UNIVERSIDAD NACIONAL
DE ROSARIO
FACULTAD DE CIENCIAS
ECONÓMICAS Y ESTADÍSTICA

Trabajo Práctico de Modelos Lineales Generalizados

Integrantes:
Candela, Ornella
Mac Kay, Agustina
Ovando, Francisco

Licenciatura en Estadística

Año 2025

Introducción

La diabetes es una enfermedad metabólica crónica caracterizada por niveles elevados de glucosa en sangre, consecuencia de una producción insuficiente de insulina o de una resistencia del organismo a su acción. Esta condición puede derivar en complicaciones cardiovasculares, renales, neurológicas y visuales, afectando de manera significativa la calidad de vida de quienes la padecen.

Dada la relevancia de esta enfermedad en términos de salud pública y su estrecha relación con factores fisiológicos y de estilo de vida, resulta fundamental comprender qué características individuales se asocian a un mayor riesgo de padecerla.

Material y métodos

El presente trabajo tiene como objetivo analizar los factores asociados a la presencia de diabetes en mujeres mayores de 21 años, a través de un modelo lineal generalizado con respuesta binaria. La variable respuesta analizada toma el valor 1 si la persona presenta diagnóstico de diabetes y 0 en caso contrario, por lo que se asume una distribución Bernoulli para cada observación.

Las variables explicativas incluidas en el análisis se describen a continuación:

- Edad: medida en años. Es una variable cuantitativa continua que permite evaluar el efecto del envejecimiento sobre la probabilidad de desarrollar diabetes.
- Embarazo: número de veces que la mujer ha estado embarazada, categorizada en tres niveles:
  - "0": ninguna gestación
  - "1": entre una y tres gestaciones
  - "2": más de tres gestaciones
- Glucosa: concentración de glucosa en ayunas. Fue categorizada en dos niveles:
  - "0": valores normales (100 mg/dL)
  - "1": valores elevados ($>$ 100 mg/dL)
- Presión arterial: presión diastólica medida en mmHg, clasificada en dos niveles:
  - "0": presión dentro del rango normal (80 mmHg)
  - "1": presión elevada ($>$ 80 mmHg)
- Obesidad: medida a través del índice de masa corporal (IMC). Se definió con dos niveles:
  - "0": IMC menor a 30
  - "1": IMC igual o superior a 30

- DPF (Diabetes Pedigree Function): índice que cuantifica la predisposición genética a la diabetes a partir del historial familiar. Fue dicotomizado en:
  - "0": valores bajos (0.5)
  - "1": valores altos (> 0.5)

Análisis descriptivo

A continuación se presenta una visión general de las características clave de la población estudiada.

Tabla 1: Frecuencias absolutas y relativas de diabetes

Diabetes	Frecuencia	Porcentaje (%)
No	261	66.75
Sí	130	33.25
Total	391	100

Tabla 2: Frecuencias absolutas y relativas de obesidad

Obesidad	Frecuencia	Porcentaje (%)
Peso Normal o sobrepeso	129	32.99
Obesidad	262	67.01
Total	391	100

Tabla 3: Frecuencias absolutas y relativas de glucosa

Glucosa	Frecuencia	Porcentaje (%)
Normal	116	29.67
Elevada	275	70.33
Total	391	100

Tabla 4: Frecuencias absolutas y relativas de presión

Presión	Frecuencia	Porcentaje (%)
Normal	313	80.05
Elevada	78	19.95
Total	391	100

Tabla 5: Frecuencias absolutas y relativas de DPF

DPF	Frecuencia	Porcentaje (%)
Riesgo bajo	223	57.03
Riesgo alto	168	42.97
Total	391	100

Tabla 6: Frecuencias absolutas y relativas de embarazo

Embarazos	Frecuencia	Porcentaje (%)
Ninguno	56	14.32
1 a 3	202	51.66
Más de 3	133	34.02
Total	391	100

Tabla 7: Estadísticas descriptivas de la variable edad

Variable	Media	Mediana	DE	Mínimo	Máximo
Edad	30.74	27	9.89	21	63

Los datos muestran que un 33.2% de la población estudiada tiene diabetes, con una alta prevalencia de obesidad (67%) y niveles elevados de glucosa (70.3%). Además, el 43% de los participantes tiene antecedentes familiares de diabetes, lo que podría indicar una predisposición genética relevante. Aunque la mayoría presenta presión arterial normal (80.1%), los factores como la obesidad, los niveles de glucosa y el historial familiar podrían ser más críticos para entender el riesgo de diabetes en esta población.

Seguidamente, se presentan gráficos que permiten observar la relación entre la variable respuesta (presencia de diabetes) y distintas variables explicativas categóricas y continuas. Esta visualización facilita la detección de posibles asociaciones y patrones relevantes para el posterior análisis.

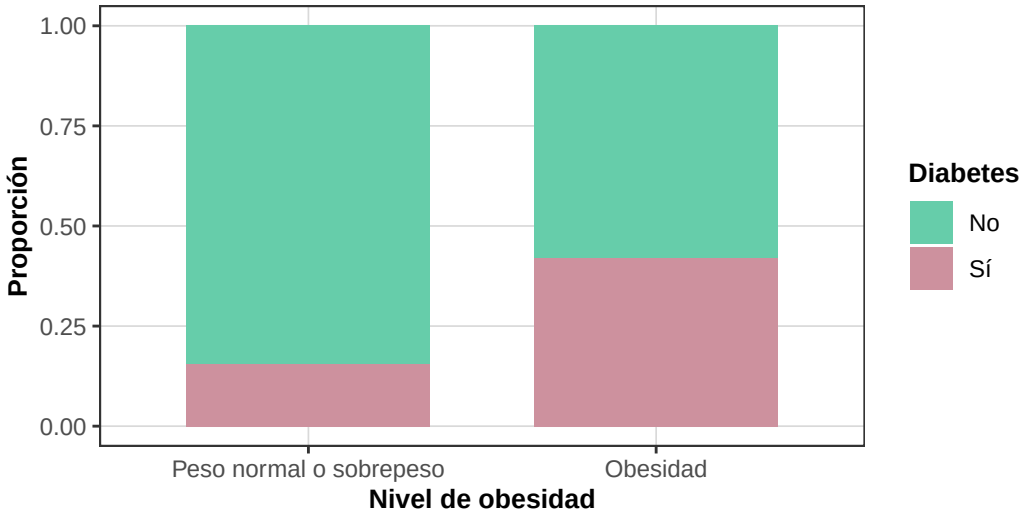


Figura 1: Proporción de personas con diabetes según nivel de obesidad

Se observa que la proporción de mujeres con diabetes es más alta para aquellas con obesidad en comparación a las pacientes sin obesidad.

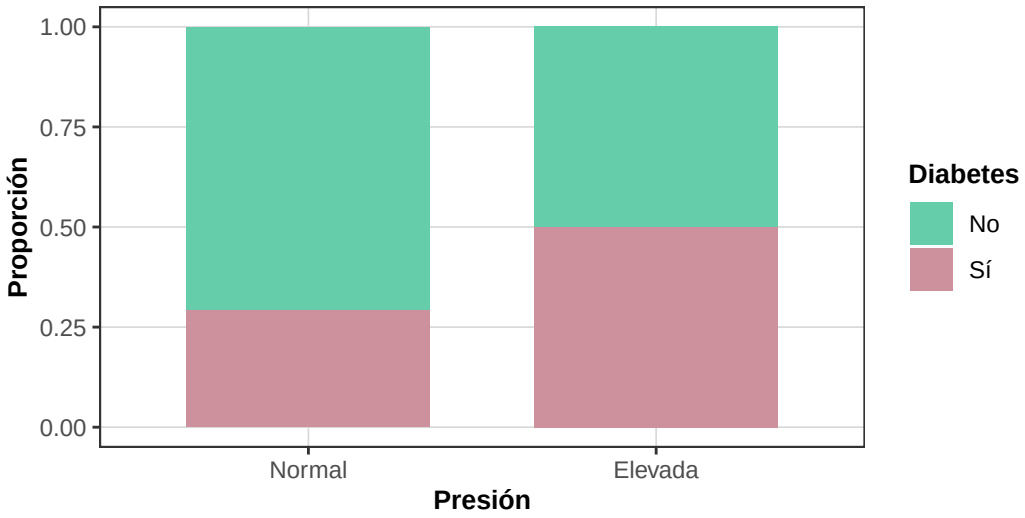


Figura 2: Proporción de personas con diabetes según presión

La presión elevada puede estar influyendo en la probabilidad de padecer diabetes, dado que se puede ver una proporción más alta de diabéticas en las mujeres con presión elevada que en aquellas con presión normal.

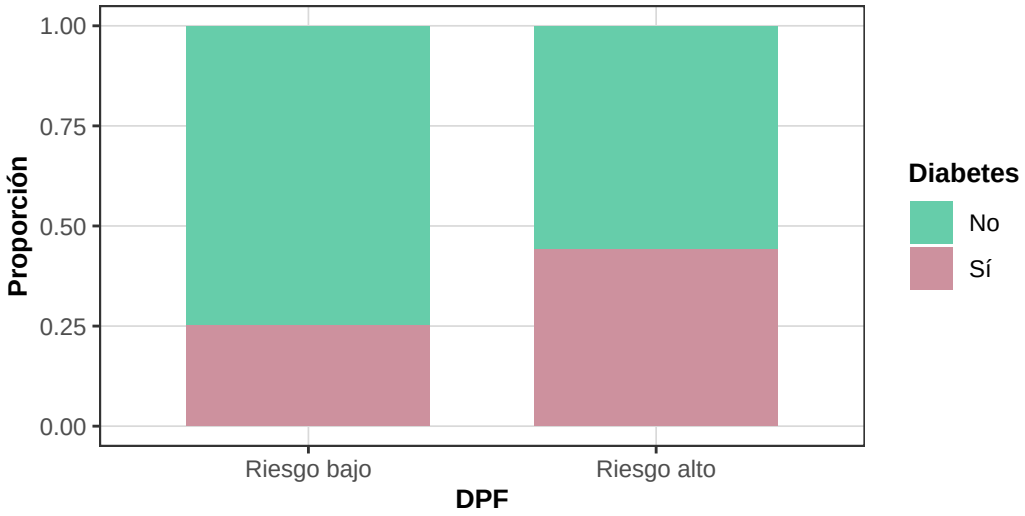


Figura 3: Proporción de personas con diabetes según DPF

Como es de esperarse, hay mayor proporción de diabéticas en aquellas mujeres con un DPF más alto comparando con las pacientes con DPF bajo.

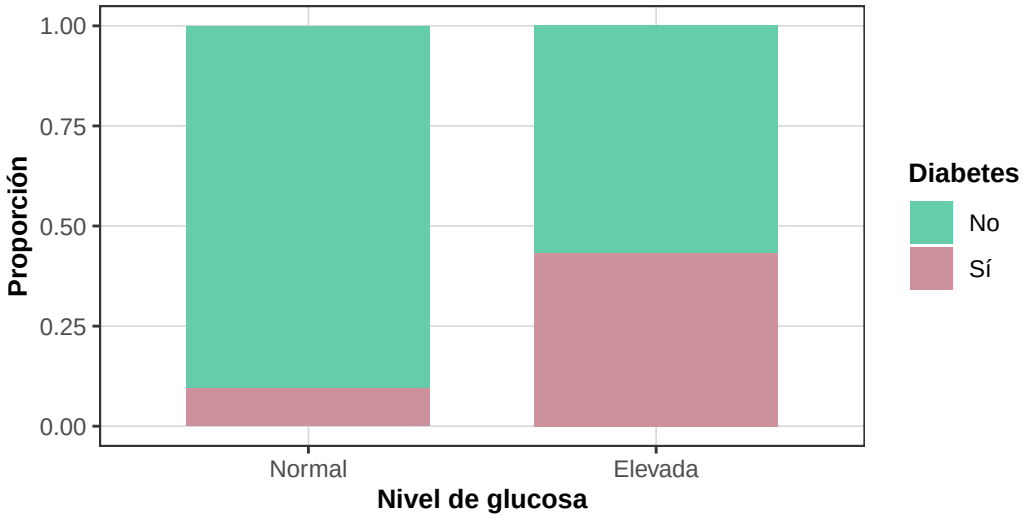


Figura 4: Proporción de personas con diabetes según nivel de glucosa

La relación entre la glucosa y la diabetes es directa por la naturaleza de la enfermedad, hay mayor proporción de diabéticas en pacientes con glucosa elevada que en aquellas con niveles de glucosa normales.

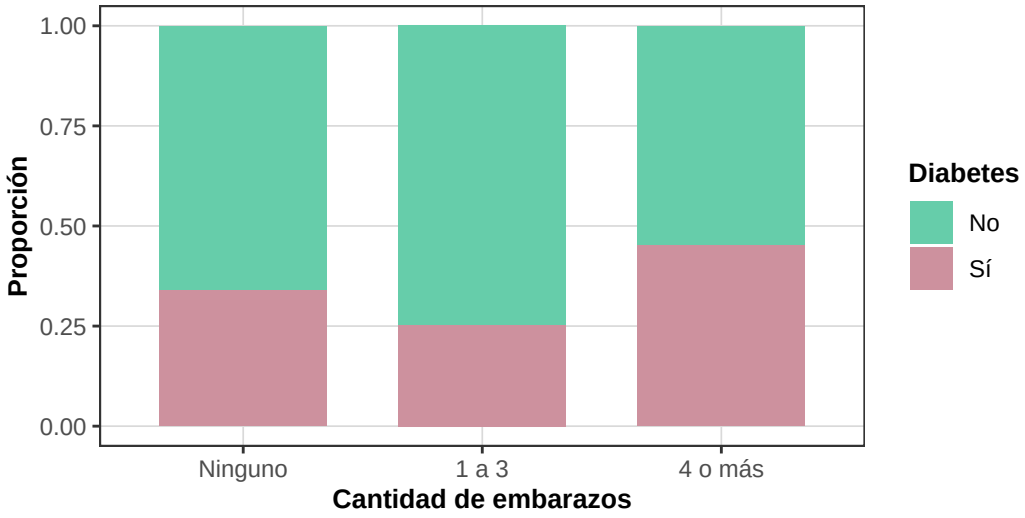


Figura 5: Proporción de personas con diabetes según cantidad de embarazos

Parece haber un aumento en la proporción de mujeres con diabetes en aquellas que tuvieron 4 o más embarazos, y un descenso para las que tuvieron entre 1 y 3, en comparación a las mujeres que no tuvieron ningún embarazo.

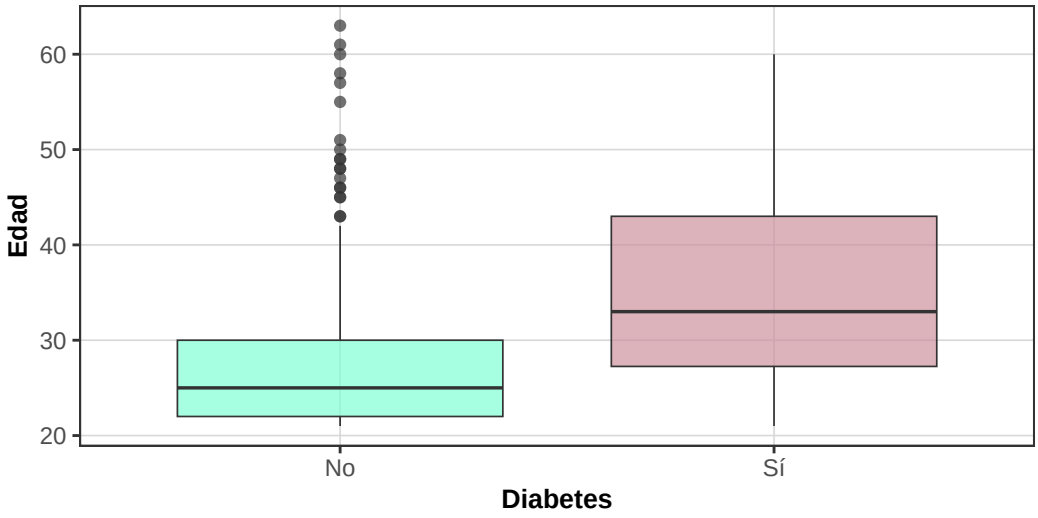


Figura 6: Distribución de la edad según presencia de diabetes

El 50% de las mujeres sin diabetes parece tener entre 23 y 30 años aproximadamente, mientras que el 50% de las mujeres con diabetes tienen entre 27 y 43 años aproximadamente. Esto habla de que las mujeres con diabetes en general tienen una edad mayor a las mujeres sin diabetes.

Modelado estadístico

El trabajo se realizará bajo el enfoque de modelos lineales generalizados, utilizando lo desarrollado por Nelder y Wedderburn. La función de enlace a utilizar es la función *logit*.

Selección del modelo

Para modelar la probabilidad de presentar diabetes se empleó un modelo lineal generalizado con función de enlace logit, dado que la variable respuesta es dicotómica (1 = presencia de diabetes, 0 = ausencia).

En primer lugar, se ajustó un modelo nulo, que incluye únicamente el intercepto, y un modelo completo que incorpora todas las variables explicativas consideradas: edad, glucosa, presión arterial, DPF, embarazo y obesidad.

Posteriormente, se aplicaron procedimientos automáticos de selección de variables mediante la medida de criterio de la información de Akaike eligiendo en cada paso el modelo con menor AIC, utilizando los métodos de selección hacia adelante y hacia atrás. En ambos casos se obtuvo el mismo modelo final, compuesto únicamente por efectos principales.

A continuación, se exploró la posibilidad de incorporar interacciones de segundo orden entre las variables explicativas, ajustando un modelo ampliado y comparándolo con el modelo sin interacciones. Dado que la inclusión de interacciones no mejoró el ajuste, se mantuvo el modelo con efectos simples.

El modelo lineal generalizado seleccionado se expresa como:

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\,\text{Edad}_i + \beta_2\,\text{Glucosa}_i + \beta_3\,\text{DPF}_i + \beta_4\,\text{Obesidad}_i$$

Para evaluar la bondad de ajuste del modelo seleccionado, se aplicó la prueba de Hosmer y Lemeshow. Esta prueba compara las frecuencias observadas y esperadas del evento de interés en grupos formados a partir de los valores predichos por el modelo.

Dado el tamaño muestral disponible, se optó por dividir las observaciones en 10 grupos de igual tamaño según los valores de probabilidad predicha.

Finalmente, se concluye que el modelo presenta un ajuste adecuado a los datos ($p - value = 0.7069$).

Además, se evaluó la adecuación de la función de enlace logit utilizada en el modelo. Para ello, se aplicó el test de especificación de la función de enlace, que consiste en incorporar al modelo un término adicional cuadrático del predictor lineal estimado:

$$M1) \quad \text{logit}(\pi_i) = \beta_0 + \beta_1\,\text{Edad}_i + \beta_2\,\text{Glucosa}_i + \beta_3\,\text{DPF}_i + \beta_4\,\text{Obesidad}_i$$
$$M2) \quad \text{logit}(\pi_i) = \beta_0 + \beta_1\,\text{Edad}_i + \beta_2\,\text{Glucosa}_i + \beta_3\,\text{DPF}_i + \beta_4\,\text{Obesidad}_i + \beta_5\,\hat{\eta}_i^2$$

El procedimiento implica ajustar un modelo extendido que incluye esta nueva variable y comparar su ajuste con el modelo original mediante una prueba de razón de verosimilitud.

En este análisis, la comparación entre ambos modelos no mostró evidencia suficiente para rechazar la adecuación del enlace logit, por lo que se considera apropiado mantenerlo en el modelo final ($p - value = 0.6489$).

Con el objetivo de verificar el supuesto de linealidad entre la variable continua edad y el logit de la probabilidad de presentar diabetes, se implementó una comparación entre dos modelos alternativos.

En primer lugar, se construyó un modelo que incorpora la variable edad como ordinal, dividiendo su rango en cinco categorías según los cuantiles 0.2, 0.4, 0.6 y 0.8. Posteriormente, se ajustó un segundo modelo en el que la edad se representó mediante variables indicadoras (dummies), permitiendo una relación no necesariamente lineal con la respuesta.

El modelo ordinal se puede expresar como:

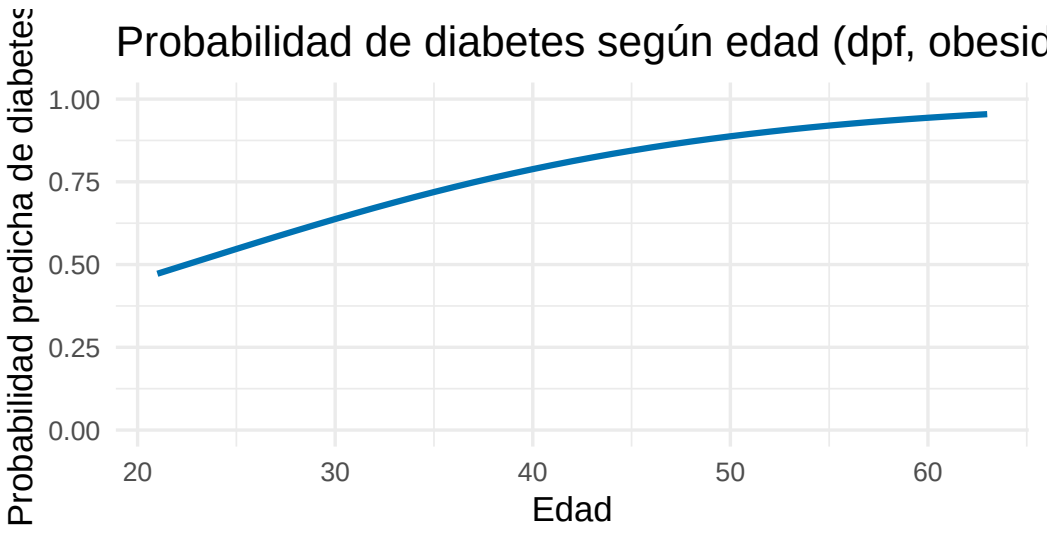
$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \,\text{Glucosa}_i + \beta_2 \,\text{Obesidad}_i + \beta_3 \,\text{DPF}_i + \beta_4 \,\text{EdadORDEN}_i$$

Mientras que el modelo con variables indicadoras se define como:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \,\text{Glucosa}_i + \beta_2 \,\text{Obesidad}_i + \beta_3 \,\text{DPF}_i + \beta_4 \,\text{Edad2}_i + \beta_5 \,\text{Edad3}_i + \beta_6 \,\text{Edad4}_i + \beta_7 \,\text{Edad5}_i$$

Ambos modelos fueron comparados mediante la prueba de razón de verosimilitud. Dado que la diferencia en el ajuste entre el modelo con edad tratada como ordinal y el modelo con las variables dummies no resultó estadísticamente significativa, se concluye que la relación entre la edad y el logit de la probabilidad de diabetes puede considerarse lineal ($p - value = 0.6408$). Por tanto, se mantiene la variable edad como cuantitativa continua en el modelo final.

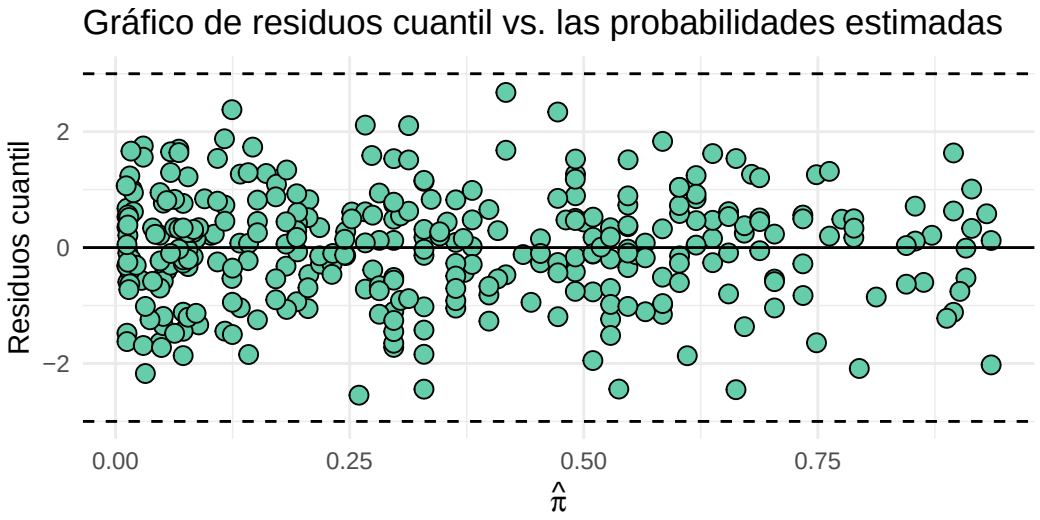
Curvas ajustadas



Análisis de residuos

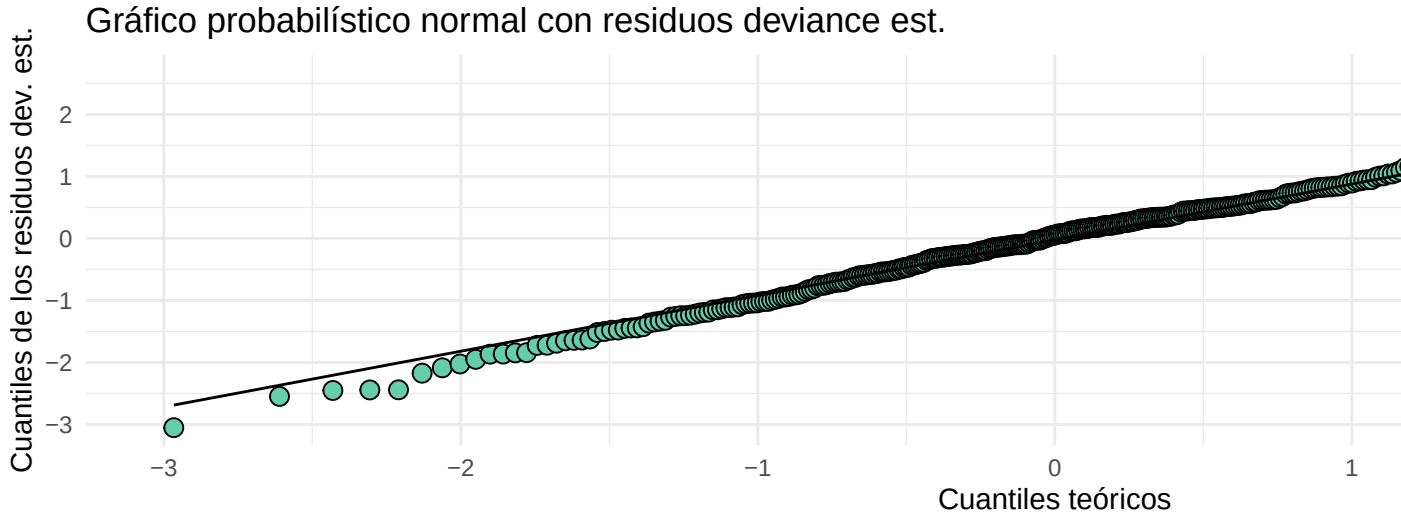
Evaluación de la componente sistemática

Observe como se ve el gráfico de los residuos vs. las medias estimadas bajo el enlace logit:



Evaluación de la componente aleatoria

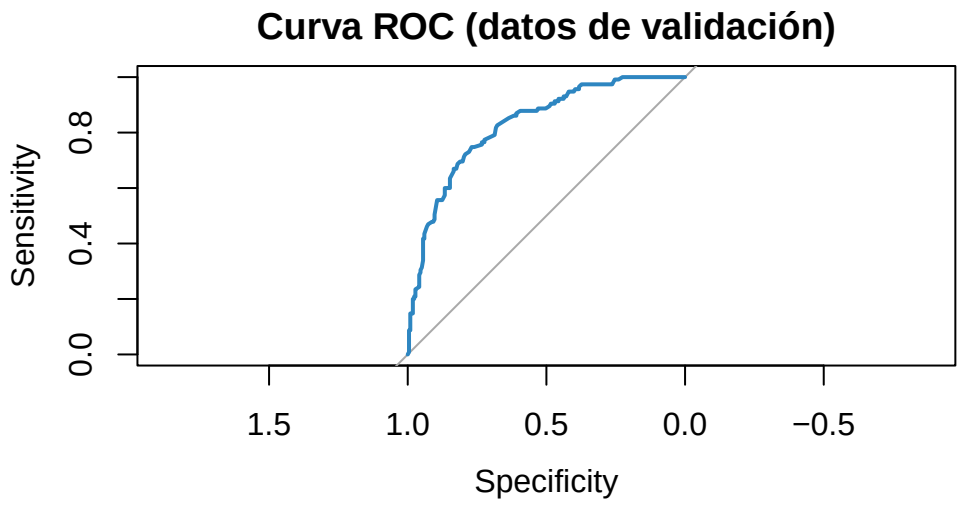
Mediante un gráfico QQ puede evaluarse si la distribución supuesta es adecuada.



Capacidad predictiva del modelo

Con el objetivo de evaluar la capacidad predictiva del modelo ajustado, se procedió a dividir la base de datos en dos subconjuntos: un conjunto de entrenamiento, utilizado para estimar los parámetros del modelo, y un conjunto de validación, empleado para evaluar su desempeño sobre observaciones no utilizadas en el ajuste. Esta separación permite obtener una medida más realista de la capacidad del modelo para generalizar a nuevos datos.

A partir de las probabilidades estimadas en el conjunto de validación, se construyó la curva ROC, que representa la sensibilidad y especificidad frente a para distintos puntos de corte. Este gráfico permite analizar el equilibrio entre verdaderos positivos y falsos positivos, siendo un instrumento útil para determinar el umbral óptimo de clasificación.




```
##   threshold
## 1 0.3767094
```

El punto de corte seleccionado fue de 0.377, correspondiente al valor que maximiza simultáneamente la sensibilidad y la especificidad del modelo. Con este umbral se obtuvo la siguiente matriz de confusión, que resume el desempeño del modelo en la clasificación de los casos:

Predicho / Observado	**No diabetes**	**Diabetes**
No diabetes	31	7
Diabetes	10	11

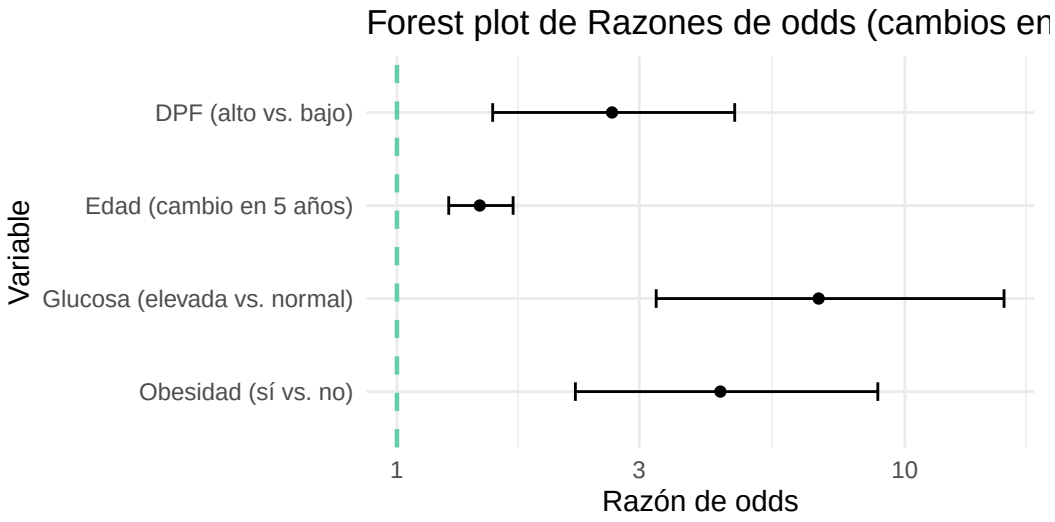
A partir de esta matriz, el modelo logró clasificar correctamente al 71% de las observaciones del conjunto de validación, con una sensibilidad del 61% (proporción de diabéticas correctamente identificadas) y una especificidad del 75.6% (proporción de no diabéticas correctamente clasificadas).

Resultados

Razones de odds

Tabla 9: Razones de Odds estimadas por el modelo

Variable	$\widehat{RO}$	IC de $\widehat{RO}$	
		Límite inferior	Límite superior
Glucosa (elevada vs. normal)	6.765346	3.239073	15.676124
Obesidad (sí vs. no)	4.333456	2.245384	8.846332
DPF (alto vs. bajo)	2.651675	1.543399	4.623491
Edad (cambio en 5 años)	1.455573	1.264587	1.693026



Conclusión