

Facultad de Ciencias

UNAM

ESTADÍSTICA 3. MODELOS DE SUPERVIVENCIA Y SERIES DE TIEMPO

Proyecto final

ANÁLISIS DE SUPERVIVENCIA DE CLIENTES BANCARIOS

Enríquez Hernández Leobardo

14 de junio de 2024

Índice

Introducción.	1
1. Estadística descriptiva y procesamiento de datos.	1
2. El problema de supervivencia para la salida de los clientes.	5
3. Un modelo logit para modelar la salida de los clientes.	10
Conclusiones	15
Referencias	16

Introducción.

Este documento tiene como principal objetivo el análisis de supervivencia con la información de los clientes de un banco (KAGGLE (2022)), particularmente analizar la variable que indica si el cliente abandona o no abandona el banco (Exited), tomando como variable de tiempo los años de tenencia de la cuenta (Tenure). Complementamos el análisis con un modelo de regresión probabilístico logit para entender qué variables aumentan la probabilidad de salida del cliente y clasificar cierto perfil de clientes y su probabilidad de salir, como un breve planteamiento que se pudiera profundizar en algún trabajo futuro.

En la primera sección se hace la estadística descriptiva de los datos con los que se trabajará, para dar un contexto y un panorama general de la naturaleza y características de las variables. En esta sección, haremos procesamiento de los datos en caso de que sea necesario, por ejemplo tratar con valores perdidos, valores atípicos, etc. En la segunda sección se plantea y desarrolla el problema de supervivencia. En la tercera sección se analiza un modelo logit, con la variable Exited como variable explicada para complementar el análisis de supervivencia. Y finalmente se presentan las principales conclusiones del trabajo.

1. Estadística descriptiva y procesamiento de datos.

El conjunto de datos es de clientes de un banco con las siguientes variables:

- id: número de fila de la observación, comenzando por el 0.
- CustomerId: número de cuenta del cliente.
- Surname: apellido.
- CreditScore: puntaje de crédito.
- Geography: país de residencia.
- Gender: género del cliente.
- Age: edad del cliente.
- Tenure: cuántos años ha tenido cuenta bancaria en el Banco.
- Balance: saldo de la cuenta.
- NumOfProducts: número de productos bancarios en el Banco.
- HasCrCard: si tiene o no tarjeta de crédito (sí=1).
- IsActiveMember: si es miembro activo del banco (sí=1).
- EstimatedSalary: salario estimado.
- Exited: si el cliente ha dejado el banco por algún periodo (sí=1).

Primero tomaremos un subconjunto del conjunto total de variables, omitiremos variables que no utilizaremos en el análisis tales como id, CustomerId, y Surname. Luego mostraremos en el siguiente cuadro que no hay datos faltantes (NA's) para las variables elegidas.

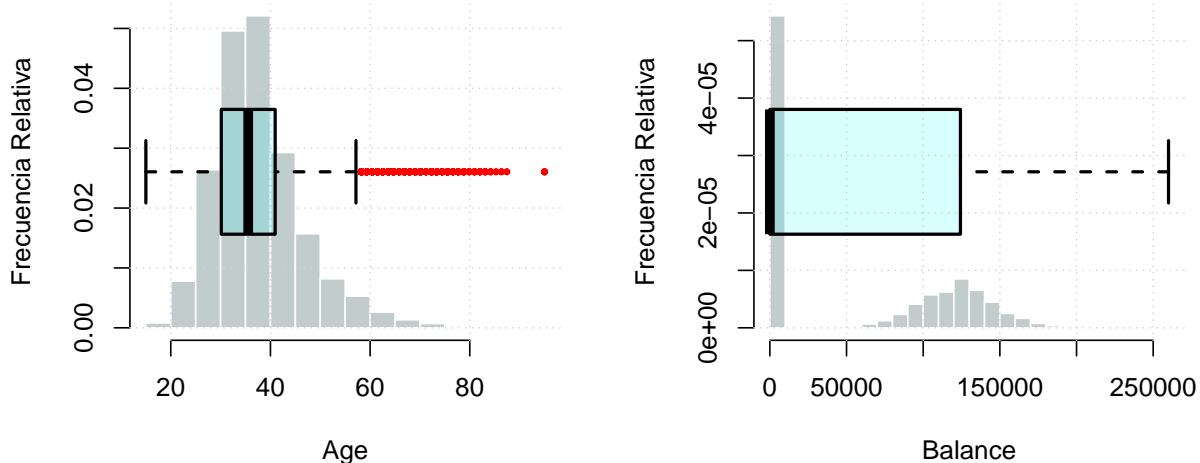
CreditScore : 0	Geography: 0	Gender: 0	Age: 0
Tenure : 0	Balance: 0	NumOfProducts: 0	HasCrCard: 0
IsActiveMember : 0	EstimatedSalary: 0	Exited: 0	

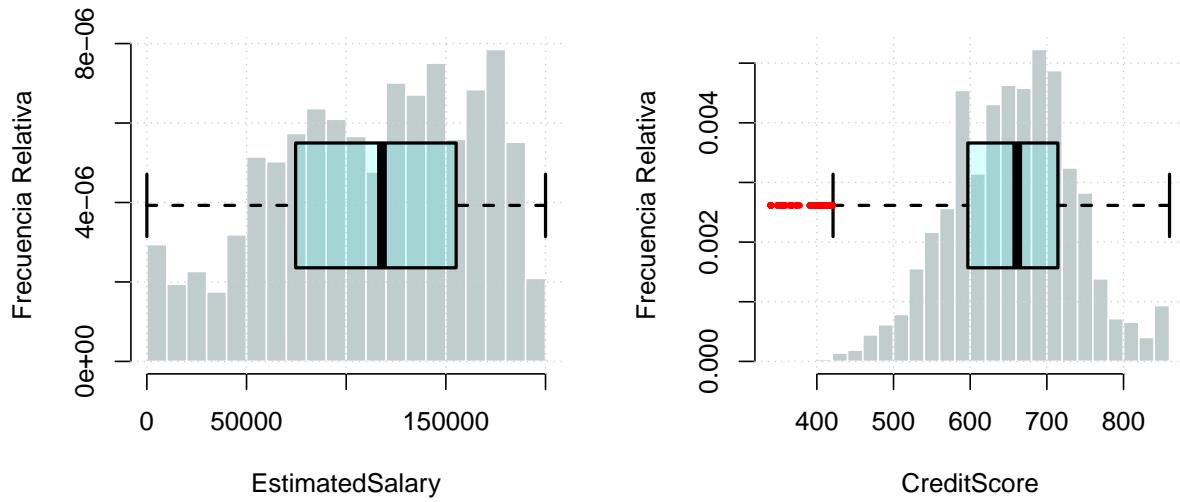
A continuación se muestra la estadística descriptiva de los valores numéricos relevantes. Son 165,034 observaciones, con edades entre 18 y 92 años, con un balance de 0 a 250,898 unidades monetarias, con salario estimado de entre 11.58 a 199,992.5, un score de crédito de 350 a 850, y tenencia de cuenta bancaria de 0 a 10 años. El promedio de edad es de 38 años, con un balance promedio de 55,478 unidades monetarias, un promedio de salario estimado de 112,575.8, un promedio de credit score de 656.45, y un promedio de tenencia de cuenta de 5 años.

Statistic	N	Mean	St. Dev.	Min	Max
Age	165,034	38.126	8.867	18.000	92.000
Balance	165,034	55,478.090	62,817.660	0.000	250,898.100
EstimatedSalary	165,034	112,574.800	50,292.870	11.580	199,992.500
CreditScore	165,034	656.454	80.103	350	850
Tenure	165,034	5.020	2.806	0	10

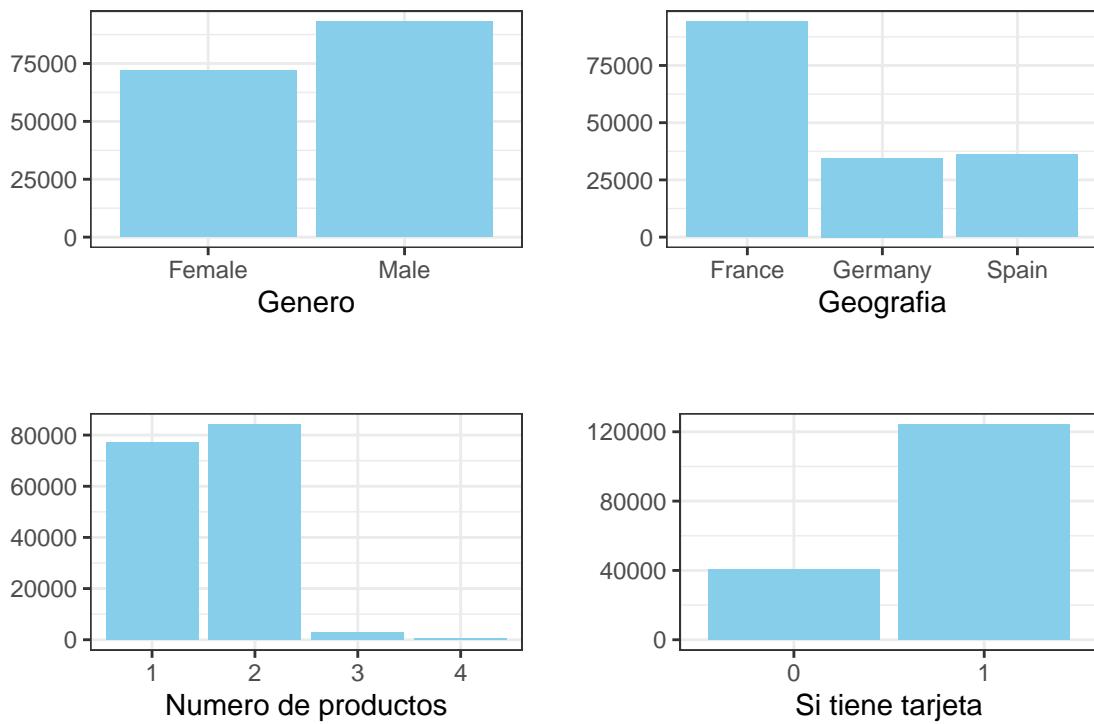
A continuación se muestran los histogramas de frecuencias relativas y boxplot para las variables numéricas, a excepción de `Tenure` que es nuestra variable de tiempo que analizaremos con más detalle más adelante. Por ejemplo, el balance tiene un sesgo muy notable a la izquierda ya que hay muchos valores nulos para esta variable, mientras que el salario estimado presenta una distribución multimodal muy irregular, y el credit score por ser una variable discreta se percibe con varios saltos.

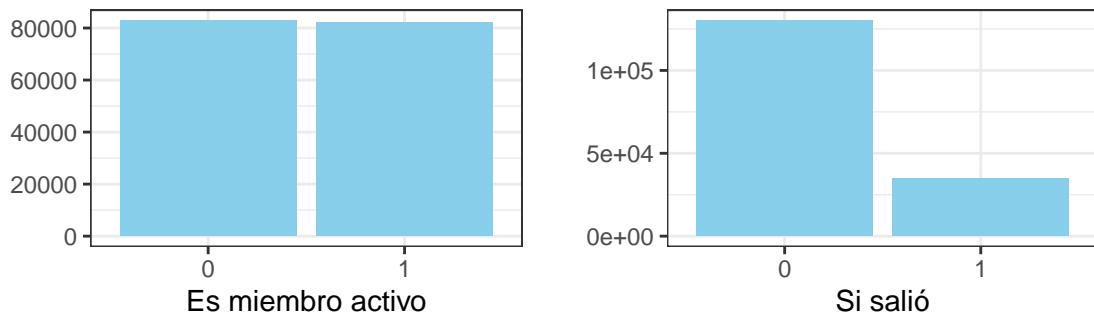
Se observan outliers en la variable de la edad (Age) después de los 55 años, además, después de 85 años hay un brinco de valores de 92 años que corresponde a 11 observaciones. En la variable score de créditos (CreditScore) también hay outliers, los cuales son 80 observaciones con valores menores a 410.





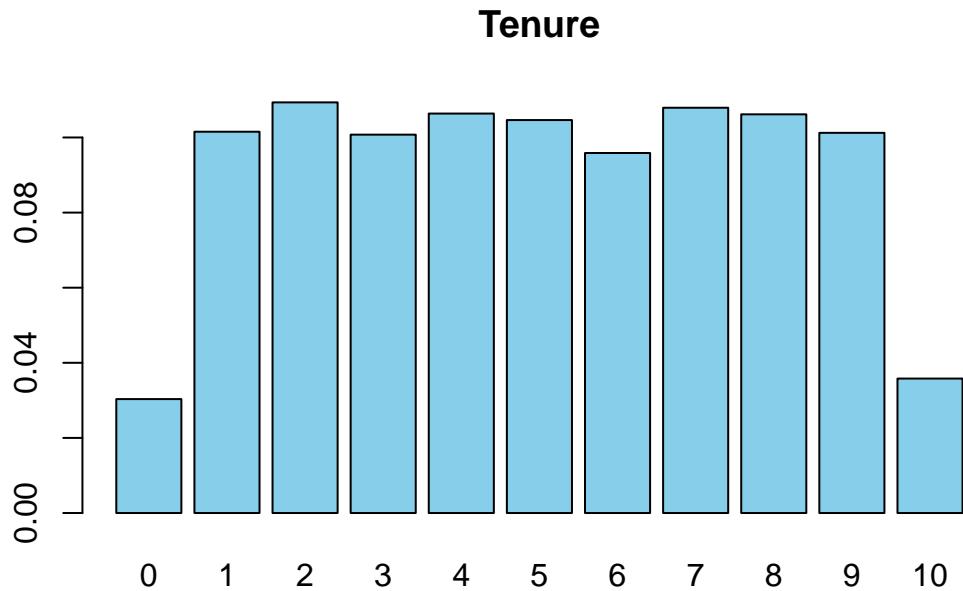
A continuación presentaremos las datos categóricos, por su frecuencia absoluta en el conjunto de datos. Se tienen 71,884 mujeres y 93,150 hombres; 94,215 son de Francia, 34,606 de Alemania y 36,213 de España; 77,374 tiene un solo producto bancario, 84,291 tienen dos, mientras que 2,894 tienen tres productos, y 475 tienen 4 productos; 40,606 no tiene tarjeta de crédito y 124,428 sí tiene; 82,885 no es miembro activo y 82,149 son miembros activos; y 130,113 permanecen con el banco, mientras que 34,921 salieron del banco en algún periodo, es decir, permanecieron el 79 % de los clientes.





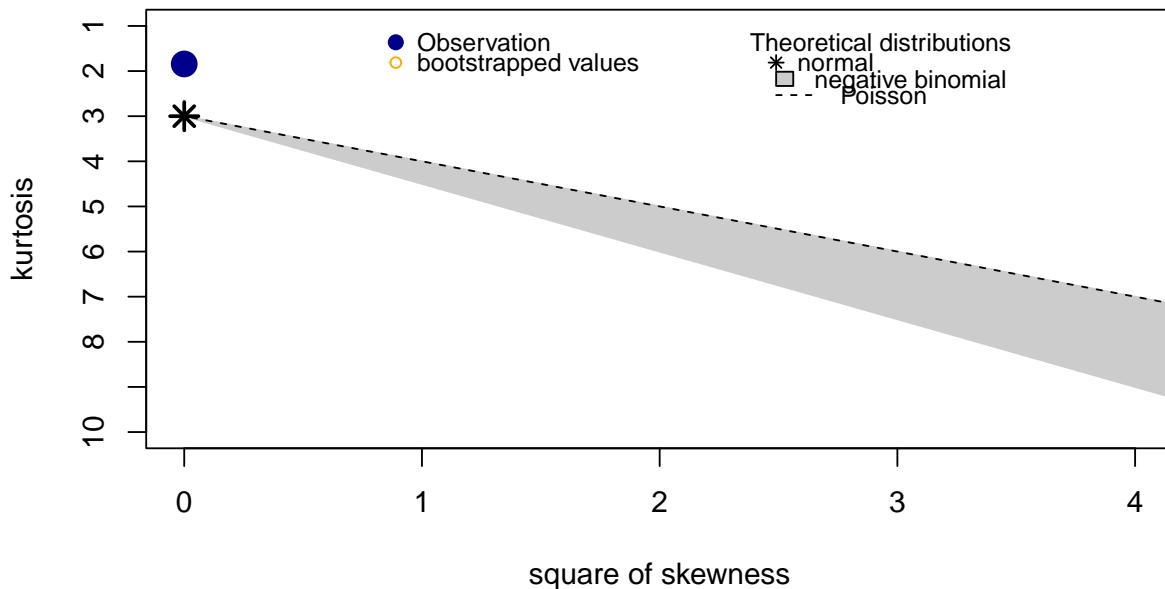
Haremos dos principales cambios en el conjunto de datos, primero quitaremos los outliers de la edad 92 años y luego modificaremos la variable del número de productos para tener una binaria que indique que se tiene 1 producto contratado o se tienen 2 o más productos, pues hay muy pocas observaciones que tienen 3 o 4 productos.

La variable de tiempo en este trabajo es **Tenure**, y es una variable aleatoria discreta por tener valores discretos de 0 a 10 años, a continuación mostramos el histograma de frecuencias relativas.



A continuación, mostramos la gráfica de Cullen y Frey para **Tenure**, lo que permite eximir algunas distribuciones mediante los parámetros de asimetría y curtosis utilizando la función `descdist` para datos discretos; los valores de arranque provienen de muestras aleatorias con reemplazo (bootstrap) de los datos. A partir de este gráfico, nuestras opciones para ajustes parecerían no estar cercanas a ninguna distribución disponible para esta prueba.

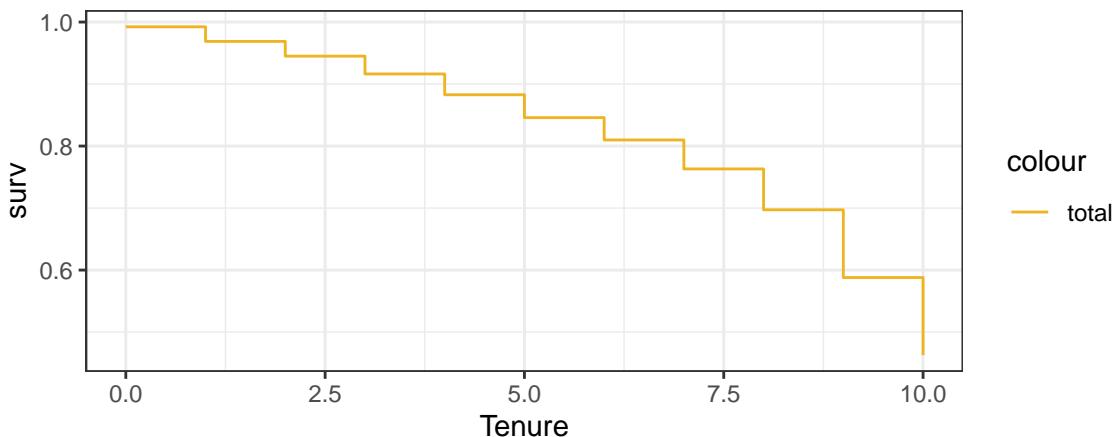
Cullen and Frey graph



2. El problema de supervivencia para la salida de los clientes.

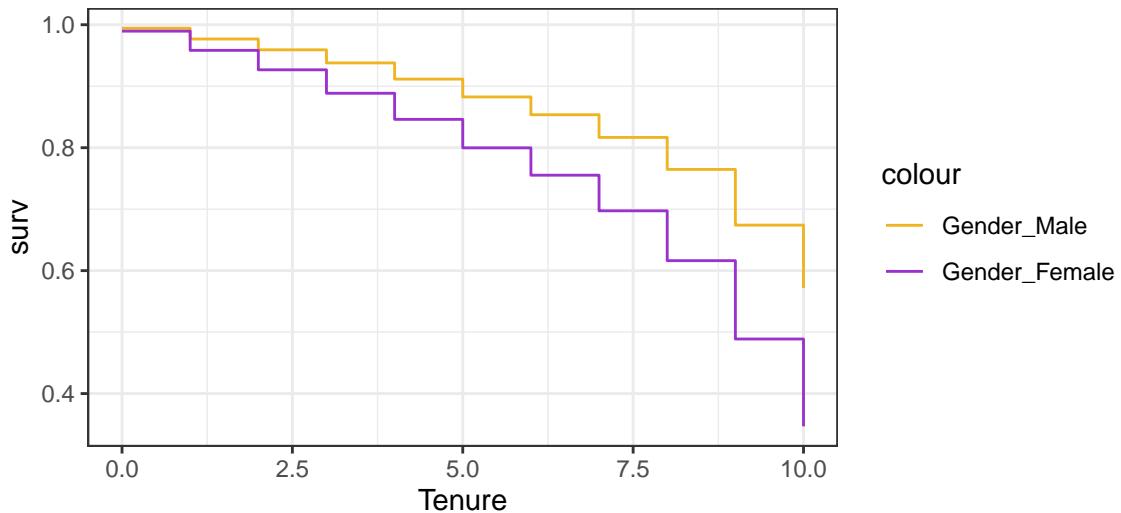
Para esta sección haremos un análisis de supervivencia que se basa en el estudio del tiempo en la ocurrencia de un evento, donde el tiempo de supervivencia o falla se define como el tiempo transcurrido desde el estado inicial hasta la ocurrencia de un evento dado (Villers (2023)).

Usaremos `survfit()` y `Surv()` para construir el objeto de supervivencia estándar, usando Kaplan Meier. Usamos la fórmula `survfit(Surv(Tenure, Exited) ~ 1)` para producir las estimaciones de Kaplan-Meier de la probabilidad de supervivencia en el tiempo para cada una de las categorías de Gender. A continuación, mostramos la curva de supervivencia obtenida, la cual podemos interpretar como la probabilidad de permanecer en el banco más allá de cierto tiempo. La curva es decreciente, lo que implica que conforme pasan los años con la cuenta, la probabilidad de permanecer disminuye.

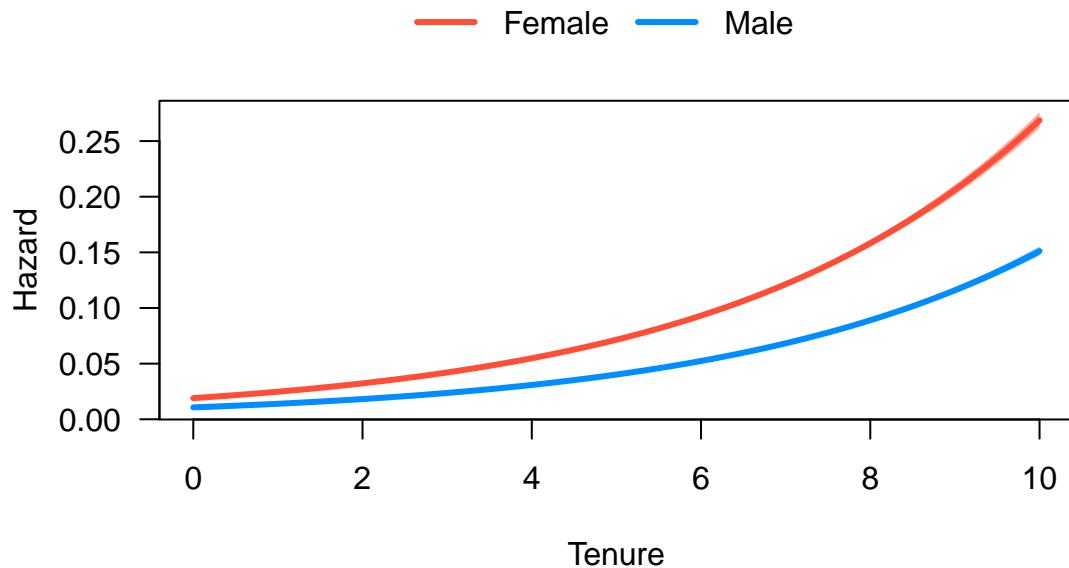


En la siguiente gráfica, podemos observar que por cada uno de los dos grupos en que se clasificaron como

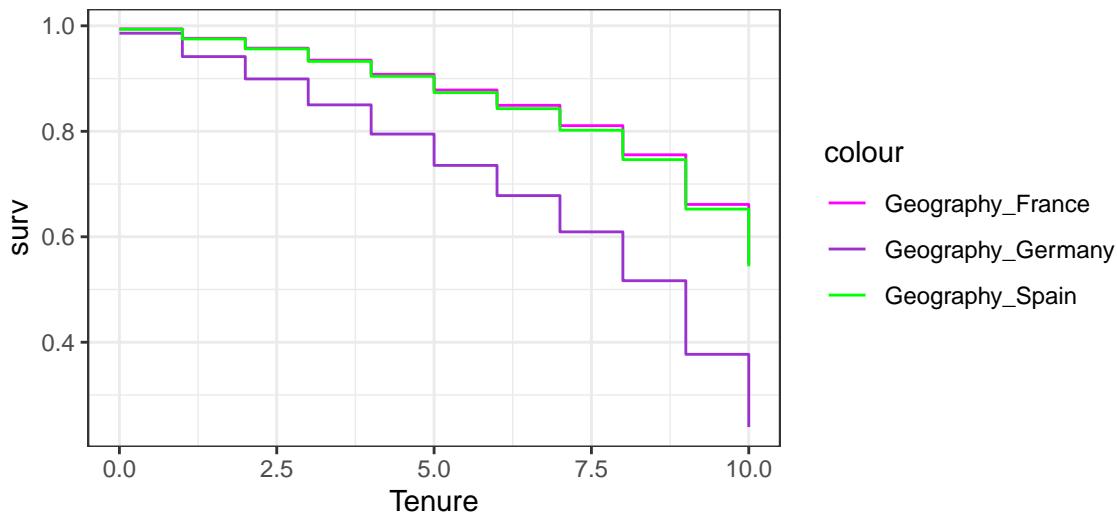
hombre y mujer, la tasa de permanencia es menor para las mujeres comparado con los hombres.



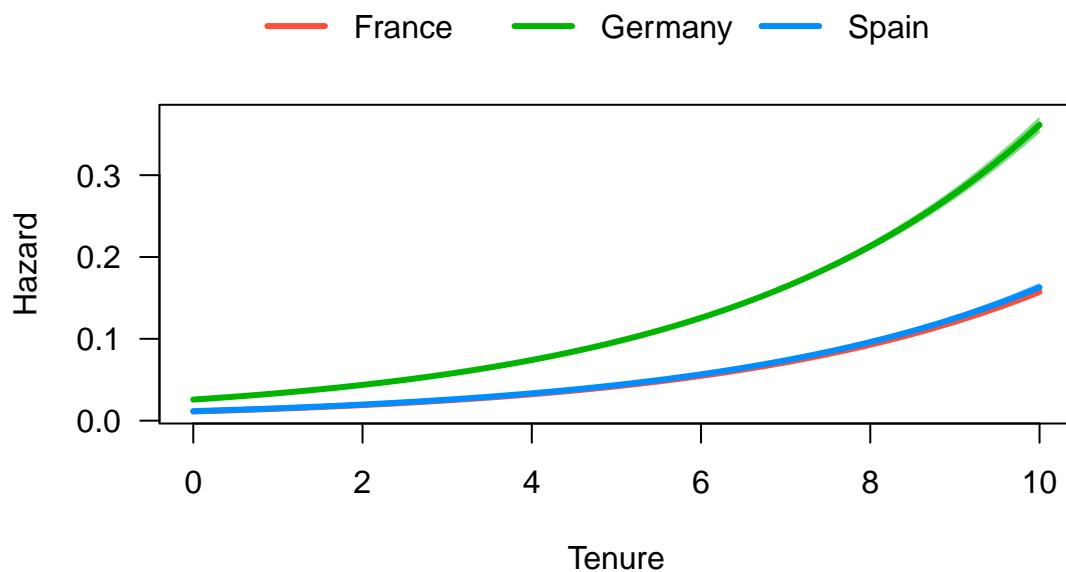
A continuación podemos ver la función de riesgo asociada, la cual es convexa y creciente. Esta función de densidad condicional, llamada también tasa de falla condicional, es la probabilidad de falla (dejar el banco) durante un intervalo de tiempo muy pequeño dado que el individuo ha permanecido hasta el inicio del intervalo. Observamos entonces que la probabilidad condicional de abandonar el banco, dado que hasta ahora no se ha sucedido, aumentará con la variable de tiempo. Esta probabilidad es mayor e incrementa más rápido para las mujeres que para los hombres.



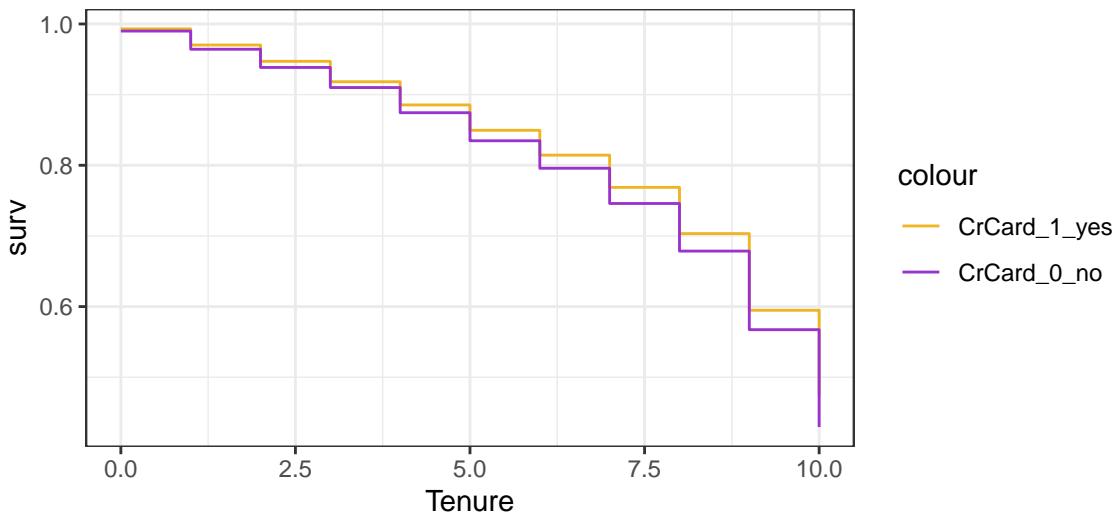
Se realizó el mismo análisis para la variable Geography, que integra a Francia, Alemania y España. A continuación se muestran las funciones de supervivencia. Se puede observar que la tasa de supervivencia o permanencia en el banco para los residentes de Alemania es menor que los residentes de Francia y España. Estos dos últimos muestran la misma tasa de supervivencia.



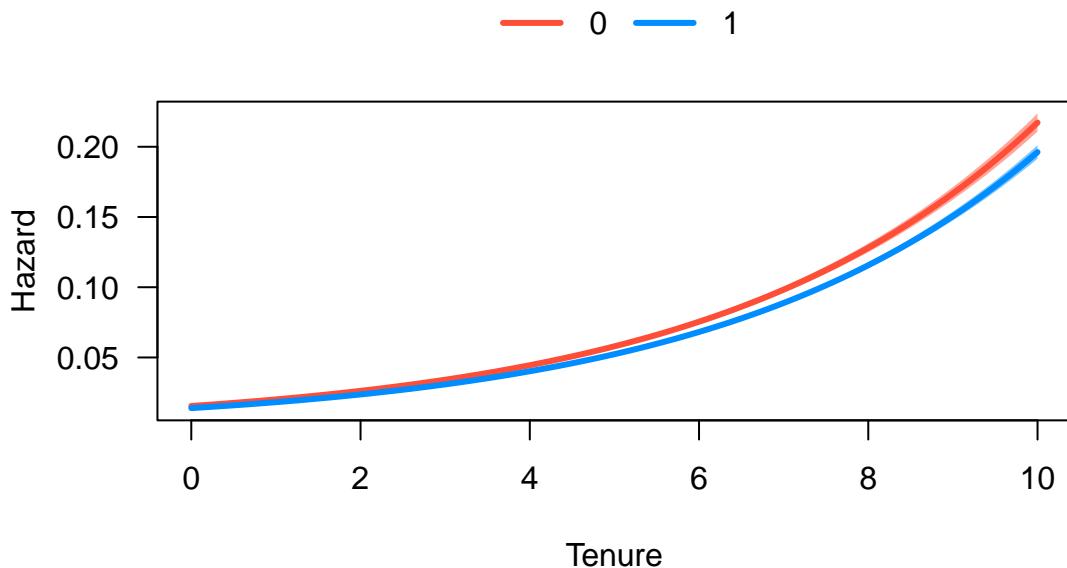
A continuación podemos ver la función de riesgo asociada, la probabilidad condicional de abandonar el banco, dado que hasta ahora no se ha sucedido, aumentará con la variable de tiempo. Esta probabilidad es mayor e incrementa más rápido para Alemania que para Francia o España.



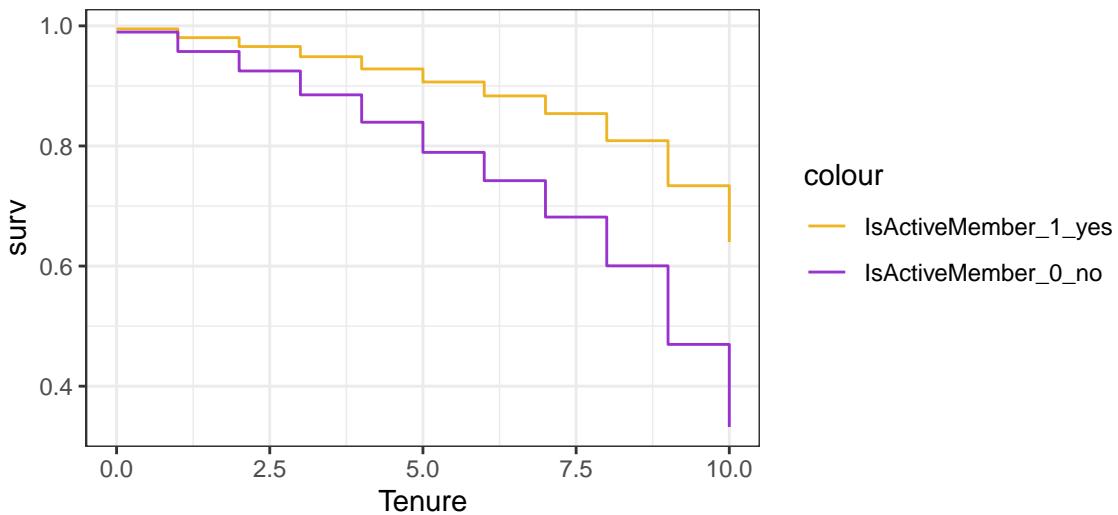
Para clientes con y sin tarjeta de crédito (HasCrCard). Se observa que la tasa de supervivencia no es tan diferente si se tiene tarjeta de crédito o no, aunque sí se observa una tasa de supervivencia ligeramente mayor si se tiene este producto.



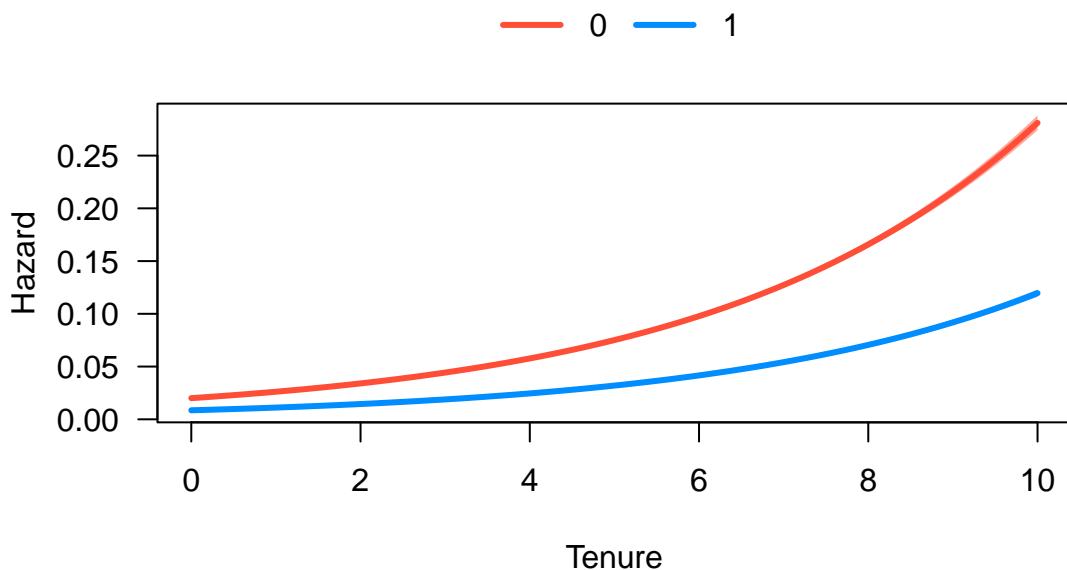
A continuación podemos ver la función de riesgo asociada, la probabilidad condicional de abandonar el banco, dado que hasta ahora no se ha sucedido, aumentará con la variable de tiempo y es mayor si no se tiene tarjeta de crédito.



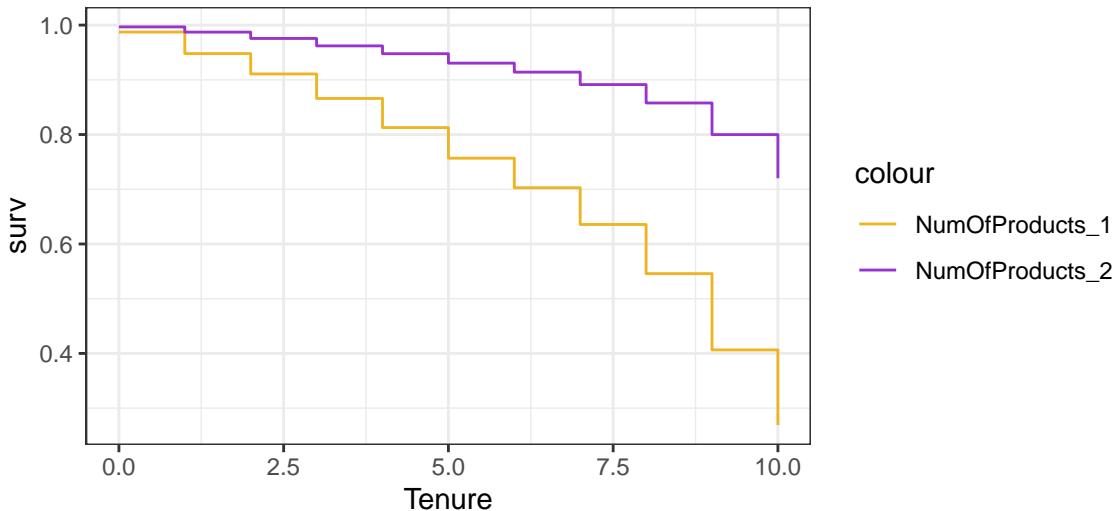
Para clientes que son miembros activos (IsActiveMember). Se observa que la tasa de supervivencia es mayor si se es un miembro activo.



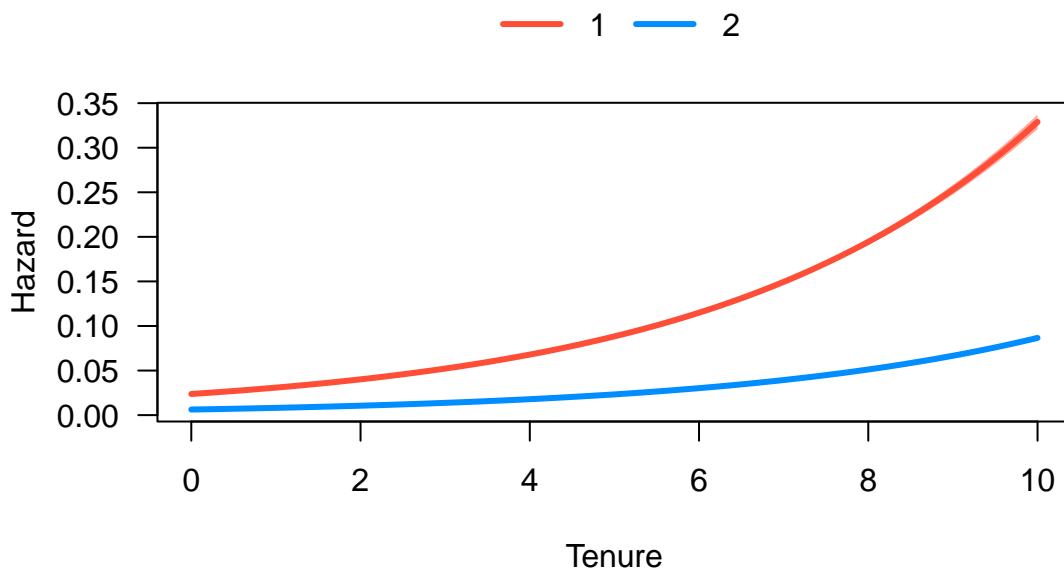
A continuación podemos ver la función de riesgo asociada, la probabilidad condicional de abandonar el banco, dado que hasta ahora no se ha sucedido, aumentará con la variable de tiempo y es mayor si no es miembro activo.



Para clientes que son miembros con un producto o dos o más productos (NumOfProducts). La tasa de supervivencia es mayor si se cuenta con dos o más productos.



A continuación podemos ver la función de riesgo asociada, la probabilidad condicional de abandonar el banco, dado que hasta ahora no se ha sucedido, aumentará con la variable de tiempo y es mayor si solamente se tiene contratado un solo producto bancario.



3. Un modelo logit para modelar la salida de los clientes.

Ajustaremos un modelo lineal generalizado binomial con liga logit, es decir, una regresión logística con la variable dependiente binaria `Exited` y las covariables del conjunto de datos. Esto es, plantearemos un modelo para explicar la probabilidad de salirse o no salirse del banco en función de las variables independientes proporcionadas. Podemos observar en el cuadro MODELO, que dadas las demás variables en el modelo, parece ser que `GeographySpain` es la única que ya no agrega más información al modelado, tomando en consideración que el país de referencia para la variable `Geography` es Francia. Los demás coeficientes estimados son estadísticamente significativos con un nivel de confianza del 95 % para las pruebas de hipótesis, por lo que si se cumplen los supuestos del modelo, podemos interpretar sus signos y valores más adelante y hacer

inferencia sobre ellos. Por esto, primero revisaremos el cumplimiento de los supuestos del modelado, antes de empezar a inferir e interpretar.

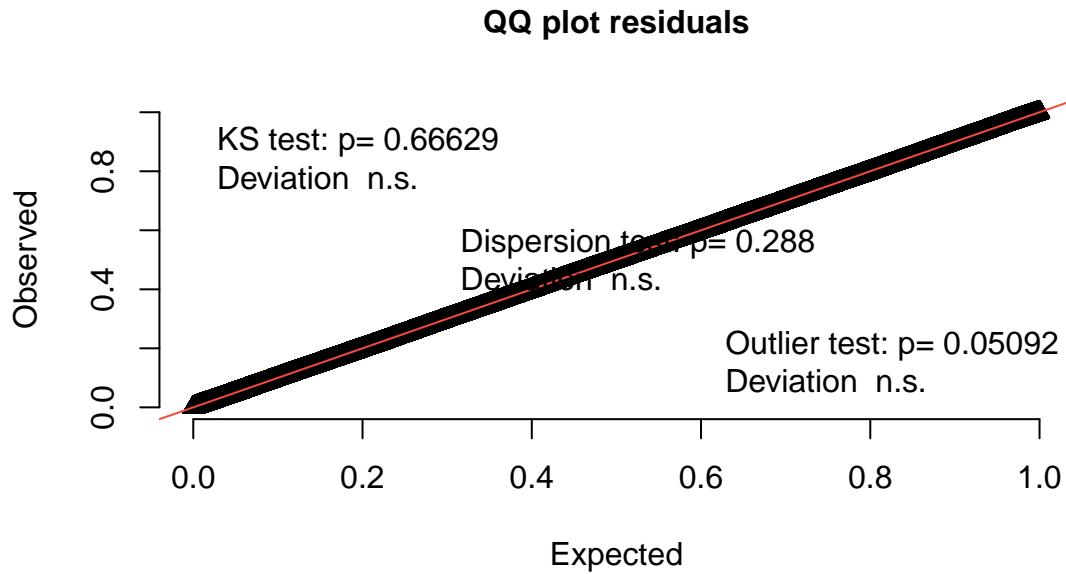
MODELO

<i>Dependent variable:</i>	
	Exited
CreditScore	−0.001*** (0.0001)
GeographyGermany	0.944*** (0.017)
GeographySpain	0.027 (0.019)
GenderMale	−0.676*** (0.014)
Age	0.093*** (0.001)
Tenure	−0.016*** (0.003)
HasCrCard1	−0.157*** (0.016)
IsActiveMember1	−1.281*** (0.015)
EstimatedSalary	0.00000*** (0.00000)
NumOfProducts2.L	−1.075*** (0.011)
Constant	−3.941*** (0.072)
Observations	165,023
Log Likelihood	−62,160.540
Akaike Inf. Crit.	124,343.100

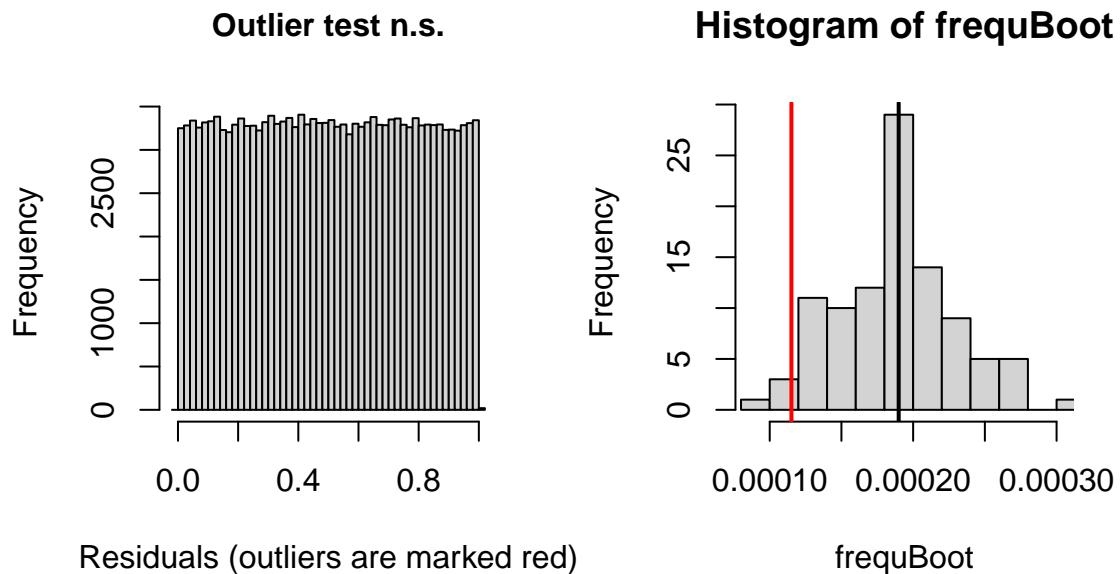
Note: *p<0.1; **p<0.05; ***p<0.01

Hacemos una prueba de hipótesis similar a la prueba F asociada a la tabla ANOVA, como la variable dependiente es binaria ésta es la lineal general con la Chi-cuadrada. La hipótesis nula es la misma, que los estimadores $\hat{\beta}_i = 0$, $\forall i = 1, \dots, p$, contra la alternativa de que al menos una $\hat{\beta}_i \neq 0$. Obtuvimos que el p-value es menor a 0.05, por lo que rechazamos la hipótesis nula con un nivel de confianza del 95 %. Entonces podemos continuar con la revisión de los supuestos del modelo planteado, para su posterior interpretación.

En la siguiente gráfica se muestran en general las pruebas KS test, Dispersion test y Outlier test, con p-values mayores a 0.05, por lo que no podemos rechazar los supuestos de normalidad, homocedasticidad y no presencia de outliers influyentes. Con la prueba de uniformidad **Asymptotic one-sample Kolmogorov-Smirnov test** se obtuvo un p-value de 0.666629 por lo que la distribución general se ajusta a las expectativas. Por otra parte, la prueba DHARMA nonparametric dispersion test via sd of residuals fitted vs. simulated tiene un p-value de 0.288, por lo que la dispersión simulada es igual a la dispersión observada. Por último, con la prueba DHARMA outlier test based on exact binomial test with approximate expectations se tiene un p-value de 0.05092, por lo que no podemos afirmar que haya más valores atípicos de simulación de los esperados.

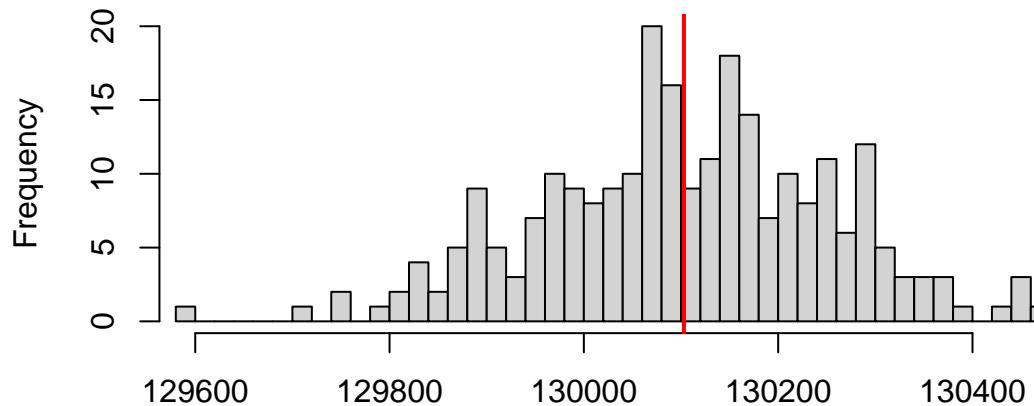


En las siguientes gráficas, podemos observar la prueba de si hay más valores atípicos de simulación que los esperados considerando el método de bootstrap. En este caso, no se detectó un problema con los outliers.



Podemos observar que no tenemos un problema de balanceo, el p-value asociado al **zero-inflation test** es mayor que 0.05 con un nivel de confianza del 95 %, esto es, los ceros esperados son muy similares a los simulados. Por lo tanto, no es necesario ajustar un modelo de balanceo.

DHARMA zero-inflation test via comparison to expected zeros with simulation under H0 = fitted model

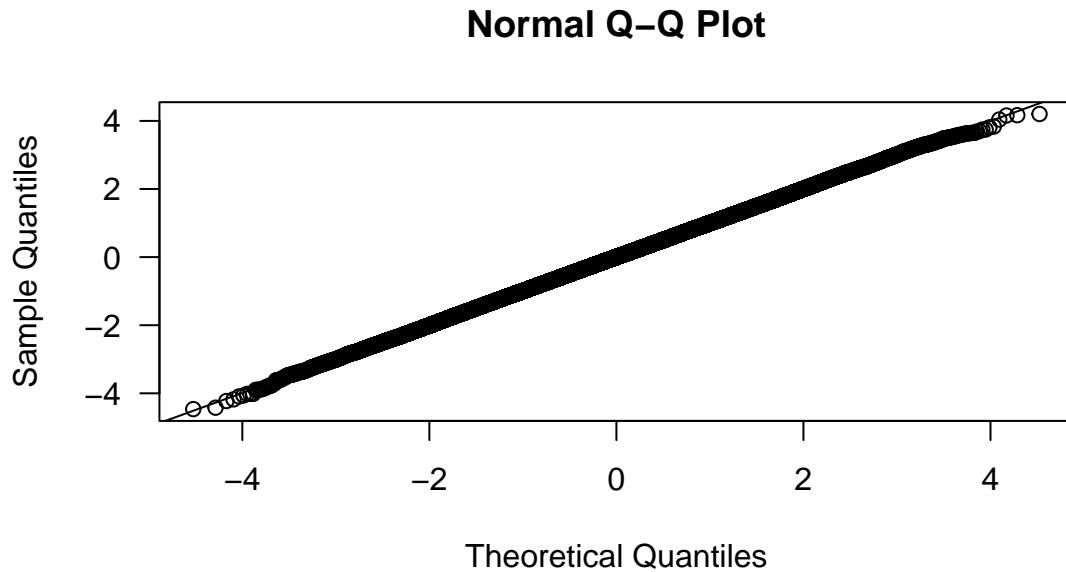


Simulated values, red line = fitted model. p-value (two.sided) = 0.992

Recordemos que la dispersión excesiva puede causar esto, por lo que para complementar el análisis lo verificaremos.

La regla de dedo para verificar si el **parámetro de dispersión** es de 1, con la devianza de residuales entre los grados de libertad, muestra un valor de 0.7534063, lo cual se acerca a 1, por lo que no tenemos problemas de la dispersión o varianza. Usando el estimador del parámetro de dispersión ϕ tenemos un valor de 1.0348873 que es muy cercano a uno, lo cual refuerza la hipótesis de una varianza constante.

A continuación se muestran la gráfica **Normal Q-Q Plot**.



Además, en la prueba de normalidad **Lilliefors** (Kolmogorov-Smirnov) **normality test** tenemos que el p-value es de 0.888400227648293, por lo que no se puede rechazar la hipótesis nula de normalidad. La prueba

Shapiro-Wilk refuerza este resultado con un p-value de 0.6946236.

Como el modelo cumple todos los supuestos, podemos hacer estimación e inferencia, es decir, podemos interpretar los coeficientes estimados $\hat{\beta}_i$. Recordemos que la única variable que ya no agrega más información al modelado, dado que están en el modelo las otras variables, es GeographySpain. Esto se puede observar en los intervalos de confianza, pues en este caso incluye al cero, en los demás casos no. Como podemos observar, los signos indican que los factores que disminuyen la probabilidad de dejar el banco son el score crediticio , ser hombre (en comparación a ser mujer), los años de tenencia de la cuenta, tener tarjeta de crédito, ser un miembro muy activo, y tener dos o más productos. Por otro lado, los factores que aumentan la probabilidad de dejar el banco son el ser residente aleman (comparado a ser francés), la edad, y el salario.

Estimación puntual

	x
(Intercept)	-3.9411179
CreditScore	-0.0007873
GeographyGermany	0.9444493
GeographySpain	0.0270883
GenderMale	-0.6758896
Age	0.0931058
Tenure	-0.0155609
HasCrCard1	-0.1574090
IsActiveMember1	-1.2806593
EstimatedSalary	0.0000009
NumOfProducts2.L	-1.0747343

Adicionalmente, se muestran los intervalos de confianza de los estimadores de los parámetros del modelo.

Intervalos de confianza

	2.5 %	97.5 %
(Intercept)	-4.0816805	-3.8008248
CreditScore	-0.0009614	-0.0006132
GeographyGermany	0.9117832	0.9771233
GeographySpain	-0.0099506	0.0640420
GenderMale	-0.7040248	-0.6477781
Age	0.0915284	0.0946879
Tenure	-0.0205399	-0.0105828
HasCrCard1	-0.1894611	-0.1253079
IsActiveMember1	-1.3106723	-1.2507311
EstimatedSalary	0.0000007	0.0000012
NumOfProducts2.L	-1.0960399	-1.0534928

Para una interpretación más directa, en términos de las probabilidades de irse del banco o no irse, hacemos la transformación correspondiente al modelo logit, en este caso recordemos que el componente lineal se plantea como $\eta_i = \eta(\beta, x_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$ para $i = 1, \dots, p$ y la función liga que en este caso es monótona creciente $g(\mu_i) = \eta_i =$, donde $E(y_i; x_i) = E(y_i) = \mu_i$. Así, para la distribución Bernoulli y Binomial, con liga logit tenemos que $g(\mu_i) = \eta_i = \ln(\frac{\mu_i}{1-\mu_i})$ y $\mu_i = g^{-1}(\eta_i) = \frac{exp\{\eta_i\}}{1+exp\{\eta_i\}}$. Por lo tanto, mostramos a continuación las estimaciones puntuales y sus intervalos de confianza para $\mu_i = E(y_i; x_i) = P(Exited = 1|x_i)$, es decir, la probabilidad de salirse del banco, en función de las variables explicativas x_i .

Salvo GeographySpain, podemos interpretar los intervalos de confianza de los coeficientes estimados.

Si un cliente tiene un score crediticio mínimo de 350, es francés, es mujer, con 18 años de edad, con cero años de tenencia de la cuenta, sin tarjeta de crédito, no es una cliente activa, con salario estimado mínimo de 11.58 y con un único producto en el banco, su probabilidad de dejar el banco se calcula como (Ver Hosmer (2013), pág. 7, para la transformación; y Chunk Perfil1, linea 758 para el cálculo):

$$\frac{e^{\beta_0 + \beta_1(350) + \beta_5(18) + \beta_9(11.58)}}{1+e^{\beta_0 + \beta_1(350) + \beta_5(18) + \beta_9(11.58)}} = \frac{e^{-3.941118 - 0.0007873070(350) + 0.09310582(18) + 0.0000009351908(11.58)}}{1+e^{-3.941118 - 0.0007873070(350) + 0.09310582(18) + 0.0000009351908(11.58)}} = 7.3049711\%.$$

Si un cliente tiene un score crediticio máximo de 850, es francés, es mujer, con 40 años de edad, con cero años de tenencia de la cuenta, sin tarjeta de crédito, no es una cliente activa, con salario estimado máximo de 199992.5 y con dos o más productos en el banco, su probabilidad de dejar el banco se calcula como:

$$\frac{e^{\beta_0 + \beta_1(850) + \beta_5(40) + \beta_9(199992.5) + \beta_{10}}}{1+e^{\beta_0 + \beta_1(850) + \beta_5(40) + \beta_9(199992.5) + \beta_{10}}} = \frac{e^{-3.941118 - 0.0007873070(850) + 0.09310582(40) + 0.0000009351908(199992.5) - 1.074734}}{1+e^{-3.941118 - 0.0007873070(850) + 0.09310582(40) + 0.0000009351908(199992.5) - 1.074734}} = 40.1801603\%.$$

En general, podemos clasificar varios perfiles de clientes y obtener la probabilidad de que dejen el banco, lo cual se puede abordar con mayor detalle en algún análisis posterior más exhaustivo.

Conclusiones

Las curvas de supervivencia muestran que la probabilidad de permanecer en el banco más allá de cierto tiempo de permanencia con la cuenta disminuye, la tasa de supervivencia o permanencia en el banco es menor para las mujeres en comparación con los hombres, al igual que con los residentes de Alemania en comparación con los de Francia y España. Además, la tasa de supervivencia es mayor si se es un miembro activo que si no es un cliente activo, así como para clientes con dos o más productos comparado con clientes que solamente tienen un único producto en el banco.

Con el modelo logit, pudimos observar que los factores que disminuyen la probabilidad de dejar el banco son el score crediticio, ser hombre (en comparación a ser mujer), los años de tenencia de la cuenta, tener tarjeta de crédito, ser un miembro muy activo, y tener dos o más productos. Por otro lado, los factores que aumentan la probabilidad de dejar el banco son el ser residente alemán (comparado a ser francés), la edad, y el salario.

Referencias

- Hosmer, D. (2013). *Applied logistic regression*. Wiley.
- KAGGLE. (2022). *Bank customer churn dataset*. <https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset>
- Villers. (2023). *Modelos de supervivencia*. <https://svg18.github.io/Supervivencia/>