TRABAJO DE FIN DE GRADO

Grado en Estadística y Economía

Título: Predicción de las rentas de un censo mediante Regresión Logística y Regresión Logística Robusta

Autor: Jinduo Zang

Director: Montserrat Guillén, Ana María Pérez

Departamento: Departamento de Econometría, Estadística y Economía Aplicada

Convocatoria: 2019-2020

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resumen

En la regresión logística simple, las observaciones que son consideradas atípicas influyen en gran medida los resultados que arroja el modelo por los fundamentos en los cuales se rige, haciendo que éstos sean imprecisos, poco fiables y, en consecuencia, las conclusiones que se puede extraer también sean poco fiables. Con el presente trabajo, se pretende aplicar una extensión robusta llamada regresión logística robusta para tratar de corregir dicho problema.

Palabras clave

Regresión logística, regresión logística robusta

Abstract

In simple logistic regression, outliers greatly influence the results produced by the model due to its fundamental assumptions, making the results imprecise, unreliable and, consequently, the conclusions that can be drawn are also unreliable. With the present work, we intend to apply a robust extension called robust logistic regression to try to correct this problem.

Key words

Logistic regression, robust logistic regression

Clasificación AMS

62F35 Robustness and adaptive procedures

62J12 Generalized linear models

62-07 Data analysis

Índice

1. Introducción 8
2. Antecedentes 9
3. Metodología 11
3.1. Regresión logística simple 12
3.2. Regresión logística robusta 12
3.3. Matriz de confusión 13
3.4. Balance de los datos 15
4. Descripción de las bases de datos 16
4.1. Base de datos 1 16
4.2. Base de datos 2 16
5. Resultados 18
5.1. Base de datos 1 18
5.1.1. Imputación de los *missings* 18
5.1.2. Estadística descriptiva 18
5.1.3. Regresión logística simple sin introducir observaciones atípicas 22
5.1.4. Regresión logística robusta sin introducir observaciones atípicas 26
5.1.5. Regresión logística simple introduciendo observaciones atípicas 27
5.1.6. Regresión logística robusta introduciendo observaciones atípicas 30
5.1.7. Comparativa de los modelos 30
5.2. Base de datos 2 31
5.2.1. Imputación de los *missings* e incongruencias 31
5.2.2. Estadística descriptiva 32
5.2.3. Regresión logística simple 37
5.2.4. Regresión logística robusta 39
5.2.5. Comparativa de los modelos 40
6. Conclusión 41
7. Bibliografía 42
8. Anexo 43

Ilustraciones

Ilustración 1. Ajuste del modelo simple y robusto. 9
Ilustración 2. Rango de la regresión lineal y de la regresión logística. 11
Ilustración 3. Summary del modelo simple sin introducir observaciones atípicas. 22
Ilustración 4. Matriz de confusión del modelo simple sin introducir observaciones atípicas. .25
Ilustración 5. Summary del modelo robusto sin introducir observaciones atípicas. 26
Ilustración 6. Matriz de confusión del modelo robusto sin introducir observaciones atípicas. 27
Ilustración 7. Summary del modelo simple introduciendo observaciones atípicas....29
Ilustración 8. Matriz de confusión del modelo simple introduciendo observaciones atípicas. 29
Ilustración 9. Summary del modelo robusto introduciendo observaciones atípicas. 30
Ilustración 10. Matriz de confusión del modelo robusto introduciendo observaciones atípicas. 30
Ilustración 11. Ejemplo de una variable con missings. 31
Ilustración 12. Summary del modelo simple, parte I.....37
Ilustración 13. Summary del modelo simple, parte II. 37
Ilustración 14. Summary del modelo simple, parte III. 38
Ilustración 15. Matriz de confusión del modelo simple. 38
Ilustración 16. Summary del modelo robusto. 39
Ilustración 17. Matriz de confusión del modelo robusto. 39

Gráficos

Gráfico 1. Correlaciones de las variables numéricas.....18
Gráfico 2. Boxplot e histograma de V2 y V3 por V15. 19
Gráfico 3. Boxplot e histograma de V7 y V10 por V15. 20
Gráfico 4. Boxplot e histograma de V13 y V14 por V15. 20
Gráfico 5. Gráfico circular de algunas variables categóricas. 21
Gráfico 6. Histograma de V5 y V6.....21
Gráfico 7. Curva ROC del modelo simple sin observaciones atípicas. 25
Gráfico 8. Curva ROC del modelo simple introduciendo observaciones atípicas. 29
Gráfico 9. Correlaciones de las variables numéricas.....32
Gráfico 10. Boxplot de Age por Y. 32
Gráfico 11. Histograma de Age.....32
Gráfico 12. Boxplot de EducYear por Y. 33
Gráfico 13. Histograma de EducYear.....33
Gráfico 14. Boxplot de HperWeek por Y. 34
Gráfico 15. Histograma de HperWeek. 34
Gráfico 16. Curva ROC del modelo simple. 39

Tablas

Tabla 1. Matriz de confusión. 13
Tabla 2. Estadística descriptiva de las variables numéricas. 19
Tabla 3. Estadística descriptiva de las variables numéricas con observaciones atípicas. 28
Tabla 4. Comparativa de los modelos de la base de datos 1. RLS es regresión logística simple, RLR es regresión logística robusta. 30
Tabla 5. Estadística descriptiva de Age. 32
Tabla 6. Estadística descriptiva de EducYear. 33
Tabla 7. Estadística descriptiva de HperWeek. 34
Tabla 8. Estadística descriptiva de Educ. 34
Tabla 9. Estadística descriptiva de MaritalStatus. 34
Tabla 10. Tabla de frecuencia de NatCountry. 35
Tabla 11. Estadística descriptiva de Race. 35
Tabla 12. Estadística descriptiva de Sex. 35
Tabla 13. Estadística descriptiva de WorkClass. 36
Tabla 14. Comparativa de los modelos de la base de datos 2. 40

1. Introducción

El presente trabajo surge de aplicar los conocimientos y técnicas adquiridos en el Grado de Estadística a bases de datos reales y a la implementación de métodos más innovadores que no se han estudiado a lo largo de las diferentes asignaturas del Grado.

Me fue introducido el concepto de regresión logística en las asignaturas de Econometría del Grado de Economía y recuerdo que una de las características del modelo era lo sensible que puede llegar a ser a los *outliers*, que podían llegar a alterar el resultado final. Las tutoras del presente trabajo me mostraron el fascinante campo de los métodos robustos. Las técnicas y métodos pertenecientes a esta área de la estadística ofrecen, en general, resultado más consistentes y fiables que los métodos más clásicos dado que es una aproximación en la cual se pretende que las pequeñas discrepancias que pueda haber en los datos no afecten al modelo debido a sus asunciones iniciales. Estos métodos robustos son una herramienta útil para tratar los datos reales porque tienen en cuenta las imperfecciones que pueden presentar los datos, aún después de realizar una depuración previa.

Este trabajo se divide en 6 apartados. Un primer apartado contiene la introducción y la justificación del trabajo. Seguidamente se procederá a explicar de forma detallada los fundamentos matemáticos de la regresión logística y de la extensión robusta ideada por She y Owen en 2011. También se definirá la matriz de confusión y los cálculos que se derivada de este, así como su significado e interpretación. Seguidamente se procederá a una introducción de las dos bases de datos con las que se implementarán los modelos a lo largo de este trabajo. Comenzando por explicar en qué se basa cada una de ellas, la procedencia de estas, el significado de sus variables y un análisis descriptivo mediante tablas y gráficos. Posteriormente se mostrará los diferentes resultados obtenidos, en forma de salidas del programa y representaciones gráficas, mediante las dos aproximaciones expuestas en la parte de metodología. Finalmente se explicará las conclusiones obtenidas.

Después de la realización del presente TFG, me ha permitido conocer más profundidad los modelos lineales generalizados, tener la ocasión de trabajar y familiarizarme con el análisis de grandes bases de datos del ámbito económico. También me ha brindado la oportunidad de conocer con una mayor profundidad los fundamentos de los métodos robustos, el por qué es necesaria su implementación.

Para terminar con esta introducción, quiero agradecer a Montserrat y a Ana por brindarme su ayuda y por guiarme a lo largo de este último semestre.

2. Antecedentes

En 1958, el estadístico David Cox propuso la regresión logística simple para predecir variables categóricas, pero en esta ocasión, se centró en los casos de regresión logística binaria. El modelo propuesto por Cox era simple y efectivo, pero no contemplaba contratiempos como puede ser errores de etiquetado, que se manifestaba en la base de datos como observaciones atípicas.

Brodley y Friedl (1999) propusieron un método nuevo para identificar y eliminar las instancias mal etiquetadas de tal forma que, como resultado final, se obtenía una base de datos de mayor calidad para entrenar el modelo. Utilizaron una especie de filtro el cual solamente permitía pasar las observaciones que eran adecuadas. Mediante un experimento empírico, pudieron probar la efectividad de su método propuesto y que este ofrecía unas ventajas significativas.

El término de procedimientos robusto de la estadística lo empezó a utilizar Huber y Ronchetti (2009). Las suposiciones iniciales como pueden ser la independencia, la distribución a la que pertenecen los datos o la aleatoriedad no siempre son ciertas. Este campo de la estadística tiene como objetivo idear estimadores que sean significativamente poco sensibles a las pequeñas desviaciones de las suposiciones iniciales.

Una forma habitual de abordar el problema de las observaciones atípicas es utilizar una función de pérdida el cual otorga poca importancia a los puntos que se encuentren lejos de la frontera o límite.

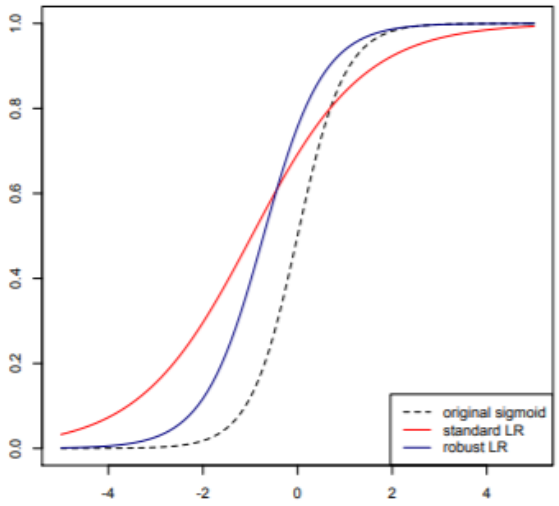

Ilustración 1. Ajuste del modelo simple y robusto. Fuente: She y Owen (2011).

She y Owen (2011) demostraron que la introducción de un parámetro de cambio podía aumentar la robusticidad de la regresión.

El aprendizaje supervisado se ha convertido en un reto debido a los retos y problemas que presentan las imperfecciones inherentes de las bases de datos que se utiliza para este fin. Bootkrajang (2016) propuso un nuevo modelo que es capaz de soportar los efectos no deseados de ruidos de etiquetas no aleatorios. Mediante estudios y ensayos de carácter empírico,

demostró que el modelo propuesto mejoraba, en relación a los modelos existentes que trataban el ruido de etiquetado, la precisión de las clasificaciones.

Existe una relación directa entre el conocimiento que se obtiene de los datos con la calidad de estos últimos en el campo de minería de datos. Esto hace que el preprocesado sea un paso realmente importante. Frecuentemente, la presencia de ruidos en la base de datos a tratar reduce la calidad de esta última. Para solucionar este problema, Morales et al. (2017) presentaron el paquete de R llamado *NoiseFiltersR*. Este contiene las técnicas más conocidas y utilizadas hasta ahora para el preprocesado de los ruidos de etiqueta, además, los algoritmos utilizados están adaptados para poder ser utilizados por algunas funciones existentes del programa para visualizar el resultado de los modelos construidos.

3. Metodología

En el ámbito estadístico, se trabaja frecuentemente con bases de datos que contienen informaciones de diferentes tipologías, como, por ejemplo, censos de la población de un determinado lugar, ensayos clínicos de alguna enfermedad o vacuna, los perfiles de los clientes que utilizan las aseguradoras para calcular el precio a cobrar, o diversas transacciones financieras.

La regresión lineal es una herramienta simple, útil y al mismo tiempo, su implementación no ofrece dificultad alguna. A menudo se requiere estudiar la relación de una variable respuesta, siempre y cuando esta sea continua, con las variable o variables predictoras. Después de construir el modelo en el programa escogido, se obtendrá, entre otros, el *intercept* y los coeficientes asociados a la variable o variables predictoras, con estos coeficientes se podrá predecir el valor de la variable respuesta de una observación introduciendo el valor de su variable o variables predictoras.

Así como posee bondades, también presenta algunas limitaciones. Tal como su propio nombre indica, describe la relación lineal entre las variables, por lo que el modelo no será adecuado cuando no existe una relación lineal. La regresión lineal tampoco es ideal cuando lo que se pretende es clasificar una observación entre dos clases, debido a que los valores predichos por el modelo pueden hallarse fuera del rango.

La regresión logística ofrece solución para este último caso debido a que predice la probabilidad de ocurrencia de un evento determinado y dicha probabilidad se hallará siempre dentro del rango.

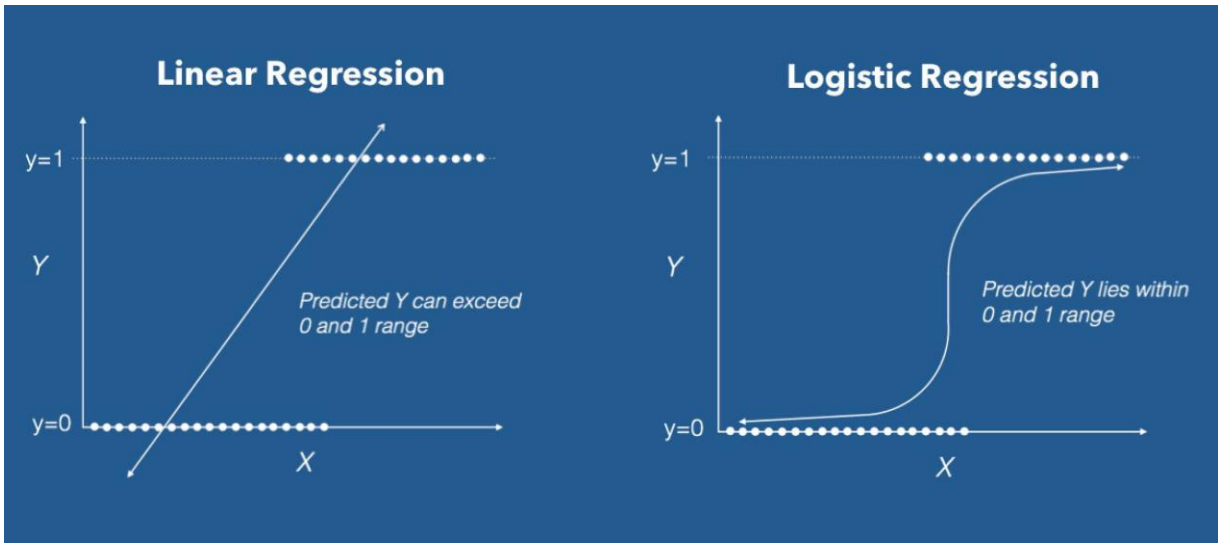

Ilustración 2. Rango de la regresión lineal y de la regresión logística. Fuente: https://www.datacamp.com.

Cuando la variable respuesta posee dos categorías, entonces se estará delante de una regresión logística binaria. En cambio, si la variable respuesta posee más de dos categorías, se usará la regresión logística multinomial.

Antes de explicar los fundamentos de la regresión logística y de su extensión robusta, conviene detallar a continuación las condiciones que se precisan para poder construir el modelo.

- Las observaciones han de ser independientes entre sí.
- La multicolinealidad entre las variables ha de ser prácticamente nula o nula.
- Requiere una cantidad mínima de observaciones.
- No requiere que las variables continuas independientes provengan de una distribución normal.
- La variable respuesta ha de ser binaria.

3.1.Regresión logística simple

La regresión logística simple, también conocida como regresión *logit*, es una de las herramientas que ofrece los Modelos Lineales Generalizados, GLM por sus siglas en inglés.

Sea k la cantidad de variables independientes del modelo. Sea Y la variable respuesta o dependiente de tipo binaria, donde cada componente de Y se distribuye mediante una distribución de Bernoulli. Sea n el número de observaciones. Sea $X = (x_1, \dots, x_n)^T$ el conjunto de variable independientes. Sea θ el vector de parámetros asociado al modelo, de forma que $\theta \in \mathbb{R}^{k+1}$. Sea $\pi(\theta^T x_i)$ la probabilidad de que Y_i tome un valor igual a 1, entonces su modelo se puede escribir como

$$\pi(\theta^T x_i) = P(Y = 1|X = x) = \frac{1}{1 + e^{-\theta^T x_i}}$$

Si $\theta^T x_i$ toma valores elevados y positivos, entonces $e^{-\theta^T x_i}$ se aproximará a 0 y, en consecuencia, el valor de la función anterior será igual a 1. En caso de que $\theta^T x_i$ tome valores elevados pero negativos, entonces el valor de la función será 0 dado que $e^{-\theta^T x_i}$ tenderá a infinito.

Es fácil darse cuenta de que la probabilidad de que Y_i tome un valor igual a 0 es el complementario de la ecuación anterior.

$$P(Y = 0|X = x) = 1 - P(Y = 1|X = x)$$

Efectuando la transformación *logit* a la expresión inicial, se obtiene

$$logit\big(\pi(\theta^T x_i)\big) = \ln\left(\frac{\pi(\theta^T x_i)}{1 - \pi(\theta^T x_i)}\right)$$

3.2.Regresión logística robusta

La siguiente extensión robusta de la regresión logística fue propuesta por She y Owen en 2011.

Se propone introducir para cada observación $i = 1, \dots, n$, un parámetro γ_i que cambia el valor del modelo, de modo que el nuevo modelo se puede escribir como

$$\pi(\theta^T x_i + \gamma_i) = P(Y = 1|X = x) = \frac{1}{1 + e^{-\theta^T x_i - \gamma_i}}$$

Se ha practicado una regularización del tipo L_1 a los parámetros γ_i para incentivar la escasez de estos dado que se parte de la creencia de que la mayoría de las observaciones no cuentan con errores que se han producido durante su anotación o etiquetado. Si se fija $\lambda \geq 0$, entonces el objetivo viene dado por

$$l(\theta, \gamma) = \sum_{i=1}^{n} [y_i \log \pi(\theta^T x_i + \gamma_i) + (1 - y_i) \log (1 - \pi(\theta^T x_i + \gamma_i))] - \lambda \sum_{i=1}^{n} |\gamma_i|$$

Se espera que el parámetro γ_i sea cero si efectivamente la observación i se encuentra etiquetada correctamente, en este caso γ_i no afectaría al modelo. Si se diera el caso en el cual todas γ_i que se dispone toman cero como valor, entonces se tendría una regresión logística usual. En el caso de que la observación i pertenezca a la clase negativa pero debido a errores, se ha anotado como miembro de la clase positiva, en tal caso, γ_i podría tomar un valor negativo. Analógicamente, se puede aplicar la misma lógica para el caso contrario.

Si se eligiera una penalización de tipo L_1, el nuevo objetivo se puede escribir como

$$l(\theta, \gamma) = \sum_{i=1}^{n} [y_i \log \pi(\theta^T x_i + \gamma_i) + (1 - y_i) \log (1 - \pi(\theta^T x_i + \gamma_i))] - \kappa \sum_{j=1}^{m} |\theta_j| - \lambda \sum_{i=1}^{n} |\gamma_i|$$

3.3.Matriz de confusión

Una vez que se ha obtenido el modelo deseado, es conveniente utilizar una base de datos, diferente a la utilizada para obtener el modelo para realizar predicciones. Con las predicciones en mano, es posible crear la denominada matriz de confusión, que se obtiene al realizar la validación cruzada. La matriz de confusión es una matriz que cuenta con dos filas y dos columnas en los cuales se comparan las observaciones con las predicciones, tal como se muestra en la siguiente tabla.

	Observaciones negativas	*Observaciones positivas*	*Total*
Predicciones negativas	Verdadero Negativo (VN)	Falso Negativo (FN)	VN+FN
Predicciones positivas	Falso Positivo (FP)	Verdadero Positivo (VP)	FP+VP
	VN+FP	FN+VP	Número individuos

Tabla 1. Matriz de confusión. Fuente: Elaboración propia.

Los falsos positivos corresponden al error de tipo I y los falsos negativos corresponden al error de tipo II.

A partir de la matriz de confusión se puede realizar diversos cálculos que serán de ayuda para comprobar el ajuste del modelo.

La sensibilidad (*Sensitivity*) es la proporción de los positivos predichos dentro de las observaciones positivas.

$$Sensitivity = \frac{VP}{FN + VP}$$

La especificidad (*Specificity*) es la proporción de los negativos predichos dentro de las observaciones negativas.

$$Specificity = \frac{VN}{VN + FP}$$

El valor positivo predicho (*Positive Predicted Value*) es la proporción de las observaciones positivas dentro de los positivos predichos.

$$PPV = \frac{VP}{FP + VP}$$

El valor negativo predicho (*Negative Predicted Value*) es la proporción de las observaciones negativas dentro de los negativos predichos.

$$NPV = \frac{VN}{VN + FN}$$

La prevalencia (*Prevalence*) es la proporción de las observaciones positivas dentro de las observaciones totales.

$$Prevalence = \frac{FN + VP}{VN + FN + FP + VP}$$

La ratio de detección (*Detection Rate*) es la proporción de las observaciones positivas correctamente predichas dentro de las observaciones totales.

$$Detection\ Rate = \frac{VP}{VN + FN + FP + VP}$$

La ratio de verosimilitud positivo (*Positive Likelihood Ratio*) representa, en una observación predicha como positiva, cuánto más probable es una respuesta positiva respecto a respuesta negativa.

$$PLR = \frac{\frac{VP}{FP + VP}}{1 - \frac{VN}{VN + FN}} = \frac{PPV}{1 - NPV}$$

La ratio de verosimilitud negativo (*Negative Likelihood Ratio*) representa, en una observación predicha como negativa, cuánto más probable es una respuesta negativa respecto a una respuesta positiva.

$$NLR = \frac{\frac{VN}{VN + FN}}{1 - \frac{VP}{FP + VP}} = \frac{NPV}{1 - PPV}$$

La precisión (*Accuracy*) es la proporción de las observaciones correctamente predichas dentro de las observaciones totales.

$$Accuracy = \frac{VN + VP}{VN + FN + FP + VP}$$

La precisión balanceada (*Balanced Accuracy*) es la media aritmética de sensibilidad y especificidad.

$$Balanced\ Accuracy = \frac{\frac{VP}{FN + VP} + \frac{VN}{VN + FP}}{2} = \frac{Sensitivity + Specifity}{2}$$

3.4.Balance de los datos

Al trabajar con métodos de clasificación supervisada como lo es la regresión logística, siempre es recomendable utilizar una base de datos en la cual los datos se encuentren balanceados. Si el número de observaciones para cada valor que puede tomar la variable respuesta son similares, quiere decir que se dispone de unos datos balanceados.

Sabiendo la cantidad de observaciones que toman un valor u otro para la variable respuesta, se definirá el denominado *threshold*, que es el umbral de clasificación o umbral de decisión. De este valor dependerá si una nueva observación pertenecerá a una clase o a otra. El *threshold* no es una constante, dependerá de las características de cada base de datos con los que se va a trabajar. Un valor alto (bajo) de *threshold* comportará una elevada (baja) especificidad y baja (elevada) sensibilidad.

En la práctica, la mayoría de base de datos serán del tipo no balanceado o desbalanceado. Muchas causas pueden derivar a este resultado, pero una de las causas más comunes es debido al tipo de problema al que se enfrenta. Por ejemplo, si se desea estudiar si una operación financiera es un fraude o no. Entonces la variable respuesta será de tipo binaria, tomando 1 si es fraude y 0 en caso contrario, por ejemplo. Seguramente en dicha base de datos predominen las observaciones cuyos valores toman 0 en la variable respuesta pues es lógico y razonable partir de la hipótesis de que la gran mayoría de las operaciones son lícitas.

4. Descripción de las bases de datos

4.1.Base de datos 1

La base de datos ha sido facilitada por Ross Quinlan y se puede descargar en *https://archive.ics.uci.edu/ml/datasets/Statlog+%28Australian+Credit+Approval%29*.
Contiene información sobre unas aplicaciones de las tarjetas de crédito que tuvo lugar en Australia. Se ha optado por dotar a las variables explicativas nombres sin significado alguno debido a la información sensible que contiene la base de datos.

Cuenta con 690 observaciones en total. Excluyendo la variable respuesta, se puede encontrar un total de 14 variables explicativas, de las cuales 6 son numéricas y 8 son categóricas. Contiene una mezcla de variables continuas, variables nominales con pocas categorías y variables nominales con muchas categorías.

Previamente se ha codificado las variables categóricas para facilitar su posterior tratamiento estadístico. Hubo 37 casos de valores *missings*.

4.2.Base de datos 2

La base de datos ha sido donada por Ronny Kohavi y Barry Becker, se puede descargar en *https://archive.ics.uci.edu/ml/datasets/Adult*. Se trata de un censo que tuvo lugar en los Estados Unidos en el año 1994. Es necesario destacar que dicha base de datos fue depurada previamente y todos los individuos cumplían simultáneamente los siguientes requisitos

- Mayor de 16 años.
- Trabajar más de 0 horas a la semana.

Previamente, Kohavi y Becker ya habían definido la base de datos que se usará para entrenar y la base de datos para testear los modelos. Después de la depuración de las bases de datos, la primera de ellas cuenta con 29818 individuos mientras que la segunda de ellas cuenta con un total de 14880 individuos. Las dos bases de datos contienen las mismas variables.

Se pueden encontrar un total de 14 variables, las cuales 5 son variables numéricas y continuas, y 9 son variables categóricas entre los cuales se encuentra la variable respuesta a la que se le ha nombrado *Y*.

Se ha asignado a cada variable un nombre que está relacionado con su significado y permitirá su rápida identificación.

- **Age**: variable numérica que indica la edad del individuo en años.
- **CapGain**: variable numérica que indica los ingresos percibidos del individuo durante el último año.
- **CapLoss**: variable numérica que indica las pérdidas sufridas del individuo durante el último año.

- **EducYear**: variable numérica que indica los años de formación que ha recibido el individuo.
- **HperWeek**: variable numérica que indica las horas trabajadas por el individuo.
- **Educ**: variable categórica que indica el nivel máximo de educación del individuo. *Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool*.
- **MaritalStatus**: variable categórica que indica el estado civil del individuo. *Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse*.
- **NatCountry**: variable categórica que indica el país de nacimiento del individuo. *United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US (Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&Tobago, Peru, Hong, Holland-Netherlands*.
- **Occup**: variable categórica que indica la profesión del individuo. *Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Froces.*
- **Race**: variable categórica que indica la raza del individuo. *White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.*
- **Relationship**: variable categórica que indica el parentesco del individuo. *Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.*
- **Sex**: variable categórica que indica el sexo del individuo. *Female, Male*.
- **WorkClass**: variable categórica que indica el ámbito en el cual trabaja el individuo. *Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked*.
- **Y**: variable categórica, binaria, que indica si las rentas del individuo es igual o mayor a 50.000 dólares estadounidenses.

5. Resultados

5.1.Base de datos 1

En este apartado, se ha optado por dividir la base de datos inicial en dos partes. Una primera con 460 observaciones que servirá para entrenar el modelo de cara a su construcción, mientras que la segunda cuenta con las 230 observaciones restantes y servirá para probar la fiabilidad del modelo obtenido.

Se tiene como objetivo comprobar cómo afecta a la regresión logística simple y a la regresión logística robusta el hecho de introducir manualmente algunas observaciones atípicas de tipo numérica.

5.1.1. Imputación de los *missings*

Esta base de datos en particular ya ha sido tratada previamente de la siguiente forma, tal como lo notifica el documento que describe la base de datos. En concreto, había 37 observaciones, que representan el 5% de las observaciones totales, que contenían 1 o más *missings*, que han sido sustituidos por la media de la variable si esta era numérica, y por la moda de la variable si esta última era categórica.

5.1.2. Estadística descriptiva

En primer lugar, conviene visualizar las correlaciones existentes entre las variables numéricas mediante la función *corrplot* del paquete con el mismo nombre.

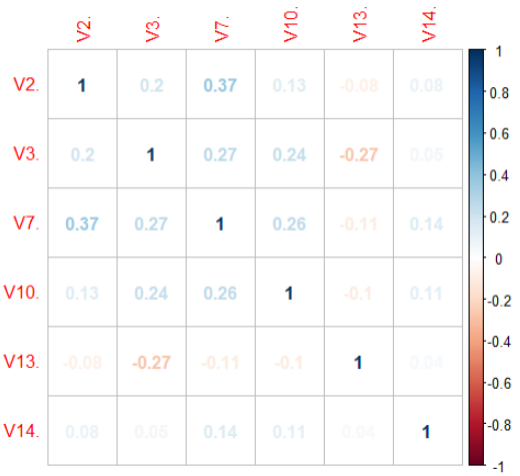

Gráfico 1. Correlaciones de las variables numéricas. Fuente: Elaboración propia.

En general, no existen una elevada correlación excepto para el caso de la variable *V2* con la variable *V7*. Las variables numéricas no parecen estar correlacionadas entre ellas.

Dado que se desconoce el significado de las variables de esta base de datos, se ha optado por realizar una descriptiva sin profundizar en todas las variables.

```
---------------------------------------------------------------------------------------------------------------------------
V2.
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     460        0      286        1    31.52    12.94    18.08    19.42    22.50    28.21    37.56    48.58    55.95

lowest : 15.75 15.92 16.00 16.08 16.17, highest: 69.17 71.58 73.42 74.83 76.75
---------------------------------------------------------------------------------------------------------------------------
V3.
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     460        0      168        1     4.85    5.256   0.1698   0.3710   0.8750   2.7500   8.2425  11.7750  14.5042

lowest :  0.000  0.040  0.085  0.125  0.165, highest: 21.000 22.290 25.085 25.125 26.335
---------------------------------------------------------------------------------------------------------------------------
V7.
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     460        0      105    0.998    2.081    2.695    0.000    0.000    0.165    1.000    2.500    5.500    8.500

lowest :  0.000  0.040  0.085  0.125  0.165, highest: 15.000 15.500 16.000 17.500 20.000
---------------------------------------------------------------------------------------------------------------------------
V10.
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     460        0       21    0.812    2.524    3.889        0        0        0        0        3        9       12

lowest :  0  1  2  3  4, highest: 16 17 19 20 67
---------------------------------------------------------------------------------------------------------------------------
V13.
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     460        0      128    0.991    174.2    167.6      0.0      0.0     61.5    147.5    250.5    380.0    420.1

lowest :    0   17   20   21   22, highest:  680  720  928 1160 2000
---------------------------------------------------------------------------------------------------------------------------
V14.
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     460        0      170    0.915    956.1     1729      1.0      1.0      1.0      4.0    351.2   1963.2   4004.5

lowest :     1     2     3     4     5, highest: 15109 26727 31286 50001 51101
---------------------------------------------------------------------------------------------------------------------------
```

Tabla 2. Estadística descriptiva de las variables numéricas. Fuente: Elaboración propia.

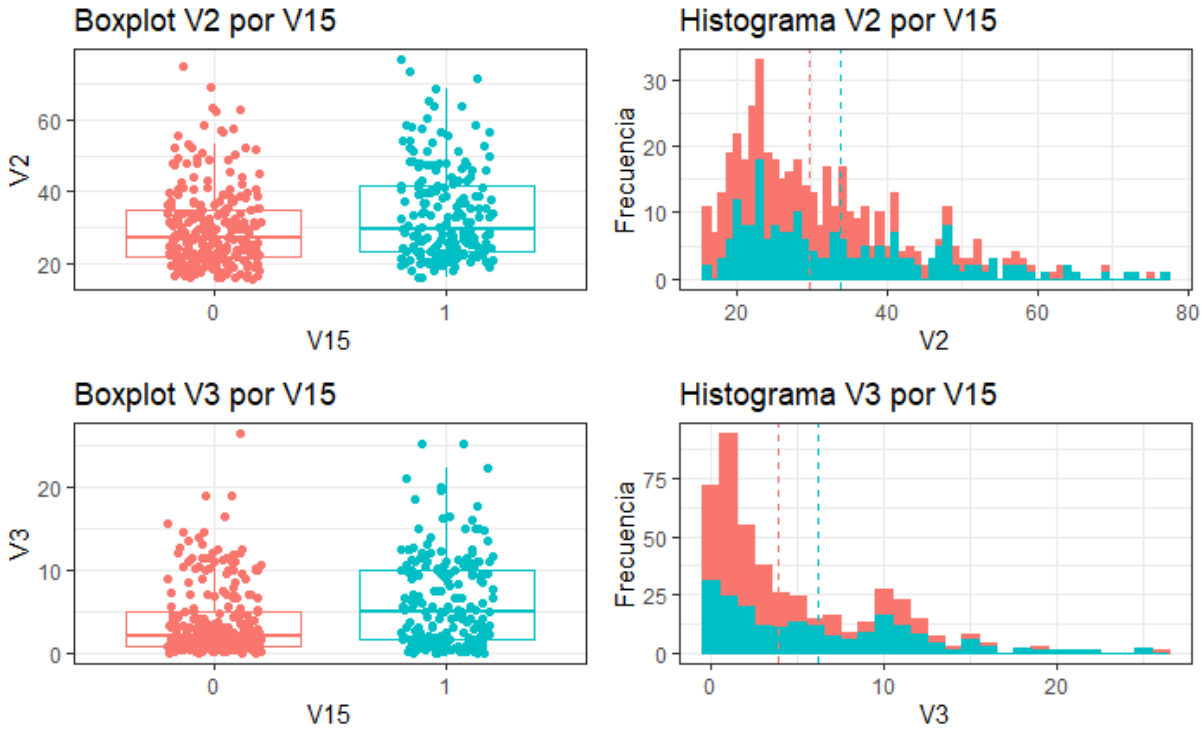

Gráfico 2. Boxplot e histograma de V2 y V3 por V15. Fuente: Elaboración propia.

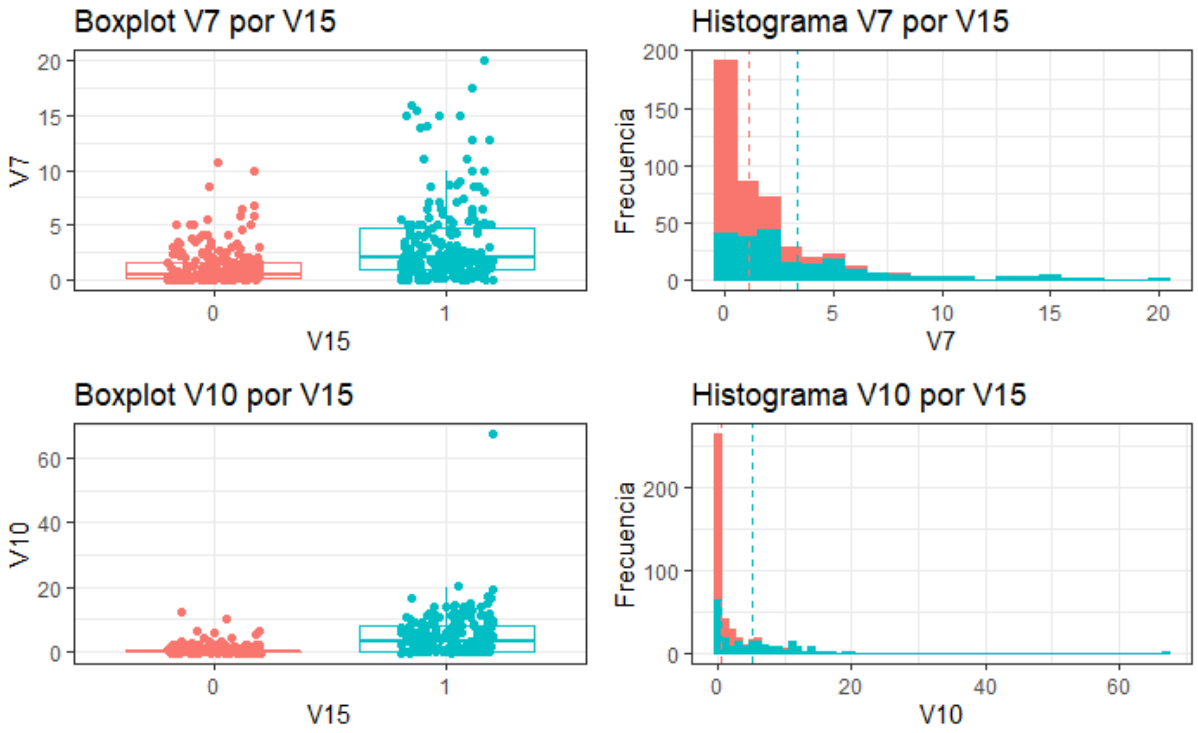

Gráfico 3. Boxplot e histograma de V7 y V10 por V15. Fuente: Elaboración propia.

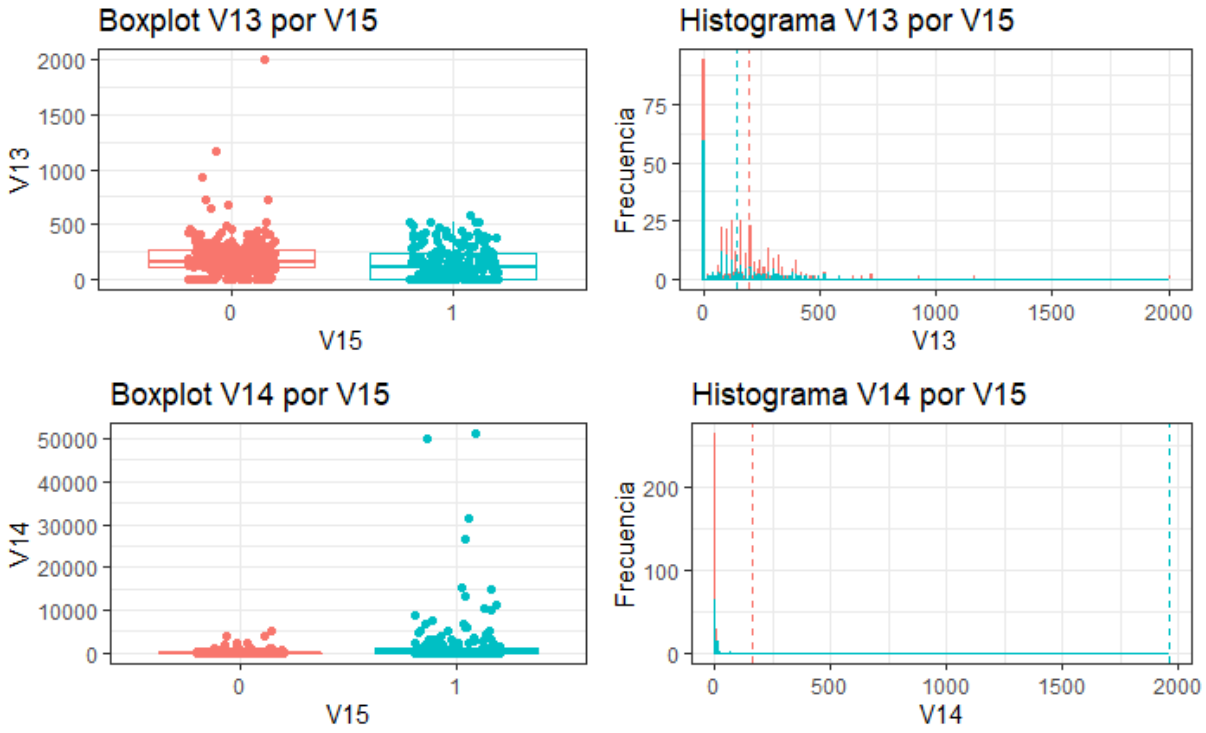

Gráfico 4. Boxplot e histograma de V13 y V14 por V15. Fuente: Elaboración propia.

En cuanto a las variables numéricas, todas presentan valores iguales a ceros o positivos, nunca negativos. Observando los distintos boxplots e histogramas, se puede apreciar cómo, en general, las observaciones pertenecientes al grupo 0 de la variable respuesta presentan unos valores

menores y más concentrados que el grupo 1, que tienden a presentar unos valores más amplios. También se puede detectar algún que otro valor atípico para algunas observaciones mediante los boxplots.

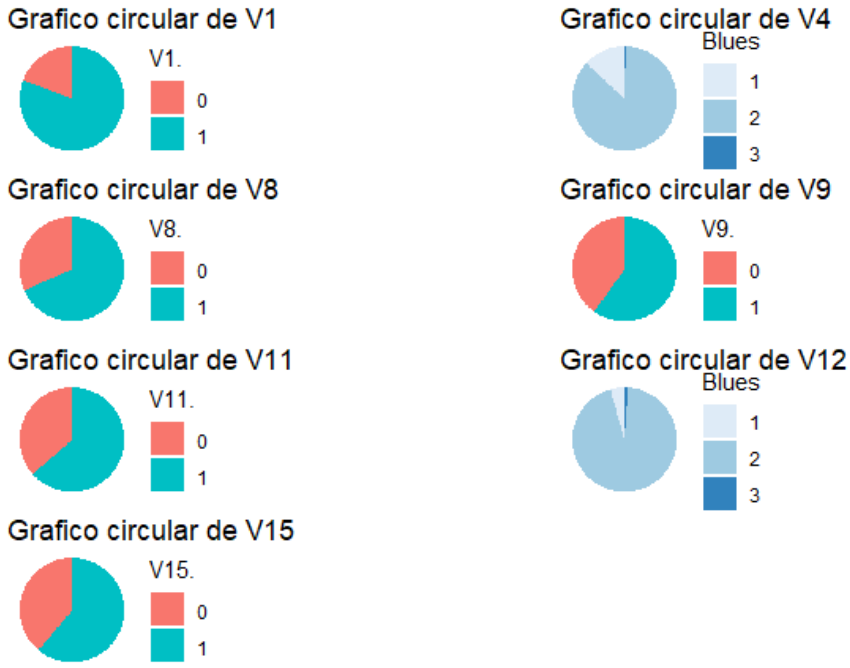

Gráfico 5. Gráfico circular de algunas variables categóricas. Fuente: Elaboración propia.

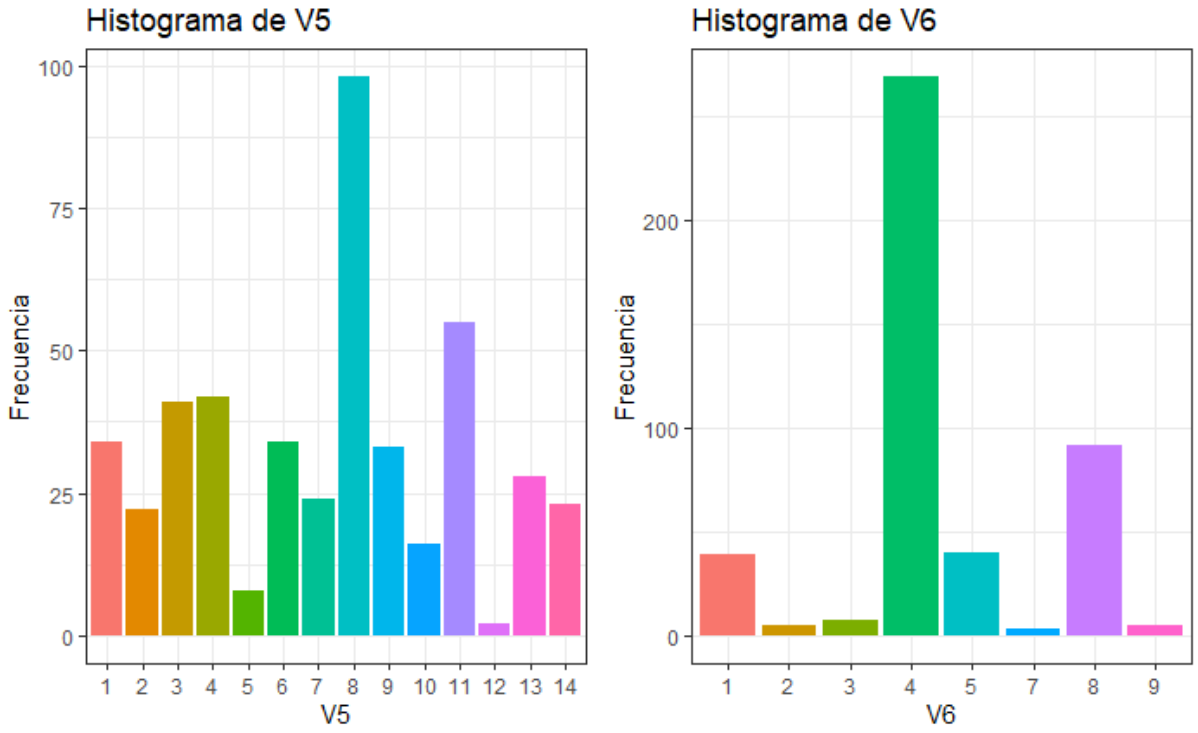

Gráfico 6. Histograma de V5 y V6. Fuente: Elaboración propia.

Dado que el color de algunos de los gráficos circulares coincide con los gráficos de las variables numéricas, las variables categóricas no se ha separado según la clase que pertenecían para su análisis descriptivo. La proporción de las variables que cuentan solamente con dos clases se asemejan bastante a la proporción de la variable respuesta. Para las variables que cuentan con tres clases, ambos siguen el mismo patrón: una clase claramente predominante y otra claramente no dominante. En cuanto a las variables que cuentan con más de 3 categorías, hay 1 categoría que tiene una frecuencia muy elevada, algunas categorías con una frecuencia cercana al cero y el resto de categorías tienen una frecuencia intermedia.

5.1.3. Regresión logística simple sin introducir observaciones atípicas

```
Call:
glm(formula = V15. ~ V1. + V2. + V3. + V4. + V5. + V6. + V7. +
    V8. + V9. + V10. + V11. + V12. + V13. + V14., family = "binomial",
    data = dd1train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.00978  -0.28175  -0.09279   0.25012   2.93801

Coefficients:
                Estimate    Std. Error z value             Pr(>|z|)
(Intercept)    -6.6637305     1.4908087  -4.470 0.000007826483040746 ***
V1.1            0.0027649     0.4155753   0.007              0.99469
V2.             0.0087432     0.0180933   0.483              0.62893
V3.            -0.0219671     0.0408245  -0.538              0.59052
V4.2            1.3996949     0.4640500   3.016              0.00256 **
V4.3           18.3883911  2399.5462860   0.008              0.99389
V5.2            2.8381601     2.6313814   1.079              0.28077
V5.3            2.3534050     2.5883321   0.909              0.36323
V5.4            2.7520910     2.4475051   1.124              0.26082
V5.5           -2.1361947     2.6047057  -0.820              0.41214
V5.6            2.5218573     2.5495171   0.989              0.32259
V5.7            2.4444529     2.6408605   0.926              0.35464
V5.8            3.4715031     2.5022431   1.387              0.16533
V5.9            3.6111915     2.5840112   1.398              0.16226
V5.10           2.3133207     3.3341149   0.694              0.48779
V5.11           3.0107519     2.5172301   1.196              0.23167
V5.12           1.4258383 10328.5482237   0.000              0.99989
V5.13           4.7314723     2.6390229   1.793              0.07299 .
V5.14           5.8113679     2.7004362   2.152              0.03140 *
V6.2            0.0702727     3.6223702   0.019              0.98452
V6.3            3.3971669     2.5246489   1.346              0.17843
V6.4           -1.1067681     2.3142420  -0.478              0.63248
V6.5           -1.7126842     2.4150835  -0.709              0.47822
V6.7          -14.4726381  1228.0250552  -0.012              0.99060
V6.8           -1.1862533     2.3544516  -0.504              0.61438
V6.9           -0.7551552     4.2499672  -0.178              0.85897
V7.             0.1586224     0.0905393   1.752              0.07978 .
V8.1            3.8982350     0.4782728   8.151 0.000000000000000362 ***
V9.1           -0.3617176     0.5550689  -0.652              0.51462
V10.            0.2987604     0.0971182   3.076              0.00210 **
V11.1          -0.5029317     0.3867316  -1.300              0.19344
V12.2           0.7367914     0.7169252   1.028              0.30409
V12.3          19.8131682  1380.4458633   0.014              0.98855
V13.           -0.0037461     0.0015658  -2.392              0.01674 *
V14.            0.0006355     0.0003074   2.067              0.03871 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 631.34  on 459  degrees of freedom
Residual deviance: 221.03  on 425  degrees of freedom
AIC: 291.03
```

Ilustración 3. Summary del modelo simple sin introducir observaciones atípicas. Fuente: Elaboración propia.

En este apartado se procederá a realizar la construcción del modelo, mediante la función *glm()* de la librería *stats* de R. Seguidamente se comentará, en detalle, la salida que ofrece la función *summary()* cuando se introduce como argumento el modelo construido.

En la parte superior, se puede apreciar los argumentos que se han introducido a la función *glm()*: la fórmula utilizada, la familia de distribución escogida y los datos con los que se quiere trabajar. Seguidamente, se halla la *Deviance Residuals*, se define como: "The residual deviance is a measure of the residual variability across *n* observations after fitting the model" (Dunn & Smyth, 2018, pág. 248). Su mediana toma un valor cercano a cero, esto significa que el modelo no se encuentra sesgado en una única dirección.

En tercer lugar, se puede encontrar la información relacionada con las variables, que aparece como *Coefficients*.

- La primera columna muestra el *intercept* y los nombres de las variables. En algunas variables tienen un número después del punto y en otras no. Esto se debe a que los que no poseen números después del punto son variables numéricas. Por ejemplo, la variable *V2* y *V3* son variables numéricas. En cambio, los que poseen números después del punto corresponden a variables categóricas. Por ejemplo, *V1.1* corresponde a la categoría 1 de la variable *V1* y *V4.2* corresponde a la categoría 2 de la variable *V4*.
- La segunda columna, denominada *Estimate*, indica el valor numérico de los coeficientes asociados al *intercept* y a las variables predictoras. Su interpretación depende de si la variable a la cual se encuentra asociada es numérica o categórica.
  Para el primer caso, y tomando *V2* como ejemplo, el cual su coeficiente toma 0.0087432, se puede decir que por cada unidad que incremente dicha variable y sin variar las variables restantes, el *log odds* se incremente, en promedio, 0.0087432. Efectuando la inversa del logaritmo natural, $e^{0.0087432} = 1.008782$, por cada unidad extra en la variable *V2* hará que los *odds* de pertenecer a la categoría 1 de la variable *V15* se incremente, en promedio, 1.008782 veces, es decir un 0.8782%.
  Para el segundo caso, y tomando la categoría 1 de la variable *V8* como ejemplo, si se compara un individuo que pertenezca a dicha categoría con otro individuo de idénticas características, pero siendo este último parte de la categoría 0 para la variable *V8*, el *log odds* del primer individuo frente el *log* odds del segundo individuo de pertenecer a la categoría 1 de la variable *V15* es $e^{3.898235} = 49.31533$.
- La tercera columna, denominada *Std. Error*, muestra el error estándar de cada coeficiente estimado. Representa la precisión de los coeficientes. Lo deseable es que el error estándar sea lo más pequeño posible, como por ejemplo la de la variable *V14*, que es prácticamente igual a cero. Algunas categorías de algunas variables, como por ejemplo la categoría 12 de la variable *V5*, presentan un error estándar anormalmente elevado, debido a que hay pocas observaciones que pertenecen a esta categoría, tal como se ha podido comprobar cuando se efectuó la estadística descriptiva.
- La cuarta columna, denominada *z value*, corresponde al estadístico Z. Es resultado de dividir la columna *Estimate* con la columna *Std. Error*.
- La quinta y última columna, denominada *Pr(>|z|)*, es el p-valor del estadístico Z. Si p-valor es menor que el nivel de significación escogido, entonces el efecto que tiene la

variable o categoría de una variable sobre la probabilidad de la respuesta igual a 1 será estadísticamente significativo. Si el nivel de significación es del 0.05, entonces la categoría 1 de la variable *V8* y la variable *V8* tienen un efecto que es estadísticamente significativo.

En la parte inferior, se encuentran *Null deviance* y *Residual deviance*, que son una medida para poder comparar el ajuste entre diferentes modelos. Valores elevados indican un mal ajuste. El primero de ellos indica cuán bien la variable respuesta (probabilidad de 1) es predicha por un modelo que solamente incluya el término *intercept*, mientras que el segundo indica cuán bien la variable respuesta (probabilidad de 1) es predicha por un modelo que incluya todas las variables. El modelo nulo tiene 631.34 como Null *deviance* y cuenta con 460 - 1 = grados de libertad. En cambio, el modelo propuesto tiene 221.03 como Residual *deviance* y cuenta con 460 - 35 = 425. Esto quiere decir que *Residual deviance* se ha reducido en 631.34 - 221.03 = 410.21 con una pérdida de 450 - 425 = 34 grados de libertad, cantidad que coincide con el número de filas perteneciente al recuadro *Coefficients*, excluyendo el término *intercept*. Si se utiliza la función *pchisq(q = 410.21, df = 34, lowe.tail = FALSE)*, también se llegará a la misma conclusión, el modelo propuesto es significativo.

Una vez que ya se ha obtenido el resultado de la estimación del modelo, es conveniente realizar la validación cruzada. En R, no hay dificultad alguna en programarlo manualmente, pero la función *confusionMatrix()* del paquete *caret*, introduciendo la base de datos de testeo, proporciona toda la información estadística necesaria en relación a la matriz de confusión. Como los datos que se han utilizado para construir el modelo contenían 257 observaciones pertenecientes a la categoría 0 y 460 observaciones en total, se ha fijado un valor de *threshold* igual a 0.558657. El *threshold* define el límite donde a partir del cual, si el *log odds* de una observación es mayor que este valor, entonces pertenecerá a la categoría 1, para este caso.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 106  20
         1  21  83

               Accuracy : 0.8217
                 95% CI : (0.766, 0.8689)
    No Information Rate : 0.5522
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6399

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8346
            Specificity : 0.8058
         Pos Pred Value : 0.8413
         Neg Pred Value : 0.7981
             Prevalence : 0.5522
         Detection Rate : 0.4609
   Detection Prevalence : 0.5478
      Balanced Accuracy : 0.8202

       'Positive' Class : 0
```

Ilustración 4. Matriz de confusión del modelo simple sin introducir observaciones atípicas. Fuente: Elaboración propia.

Se ha obtenido una precisión del 82.17% utilizando una base de datos distinta a la inicial. Al menos cada 4 de 5 predicciones son correctas. El modelo es bastante fiable. Los valores de los falsos positivos y los falsos negativos son casi idénticos.

La curva *Receiver Operating Characteristic* o ROC es una representación gráfica de la especificidad frente a la sensibilidad. Para comparar dos modelos se emplea la conocida *Area Under the Curve* o AUC. Este valor se comprende entre 0.5 y 1. Siendo el primero un valor que indica que el modelo no posee capacidad discriminatoria alguna, mientras que el segundo valor corresponde a un modelo que discrimina perfectamente.

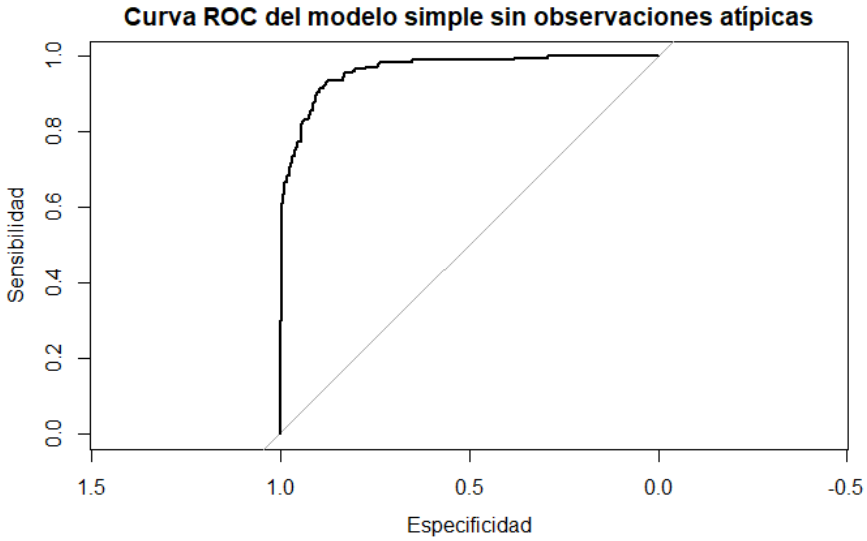

Gráfico 7. Curva ROC del modelo simple sin observaciones atípicas. Fuente: Elaboración propia.

El valor de AUC es de 0.9634, que se podría extender como un valor casi excelente.

5.1.4. Regresión logística robusta sin introducir observaciones atípicas

Para la construcción del modelo de extensión robusta, se le ha proporcionado las variables predictoras, la variable respuesta y la familia de distribución, que en este caso es binomial. Es interesante señalar que, si las variables predictoras son todas numéricas, no hay dificultad alguna en la construcción del modelo. Pero si existen algunas variables categóricas, no se pueden introducir directamente a la función como si fueran variables predictoras numéricas. Es necesario primero convertir estas variables categóricas en variables *dummies*. En R, se puede lograr el resultado deseado usando la función *model.matrix()*.

También es necesario cerciorarse de que la matriz devuelta por la función *model.matrix()* tengan las mismas variables tanto para la base de datos de entreno como para la base de datos de testeo, dado que puede darse el caso en que haya una categoría que solamente existe en la base de datos de entreno y no en la de testeo, o viceversa. El punto anterior es clave cuando se quiere realizar predicciones a partir del modelo construido.

```
           Length Class     Mode
a0             91 -none-    numeric
beta         3185 dgCMatrix S4
df             91 -none-    numeric
dim             2 -none-    numeric
lambda         91 -none-    numeric
dev.ratio      91 -none-    numeric
nulldev         1 -none-    numeric
npasses         1 -none-    numeric
jerr            1 -none-    numeric
offset          1 -none-    logical
classnames      2 -none-    character
call            4 -none-    call
nobs            1 -none-    numeric
```

Ilustración 5. Summary del modelo robusto sin introducir observaciones atípicas. Fuente: Elaboración propia.

El *summary* del modelo no ofrece mucha información. La fila *lambda* indica el número de lambdas que tiene el modelo. En este caso, hay un total de 91 lambdas. A la hora de realizar predicciones, se deberá de introducir un valor de lambda. Se ha escogido la lambda que mejor precisión ofrece.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 99 27
         1  8 96

               Accuracy : 0.8478
                 95% CI : (0.7948, 0.8917)
    No Information Rate : 0.5348
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6977

 Mcnemar's Test P-Value : 0.002346

            Sensitivity : 0.9252
            Specificity : 0.7805
         Pos Pred Value : 0.7857
         Neg Pred Value : 0.9231
             Prevalence : 0.4652
         Detection Rate : 0.4304
   Detection Prevalence : 0.5478
      Balanced Accuracy : 0.8529

       'Positive' Class : 0
```

Ilustración 6. Matriz de confusión del modelo robusto sin introducir observaciones atípicas. Fuente: Elaboración propia.

Las predicciones del modelo robusto presentan una tasa de acierto de 84.78%, que es ligeramente más alta a la que arroja el modelo simple. Pero en este caso se puede apreciar cómo los casos de falsos negativos representan más de tres veces los casos de falsos positivos.

5.1.5. Regresión logística simple introduciendo observaciones atípicas

Para este y el siguiente apartado, se ha introducido manualmente 3 observaciones atípicas. No se ha añadido nuevas categorías en las variables categóricas, sino que se ha introducido valores elevados para las variables numéricas de forma que todas las variables de todas las observaciones siguen teniendo valores positivos, igual que la base de datos original.

```
---------------------------------------------------------------------------------------------------------------------------
V2.
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     463        0      289        1    32.26    14.25    18.09    19.44    22.54    28.25    38.21    48.95    56.80

lowest :  15.75  15.92  16.00  16.08  16.17, highest:  74.83  76.75 120.00 150.00 168.00
---------------------------------------------------------------------------------------------------------------------------
V3.
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     463        0      171        1    5.748    6.977   0.1735   0.3750   0.8750   2.8750   8.5000  12.0000  14.9790

lowest :   0.000   0.040   0.085   0.125   0.165, highest:  25.125  26.335 100.000 130.000 200.000

Value          0     2     4     6     8    10    12    14    16    18    20    22    26   100   130   200
Frequency    125   107    65    31    26    38    40     7    10     2     5     1     3     1     1     1
Proportion 0.270 0.231 0.140 0.067 0.056 0.082 0.086 0.015 0.022 0.004 0.011 0.002 0.006 0.002 0.002 0.002

For the frequency table, variable is rounded to the nearest 2
---------------------------------------------------------------------------------------------------------------------------
V7.
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     463        0      108    0.998    2.733    3.959    0.000    0.000    0.165    1.000    2.500    5.500    8.500

lowest :   0.000   0.040   0.085   0.125   0.165, highest:  17.500  20.000  87.000 100.000 121.000
---------------------------------------------------------------------------------------------------------------------------
V10.
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     463        0       24    0.816    4.084    6.951        0        0        0        0        3        9       12

lowest :   0   1   2   3   4, highest:  20  67 100 230 400
---------------------------------------------------------------------------------------------------------------------------
V13.
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     463        0      131    0.991    207.9    232.7      0.0      0.0     66.0    150.0    252.5    380.8    439.4

lowest :    0   17   20   21   22, highest: 1160 2000 3400 5060 7680

Value          0   100   200   300   400   500   600   700   900  1200  2000  3400  5100  7700
Frequency    108   122   115    59    38    10     2     3     1     1     1     1     1     1
Proportion 0.233 0.263 0.248 0.127 0.082 0.022 0.004 0.006 0.002 0.002 0.002 0.002 0.002 0.002

For the frequency table, variable is rounded to the nearest 100
---------------------------------------------------------------------------------------------------------------------------
V14.
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     463        0      172    0.917    971.5     1751      1.0      1.0      1.0      4.0    358.5   2168.2   4001.0

lowest :     1     2     3     4     5, highest: 15109 26727 31286 50001 51101
---------------------------------------------------------------------------------------------------------------------------
```

Tabla 3. Estadística descriptiva de las variables numéricas con observaciones atípicas. Fuente: Elaboración propia.

Debido a que solamente se ha introducido 3 observaciones, se puede observar cómo los últimos elementos del apartado *highest* de cada variable ha sufrido un aumento considerable. Obviamente, otros valores como las medias y los cuartiles también han aumentado considerablemente.

```
Call:
glm(formula = V15. ~ V1. + V2. + V3. + V4. + V5. + V6. + V7. +
    V8. + V9. + V10. + V11. + V12. + V13. + V14., family = "binomial"
    data = dd1train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.94238  -0.28193  -0.08876   0.25240   2.94657

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.565e+00  1.487e+00  -4.416 1.01e-05 ***
V1.1         1.502e-02  4.171e-01   0.036  0.97127
V2.          1.056e-02  1.799e-02   0.587  0.55739
V3.         -3.899e-02  3.529e-02  -1.105  0.26918
V4.2         1.408e+00  4.629e-01   3.042  0.00235 **
V4.3         2.376e+01  2.923e+04   0.001  0.99935
V5.2         2.927e+00  2.591e+00   1.130  0.25863
V5.3         2.436e+00  2.548e+00   0.956  0.33907
V5.4         2.826e+00  2.404e+00   1.176  0.23976
V5.5        -1.831e+00  2.536e+00  -0.722  0.47028
V5.6         2.689e+00  2.502e+00   1.075  0.28254
V5.7         2.625e+00  2.590e+00   1.013  0.31085
V5.8         3.583e+00  2.459e+00   1.457  0.14502
V5.9         3.704e+00  2.541e+00   1.457  0.14501
V5.10        2.679e+00  3.301e+00   0.812  0.41707
V5.11        3.120e+00  2.476e+00   1.260  0.20755
V5.12        4.964e+00  6.079e+04   0.000  0.99993
V5.13        4.793e+00  2.604e+00   1.840  0.06570 .
V5.14        6.115e+00  2.636e+00   2.320  0.02036 *
V6.2        -2.095e-01  3.621e+00  -0.058  0.95385
V6.3         3.081e+00  2.447e+00   1.259  0.20801
V6.4        -1.141e+00  2.269e+00  -0.503  0.61515
V6.5        -1.768e+00  2.372e+00  -0.745  0.45605
V6.7        -2.149e+01  3.505e+04  -0.001  0.99951
V6.8        -1.125e+00  2.310e+00  -0.487  0.62625
V6.9        -1.199e+00  4.077e+00  -0.294  0.76860
V7.          1.250e-01  7.844e-02   1.593  0.11115
V8.1         3.981e+00  4.713e-01   8.447  < 2e-16 ***
V9.1        -2.132e-01  5.253e-01  -0.406  0.68484
V10.         2.679e-01  8.787e-02   3.049  0.00230 **
V11.1       -4.421e-01  3.796e-01  -1.165  0.24416
V12.2        6.633e-01  7.092e-01   0.935  0.34968
V12.3        1.943e+01  1.187e+03   0.016  0.98694
V13.        -4.679e-03  1.148e-03  -4.075 4.61e-05 ***
V14.         6.369e-04  3.124e-04   2.039  0.04148 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 635.77  on 462  degrees of freedom
Residual deviance: 221.80  on 428  degrees of freedom
AIC: 291.8
```

Ilustración 7. Summary del modelo simple introduciendo observaciones atípicas. Fuente: Elaboración propia.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 107  23
         1  19  81

               Accuracy : 0.8174
                 95% CI : (0.7613, 0.8651)
    No Information Rate : 0.5478
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6302

 Mcnemar's Test P-Value : 0.6434

            Sensitivity : 0.8492
            Specificity : 0.7788
         Pos Pred Value : 0.8231
         Neg Pred Value : 0.8100
             Prevalence : 0.5478
         Detection Rate : 0.4652
   Detection Prevalence : 0.5652
      Balanced Accuracy : 0.8140

       'Positive' Class : 0
```

Ilustración 8. Matriz de confusión del modelo simple introduciendo observaciones atípicas. Fuente: Elaboración propia.

En primer lugar, se puede apreciar cómo las variables cuyos coeficientes eran estadísticamente significativos antes lo son ahora también, aunque por ejemplo el coeficiente de la variable *V13* ha ganado significancia al presentar un p-valor asociado mucho menor. Ahora las predicciones correctas del modelo simple han disminuido hasta 81.74%, cuando sin las observaciones atípicas el modelo tenía una precisión del 82.17%. Se puede afirmar que los *outliers* afectan negativamente al modelo.

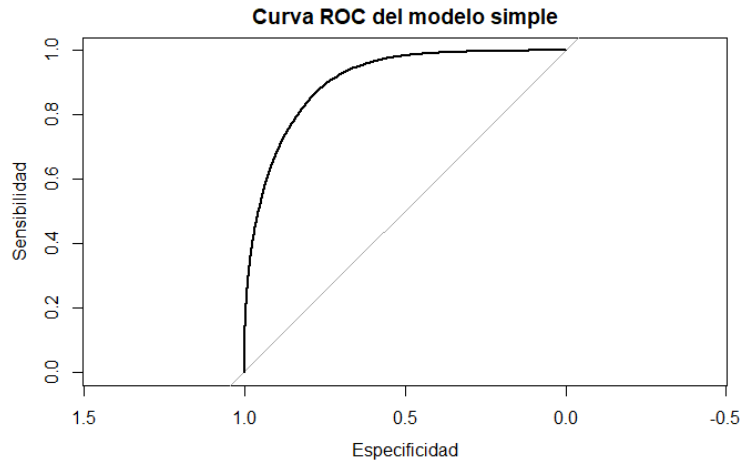

Gráfico 8. Curva ROC del modelo simple introduciendo observaciones atípicas. Fuente: Elaboración propia.

El valor de AUC con las observaciones atípicas es de 0.9634, no ha variado.

5.1.6. Regresión logística robusta introduciendo observaciones atípicas

```
           Length Class      Mode
a0            91  -none-     numeric
beta        3185  dgCMatrix  S4
df            91  -none-     numeric
dim            2  -none-     numeric
lambda        91  -none-     numeric
dev.ratio     91  -none-     numeric
nulldev        1  -none-     numeric
npasses        1  -none-     numeric
jerr           1  -none-     numeric
offset         1  -none-     logical
classnames     2  -none-     character
call           4  -none-     call
nobs           1  -none-     numeric
```

Ilustración 9. Summary del modelo robusto introduciendo observaciones atípicas. Fuente: Elaboración propia.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 105  13
         1  21  91

               Accuracy : 0.8522
                 95% CI : (0.7996, 0.8954)
    No Information Rate : 0.5478
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7036

 Mcnemar's Test P-Value : 0.2299

            Sensitivity : 0.8333
            Specificity : 0.8750
         Pos Pred Value : 0.8898
         Neg Pred Value : 0.8125
             Prevalence : 0.5478
         Detection Rate : 0.4565
   Detection Prevalence : 0.5130
      Balanced Accuracy : 0.8542

       'Positive' Class : 0
```

Ilustración 10. Matriz de confusión del modelo robusto introduciendo observaciones atípicas. Fuente: Elaboración propia.

El método robusto ofrece una precisión de más del 3 por ciento respecto al método clásico. La fiabilidad del modelo es muy buena dado que es capaz de predecir correctamente 5 de cada 6 predicciones.

5.1.7. Comparativa de los modelos

	RLS	*RLR*	*RLS con outliers*	*RLR con outliers*
Precisión	82.17%	84.74%	81.74%	85.22%
AIC	291.03	-	291.08	-

Tabla 4. Comparativa de los modelos de la base de datos 1. RLS es regresión logística simple, RLR es regresión logística robusta. Fuente: Elaboración propia.

En el recuadro anterior se puede apreciar las precisiones arrojadas por cada modelo según si contiene observaciones atípicas introducidas manualmente o no. En primer lugar, para los dos modelos sin *outliers* introducidos manualmente, la regresión logística simple ofrece de entrada una buena precisión, pero si se observa la extensión robusta, este último presenta una precisión mayor, llegando al 84.74%, siendo más del dos por ciento porcentuales. En segundo lugar, para los dos modelos con *outliers* introducidos manualmente, el modelo simple arroja una precisión menor, en concreto, la precisión se ha visto reducida en un 0.43 por ciento, es decir 1 sobre 230. En cambio, el modelo robusto con observaciones atípicas ofrece una precisión más elevada que el modelo simple. Curiosamente, parece ser que los *outliers* no han mermado la capacidad predictiva del modelo, sino que la han reforzado. De hecho, el modelo de regresión logística

robusta con presencia de 3 *outliers* es el modelo que ofrece la mejor tasa de acierto de todos cuatro los modelos construidos.

Es posible comparar los modelos sin la extensión robusta mediante el *Akaike Information Criterion* o AIC. Sea k el número de parámetros del modelo y L el valor máximo de la función de verosimilitud del modelo. El AIC se calcula de la siguiente forma

$$AIC = 2k - 2\ln(L)$$

Su interpretación es sencilla, a menor AIC mejor es el modelo.

Para este caso solamente se dispone el AIC de los modelos simples dado que la función *summary()* no ofrece dicha salida para las extensiones robustas. Al igual que la precisión, el modelo sin *outliers* ofrece un mejor ajuste. Aunque, nuevamente, la diferencia existente entre ambos modelos es mínima.

5.2.Base de datos 2

En este apartado, se construirá los dos modelos de regresión logística y de regresión logística robusta, respectivamente. A diferencia de la base de datos anterior, esta base de datos presenta más observaciones y se conoce el significado de las variables que lo conforman, por lo que una vez se obtenga los resultados deseados, se podrá realizar una interpretación económica, además de la estadística.

5.2.1. Imputación de los *missings* e incongruencias

Primeramente, se ha identificado las observaciones que contenían algún elemento codificado como un *missing*, pero estos elementos estaban codificados como " ?", tal como se muestra a continuación la tabla de frecuencia para la variable *WorkClass*.

?	Federal-gov	Local-gov	Never-worked	Private	Self-emp-inc	Self-emp-not-inc	State-gov	Without-pay
1836	960	2093	7	22696	1116	2541	1298	14

Ilustración 11. Ejemplo de una variable con missings. Fuente: Elaboración propia.

Posteriormente, se ha detectado algunas incongruencias en algunas variables. Una de ellas corresponde a la variable *HperWeek,* en el cual algunos individuos afirmaban trabajar 99 horas semanales, hecho que parece poco realista. Se ha optado por eliminar todas las observaciones con más de 80 horas semanales trabajadas.

Se ha optado por eliminar la variable denominada *FinalWeight* que fue creada a partir de un recuento en el cual se pondera las características socioeconómicas del individuo. El motivo es debido a la naturaleza de esta variable artificial, es considerada como variable de ponderación y no sería correcto introducirla al modelo directamente.

Finalmente, dado que se dispone de muchísimas observaciones, concretamente 32561, se ha optado por eliminar las que contenían uno o más *missings*, o que presentaban incongruencias en alguna de sus variables.

5.2.2. Estadística descriptiva

A continuación, se realizará una descriptiva para las variables más importantes, de forma que informe correctamente los rasgos o características de las observaciones que conforman esta base de datos.

Primero de todo se graficará la correlación para las variables numéricas, de forma que se averiguará qué variables varían al mismo tiempo.

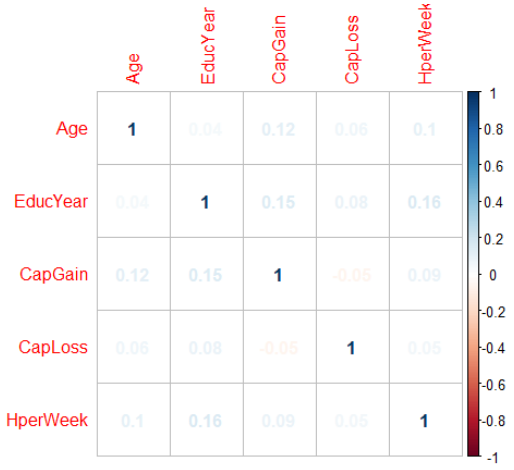

Gráfico 9. Correlaciones de las variables numéricas. Fuente: Elaboración propia.

A simple vista, las correlaciones de Pearson de las distintas variables son cercanas al 0. No hay ninguna correlación destacable.

Variable *Age*

```
Age
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   29819        0       72    0.999    38.38    14.85       20       22       28       37       47       57       62

lowest : 17 18 19 20 21, highest: 84 85 86 88 90
```

Tabla 5. Estadística descriptiva de Age. Fuente: Elaboración propia.

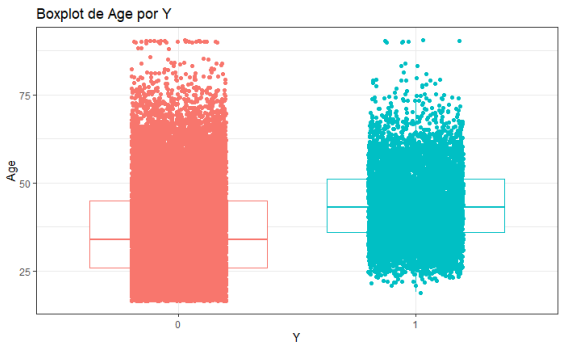

Gráfico 10. Boxplot de Age por Y. Fuente: Elaboración propia.

Gráfico 11. Histograma de Age. Fuente: Elaboración propia.

El rango de esta variable comienza en los 17 años y acaba en los 90 años, por lo que se tiene individuos de prácticamente todas las edades. La mediana toma valor de 37 años, que es muy parecida a la media, de 38.38 años. El boxplot parece indicar que, a más edad, más probable

que el individuo tenga una renta anual igual o más de 50.000 dólares. El histograma indica que casi la totalidad de los individuos tienen entre 20 y 50 años. Los individuos que tienen 65 años o más son una minoría.

Variable *EducYear*

```
EducYear
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   29819        0       16    0.949    10.11    2.712        6        7        9       10       12       13       14

lowest :  1  2  3  4  5, highest: 12 13 14 15 16

Value          1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16
Frequency     45   149   288   549   452   812  1044   373  9746  6631  1296  1003  4989  1600   487   355
Proportion 0.002 0.005 0.010 0.018 0.015 0.027 0.035 0.013 0.327 0.222 0.043 0.034 0.167 0.054 0.016 0.012
```

Tabla 6. Estadística descriptiva de EducYear. Fuente: Elaboración propia.

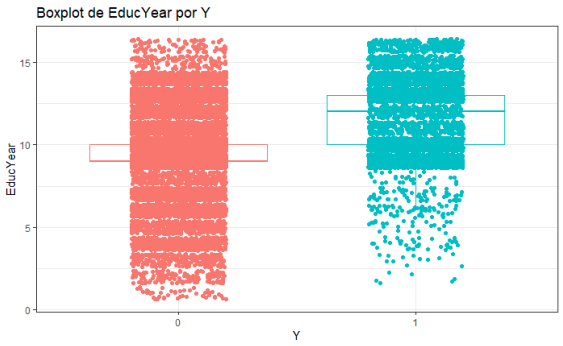

Gráfico 12. Boxplot de EducYear por Y. Fuente: Elaboración propia.

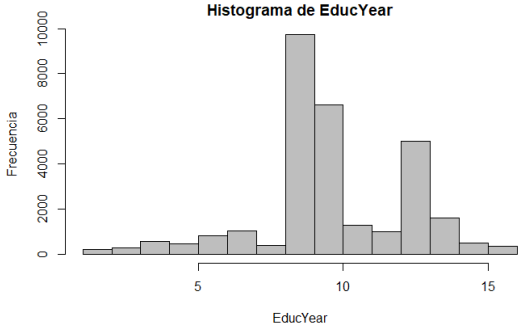

Gráfico 13. Histograma de EducYear. Fuente: Elaboración propia.

De nuevo hay mucha diversidad respecto a esta variable. Hay individuos que han estudiado solamente 1 año y otros que han llegado a estudiar un total de 16 años. Nuevamente la media y la mediana toman valores muy similares. Por el gráfico de boxplot, parece ser que los años de formación influye en la renta anual. En este caso, contra mayor sea esta variable, más probable es que el individuo tenga una renta anual igual o superior a los 50.000 dólares.

Variable *HperWeek*

```
HperWeek
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   29819        0       77     0.89    40.55     11.3       20       25       40       40       45       55       60

lowest :  1  2  3  4  5, highest: 75 76 77 78 80
```

Tabla 7. Estadística descriptiva de HperWeek. Fuente: Elaboración propia.

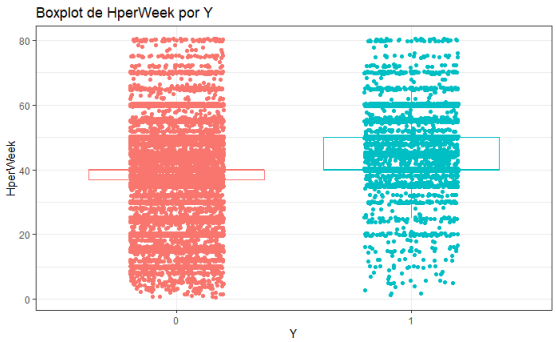

Gráfico 14. Boxplot de HperWeek por Y. Fuente: Elaboración propia.

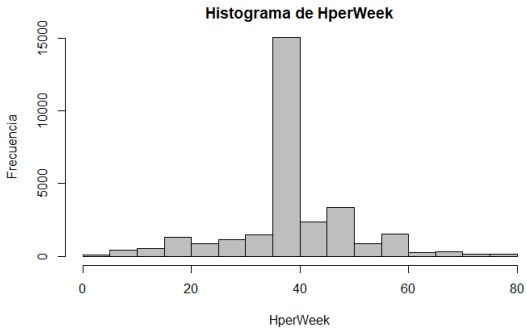

Gráfico 15. Histograma de HperWeek. Fuente: Elaboración propia.

En cuanto a las horas semanales trabajadas, la media y mediana toman el valor de 40. El histograma confirma que lo dicho anteriormente. Esta variable, a priori, no parece ser relevante para determinar si el individuo tiene una renta anual de 50.000 o más dólares. Hay personas que solamente trabajan 1 hora a la semana, probablemente algún trabajo ocasional. En cambio, existe algunos individuos los cuales trabajan 80 horas semanales.

Variable *Educ*

```
Educ
       n  missing distinct
   29819        0       16

lowest :  10th         11th         12th         1st-4th      5th-6th      , highest:  HS-grad      Masters      Preschool    Prof-school  Some-college

Value             10th         11th         12th         1st-4th      5th-6th      7th-8th          9th   Assoc-acdm   Assoc-voc    Bachelors     Doctorate       HS-grad
Frequency          812         1044          373          149          288          549          452         1003         1296         4989          355         9746
Proportion       0.027        0.035        0.013        0.005        0.010        0.018        0.015        0.034        0.043        0.167        0.012        0.327

Value          Masters    Preschool  Prof-school Some-college
Frequency         1600           45          487         6631
Proportion       0.054        0.002        0.016        0.222
```

Tabla 8. Estadística descriptiva de Educ. Fuente: Elaboración propia.

Cerca del 23% de los individuos tienen estudios universitarios. El 16.7% tienen la formación equivalente a licenciatura o grado, el 5.4% poseen una maestría, y solamente el 1.2% de los encuestados son doctores. El 32.7% tienen el denominado *high school*, que actualmente en España sería el equivalente a bachillerato. El último grupo significativo tienen el graduado escolar y representan el 22.2% del total.

Variable *MaritalStatus*

```
MaritalStatus
       n  missing distinct
   29819        0        7

lowest :  Divorced              Married-AF-spouse     Married-civ-spouse    Married-spouse-absent  Never-married
highest:  Married-civ-spouse    Married-spouse-absent Never-married         Separated              Widowed

Value                 Divorced     Married-AF-spouse    Married-civ-spouse Married-spouse-absent        Never-married            Separated              Widowed
Frequency                 4181                   20                13813                   366                 9682                  933                   824
Proportion               0.140                0.001                0.463                 0.012                0.325                0.031                 0.028
```

Tabla 9. Estadística descriptiva de MaritalStatus. Fuente Elaboración propia.

En cuanto al estado civil de los encuestados, el 32.5% no han estado casados nunca mientras que un 46.3% sí que tienen cónyuges. El 20% restante lo forman los divorciados, los que están separados, los viudos, entre otros.

Variable *NatCountry*

Cambodia	Canada	China	Columbia	Cuba	Dominican-Republic
18	106	68	56	92	65
Ecuador	El-Salvador	England	France	Germany	Greece
27	100	84	26	128	29
Guatemala	Haiti	Holand-Netherlands	Honduras	Hong	Hungary
63	42	1	12	19	13
India	Iran	Ireland	Italy	Jamaica	Japan
96	42	24	68	80	56
Laos	Mexico	Nicaragua	Outlying-US(Guam-USVI-etc)	Peru	Philippines
17	605	33	14	30	186
Poland	Portugal	Puerto-Rico	Scotland	South	Taiwan
56	33	109	11	70	40
Thailand	Trinadad&Tobago	United-States	Vietnam	Yugoslavia	
17	18	27185	64	16	

Tabla 10. Tabla de frecuencia de NatCountry. Fuente: Elaboración propia.

Aunque se dispone de personas de nacionalidades muy diversas y de diferentes continentes, prácticamente la totalidad de los encuestados son estadounidenses de nacimiento.

Variable *Race*

Race

n	missing	distinct
29819	0	5

lowest :	Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White
highest:	Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White

Value	Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White
Frequency	283	881	2800	226	25629
Proportion	0.009	0.030	0.094	0.008	0.859

Tabla 11. Estadística descriptiva de Race. Fuente: Elaboración propia.

La raza blanca es claramente la más dominante, seguida de la raza negra.

Variable *Sex*

Sex

n	missing	distinct
29819	0	2

Value	Female	Male
Frequency	9736	20083
Proportion	0.327	0.673

Tabla 12. Estadística descriptiva de Sex. Fuente: Elaboración propia.

Hay una clara dominancia de los hombres dado que estos representan dos terceras partes de las observaciones totales.

Variable *WorkClass*

```
WorkClass
       n  missing distinct
   29819        0        7

lowest :  Federal-gov       Local-gov         Private           Self-emp-inc      Self-emp-not-inc
highest:  Private           Self-emp-inc      Self-emp-not-inc  State-gov         Without-pay

Value          Federal-gov       Local-gov         Private     Self-emp-inc Self-emp-not-inc        State-gov      Without-pay
Frequency              941            2053           22106             1023             2408             1274               14
Proportion           0.032           0.069           0.741            0.034            0.081            0.043            0.000
```

Tabla 13. Estadística descriptiva de WorkClass. Fuente: Elaboración propia.

Tres cuartas partes del total de las observaciones son trabajadores del sector privado. Alrededor del 15% son funcionarios, ya sea a nivel local, federal, o estatal. El porcentaje restante lo conforma los que trabajan por cuenta propia, es decir, los autónomos, de los cuales el 8% no obtienen ningún ingreso por esta vía.

5.2.3. Regresión logística simple

```
Call:
glm(formula = Y ~ Age + WorkClass + Educ + EducYear + MaritalStatus +
    Occup + Relationship + Race + Sex + CapGain + CapLoss + HperWeek +
    NatCountry, family = "binomial", data = dd2train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.1229  -0.5147  -0.1896  -0.0190   3.7725

Coefficients: (1 not defined because of singularities)
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -6.523e+00  7.660e-01  -8.516  < 2e-16 ***
Age                              2.613e-02  1.727e-03  15.128  < 2e-16 ***
WorkClass Local-gov             -7.016e-01  1.133e-01  -6.192 5.93e-10 ***
WorkClass Private               -4.975e-01  9.401e-02  -5.292 1.21e-07 ***
WorkClass Self-emp-inc          -3.426e-01  1.246e-01  -2.750  0.00596 **
WorkClass Self-emp-not-inc      -9.739e-01  1.105e-01  -8.813  < 2e-16 ***
WorkClass State-gov             -8.168e-01  1.259e-01  -6.486 8.83e-11 ***
WorkClass Without-pay           -1.328e+01  1.986e+02  -0.067  0.94668
Educ 11th                        9.336e-02  2.150e-01   0.434  0.66417
Educ 12th                        4.346e-01  2.825e-01   1.538  0.12398
Educ 1st-4th                    -4.086e-01  4.979e-01  -0.821  0.41183
Educ 5th-6th                    -4.315e-01  3.605e-01  -1.197  0.23133
Educ 7th-8th                    -5.266e-01  2.453e-01  -2.146  0.03184 *
Educ 9th                        -2.106e-01  2.707e-01  -0.778  0.43659
Educ Assoc-acdm                  1.276e+00  1.811e-01   7.047 1.83e-12 ***
Educ Assoc-voc                   1.267e+00  1.742e-01   7.273 3.53e-13 ***
Educ Bachelors                   1.889e+00  1.622e-01  11.652  < 2e-16 ***
Educ Doctorate                   2.971e+00  2.258e-01  13.158  < 2e-16 ***
Educ HS-grad                     7.708e-01  1.576e-01   4.889 1.01e-06 ***
Educ Masters                     2.254e+00  1.734e-01  13.003  < 2e-16 ***
Educ Preschool                  -1.875e+01  1.013e+02  -0.185  0.85316
Educ Prof-school                 2.859e+00  2.093e-01  13.660  < 2e-16 ***
Educ Some-college                1.103e+00  1.600e-01   6.894 5.43e-12 ***
EducYear                                NA         NA      NA       NA
MaritalStatus Married-AF-spouse  2.931e+00  5.929e-01   4.943 7.68e-07 ***
MaritalStatus Married-civ-spouse 2.100e+00  2.772e-01   7.576 3.57e-14 ***
MaritalStatus Married-spouse-absent 3.530e-02 2.425e-01  0.146  0.88425
MaritalStatus Never-married     -4.567e-01  8.971e-02  -5.091 3.56e-07 ***
MaritalStatus Separated         -4.011e-02  1.657e-01  -0.242  0.80873
MaritalStatus Widowed            2.299e-01  1.584e-01   1.451  0.14686
Occup Armed-Forces              -1.098e+00  1.511e+00  -0.727  0.46738
Occup Craft-repair               5.400e-02  8.100e-02   0.667  0.50497
Occup Exec-managerial            7.888e-01  7.817e-02  10.090  < 2e-16 ***
Occup Farming-fishing           -9.831e-01  1.433e-01  -6.858 6.97e-12 ***
Occup Handlers-cleaners         -6.979e-01  1.450e-01  -4.814 1.48e-06 ***
Occup Machine-op-inspct         -2.776e-01  1.031e-01  -2.692  0.00710 **
Occup Other-service             -8.396e-01  1.203e-01  -6.982 2.91e-12 ***
Occup Priv-house-serv           -3.848e+00  1.961e+00  -1.963  0.04967 *
Occup Prof-specialty             5.209e-01  8.275e-02   6.296 3.06e-10 ***
Occup Protective-serv            6.476e-01  1.271e-01   5.094 3.50e-07 ***
Occup Sales                      2.730e-01  8.348e-02   3.270  0.00107 **
```

Ilustración 12. Summary del modelo simple, parte I. Fuente: Elaboración propia.

```
Occup Sales                      2.730e-01  8.348e-02   3.270  0.00107 **
Occup Tech-support               6.782e-01  1.121e-01   6.049 1.45e-09 ***
Occup Transport-moving          -9.938e-02  1.007e-01  -0.987  0.32368
Relationship Not-in-family       4.300e-01  2.742e-01   1.568  0.11680
Relationship Other-relative     -3.989e-01  2.495e-01  -1.598  0.10996
Relationship Own-child          -7.174e-01  2.727e-01  -2.630  0.00853 **
Relationship Unmarried           3.039e-01  2.899e-01   1.048  0.29453
Relationship Wife                1.376e+00  1.063e-01  12.949  < 2e-16 ***
Race Asian-Pac-Islander          8.436e-01  2.876e-01   2.933  0.00335 **
Race Black                       5.400e-01  2.406e-01   2.244  0.02482 *
Race Other                       2.717e-01  3.782e-01   0.718  0.47248
Race White                       6.450e-01  2.295e-01   2.810  0.00495 **
Sex Male                         8.593e-01  8.127e-02  10.574  < 2e-16 ***
CapGain                          3.233e-04  1.081e-05  29.909  < 2e-16 ***
CapLoss                          6.371e-04  3.868e-05  16.469  < 2e-16 ***
HperWeek                         3.519e-02  1.897e-03  18.547  < 2e-16 ***
NatCountry Canada               -8.718e-01  6.895e-01  -1.264  0.20607
NatCountry China                -1.930e+00  7.031e-01  -2.745  0.00604 **
NatCountry Columbia             -3.355e+00  1.037e+00  -3.235  0.00122 **
NatCountry Cuba                 -7.832e-01  7.026e-01  -1.115  0.26500
NatCountry Dominican-Republic   -2.881e+00  1.222e+00  -2.358  0.01836 *
NatCountry Ecuador              -1.461e+00  9.566e-01  -1.528  0.12662
NatCountry El-Salvador          -1.767e+00  7.951e-01  -2.223  0.02625 *
NatCountry England              -9.273e-01  7.036e-01  -1.318  0.18754
NatCountry France               -6.513e-01  8.144e-01  -0.800  0.42389
NatCountry Germany              -7.421e-01  6.783e-01  -1.094  0.27396
NatCountry Greece               -2.247e+00  8.377e-01  -2.682  0.00731 **
NatCountry Guatemala            -1.392e+00  9.762e-01  -1.426  0.15395
NatCountry Haiti                -1.247e+00  9.295e-01  -1.342  0.17960
NatCountry Honduras             -2.445e+00  2.803e+00  -0.872  0.38299
NatCountry Hong                 -1.349e+00  9.012e-01  -1.497  0.13429
NatCountry Hungary              -1.325e+00  9.924e-01  -1.336  0.18169
NatCountry India                -1.709e+00  6.689e-01  -2.555  0.01062 *
NatCountry Iran                 -1.214e+00  7.580e-01  -1.602  0.10916
NatCountry Ireland              -7.143e-01  8.895e-01  -0.803  0.42199
NatCountry Italy                -4.002e-01  7.099e-01  -0.564  0.57292
NatCountry Jamaica              -1.210e+00  7.711e-01  -1.569  0.11665
NatCountry Japan                -1.117e+00  7.364e-01  -1.517  0.12916
NatCountry Laos                 -1.846e+00  1.042e+00  -1.771  0.07651 .
NatCountry Mexico               -1.596e+00  6.644e-01  -2.402  0.01632 *
NatCountry Nicaragua            -1.782e+00  1.014e+00  -1.758  0.07874 .
NatCountry Outlying-US(Guam-USVI-etc) -1.343e+01 2.110e+02 -0.064  0.94927
NatCountry Peru                 -1.950e+00  1.057e+00  -1.845  0.06498 .
NatCountry Philippines          -8.551e-01  6.436e-01  -1.329  0.18394
NatCountry Poland               -1.197e+00  7.447e-01  -1.607  0.10808
NatCountry Portugal             -1.108e+00  8.895e-01  -1.246  0.21290
NatCountry Puerto-Rico          -1.501e+00  7.380e-01  -2.034  0.04193 *
NatCountry Scotland             -1.490e+00  1.089e+00  -1.369  0.17111
NatCountry South                -2.361e+00  7.332e-01  -3.220  0.00128 **
NatCountry Taiwan               -1.468e+00  7.622e-01  -1.926  0.05416 .
```

Ilustración 13. Summary del modelo simple, parte II. Fuente: Elaboración propia.

```
NatCountry Taiwan                  -1.468e+00  7.622e-01  -1.926  0.05416 .
NatCountry Thailand                -1.981e+00  1.021e+00  -1.941  0.05229 .
NatCountry Trinadad&Tobago         -1.659e+00  1.059e+00  -1.566  0.11745
NatCountry United-States           -1.015e+00  6.302e-01  -1.611  0.10727
NatCountry Vietnam                 -2.451e+00  8.488e-01  -2.887  0.00389 **
NatCountry Yugoslavia              -5.237e-01  9.215e-01  -0.568  0.56982
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33187  on 29817  degrees of freedom
Residual deviance: 19264  on 29724  degrees of freedom
AIC: 19452

Number of Fisher Scoring iterations: 13
```

Ilustración 14. Summary del modelo simple, parte III. Fuente: Elaboración propia.

Hay variables como *Age* que sí tienen un coeficiente que es estadísticamente significativo y otras como podría ser la variable *NatCountry* en el cual una gran parte de sus categorías no lo son, hecho razonable si se tiene en cuenta que prácticamente la totalidad de los individuos son estadounidenses de nacimiento.

Antes del *intercept* se puede observar "*Coefficients: (1 not defined because of singularities)*". Luego de intentar construir el modelo utilizando diferentes combinaciones de las variables disponibles, se ha llegado a la conclusión que este error se produce cuando se utilizan al mismo tiempo *Educ* y *EducYear*. Cabe señalar que se puede visualizar el coeficiente de todas las combinaciones de diferentes modelos excepto de la combinación anteriormente mencionada. Aun así, se ha optado por no eliminar *EducYear* dado que en los procedimientos posteriores no ofrece ningún problema.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 11126   169
         1  2452  1133

               Accuracy : 0.8239
                 95% CI : (0.8176, 0.8299)
    No Information Rate : 0.9125
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3847

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8194
            Specificity : 0.8702
         Pos Pred Value : 0.9850
         Neg Pred Value : 0.3160
             Prevalence : 0.9125
         Detection Rate : 0.7477
   Detection Prevalence : 0.7591
      Balanced Accuracy : 0.8448

       'Positive' Class : 0
```

Ilustración 15. Matriz de confusión del modelo simple. Fuente: Elaboración propia.

De entrada, se puede ver cómo las predicciones son correctas en un 82.39%, que es bueno. Es interesante observar cómo el modelo solamente cuenta con 160 casos de falsos negativos. La

gran parte de los errores provienen de la parte de los falsos positivos, que han sido 2442 casos en total.

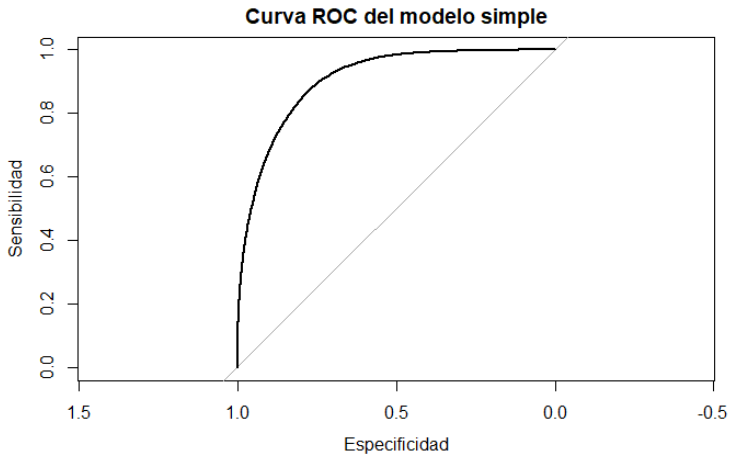

Gráfico 16. Curva ROC del modelo simple. Fuente: Elaboración propia.

El valor del AUC es de 0.9061, cosa que indica que el modelo es muy bueno.

5.2.4. Regresión logística robusta

	Length	Class	Mode
a0	79	-none-	numeric
beta	7505	dgCMatrix	S4
df	79	-none-	numeric
dim	2	-none-	numeric
lambda	79	-none-	numeric
dev.ratio	79	-none-	numeric
nulldev	1	-none-	numeric
npasses	1	-none-	numeric
jerr	1	-none-	numeric
offset	1	-none-	logical
classnames	2	-none-	character
call	4	-none-	call
nobs	1	-none-	numeric

Ilustración 16. Summary del modelo robusto. Fuente: Elaboración propia.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 11135   160
         1  2467  1118

               Accuracy : 0.8235
                 95% CI : (0.8172, 0.8296)
    No Information Rate : 0.9141
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3815

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8186
            Specificity : 0.8748
         Pos Pred Value : 0.9858
         Neg Pred Value : 0.3119
             Prevalence : 0.9141
         Detection Rate : 0.7483
   Detection Prevalence : 0.7591
      Balanced Accuracy : 0.8467

       'Positive' Class : 0
```

Ilustración 17. Matriz de confusión del modelo robusto. Fuente: Elaboración propia.

Para la extensión robusta, la precisión ha sido del 82.35%, que es muy similar al de la regresión logística simple. El número de casos de verdaderos negativos, verdaderos positivos, falso negativo y falsos positivos son muy similares con los que arroja la aproximación clásica. Por

esta razón, la precisión y valores como la sensibilidad y sensibilidad se asemejan tanto entre ambos modelos.

5.2.5. Comparativa de los modelos

	RLS	*RLR*
Precisión	82.39%	82.35%

Tabla 14. Comparativa de los modelos de la base de datos 2. Fuente: Elaboración propia.

En este apartado no se ha añadido ninguna observación de tipo *outlier*. Mediante las técnicas de estadística descriptiva tampoco se ha detectado alguna observación que cuente con valores anormalmente diferentes al resto. Dadas estas circunstancias, los resultados que ofrecen ambos modelos son prácticamente idénticos. Se puede afirmar que para este caso no existe diferencia alguna entre los resultados de los dos modelos.

6. Conclusión

Mediante la primera base de datos se ha podido comprobar cómo las observaciones atípicas afectaban de forma negativa a las predicciones de la regresión logística simple. Es decir, cómo unas pocas observaciones pueden tener un impacto considerable en el modelo y, en consecuencia, la fiabilidad de sus predicciones. Ocurre lo contrario para la extensión robusta: las observaciones con valores anormalmente diferentes al resto de observaciones no varían notablemente las predicciones de este modelo. De hecho, refuerzan estas predicciones dotándoles de una menor tasa de error en la matriz de confusión asociada.

Con la segunda base de datos se ha visto cómo ambos modelos, tanto la regresión logística simple como la regresión logística robusta presentan unas predicciones casi idénticas entre sí. Dada la fiabilidad, practicidad y consistencia de la extensión robusta, quizás sería un buen sustituto respecto al modelo inicial. En este caso, ambos modelos son, en la práctica, igual de fiables para predecir si las rentas del individuo en cuestión superan o no los 50.000 dólares anuales. Sabiendo los fundamentos de ambas aproximaciones y guiándose por las evidencias empíricas obtenidas a lo largo de este trabajo, se puede afirmar que la extensión robusta ofrece, con o sin observaciones atípicas, un rendimiento igual e incluso superior en algunos casos respecto al modelo clásico propuesto por Cox en 1958.

Vista la utilidad del método robusto, quizás sería conveniente que en los grados de estadística y economía además de explicar las aproximaciones clásicas, realicen una breve introducción al alumnado en el campo de la estadística robusta, campo que en las últimas décadas despierta cada vez más interés.

7. Bibliografía

Bootkrajang, J. (2016). A generalised label noise model for classification in the presence of annotation errors. *Neurocomputing*, 192, 61-71.

Brodley, C. E., & Friedl, M. A. (1999). Identifying and Eliminatinf Mislabeled Training Instances. In Proceedings of the National Conference on Artificial Intelligence (pp. 799-805).

Cox, D. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2). 215-232.

Dunn, P. K., & Smyth, G. K. (2018). *Generalized Linear Models With Examples in R.* Springer.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., & Qian, J. (acceso el 27 de Junio de 2020). *CRAN R Project, "glmnet" versión (4.0-2).* Obtenido de https://cran.r-project.org/web/packages/glmnet/glmnet.pdf

Huber, P. J., & Ronchetti, E. M. (2009). *Robust Statistics.* Wiley.

Kohavi, R., & Becker, B. *Machine Learning Repostory of University of California, Irvine*. Obtenido de https://archive.ics.uci.edu/ml/datasets/Adult (último acceso el 31 de Agosto de 2020).

Le, J. *DataCamp.* Obtenido de https://www.datacamp.com/community/tutorials/logistic-regression-R (último acceso el 31 de Agosto de 2020).

Morales, P., Luengo, J., Garcia, L. P., Lorena, A. C., Carvalho, A. C., & Herrera, F. (2017). The NoiseFiltersR Package: Label Noise Preprocessing in R. *Journal of R-Project*, 9(1), 219.-

Prabhakaran, S. *Machine Learning Plus.* Obtenido de https://www.machinelearningplus.com/machine-learning/logistic-regression-tutorial-examples-r/ (último acceso el 31 de Agosto de 2020).

Quinlan, R. *Machine Learning Repostory of University of California, Irvine*. Obtenido de https://archive.ics.uci.edu/ml/datasets/Statlog+%28Australian+Credit+Approval%29 (último acceso el 31 de Agosto de 2020).

She, Y., & Owen, A. B. (2011). Outlier Detection Using Nonconvex Penalized. *Journal of the American Statistical Association*, 106(494), 626-639.-

Tibshirani, J., & Manning, C. (2011). Robust Logistic Regression using Shift Parameters. *arXiv prepint arXiv:1305.4987*

8. Anexo

```
# Base de datos 1

library(caret)

library(corrplot)

library(ggplot2)

library(Hmisc)

library(glmnet)

library(gridExtra)

library(grid)

library(Hmisc)

library(plyr)

library(pROC)

set.seed(2020)

dd1 <- read.table("C:/Users/jindu/Desktop/TFG/ACC/australian.dat")

for (i in c(1, 4, 5, 6, 8, 9, 11, 12, 15)) {

  dd1[, i] <- as.factor(dd1[, i])

}

names(dd1) <- paste0("V", 1:15, ".")

dd1train <- dd1[muestra <- sample(1:nrow(dd1), 2 / 3 * nrow(dd1), replace = FALSE), ]

dd1test <- dd1[-muestra, ]

corrplot(cor(dd1train[, c(2, 3, 7, 10, 13, 14)]), method = "number")

describe(dd1train[, c(2, 3, 7, 10, 13, 14)])

grid.arrange(ggplot(data = dd1train, aes(x = V15., y = dd1train[, 2], color = V15.)) + xlab("V15") +
ylab("V2") + ggtitle("Boxplot V2 por V15") + geom_boxplot(outlier.shape = NA) + geom_jitter(width =
0.2) + theme_bw() + theme(legend.position = "null"), ggplot(data = dd1train, aes(x = dd1train[, 2],
color = V15., fill = V15.)) + geom_histogram(binwidth = 1) + geom_vline(data = aggregate(x =
dd1train[, 2], by = list(dd1train$V15.), FUN = mean), aes(xintercept = x, color = Group.1), linetype =
"dashed") + xlab("V2") + ylab("Frecuencia") + ggtitle("Histograma V2 por V15") + theme_bw() +
theme(legend.position = "null"), ggplot(data = dd1train, aes(x = V15., y = dd1train[, 3], color = V15.))
+ xlab("V15") + ylab("V3") + ggtitle("Boxplot V3 por V15") + geom_boxplot(outlier.shape = NA) +
geom_jitter(width = 0.2) + theme_bw() + theme(legend.position = "null"), ggplot(data = dd1train,
aes(x = dd1train[, 3], color = V15., fill = V15.)) + geom_histogram(binwidth = 1) + geom_vline(data =
aggregate(x = dd1train[, 3], by = list(dd1train$V15.), FUN = mean), aes(xintercept = x, color =
Group.1), linetype = "dashed") + xlab("V3") + ylab("Frecuencia") + ggtitle("Histograma V3 por V15") +
theme_bw() + theme(legend.position = "null"), ncol = 2)
```

```
grid.arrange(ggplot(data = dd1train, aes(x = V15., y = dd1train[, 7], color = V15.)) + xlab("V15") +
ylab("V7") + ggtitle("Boxplot V7 por V15") + geom_boxplot(outlier.shape = NA) + geom_jitter(width =
0.2) + theme_bw() + theme(legend.position = "null"), ggplot(data = dd1train, aes(x = dd1train[, 7],
color = V15., fill = V15.)) + geom_histogram(binwidth = 1) + geom_vline(data = aggregate(x =
dd1train[, 7], by = list(dd1train$V15.), FUN = mean), aes(xintercept = x, color = Group.1), linetype =
"dashed") + xlab("V7") + ylab("Frecuencia") + ggtitle("Histograma V7 por V15") + theme_bw() +
theme(legend.position = "null"), ggplot(data = dd1train, aes(x = V15., y = dd1train[, 10], color = V15.))
+ xlab("V15") + ylab("V10") + ggtitle("Boxplot V10 por V15") + geom_boxplot(outlier.shape = NA) +
geom_jitter(width = 0.2) + theme_bw() + theme(legend.position = "null"), ggplot(data = dd1train,
aes(x = dd1train[, 10], color = V15., fill = V15.)) + geom_histogram(binwidth = 1) + geom_vline(data =
aggregate(x = dd1train[, 10], by = list(dd1train$V15.), FUN = mean), aes(xintercept = x, color =
Group.1), linetype = "dashed") + xlab("V10") + ylab("Frecuencia") + ggtitle("Histograma V10 por
V15") + theme_bw() + theme(legend.position = "null"), ncol = 2)

grid.arrange(ggplot(data = dd1train, aes(x = V15., y = dd1train[, 13], color = V15.)) + xlab("V15") +
ylab("V13") + ggtitle("Boxplot V13 por V15") + geom_boxplot(outlier.shape = NA) +
geom_jitter(width = 0.2) + theme_bw() + theme(legend.position = "null"), ggplot(data = dd1train,
aes(x = dd1train[, 13], color = V15., fill = V15.)) + geom_histogram(binwidth = 1) + geom_vline(data =
aggregate(x = dd1train[, 13], by = list(dd1train$V15.), FUN = mean), aes(xintercept = x, color =
Group.1), linetype = "dashed") + xlab("V13") + ylab("Frecuencia") + ggtitle("Histograma V13 por
V15") + theme_bw() + theme(legend.position = "null"), ggplot(data = dd1train, aes(x = V15., y =
dd1train[, 14], color = V15.)) + xlab("V15") + ylab("V14") + ggtitle("Boxplot V14 por V15") +
geom_boxplot(outlier.shape = NA) + geom_jitter(width = 0.2) + theme_bw() + theme(legend.position
= "null"), ggplot(data = dd1train, aes(x = dd1train[, 10], color = V15., fill = V15.)) +
geom_histogram(binwidth = 1) + geom_vline(data = aggregate(x = dd1train[, 14], by =
list(dd1train$V15.), FUN = mean), aes(xintercept = x, color = Group.1), linetype = "dashed") +
xlab("V14") + ylab("Frecuencia") + ggtitle("Histograma V14 por V15") + theme_bw() +
theme(legend.position = "null"), ncol = 2)

describe(dd1train[, c(1, 4, 5, 6, 8)])

describe(dd1train[, c(9, 11, 12, 15)])

grid.arrange(ggplot(data = dd1train, aes(x="", y = V1., fill = V1.)) + geom_bar(stat="identity", width =
1) + coord_polar("y", start = 0) + theme_void() + ggtitle("Grafico circular de V1"), ggplot(data =
dd1train, aes(x="", y = V4., fill = V4.)) + geom_bar(stat="identity", width = 1) + coord_polar("y", start
= 0) + theme_void() + scale_fill_brewer("Blues") + ggtitle("Grafico circular de V4"), ggplot(data =
dd1train, aes(x="", y = V8., fill = V8.)) + geom_bar(stat="identity", width = 1) + coord_polar("y", start
= 0) + theme_void() + ggtitle("Grafico circular de V8"), ggplot(data = dd1train, aes(x="", y = V9., fill =
V9.)) + geom_bar(stat="identity", width = 1) + coord_polar("y", start = 0) + theme_void() +
ggtitle("Grafico circular de V9"), ggplot(data = dd1train, aes(x="", y = V11., fill = V11.)) +
geom_bar(stat="identity", width = 1) + coord_polar("y", start = 0) + theme_void() + ggtitle("Grafico
circular de V11"), ggplot(data = dd1train, aes(x="", y = V12., fill = V12.)) + geom_bar(stat="identity",
width = 1) + coord_polar("y", start = 0) + theme_void() + scale_fill_brewer("Blues") + ggtitle("Grafico
circular de V12"), ggplot(data = dd1train, aes(x="", y = V15., fill = V15.)) + geom_bar(stat="identity",
width = 1) + coord_polar("y", start = 0) + theme_void() + ggtitle("Grafico circular de V15"), ncol = 2)

grid.arrange(ggplot(data = dd1train, aes(x = V5., fill = V5.)) + geom_bar() + xlab("V5") +
ylab("Frecuencia") + ggtitle("Histograma de V5") + theme_bw() + theme(legend.position = "null"),
```

```
ggplot(data = dd1train, aes(x = V6., fill = V6.)) + geom_bar() + xlab("V6") + ylab("Frecuencia") + ggtitle("Histograma de V6") + theme_bw() + theme(legend.position = "null"), ncol = 2)

dd1RLS1 <- glm(V15. ~ V1. + V2. + V3. + V4. + V5. + V6. + V7. + V8. + V9. + V10. + V11. + V12. + V13. + V14., family = "binomial", data = dd1train)

summary(dd1RLS1)

threshold1 <- table(dd1train$V15.)[1] / sum(table(dd1train$V15.))

predRLS1 <- factor(ifelse(predict(dd1RLS1, newdata = dd1test, type = "response") > threshold1, 1, 0), levels = c(0, 1))

confusionMatrix(dd1test$V15., predRLS1)

ROC1 <- roc(dd1train[, 15] ~ predict(dd1RLS1, type = c("response")))

plot(ROC1, main = "Curva ROC del modelo simple sin observaciones atípicas", xlab = "Especificidad", ylab = "Sensibilidad")

auc(ROC1)

dd1RLR1 <- glmnet(x = model.matrix( ~ .-1, dd1train[, 1:14]), y = as.matrix(dd1train[, 15]), family = "binomial")

summary(dd1RLR1)

lambda1 <- vector(length = length(dd1RLR1$lambda))

for(i in 1:length(dd1RLR1$lambda)){

  predRLR1 <- factor(ifelse(predict(dd1RLR1, newx = model.matrix( ~ .-1, dd1test[, 1:14]), type = "response", s = dd1RLR1$lambda[i]) > threshold1, 1, 0), levels = c(0, 1))

  lambda1[i] <- confusionMatrix(predRLR1, dd1test$V15.)$overall[1]

}

predRLR1 <- factor(ifelse(predict(dd1RLR1, newx = model.matrix( ~ .-1, dd1test[, 1:14]), type = "response", s = dd1RLR1$lambda[which.max(lambda1)]) > threshold1, 1, 0))

confusionMatrix(dd1test$V15., predRLR1)

obAtip1 <- c(0, 150, 200, 1, 11, 9, 121, 0, 1, 100, 0, 1, 7680, 3000, 0)

obAtip2 <- c(1, 168, 100, 2, 9, 2, 100, 1, 1, 230, 1, 3, 5060, 3000, 1)

obAtip3 <- c(1, 120, 130, 1, 4, 7, 87, 0, 0, 400, 1, 3, 3400, 4000, 1)

dd1train <- rbind(dd1train, obAtip1, obAtip2, obAtip3)

describe(dd1train[, c(2, 3, 7, 10, 13, 14)])

dd1RLS2 <- glm(V15. ~ V1. + V2. + V3. + V4. + V5. + V6. + V7. + V8. + V9. + V10. + V11. + V12. + V13. + V14., family = "binomial", data = dd1train)

summary(dd1RLS2)

threshold2 <- table(dd1train$V15.)[1] / sum(table(dd1train$V15.))
```

```
predRLS2 <- factor(ifelse(predict(dd1RLS2, newdata = dd1test, type = "response") > threshold2, 1, 0), levels = c(0, 1))

confusionMatrix(dd1test$V15., predRLS2)

ROC2 <- roc(dd1train[, 15] ~ predict(dd1RLS2, type = c("response")))

plot(ROC2, main = "Curva ROC del modelo simple con observaciones atípicas", xlab = "Especificidad", ylab = "Sensibilidad")

auc(ROC2)

dd1RLR2 <- glmnet(x = model.matrix( ~ .-1, dd1train[, 1:14]), y = as.matrix(dd1train[, 15]), family = "binomial")

summary(dd1RLR2)

lambda2 <- vector(length = length(dd1RLR2$lambda))

for(i in 1:length(dd1RLR2$lambda)){

  predRLR2 <- factor(ifelse(predict(dd1RLR2, newx = model.matrix( ~ .-1, dd1test[, 1:14]), type = "response", s = dd1RLR2$lambda[i]) > threshold2, 1, 0), levels = c(0, 1))

  lambda2[i] <- confusionMatrix(predRLR2, dd1test$V15.)$overall[1]

}

predRLR2 <- factor(ifelse(predict(dd1RLR2, newx = model.matrix( ~ .-1, dd1test[, 1:14]), type = "response", s = dd1RLR2$lambda[which.max(lambda2)]) > threshold2, 1, 0))

confusionMatrix(dd1test$V15., predRLR2)


# Base de datos 2

dd2train <- read.table("C:/Users/jindu/Desktop/TFG/Adult/adulttrain.data", sep = ",")

dd2test <- read.table("C:/Users/jindu/Desktop/TFG/Adult/adulttest.test", sep = ",")

dd2train <- dd2train[, -3]

dd2test <- dd2test[, -3]

names(dd2train) <- names(dd2test) <- c("Age", "WorkClass", "Educ", "EducYear", "MaritalStatus", "Occup", "Relationship", "Race", "Sex", "CapGain", "CapLoss", "HperWeek", "NatCountry", "Y")

for(i in 1:ncol(dd2train)) {

  dd2train[dd2train[, i] == " ?", i] <- NA

}

for(i in 1:ncol(dd2test)) {

  dd2test[dd2test[, i] == " ?", i] <- NA

}
```

```
dd2train[dd2train$CapGain == 99999, ] <- NA

dd2test[dd2test$CapGain == 99999, ] <- NA

dd2train <- dd2train[-which(dd2train$NatCountry == " Holand-Netherlands"), ]

dd2train <- na.omit(dd2train)

for(i in c(2, 3, 5, 6, 7, 8, 9, 13, 14)) {

  dd2train[, i] <- factor(dd2train[, i])

}

dd2test <- na.omit(dd2test)

for(i in c(2, 3, 5, 6, 7, 8, 9, 13, 14)) {

  dd2test[, i] <- factor(dd2test[, i])

}

dd2train <- dd2train[-which(dd2train$HperWeek > 80), ]

dd2test <- dd2test[-which(dd2test$HperWeek > 80), ]

dd2train[, "Y"] <- factor(ifelse(dd2train[, "Y"] == " <=50K", 0, 1), levels = c(0, 1))

dd2test[, "Y"] <- factor(ifelse(dd2test[, "Y"] == " <=50K.", 0, 1), levels = c(0, 1))

corrplot(cor(dd2train[, c(1, 4, 10, 11, 12)]), method = "number")

describe(dd2train)

hist(dd2train$Age, main = "Histograma de Age", xlab = "Age", ylab = "Frecuencia", col = 8)

hist(dd2train$CapGain, main = "Histograma de CapGain", xlab = "CapGain", ylab = "Frecuencia", col =
8)

hist(dd2train$CapLoss, main = "Histograma de CapLoss", xlab = "CapLoss", ylab = "Frecuencia", col =
8)

hist(dd2train$EducYear, main = "Histograma de EducYear", xlab = "EducYear", ylab = "Frecuencia",
col = 8)

hist(dd2train$HperWeek, main = "Histograma de HperWeek", xlab = "HperWeek", ylab =
"Frecuencia", col = 8)

ggplot(data = dd2train, aes(x = Y, y = dd2train$Age, color = Y)) + xlab("Y") + ylab("Age") +
ggtitle("Boxplot de Age por Y") + geom_boxplot(outlier.shape = NA) + geom_jitter(width = 0.2) +
theme_bw() + theme(legend.position = "null")

ggplot(data = dd2train, aes(x = Y, y = dd2train$CapGain, color = Y)) + xlab("Y") + ylab("CapGain") +
ggtitle("Boxplot de CapGain por Y") + geom_boxplot(outlier.shape = NA) + geom_jitter(width = 0.2) +
theme_bw() + theme(legend.position = "null")

ggplot(data = dd2train, aes(x = Y, y = dd2train$CapLoss, color = Y)) + xlab("Y") + ylab("CapLoss") +
ggtitle("Boxplot de CapLoss por Y") + geom_boxplot(outlier.shape = NA) + geom_jitter(width = 0.2) +
theme_bw() + theme(legend.position = "null")
```

```
ggplot(data = dd2train, aes(x = Y, y = dd2train$EducYear, color = Y)) + xlab("Y") + ylab("EducYear") + ggtitle("Boxplot de EducYear por Y") + geom_boxplot(outlier.shape = NA) + geom_jitter(width = 0.2) + theme_bw() + theme(legend.position = "null")

ggplot(data = dd2train, aes(x = Y, y = dd2train$HperWeek, color = Y)) + xlab("Y") + ylab("HperWeek") + ggtitle("Boxplot de HperWeek por Y") + geom_boxplot(outlier.shape = NA) + geom_jitter(width = 0.2) + theme_bw() + theme(legend.position = "null")

dd2RLS <- glm(Y ~ Age + WorkClass + Educ + EducYear + MaritalStatus + Occup + Relationship + Race + Sex + CapGain + CapLoss + HperWeek + NatCountry, family = "binomial", data = dd2train)

summary(dd2RLS)

threshold <- table(dd2train$Y)[1] / sum(table(dd2train$Y))

predRLS <- factor(ifelse(predict(dd2RLS, newdata = dd2test, type = "response") > threshold, 1, 0), levels = c(0, 1))

confusionMatrix(dd2test$Y, predRLS)

ROC <- roc(dd2train$Y ~ predict(dd2RLS, type = c("response")))

plot(ROC, main = "Curva ROC del modelo simple", xlab = "Especificidad", ylab = "Sensibilidad")

auc(ROC)

dd2RLR <- glmnet(x = model.matrix( ~ .-1, dd2train[, 1:13]), y = as.factor(as.matrix(dd2train[, 14])), family = "binomial")

summary(dd2RLR)

lambda <- vector(length = length(dd2RLR$lambda))

for(i in 1:length(dd2RLR$lambda)){

  predRLR <- factor(ifelse(predict(dd2RLR, newx = model.matrix( ~ .-1, dd2test[, 1:13]), type = "response", s = dd2RLR$lambda[i]) > threshold, 1, 0), levels = c(0, 1))

  lambda[i] <- confusionMatrix(predRLR, dd2test$Y)$overall[1]

}

predRLR <- factor(ifelse(predict(dd2RLR, newx = model.matrix( ~ .-1, dd2test[, 1:13]), type = "response", s = dd2RLR$lambda[which.max(lambda)]) > threshold, 1, 0))

confusionMatrix(dd2test$Y, predRLR)
```