



Facultad de Ciencias

UNAM

ESTADÍSTICA 3. MODELOS DE SUPERVIVENCIA Y SERIES DE TIEMPO

Proyecto final

ANÁLISIS DE DATOS DE CLIENTES BANCARIOS

Enríquez Hernández Leobardo
Tlahuiz Tenorio Saúl Giovanni

26 de mayo de 2024

Índice

Introducción.	1
1. Estadística descriptiva y procesamiento de datos.	1
2. Un modelo logit para modelar la salida de los clientes.	5
3. Clasificación de los clientes. Componentes principales y factoriales.	10
4. El problema de supervivencia para la salida de los clientes.	12
Conclusiones	19
Referencias	20

Introducción.

Este documento tiene como objetivos generales, mostrar algunos elementos estadísticos que permitan analizar la información de los clientes de un banco (KAGGLE (2022)), así como tratar de clasificarlos y decir algo sobre la variable que indica si el cliente abandona o no abandona el banco (Exited). Para esto, plantearemos un modelo de regresión probabilístico como el modelo logit, ciertos análisis de componentes principales, factoriales y de clasificación, y el análisis de supervivencia para la variable Exited.

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 analiza un modelo logit, con la variable Exited como variable explicada. En la tercera sección aplicaremos componentes principales y análisis factorial, para las variables de la base de datos y ver si podemos clasificarlos o encontrar algunas variables latentes. En la cuarta sección se plantea y desarrolla el problema de supervivencia. Y finalmente se presentan las principales conclusiones del trabajo.

1. Estadística descriptiva y procesamiento de datos.

La base 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.

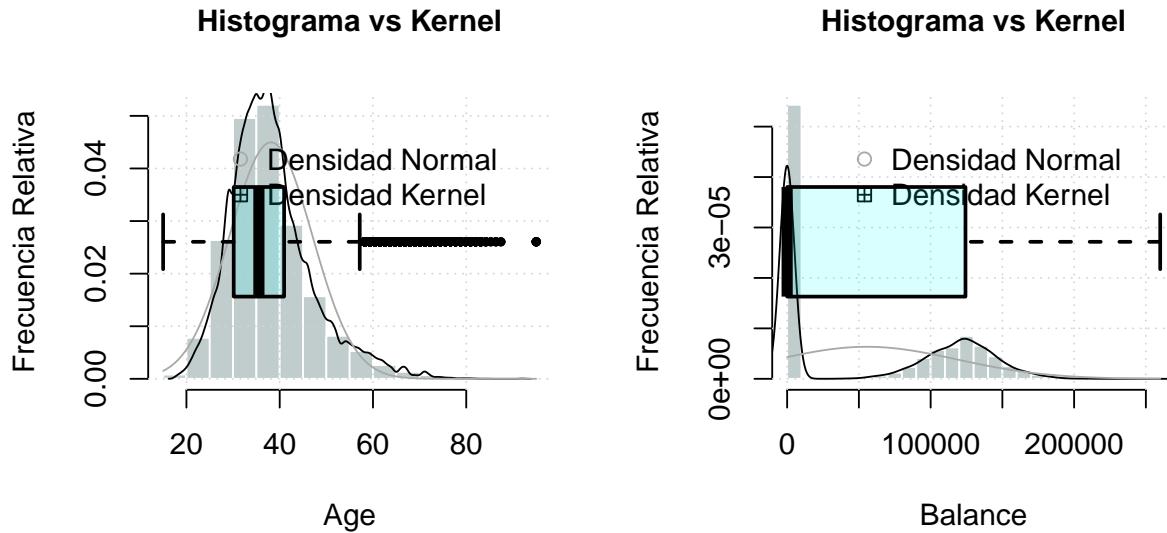
CreditScore : 0	Geography: 0	Gender: 0	Age: 0
Tenure : 0	Balance: 0	NumOfProducts: 0	HasCrCard: 0
IsActiveMember : 0	EstimatedSalary: 0	Exited: 0	

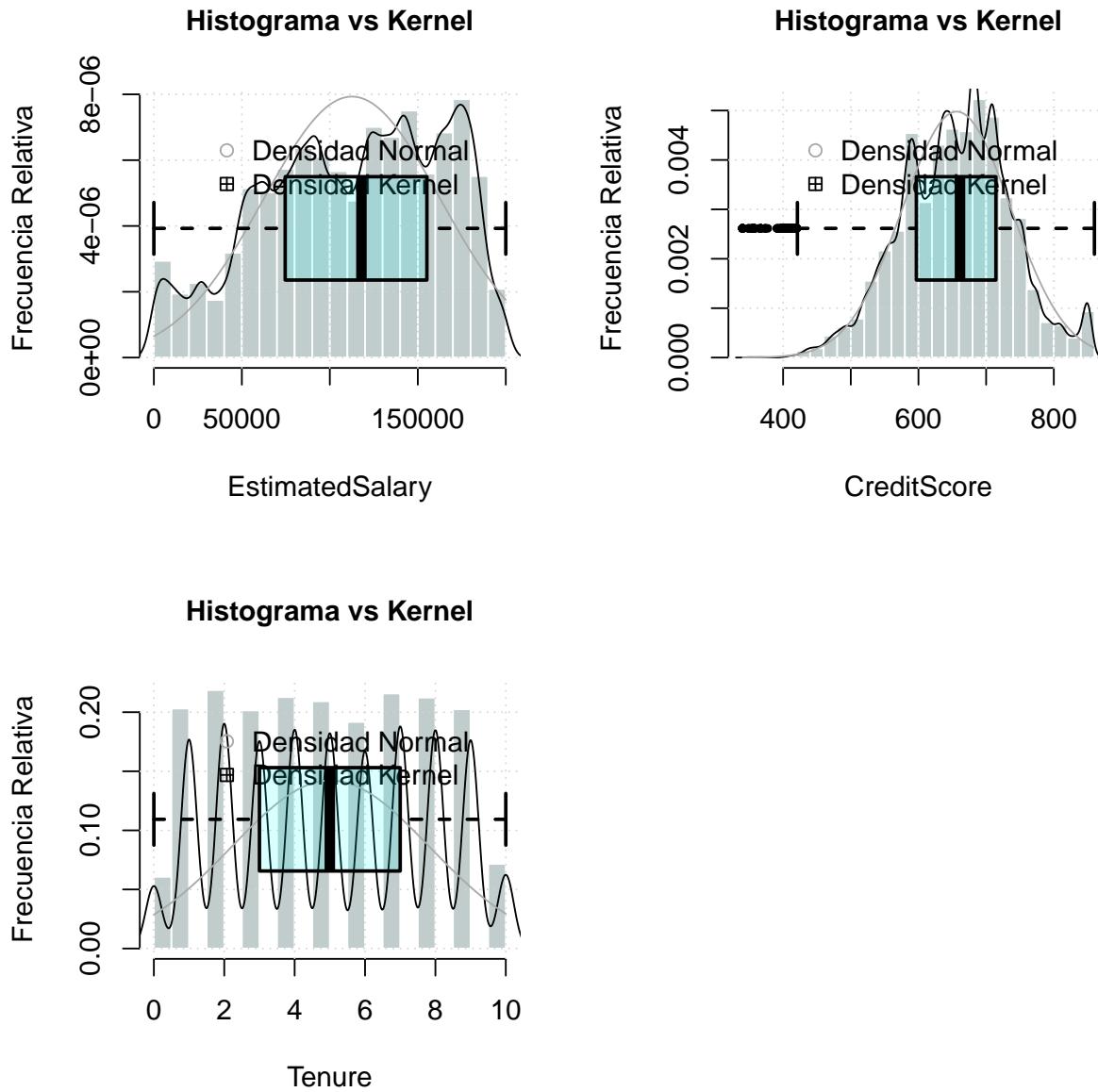
A continuación se muestra la estadística descriptiva de los valores numéricos relevantes. Son 165,034 observaciones, con edades entre 18 y 92 años, con un balance de 0 a 250,898 unidades monetarias, con salario estimado de entre 11,58 a 199,992,5, un score de crédito de 350 a 850, y tenencia de cuenta 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 112,574,800, 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

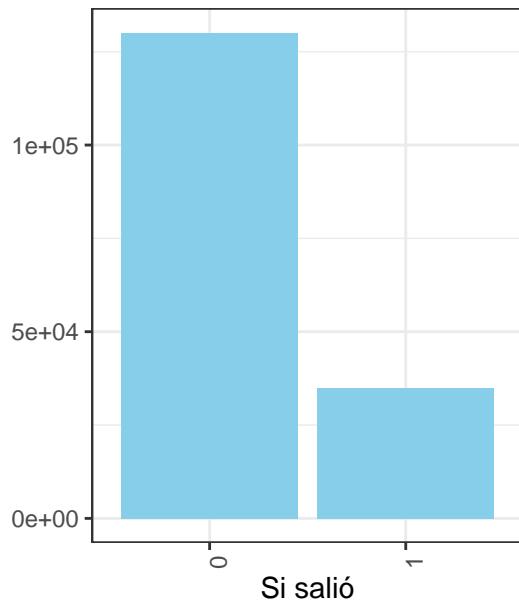
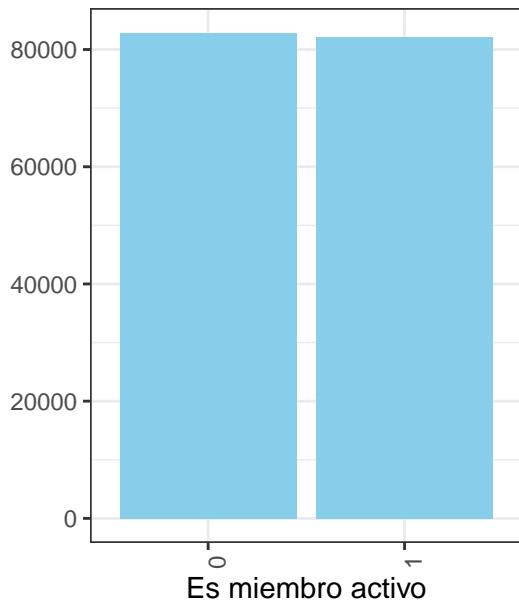
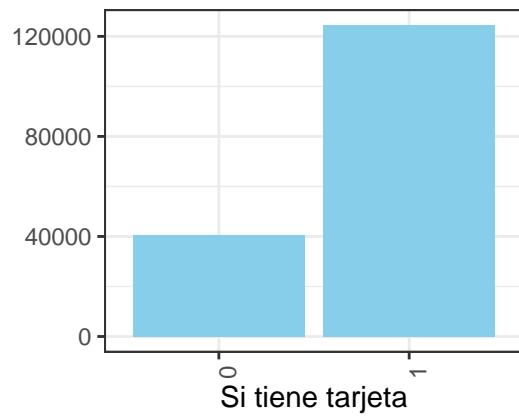
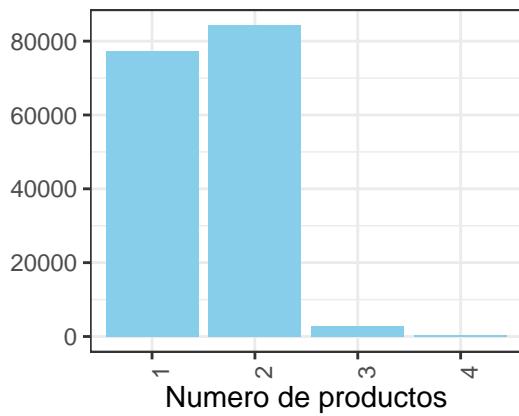
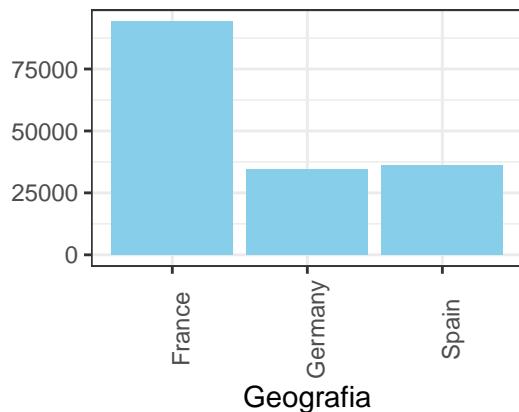
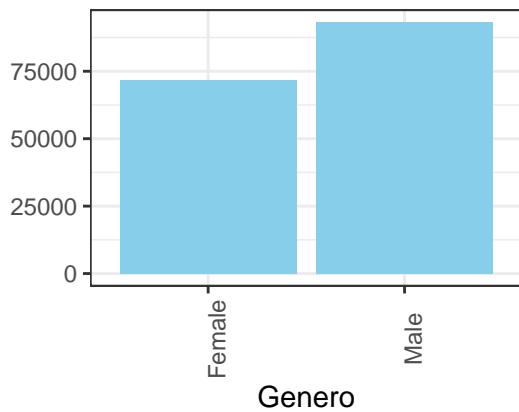
A continuación se muestran los histogramas de frecuencias relativas, distribuciones o densidades y boxplot conjuntas para las variables numéricas. Se pueden observar distribuciones multimodales, a excepción de la variable de edad. Por ejemplo, el balance tiene un sesgo muy notable a la izquierda, mientras 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 los datos categóricos, por su frecuencia absoluta en la base de datos. En la base 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 por algún periodo, es decir, permanecieron el 79 % de los clientes.



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

Haremos dos principales cambios en la base 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.

Luego ajustaremos un modelo lineal generalizado binomial con liga logit, es decir, una regresión logística con la variable dependiente binaria `Exited` y las covariables de la base 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 que dadas las demás variables en el modelo, parece ser que `GeographySpain` es la única que ya no agrega más información al modelado, con país de referencia Francia.

```
##  
## Call:  
## glm(formula = Exited ~ . - Balance, family = binomial(link = "logit"),  
##       data = datos)  
##  
## Coefficients:  
##                                     Estimate Std. Error z value Pr(>|z|)  
## (Intercept)      -3.941e+00  7.165e-02 -55.008 < 2e-16 ***  
## CreditScore     -7.873e-04  8.883e-05  -8.863 < 2e-16 ***  
## GeographyGermany 9.444e-01  1.667e-02   56.661 < 2e-16 ***  
## GeographySpain   2.709e-02  1.888e-02    1.435   0.151  
## GenderMale       -6.759e-01  1.435e-02  -47.105 < 2e-16 ***  
## Age              9.311e-02  8.060e-04  115.519 < 2e-16 ***  
## Tenure            -1.556e-02  2.540e-03   -6.126 9.00e-10 ***  
## HasCrCard1       -1.574e-01  1.637e-02   -9.618 < 2e-16 ***  
## IsActiveMember1  -1.281e+00  1.529e-02  -83.753 < 2e-16 ***  
## EstimatedSalary   9.352e-07  1.427e-07    6.555 5.57e-11 ***  
## NumOfProducts2.L -1.075e+00  1.085e-02  -99.020 < 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: 170329  on 165022  degrees of freedom  
## Residual deviance: 124321  on 165012  degrees of freedom  
## AIC: 124343  
##  
## Number of Fisher Scoring iterations: 5
```

A continuación presentamos una prueba similar a la prueba F asociada a la tabla ANOVA, pero como la variable dependiente es binaria, hacemos la prueba de hipótesis 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 $\hat{\beta}_i \neq 0$. Observamos que como el p-value es menor a 0.05, 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.

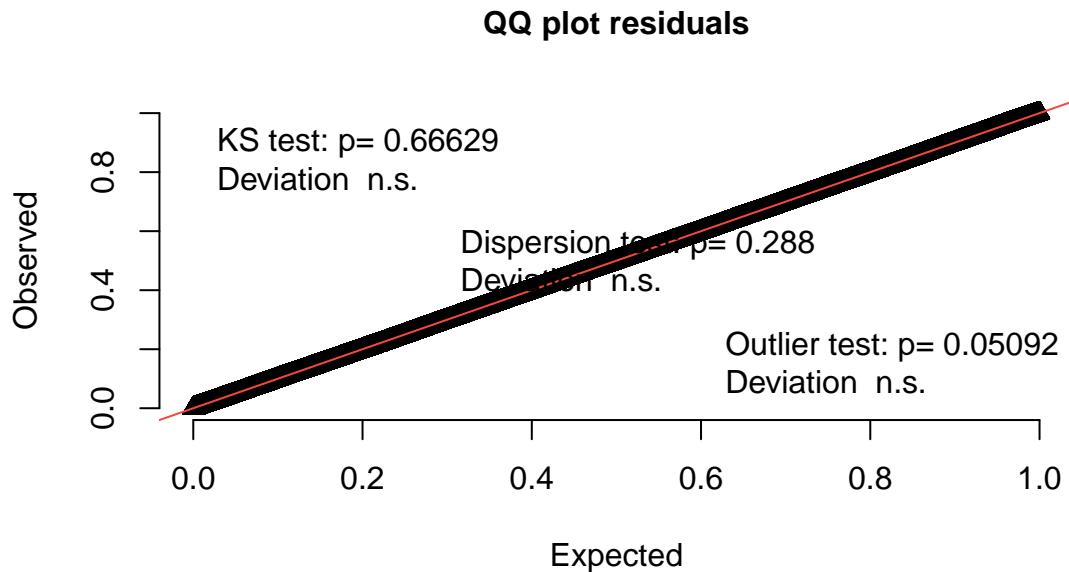
```
##  
## General Linear Hypotheses  
##  
## Linear Hypotheses:  
##                           Estimate  
## 1 == 0    -7.873e-04  
## 2 == 0     9.444e-01
```

```

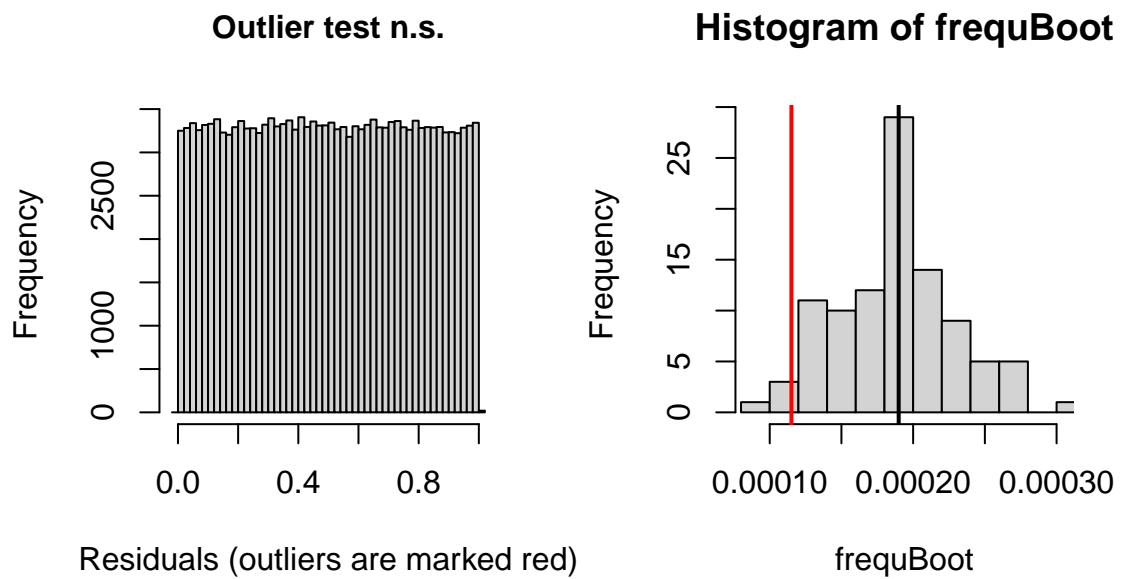
## 3 == 0  2.709e-02
## 4 == 0 -6.759e-01
## 5 == 0  9.311e-02
## 6 == 0 -1.556e-02
## 7 == 0 -1.574e-01
## 8 == 0 -1.281e+00
## 9 == 0  9.352e-07
## 10 == 0 -1.075e+00
##
## Global Test:
##   Chisq DF Pr(>Chisq)
## 1 29901  9          0

```

A continuación haremos las pruebas de los supuestos del modelo. 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. La prueba de uniformidad **Asymptotic one-sample Kolmogorov-Smirnov test** tiene un p-value de 0,66629 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.



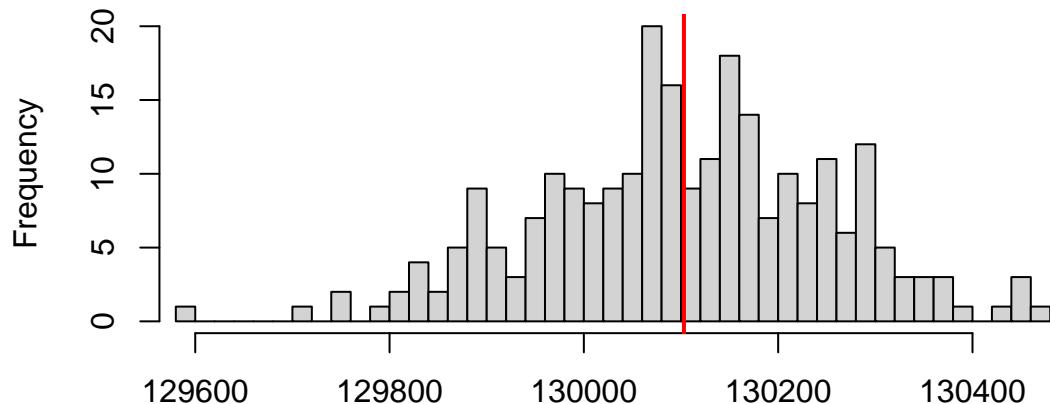
En la siguiente prueba, podemos observar la prueba de si hay más valores atípicos de simulación de los esperados considerando el método de bootstrap. En este caso, no se detectó un problema con los outliers.



```
##  
##  DHARMA bootstrapped outlier test  
##  
##  data: fitlogitres_  
##  outliers at both margin(s) = 19, observations = 165023, p-value = 0.08  
##  alternative hypothesis: two.sided  
##  percent confidence interval:  
##  0.0001151355 0.0002698109  
##  sample estimates:  
##  outlier frequency (expected: 0.000189912921229162 )  
##                                         0.0001151355
```

Podemos observar que no tenemos un problema de ceros inflados, el p-value 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 ceros inflados.

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

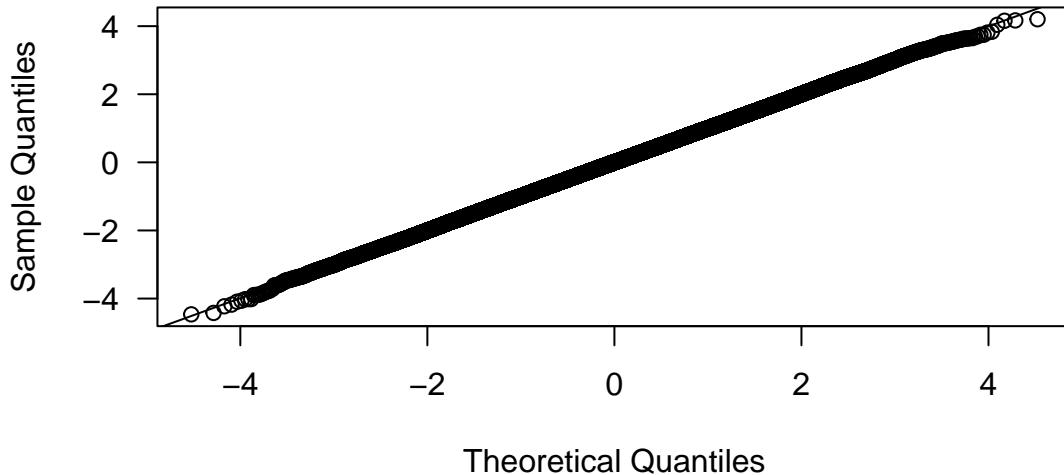
```
## 
##  DHARMA zero-inflation test via comparison to expected zeros with
##  simulation under H0 = fitted model
## 
##  data:  simulationOutput
##  ratioObsSim = 0.99999, p-value = 0.992
##  alternative hypothesis: two.sided
```

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 Plox**, la prueba de Kolmogorov-Smirnov y la prueba Shapiro-Wilk para normalidad de los residuos del modelo.

Normal Q-Q Plot



```

## 
## Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data: fitlogitqr
## D = 0.001115, p-value = 0.8884

## 
## Shapiro-Wilk normality test
## 
## data: fitlogitqr[0:500]
## W = 0.99759, p-value = 0.6946

```

En la prueba de normalidad **Lilliefors (Kolmogorov-Smirnov) normality test** tenemos que el p-value es de 0.888400227648293, por lo que no se rechaza 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, un mayor score crediticio disminuye la probabilidad de dejar el banco, ser residente aleman (comparado a ser francés) aumenta la probabilidad de salirse del banco, ser hombre disminuye la probabilidad de dejar el banco (en comparación a ser mujer), la edad aumenta la probabilidad de dejar el banco, los años de tenencia de la cuenta disminuye la probabilidad de dejar el banco, tener tarjeta de crédito disminuye la probabilidad de dejar el banco, ser un miembro muy activo disminuye la probabilidad de dejar el banco, el salario aumenta la probabilidad de dejar el banco, y tener dos o más productos disminuye la probabilidad de irse del banco.

```

## [1] "ESTIMACION PUNTUAL"

##      (Intercept)    CreditScore GeographyGermany   GeographySpain
## -3.941118e+00 -7.873070e-04  9.444493e-01  2.708834e-02
##      GenderMale       Age        Tenure      HasCrCard1
## -6.758896e-01  9.310582e-02 -1.556088e-02 -1.574090e-01
## IsActiveMember1 EstimatedSalary NumOfProducts2.L

```

```

##      -1.280659e+00      9.351908e-07     -1.074734e+00
## [1] "INTERVALOS DE CONFIANZA"

##                  2.5 %          97.5 %
## (Intercept)    -4.081680e+00  -3.800825e+00
## CreditScore    -9.614314e-04  -6.132143e-04
## GeographyGermany 9.117832e-01  9.771233e-01
## GeographySpain   -9.950640e-03  6.404204e-02
## GenderMale      -7.040248e-01  -6.477781e-01
## Age              9.152840e-02  9.468785e-02
## Tenure           -2.053992e-02  -1.058277e-02
## HasCrCard1      -1.894611e-01  -1.253079e-01
## IsActiveMember1 -1.310672e+00  -1.250731e+00
## EstimatedSalary   6.556558e-07  1.214924e-06
## NumOfProducts2.L -1.096040e+00  -1.053493e+00

```

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 $g(\mu_i) = \eta_i = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$ y $\mu_i = g^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$. 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:

$$\frac{e^{\hat{\beta}_0 + \hat{\beta}_1(350) + \hat{\beta}_5(18) + \hat{\beta}_9(11,58)}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1(350) + \hat{\beta}_5(18) + \hat{\beta}_9(11,58)}} = \frac{e^{-3,941118 - 0,0007873070(350) + 0,09310582(18) + 0,0000009351908(11,58)}}{1 + e^{-3,941118 - 0,0007873070(350) + 0,09310582(18) + 0,0000009351908(11,58)}} = 7.3049711\%.$$

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:

$$\frac{e^{\hat{\beta}_0 + \hat{\beta}_1(850) + \hat{\beta}_5(40) + \hat{\beta}_9(199992,5) + \hat{\beta}_{10}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1(850) + \hat{\beta}_5(40) + \hat{\beta}_9(199992,5) + \hat{\beta}_{10}}} = \frac{e^{-3,941118 - 0,0007873070(850) + 0,09310582(40) + 0,0000009351908(199992,5) - 1,074734}}{1 + e^{-3,941118 - 0,0007873070(850) + 0,09310582(40) + 0,0000009351908(199992,5) - 1,074734}} = 40.1801603\%.$$

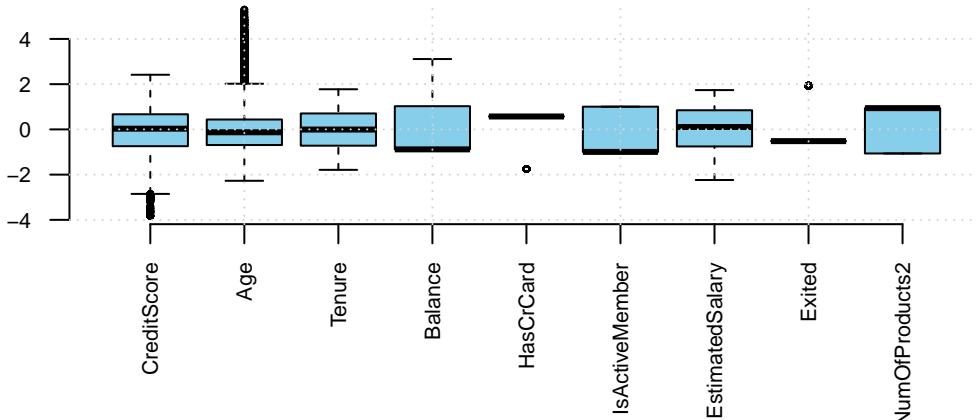
En general, podemos clasificar varios perfiles de clientes y obtener la probabilidad de que dejen el banco.

3. Clasificación de los clientes. Componentes principales y factoriales.

Primero estandarizamos los datos numéricos para tener la misma escala, además incluimos algunas variables categóricas que pueden tener un tratamiento como numéricas u ordinales y la variable de interés general en este análisis (**Exited**). La idea es ver si podemos caracterizar a los clientes, y ver qué variables nos pueden ayudar para esto y si es posible agrupar ciertas características en común, y sobre ellas dirigir estrategias a los clientes.

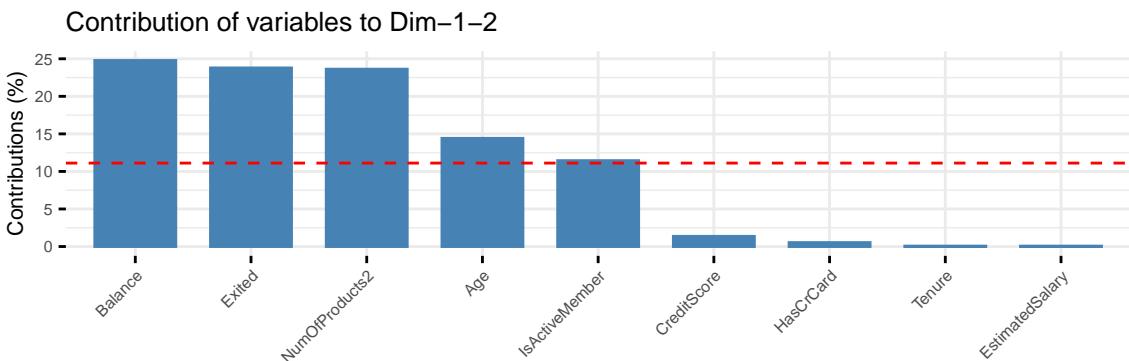
En la siguiente Gráfica se muestra que la estandarización es buena.

Grafico de caja y bigotes



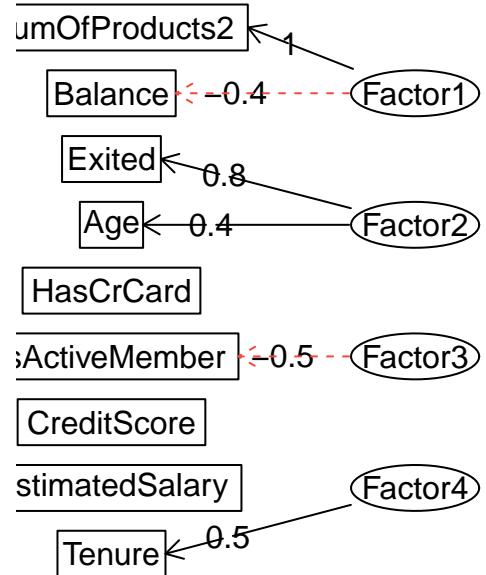
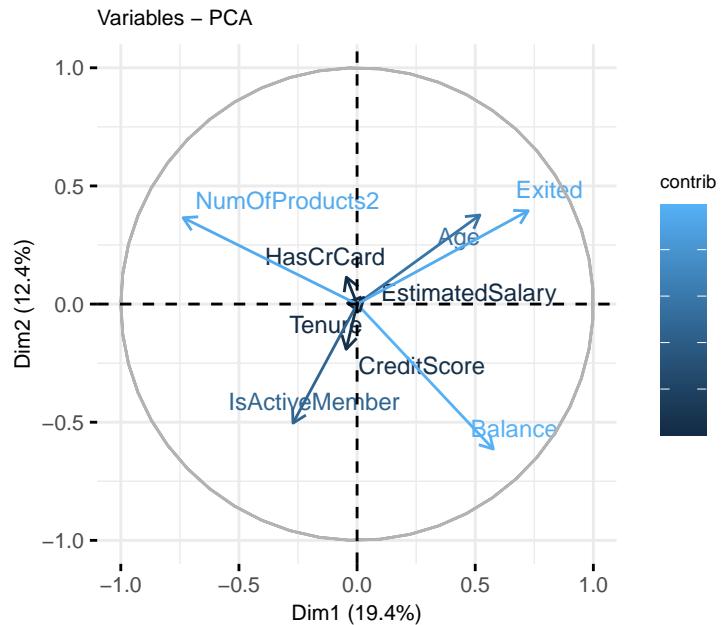
Podemos observar la varianza acumulada de los primeros componentes principales. Hasta el cuarto componente se alcanza el 54.3 %, y hasta el sexto componente el 76.4 %. Las variables con mayor peso en las primeras dos componentes son Balance, Exited, NumOfProducts2, y Age.

```
## Importance of components:
##                               Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## Standard deviation      1.322424 1.0543824 1.0139017 0.9993974 0.9979933
## Proportion of Variance 0.194313 0.1235254 0.1142225 0.1109779 0.1106663
## Cumulative Proportion  0.194313 0.3178384 0.4320610 0.5430389 0.6537052
##                               Comp.6    Comp.7    Comp.8    Comp.9
## Standard deviation      0.9967737 0.9875464 0.79198251 0.72152108
## Proportion of Variance 0.1103960 0.1083615 0.06969334 0.05784398
## Cumulative Proportion  0.7641011 0.8724627 0.94215602 1.00000000
```



En las siguientes gráficas, podemos observar estas variables de mayor peso, además podemos ver que Age y Exited están muy correlacionados, las demás variables se observan más dispersos. Con el análisis factorial, podemos ver que NumOfProductos y Balance, pueden agruparse en un factor, mientras que Exited y Age en otro, y las demás variables se encuentran de manera individual o no forman ningún factor.

Factor Analysis



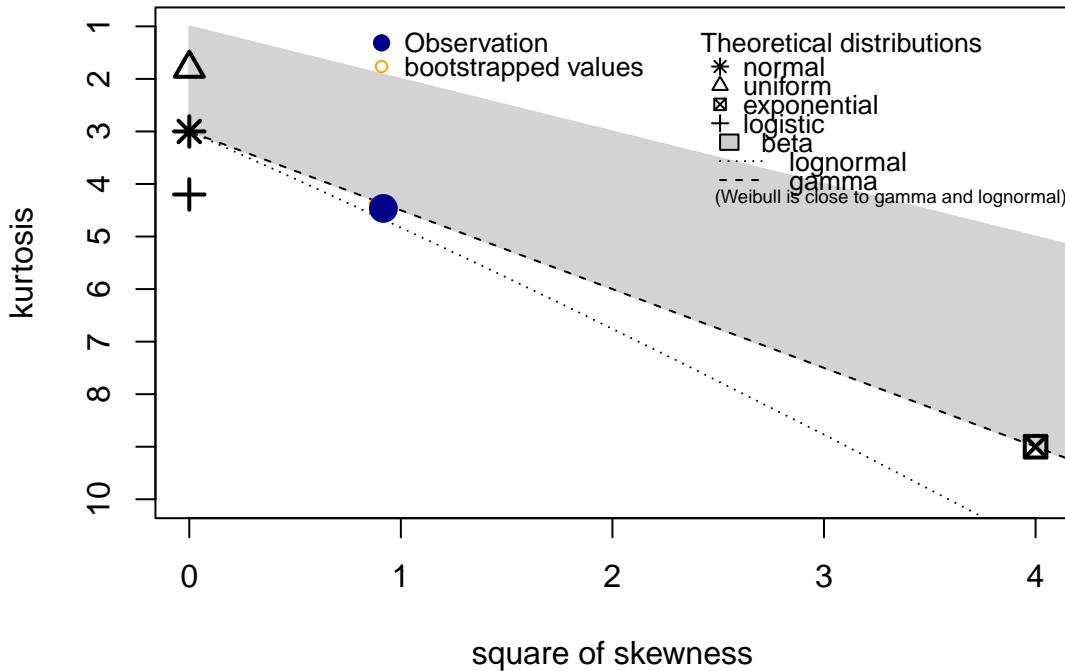
Los resultados de esta sección muestran que, si bien la estandarización de las variables es buena, los primeros componentes principales no acumulan rápidamente una gran proporción de las varianzas, lo cual se confirma con el análisis factorial. Los factores que se forman, a excepción del Factor 1 y Factor 2, están aislados o no es posible agruparlos en factores que podan ser útiles para el análisis y agrupamiento de los clientes del banco. Para el caso del Factor 1, conformado por Balance y NumOfProducts, éste podría ser un factor para alguna variable latente de qué tan profunda es la relación del cliente con el banco en términos de su monto de dinero y el número de productos con que cuenta. Por último, algo interesante de ver es que Age y Exited están muy correlacionados y forman un factor, lo que podría ser una caracterización importante para el banco, quizás esto tenga sentido si consideramos que las personas más jóvenes cuando abren una primera cuenta en el banco, es poco probable que abandone inmediatamente, mientras que una persona que va envejeciendo, es más probable que abandone el banco por cuestiones de salud, movilidad o por mortalidad.

4. 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)).

En primer lugar, mostramos la gráfica de Cullen y Frey para la variable Age, permite eximir algunas distribuciones mediante los parámetros de asimetría y curtosis utilizando la función `descdist`; los valores de arranque provienen de muestras aleatorias con reemplazo (bootstrap) de los datos. A partir de este gráfico, nuestras opciones para ajustes parecerían estar dentro de las distribuciones disponibles en el paquete `fitdistrplus`: Weibull, Gamma y Exponential.

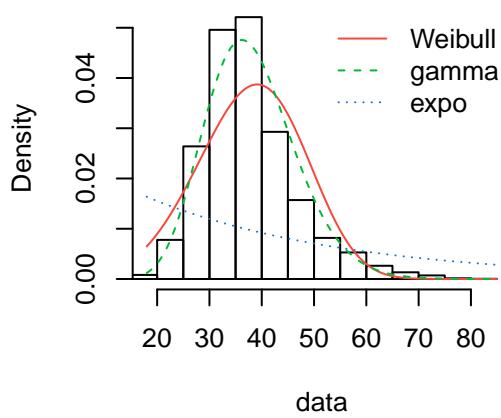
Cullen and Frey graph



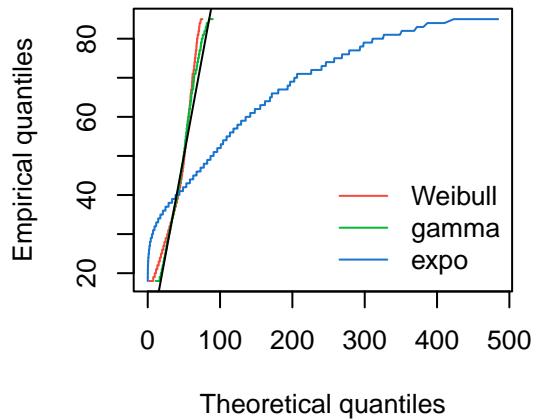
```
## summary statistics
## -----
## min: 18   max: 85
## median: 37
## mean: 38.1223
## estimated sd: 8.856584
## estimated skewness: 0.9578272
## estimated kurtosis: 4.464774
```

Estas 3 distribuciones (Weibull, Gamma y Exponential) se ajustan a cuatro parámetros de ajuste clásicos, siendo el más importante la densidad y el gráfico CDF. A partir de las métricas de ajuste trazadas a continuación, parece que Weibull y Gamma son los mejores candidatos. Observemos en la siguiente Figura que Gamma es el que mejor se ajusta.

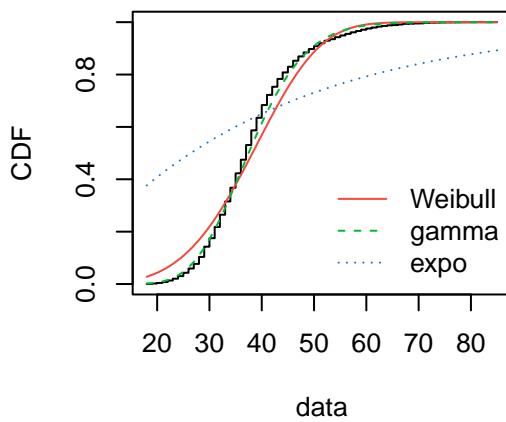
Histogram and theoretical densities



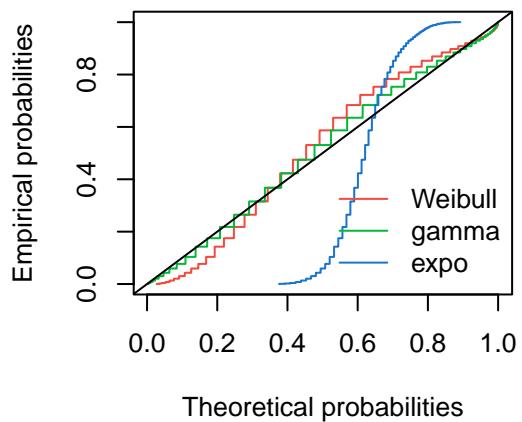
Q-Q plot



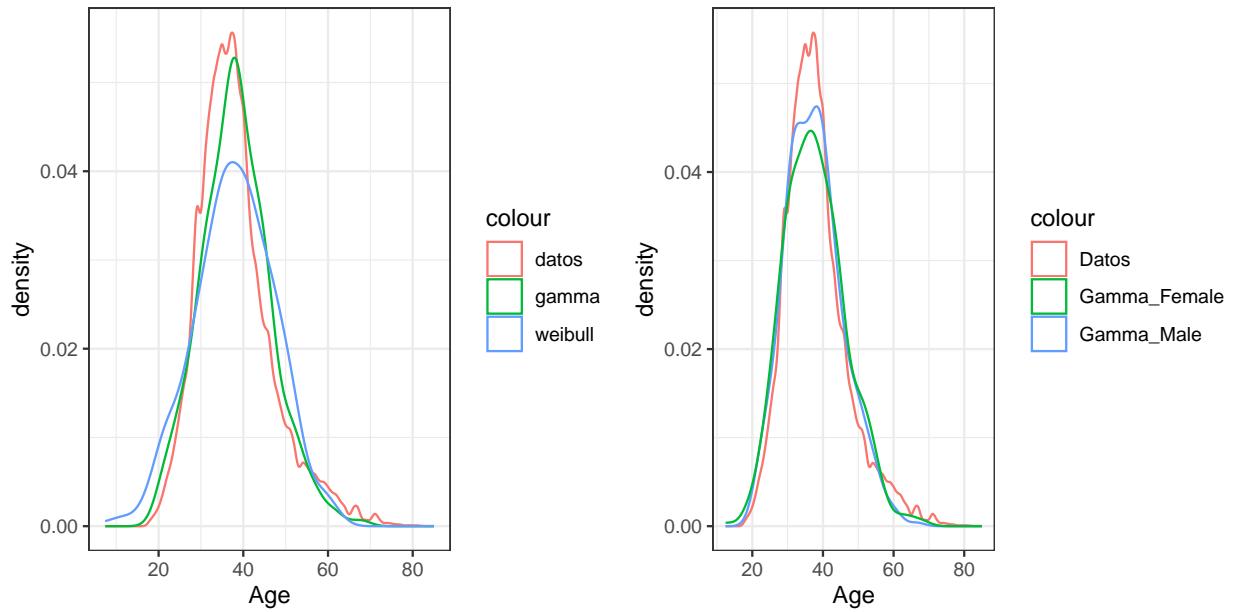
Empirical and theoretical CDFs



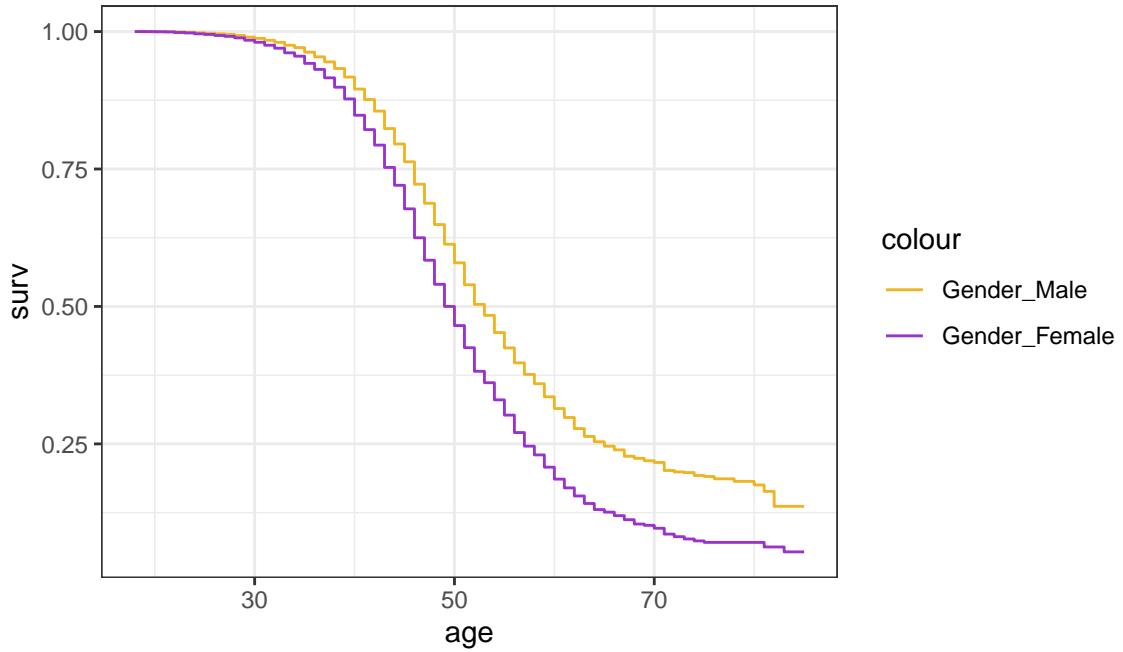
P-P plot



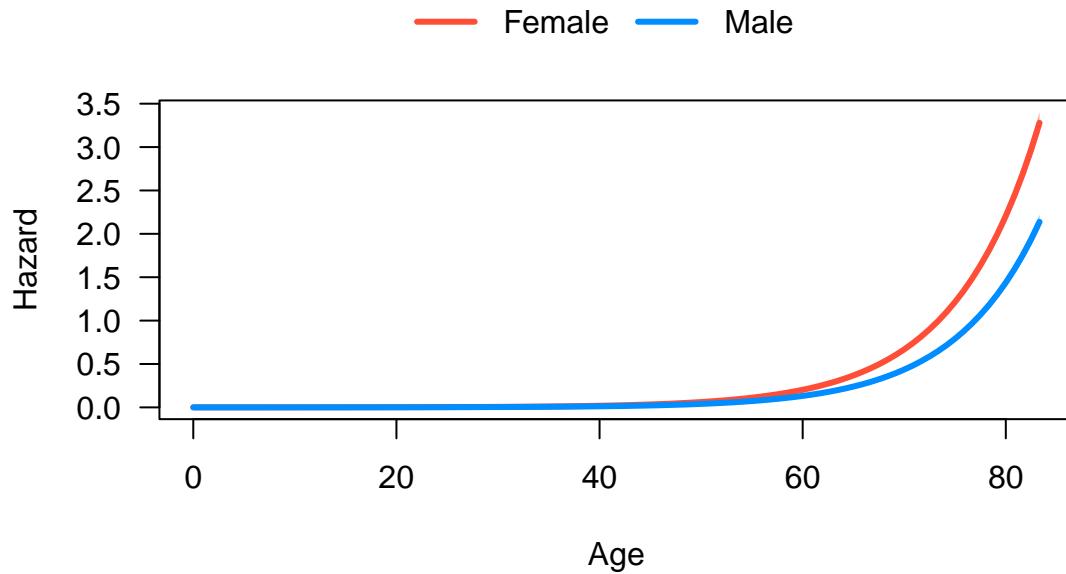
Obtenemos los parámetros estimados para las distribuciones Gama y Weibull, luego ajustamos estas distribuciones con las estimaciones y comparamos con la densidad generadas por los datos originales de Age. Podemos observar que la distribución que mejor se ajusta es la Gamma, con lo cual tomamos un ajuste para hombres y mujeres. El ajuste es un poco mejor para hombres.



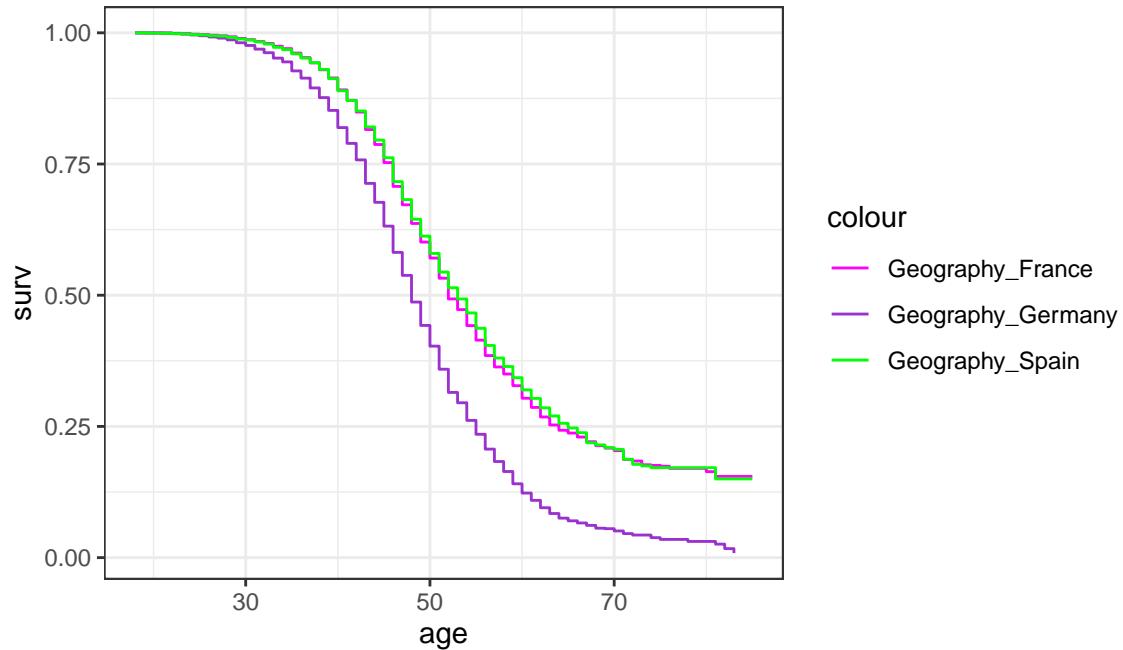
Usaremos `survfit()` y `Surv()` para construir el objeto de supervivencia estándar, usando Kaplan Meier. Usamos la fórmula $\text{survfit}(\text{Surv}(Age,Exited) \sim 1)$ para producir las estimaciones de Kaplan-Meier de la probabilidad de supervivencia en el tiempo para cada una de las categorías de Gender. A continuación, mostramos las curvas de supervivencia obtenidas, las cuales podemos interpretar como la probabilidad de permanecer en el banco más allá de cierta edad, por cada uno de los dos grupos en que se clasificaron: hombre y mujer. La tasa de supervivencia o permanencia en el banco, es menor para las mujeres en comparación con los hombres.



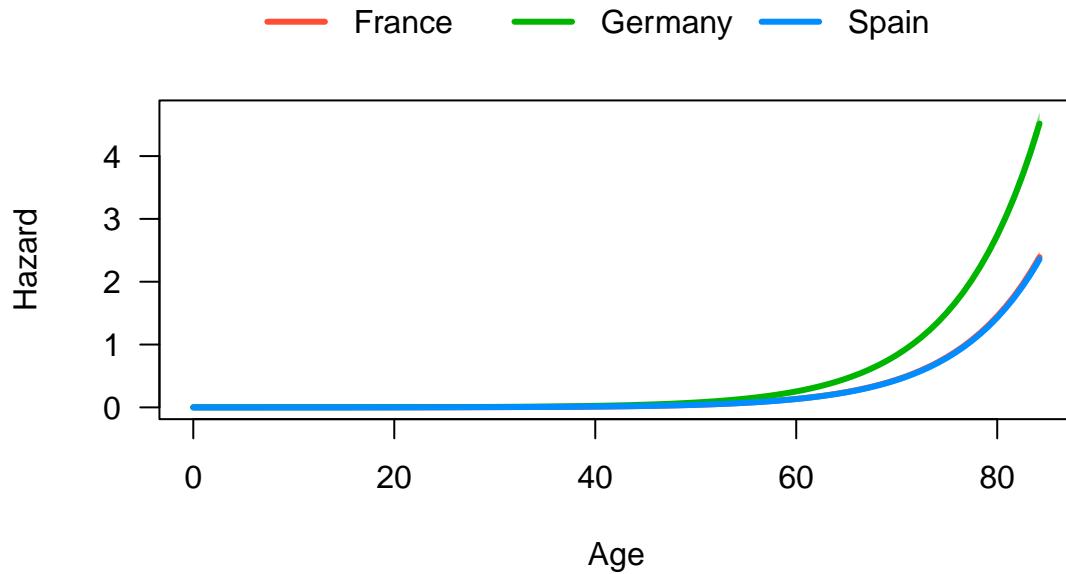
A continuación podemos ver la función de riesgo asociada.



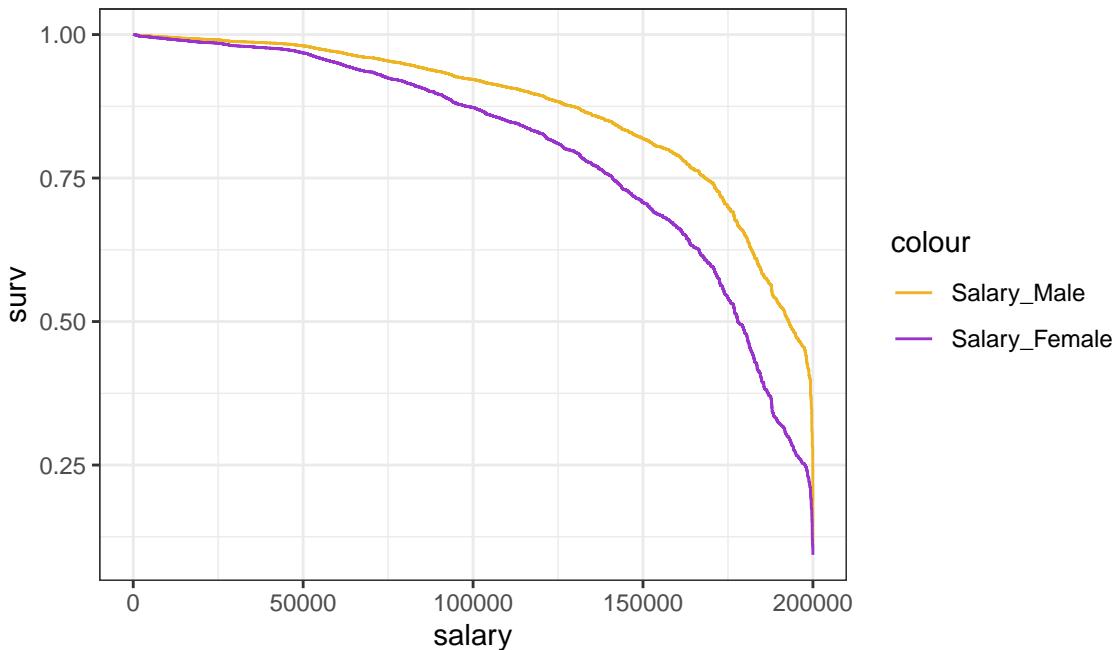
Si queremos hacer este mismo análisis para la variable Geography, que integra a Francia, Alemania y España. A continuación se muestran las funciones de supervivencia. Se puede observar que la tasa de supervivencia o permanencia en el banco para los residentes de Alemania es menor que los residentes de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



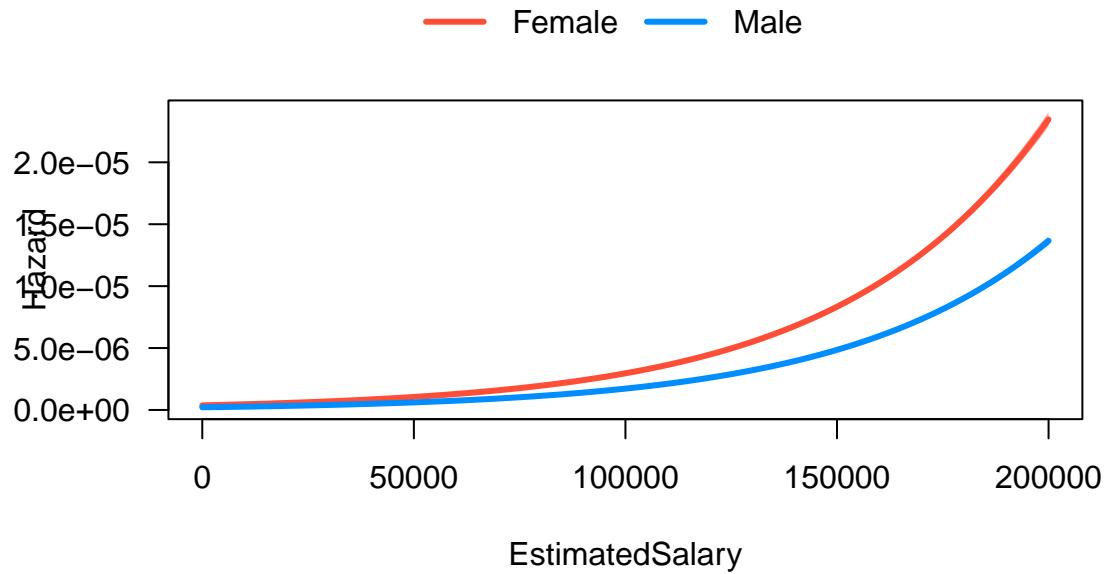
A continuación podemos ver la función de riesgo asociada.



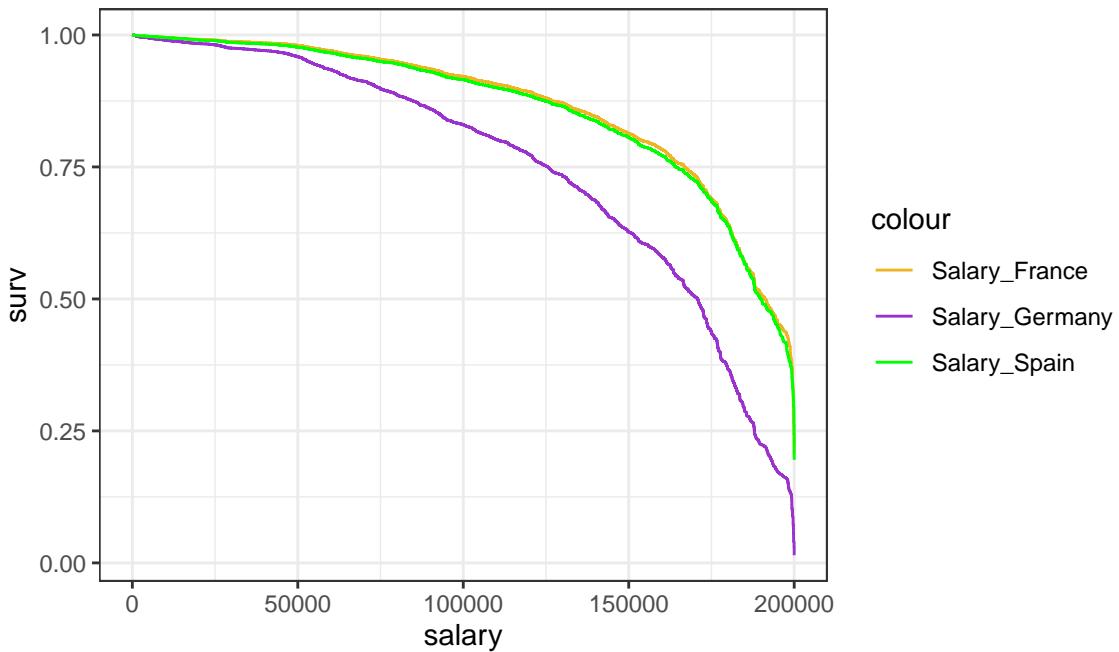
Este mismo análisis se hace para EstimatedSalary por Gender, se muestran a continuación las funciones de supervivencia, las cuales podemos interpretar como la probabilidad de permanecer en el banco más allá de cierto salario estimado por cada uno de los dos grupos en que se clasificaron: hombre y mujer. Se puede observar que a tasa de supervivencia o permanencia en el banco, es menor para las mujeres en comparación con los hombres.



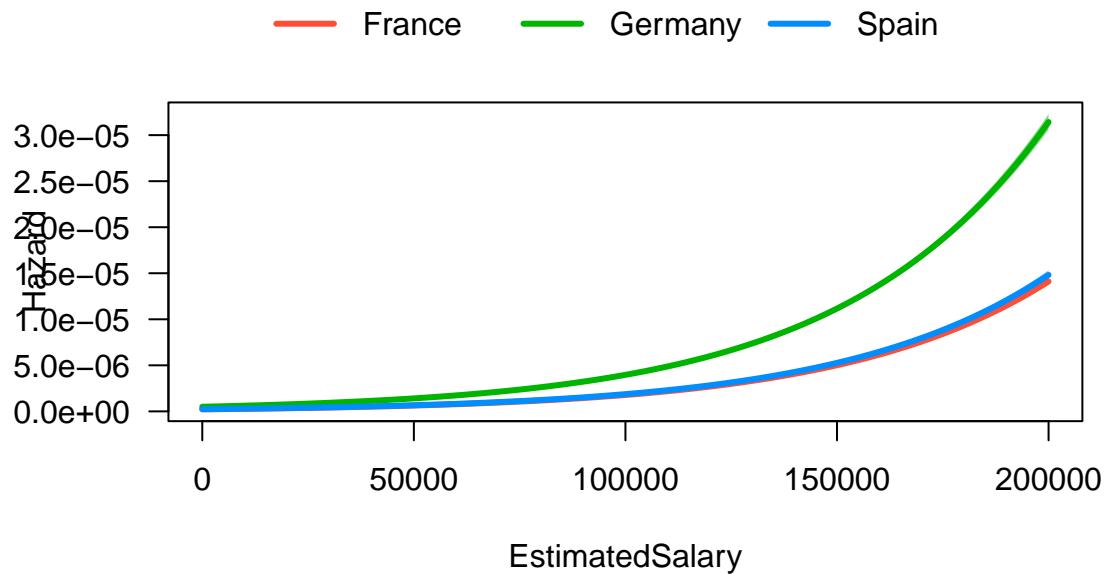
A continuación podemos ver la función de riesgo asociada.



Finalmente, en la siguiente gráfica se muestran las curvas de supervivencia considerando EstimatedSalary por país de residencia. La tasa de supervivencia o permanencia en el banco para los residentes de Alemania es menor que los residentes de Francia y España.



A continuación podemos ver la función de riesgo asociada.



Conclusiones

Con el modelo logit, pudimos observar que un mayor score crediticio disminuye la probabilidad de dejar el banco, ser residente alemán (comparado a ser francés) aumenta la probabilidad de salirse del banco, ser hombre disminuye la probabilidad de dejar el banco (en comparación a ser mujer), la edad aumenta la probabilidad de dejar el banco, los años de tenencia de la cuenta disminuye la probabilidad de dejar el banco, tener tarjeta de crédito disminuye la probabilidad de dejar el banco, ser un miembro muy activo disminuye la probabilidad de dejar el banco, el salario aumenta la probabilidad de dejar el banco, y tener dos o más productos disminuye la probabilidad de irse del banco.

Con el análisis de componentes principales y factorial, encontramos que si bien la estandarización de las variables es buena, los primeros componentes principales no acumulan rápidamente una gran proporción de las varianzas, lo cual se confirma con el análisis factorial. Los factores que se forman, a excepción del Factor 1 y Factor 2, no son útiles para el análisis y agrupamiento de los clientes del banco. Para el caso del Factor 1, conformado por Balance y NumOfProducts, éste podría ser un factor para alguna variable latente de qué tan profunda es la relación del cliente con el banco en términos de su monto de dinero y el número de productos con que cuenta. Por último, algo interesante de ver es que Age y Exited están muy correlacionados y forman un factor, lo que podría ser una caracterización importante para el banco, quizás esto tenga sentido si consideramos que las personas más jóvenes cuando abren una primera cuenta en el banco, es poco probable que abandone inmediatamente, mientras que una persona que va envejeciendo, es más probable que abandone el banco por cuestiones de salud, movilidad o por mortalidad.

Finalmente, pudimos observar que la variable de edad de los clientes (Age) se puede ajustar a una distribución Gama. Por otro lado, las curvas de supervivencia muestran que la probabilidad de permanecer en el banco más allá de cierta edad va cayendo, la tasa de supervivencia o permanencia en el banco, es menor para las mujeres en comparación con los hombres, y la tasa de supervivencia o permanencia en el banco para los residentes de Alemania es menor que los residentes de Francia y España. Además, la probabilidad de permanecer en el banco más allá de cierto salario estimado también cae, aunque menos rápidamente que como sucede con la edad.

Referencias

- KAGGLE. (2022). *Bank customer churn dataset*. <https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset>
- Villers, F., Sofía & Vazquez. (2023). *Modelos de supervivencia*. <https://svg18.github.io/Supervivencia/>