuadráticos de las variables, ayudó a reducir el valor del BIC, aunque no demasiado, existiendo un caso en el que inclusive, el mejor modelo encontrado por medio del *mejor subconjunto* considerando únicamente efectos principales es mejor que los modelos que también consideran interacciones y las variables al cuadrado (*óbservese los modelos con Liga Cloglog).*

Sin embargo, sí tuvo un mayor efecto realizar un preprocesamiento a las variables predictoras (todas siendo numéricas) aplicando logaritmo, donde justamente encontramos el mejor modelo dentro de todos los realizados por medio del criterio BIC, el cual es el siguiente:

```
Call:
glm(formula = diabetes ~ I(glucose^2) + age + mass + pedigree +
    I(age^2), family = binomial, data = datos_log)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -86.61383   22.21186  -3.899 9.64e-05 ***
I(glucose^2)   0.48788    0.06853   7.120 1.08e-12 ***
age           34.41290   12.38413   2.779 0.005456 **
mass           2.56913    0.74797   3.435 0.000593 ***
pedigree       1.96399    0.70108   2.801 0.005089 **
I(age^2)      -4.51986    1.73478  -2.605 0.009176 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 498.10  on 391  degrees of freedom
Residual deviance: 332.99  on 386  degrees of freedom
AIC: 344.99

Number of Fisher Scoring iterations: 5

[1] 368.8146
```

Figura 6: Resumen del modelo elegido

Interpretando los coeficientes del modelo que se observan en la figura 6 tenemos que:

- I(glucose^2): El efecto cuadrático de glucose es positivo y altamente significativo. Esto indica que altas concentraciones de glucosa plásmatica están fuertemente asociados con mayor probabilidad de diabetes.

  Y al ser un término cuadrático, nos indica que la probabilidad de tener diabetes podría dispararse rápidamente conforme la concentración de glucosa plasmática aumenta.

- mass (IMC) y pedigree (que investigando, es una función que cuantifica la predisposición genética de cada paciente a desarrollar diabetes, con base en el historial en la familia), nos indica que al ambas tener un valor positivo, sugiere que valores más altos en estas variables, incrementan la probabilidad de tener diabetes.

- I(age^2) y age: El signo negativo de este término cuadrático de la edad, en conjunto con el positivo de la edad simple, nos está indicando una relación parabólica invertida (forma de U invertida). Es decir, que el riesgo de diabetes aumentaría con la edad hasta un punto máximo, y luego empezaría a decaer.

  A partir de la figura 7, parece no haber evidencia en contra de los supuestos, por lo que el modelo es adecuado para utilizarse.

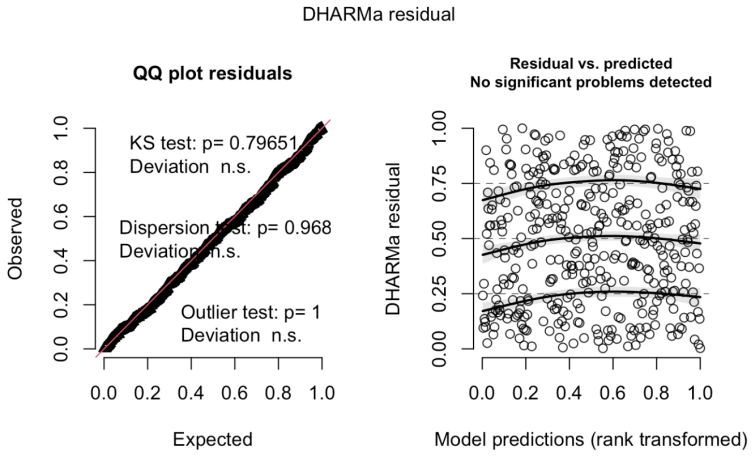

Figura 7: Verificación de Supuestos