UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales

Modelo logístico con efectos mixtos: aplicación a la detección de fraudes en seguros automotores

Tesis presentada para optar al título de Magíster de la Universidad de Buenos Aires en Estadística Matemática

Autor: Ian Evangelos Bounos
Directora: Dra. Marina Valdora
Co-Directora: Dra. Daniela Rodríguez

Buenos Aires, 14 de diciembre de 2023

Resumen

Los Modelos Lineales Generalizados (GLM) son ampliamente utilizados en diversos campos de los seguros, sobre todo en el cálculo de frecuencia y severidad de siniestros. Muchas de las covariables son cualitativas con múltiples posibles valores, por ejemplo, la marca del automóvil en el contexto de un seguro automotor. Esto es un problema para los modelos lineales generalizados con efectos fijos porque aumentan la varianza de las predicciones y dificultan la estimación cuando tenemos datos desbalanceados. Más aún, resulta imposible hacer proyecciones sin modificar el modelo en el caso en el que no se dispone de experiencia previa del comportamiento de cierto nivel de la variable cualitativa en cuestión (como podría ser una determinada marca en nuestro ejemplo). Estos problemas pueden ser abordados incorporando efectos aleatorios a los GLM, los cuales consideran el efecto de ciertas variables como algo aleatorio en lugar de un parámetro fijo a estimar como se realiza en los GLM usuales. En este trabajo se estudia la incorporación de efectos aleatorios al modelo logístico en el contexto de la detección de Fraudes en Seguros Automotores y se estudian múltiples métricas del rendimiento de la clasificación para seleccionar el modelo.

Abstract

Mixed-effects logistic model: application to fraud detection in automotive insurance

Generalized Linear Models (GLMs) are widely used in various fields of insurance, particularly in calculating the frequency and severity of losses. Many of the covariates are qualitative with multiple possible values, such as the brand of the car in the context of motor insurance. This is a problem for fixed-effect GLMs because it increases the variance of predictions and makes estimation difficult when we have unbalanced data. Furthermore, it is impossible to make projections without modifying the model in the case where there is no previous experience of the behavior of a certain level of the qualitative variable in question (such as a particular brand in our example). These problems can be addressed by incorporating random effects into GLMs, which consider the effect of certain variables as something random rather than a fixed parameter to be estimated as is done in usual GLMs. This work studies the incorporation of random effects into the logistic model in the context of detecting Fraud in Motor Insurance and studies multiple classification performance metrics to select the model.

Agradecimientos

Quiero agradecer a mis padres por siempre apoyar mis proyectos personales y profesionales. También a todos los compañeros, docentes y no docentes que permitieron la realización de la Maestría en un contexto sumamente desafiante de virtualidad a causa de la pandemia, manteniendo la calidad académica y haciendo el proceso disfrutable, además de haberse preocupado, ni bien fue posible, por organizar instancias de presencialidad. Realmente agradezco esto.

Finalmente, quiero agradecer a mi directora Marina Valdora y Co-Directora Daniela Rodríguez por mostrar siempre la mejor predisposición para acompañar y dirigir en el desarrollo de la tesis.

Índice general

Capítulo 1

Introducción

1.1. Introducción

El objetivo de este trabajo es utilizar un modelo logístico con efectos mixtos para la detección de fraudes en siniestros de seguros automotores.

Un contrato de seguro es un acuerdo bilateral en el que una parte, el **asegurador**, se compromete a proteger al otro, el **asegurado**, de ciertos riesgos o pérdidas a cambio de una contraprestación económica periódica denominada **prima**.

Un **siniestro** es un evento cubierto por el seguro que da lugar a la prestación económica por parte del asegurador. Por ejemplo, si una persona tiene un seguro de vida y fallece, el fallecimiento sería un siniestro que daría derecho a los beneficiarios del seguro a recibir la suma asegurada. En el caso de un seguro automotor, un siniestro podría ser un accidente de tráfico o un robo del vehículo, entre otros. Además, en este caso, lo que suele cubrirse no es únicamente los daños a la propiedad individual del asegurado (el automóvil, por ejemplo) sino también daños a terceros, llamados de Responsabilidad Civil (RC), los cuales resultan obligatorios en Argentina.

Como se ha mencionado, la prima es el precio que el asegurado paga al asegurador a cambio de la protección ofrecida por el seguro. La prima se calcula actuarialmente en función del riesgo que asume el asegurador al ofrecer la cobertura, y puede variar según factores como la edad del asegurado, el tipo de actividad que desempeña, el lugar donde reside o trabaja, entre otros. El primer objetivo de este cálculo es garantizar que la empresa aseguradora pueda garantizar su solvencia en sus compromisos futuros. Es por ello que se define el concepto fundamental de **prima pura**: el costo esperado de la cobertura del seguro sin tener en cuenta otros factores como los gastos administrativos, impuestos, comisiones a intermediarios y remuneración a la inversión de la empresa. Este es el pilar sobre el cual se construye la prima final que se le cobra al asegurado y permite responder a una

pregunta central en este trabajo:

¿Por qué es importante la detección de fraudes en una empresa de seguros?

El primer motivo es claro: para evitar pérdidas de la empresa aseguradora. Sin embargo, hay una segunda razón más sistémica que involucra a todos los agentes participantes: La presencia de fraudes indetectables aumenta de manera artificial el costo esperado del contrato para la aseguradora. Esto es, en otras palabras, una mayor prima pura, lo cual, manteniendo la ganancia de la empresa por contrato, gastos administrativos y demás factores constantes, redunda en una mayor prima final cobrada al asegurado. En el peor escenario, este aumento de primas puede resultar en la no realización del contrato. Por estos motivos, la presencia de fraudes indetectables resulta un problema para todo el ecosistema de seguros y no exclusivamente para un actor particular del mismo. Una segunda cuestión a considerar es:

¿Por qué utilizamos modelos lineales generalizados (GLM)? ¿Por qué efectos mixtos?

En una encuesta mundial realizada por *Akur8*, una empresa multinacional dedicada a brindar servicios de ciencias de datos a empresas de Seguros, el 90 % de los encuestados afirman que para sus modelos de *pricing* (cálculos de primas) utilizan GLM´s. El principal argumento dado por los encuestados para usar este método en lugar de otras técnicas como *Random Forest* o *Gradient Boosting* es su interpretabilidad y facilidad de comunicación de qué es lo que efectivamente realiza el modelo. Por lo tanto, incorporar modificaciones a los GLM puede ser de utilidad inmediata, lo cual puede ser una motivación para considerar **efectos mixtos**. Si bien a lo largo del trabajo se explicará qué son dichos efectos, podemos afirmar que estos son útiles en situaciones en las que hay diferencias entre individuos o grupos en la población que pueden afectar la variable respuesta y no se quieren ignorar. Por ejemplo, si se está analizando el rendimiento académico de estudiantes en varias escuelas y se sabe que hay diferencias significativas entre los distintos grupos de estudiantes (por ejemplo, los que tienen una casa con una computadora y los que no), se puede realizar un modelo de efectos mixtos en el cual la escuela de cada estudiante es un efecto aleatorio y la presencia de computadora, uno fijo.

Una gran ventaja de los efectos mixtos es que el modelo sigue teniendo capacidad predictiva aun cuando se incorporan observaciones de variables cualitativas de las cuales no se tiene experiencia previa. Esta diferencia con respecto a los GLM clásicos es crucial. Por ejemplo, para calcular la siniestralidad en los seguros de automóviles podríamos (y se suele hacer) considerar el modelo del automóvil como

covariable. Sin embargo, si aparece un modelo nuevo, el GLM o bien no serviría, o bien se tendría que realizar la predicción asumiendo que el nuevo modelo del automóvil se parece a otro del cual ya se tiene experiencia. Esto encierra cierto nivel de discrecionalidad que se desea evitar y la incorporación de efectos mixtos lo permite.

Otra desventaja de considerar el modelo del auto como covariable fija de un GLM clásico es que, al tener muchos niveles (hay innumerables modelos de autos), se incrementa en demasía el riesgo de sobreajuste y limita la capacidad de incorporar nuevas variables.

Finalmente, es lícito observar que si bien hay antecedentes de aplicar GLM con efectos mixtos, como puede observarse en el texto de Mc Neil y Wendin de 2005, esta no es una herramienta generalizada y la práctica común en estos casos es usar como covariable el modelo del auto y hacer agrupaciones de los mismos para reducir la cantidad de niveles. Por lo tanto, puede resultar relevante acumular antecedentes que incorporen efectos mixtos como metodología alternativa.

1.2. Estructura del trabajo

El trabajo se estructura de la siguiente forma: en el Capítulo 2 se introducen nociones básicas de los Modelos Lineales Generalizados con efectos mixtos. Junto a un ejemplo ilustrativo, se los define y se explican los métodos de estimación y predicción, inferencia y métricas de clasificación que serán utilizados en la el trabajo.

En la primera parte del Capítulo 3 se realiza un análisis exploratorio del dataset que disponemos para preseleccionar candidatos a variables y, en algunos casos, se generan nuevas variables a partir de las disponibles. Posteriormente, se seleccionarán las variables entre las preseleccionadas con diferentes criterios, lo que dará lugar a dos modelos. Finalmente, los resultados son analizados desde los distintos enfoques explicados en el Capítulo 3.

Finalmente, en el Capítulo 4 se exponen las conclusiones y reflexiones acerca del trabajo realizado y los resultados obtenidos. Para el tratamiento de los datos, la implementación de los modelos y la realización de gráficos fue utilizado el lenguaje de programación R.

Capítulo 2

Nociones básicas

2.1. Ejemplo Motivador: Efectos Aleatorios

En primer lugar, consideremos el siguiente ejemplo motivador para incorporar efectos aleatorios a un modelo logístico: imaginemos que se quiere estimar cuántas personas van a votar a Obama en las elecciones de 2008. Para ello, se hace una encuesta en los 50 estados de Estados Unidos y para cada uno se dispone de los siguientes datos:

- n_i: cantidad de personas encuestadas en el estado i.
- y_i: cantidad de personas que votarán a Obama en las siguientes elecciones en el estado i

Si solo se dispone de esta información, una primera aproximación es modelar el problema por medio de un **modelo lineal generalizado (GLM)** con distribución binomial y una función link logit, o, en otras palabras, un **modelo logístico**. En términos de ecuaciones puede describirse como:

$$\sum_{j=1}^{n_i} y_{ij} = y_i \sim Bin(n_i, \pi_i),$$

$$\pi_i = \frac{e^{\beta_i}}{1 + e^{\beta_i}} = g^{-1}(\beta_i),$$

donde:

- y_{ij} es el valor observado (es decir, si votará a Obama o no) el individuo j que pertenece al estado i.
- π_i es la probabilidad "real" de que un individuo del estado i vote a Obama

- la función $g(p) = log(\frac{p}{1-p})$, conocida como *logit* es nuestra función *link*.

En este caso, como puede verse en Agresti (2013), el estimador de máxima verosimilitud será la **proporción observada** de los votantes de Obama en cada estado. Es decir:

$$\hat{\pi}_i = \frac{y_i}{n_i}.$$

A primera vista, este estimador parece adecuado, pero observemos en primer lugar en el Cuadro 2.1 de las primeras diez observaciones de la encuesta:

Estado	n	y	Proporción
AK	5	3	0.60
AL	29	9	0.31
AR	17	2	0.11
AZ	35	13	0.37
CA	207	129	0.62
CO	37	16	0.43
CT	25	14	0.56
DC	4	4	1.00
DE	6	4	0.66
FL	128	73	0.57

Cuadro 2.1: Encuesta elecciones 2008. Primeras 10 observaciones

Puede advertirse cierto desequilibrio en la cantidad de observaciones. Por ejemplo. Alaska (AK) tiene 5 observaciones, mientras que California (CA) presenta 207. Este es un problema desde el punto de vista de la varianza. Para ilustrar este problema, consideremos la Figura 2.1 que simultáneamente muestra el resultado real de las elecciones y las proporciones observadas por cada estado, ordenados decrecientemente según cantidad de observaciones.

Puede apreciarse que la varianza es creciente para la proporción muestral, mientras que el resultado real muestra menor dispersión de manera consistente. Esto es un problema porque se estaría teniendo subpoblaciones con muy pocas observaciones, lo cual requiere técnicas de estimación de área pequeña. En los casos que tenemos muchas observaciones, la proporción muestral funciona de forma adecuada; mientras que en los casos de muestras pequeñas lo que hemos de hacer heurísticamente es privilegiar la información que dan los otros estados respecto de la disposición a votar a Obama, pues hay una confianza menor en los resultados de la encuesta. Esto implica realizar un *trade off* entre la información que da el Estado como subgrupo y el país en su totalidad del cual el estado forma parte. La

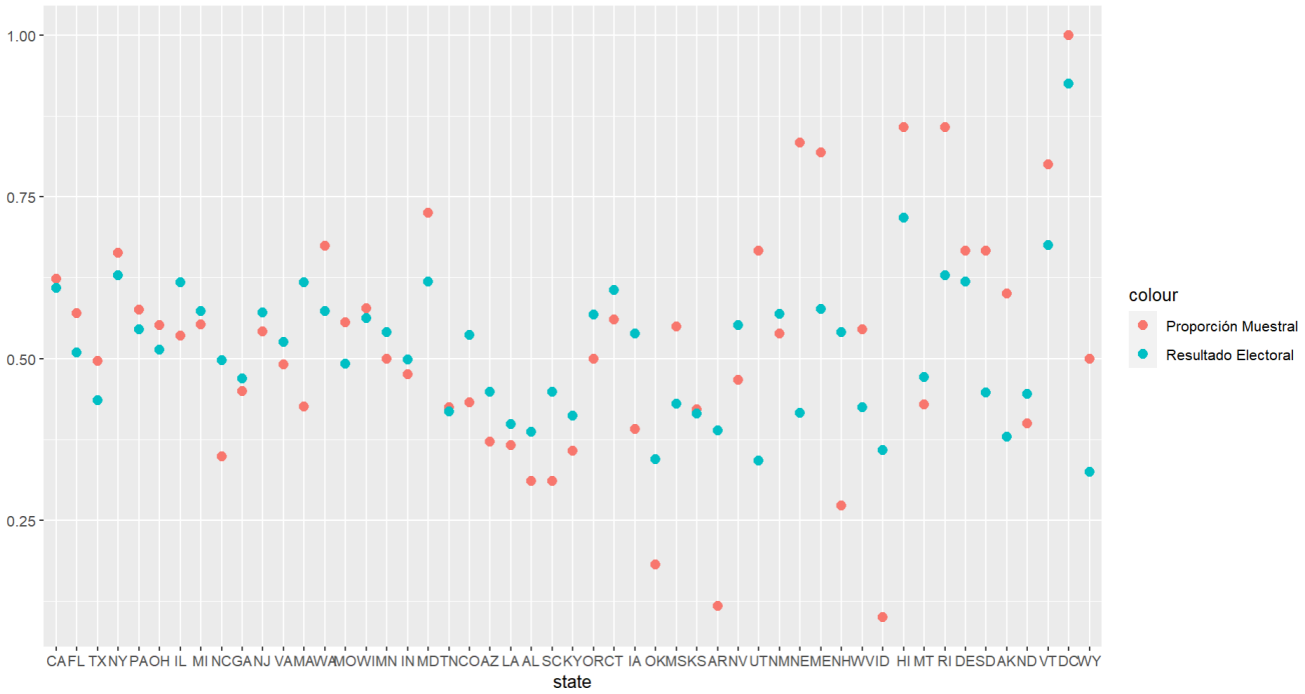


Figura 2.1: Proporciones estimadas y observadas.

forma en que haremos dicho *trade off* será por medio de los efectos aleatorios, que serán explicados en la próxima sección.

2.2. Modelo logístico con efectos mixtos

2.2.1. Definición del modelo

En un modelo logístico con efectos mixtos se tienen N observaciones de una variable aleatoria con distribución Bernoulli agrupadas en m clústers o grupos. Llamamos y_{it} a la variable número $t = 1, \ldots, n_i$ del clúster i. Además, disponemos de:

- Un vector de covariables explicativas $X_{it} \in \mathbb{R}^p$
- Un vector desconocido de parámetros de efectos fijos correspondientes a las covariables $\beta \in \mathbb{R}^p$
- Un vector de efectos aleatorios por cada clúster $u_i \sim \mathbf{N}_q(0, \Sigma)$, es decir, un vector normal multivariado.

- Un vector observado $Z_{it} \in \mathbb{R}^q$.

Recordemos que el modelo logístico con **efectos fijos** si incorporamos las covariables explicativas puede describirse como:

$$y_{it} \sim Be(\pi_{it}),$$

$$g(\pi_{it}) = X_{it}^T \beta,$$

donde g es una función link, es decir, una función derivable e invertible. Mientras que el modelo logístico con **efectos mixtos** para las covariables explicativas es el siguiente:

$$y_{it}|u_i \sim Be(\pi_{it}), \tag{2.1}$$

$$g(\pi_{it}) = X_{it}^T \beta + Z_{it}^T u_i, \tag{2.2}$$

donde g es, nuevamente, una función link.

Para entender cómo es la relación entre ambos modelos, observemos que con efectos fijos la media de cada y_{it} está en función de las covariables X_{it} y los parámetros fijos desconocidos β. Mientras que en los efectos mixtos lo que modelamos es la media condicionada al **efecto aleatorio** u_i, el cual depende del grupo al que pertenece y_{it}.

Sabemos que:

$$E[y_{it}|u_i] = g^{-1}(X_{it}^T \beta + Z_{it}^T u_i).$$

Usando la Ley de las Esperanzas iteradas obtenemos:

$$E[y_{it}] = E[E[y_{it}|u_i]] = E[g^{-1}(X_{it}^T \beta + Z_{it}^T u_i)] = \int g^{-1}(X_{it}^T \beta + Z_{it}^T u_i) f(u_i, \Sigma) du_i, \tag{2.3}$$

donde f es la densidad de una variable normal.

Consideremos el caso particular en que la función link es la identidad

$$E[y_{it}] = \int (X_{it}^T \beta + Z_{it}^T u_i) f(u_i, \Sigma) du_i,$$

$$E[y_{it}] = X_{it}^T \beta \int f(u_i, \Sigma) du_i + \int Z_{it}^T u_i f(u_i, \Sigma) du_i.$$

Sabemos que la integral de la izquierda es 1 por ser la integral de una función de densidad. Por otro lado, la de la derecha es 0 por ser la esperanza de una combinación lineal de normales con media 0, que, por lo tanto, es una normal con media 0. Se concluye que $E[y_{it}] = X_{it}^T\beta$. Cabe observar que en este caso, los estimadores de máxima verosimilitud de β serán exactamente los mismos para un modelo de efectos fijos que para uno de efectos mixtos, lo cual se debe principalmente a que la esperanza de u_i es 0. Este fenómeno no es general sino que depende de la elección de la función link. Por ejemplo, cuando el link es la función *logit* y u_i es univariada e independiente para distintas i, Zeger et al (1988) muestran la siguiente identidad:

$$E[y_{it}] = E\left[\frac{e^{X_{it}^T\beta+u_i}}{1+e^{X_{it}^T\beta+u_i}}\right].$$

Un problema aquí es que puede probarse que no vale que $E\left[\frac{e^{X_{it}^T\beta+u_i}}{1+e^{X_{it}^T\beta+u_i}}\right] = \frac{e^{X_{it}^T\beta}}{1+e^{X_{it}^T\beta}}$ con la excepción del caso degenerado donde la varianza de u_i es 0. Sin embargo, si la varianza es pequeña, en el texto de Zeger (1988) se presenta la siguiente aproximación:

$$E[y_{it}] \approx \frac{ce^{X_{it}^T\beta}}{1+ce^{X_{it}^T\beta}},$$

donde $c = (1 + 0{,}346\sigma^2)^{-1/2}$.

Finalmente, para ilustrar cómo los efectos aleatorios pueden ayudarnos a modelar el ejemplo de las elecciones de 2008 uno podría considerar el siguiente modelo:

$$logit(\pi_{it}) = \beta_0 + u_i.$$

En este caso $Z_{it} = X_{it} = 1$ y u_i es una normal univariada de media 0 y varianza σ_u desconocida que como primera aproximación podríamos suponer constante. En este caso, β_0 es interpretado como el *odds ratio* base y u_i una perturbación aleatoria que se induce por cada estado. Este modelo es logístico con **efectos aleatorios** y no mixtos porque el vector de covariables X solo tiene un intercepto. Veamos un modelo posible mixto, por ejemplo, en el cual la edad de la persona encuestada se toma como covariable de efectos fijos:

$$logit(\pi_{it}) = \beta_0 + \beta_1 Edad_{it} + u_i.$$

En este caso, $X_{it} = \binom{1}{Edad_{it}}$, $\beta = \binom{\beta_0}{\beta_1}$ y u_i sigue siendo una perturbación normal propia de cada estado, que, en este caso, no tiene en cuenta la edad. Lo que podría ocurrir es que en cada estado la edad altere de distinta manera la tendencia a votar a Obama. Eso se puede modelar de la siguiente forma:

$$logit(\pi_{it}) = \beta_0 + \beta_1 Edad_{it} + u_{1i} Edad_{it} + u_{0i}$$

Aquí, X_{it} y β son exactamente los mismos que en el anterior, pero $Z_{it} = \binom{1}{Edad_{it}}$ y $u_i = \binom{u_{0i}}{u_{1i}}$ y cada u_{ki} es una normal con su respectiva varianza. Si bien en el modelo general se puede asumir cierta correlación entre ambos de acuerdo con la matriz de varianzas y covarianzas Σ, típicamente supondremos que son independientes y que esta matriz es diagonal.

En los tres casos que hemos visto, hemos asumido una distribución para el efecto aleatorio que cada estado induce en los individuos. Esto produce inevitablemente una **correlación entre observaciones en el mismo estado**. Por ejemplo, en el primero y segundo de los modelos que vimos,

$$cov(logit(p_{it}), logit(p_{ik})) = \sigma_u^2.$$

Esto supone una diferencia porque en un modelo logístico clásico estos valores suelen suponerse independientes. Una discusión más amplia alrededor de este tópico puede encontrarse en el capítulo 13 de Agresti (2013).

2.2.2. Estimadores y Predictores

El método para la estimación de parámetros será el de **Máxima Verosimilitud** (ML por sus siglas en inglés).

Definición **2.2.2.1 (Estimador de máxima verosimilitud)** *Sean $x_1, \cdots, x_n$ observaciones independientes e idénticamente distribuidas con una densidad de probabilidad f_{θ_0} perteneciente a una familia de densidades $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$. El estimador de máxima verosimilitud $\hat{\theta}$ es aquel que maximiza la función de verosimilitud $\mathcal{L}(\theta, x_1, \cdots, x_n) := \prod_{i=1}^n f_\theta(x_i)$*

Típicamente, cuando no hay efectos aleatorios, si bien puede ser difícil de maximizar, la función de máxima verosimilitud es conocida. Sin embargo, observemos que la presencia de efectos aleatorios complejiza aún más la estimación. Consideremos el ya mencionado **modelo logístico con efectos mixtos** (Ecuaciones 2.1 y 2.2), por lo obtenido en la Ecuación 2.3:

$$P_\beta(y_{it} = 1) = E[y_{it}] = \int g^{-1}(X_{it}^T \beta + Z_{it}^T u_i) f(u_i, \Sigma) du_i.$$

Además,

$$P_\beta(y_{it} = 0) = 1 - E[y_{it}] = 1 - \int g^{-1}(X_{it}^T \beta + Z_{it}^T u_i) f(u_i, \Sigma) du_i.$$

Esto nos permitirá despejar la función de verosimilitud asumiendo independencia de las y:

$$\mathcal{L}_\beta(y,u) = P_\beta(y=\tilde{y}|u) = \prod_{it} P_\beta(y_{it}=\tilde{y}_{it}|u_i). \tag{2.4}$$

Aquí y es el vector de las y_{it} como variables aleatorias, mientras que $\tilde{y}$ y $\tilde{y}_{it}$ son las respectivas realizaciones. Observemos que $P_\beta(y_{it}=\tilde{y}_{it}|u_i)$ se puede calcular a partir de las integrales $\int g^{-1}(X_{it}^T\beta + Z_{it}^T u_i) f(u_i,\Sigma)du_i$.

El problema es que este tipo de integrales pueden ser difíciles de calcular analíticamente (incluso imposible) para funciones link generales y para la logit en particular. Sin embargo, como señalan McCullagh y Nelder (1989), cuando tenemos pocos efectos aleatorios pueden aproximarse las integrales numéricamente y luego maximizar la función de verosimilitud. Por ejemplo, para el caso de un error aleatorio unidimensional, con la cuadratura de Gauss-Hermite que consiste en encontrar pesos y puntos adecuados de modo que $\int_{\mathbb{R}} e^{-x^2}\phi(x)dx \sim \sum w_i\phi(z_i)$ para aproximar la integral de una función de la forma $e^{-x^2}\phi$. La fórmula de calcular los pesos w_i y los puntos z_i pueden encontrarse en el texto de Abramowitz y Stegun (1964).

Para encontrar los nodos z_i se utilizan las raíces de los polinomios de Hermite. Los polinomios de Hermite son una serie de polinomios ortogonales con respecto a la función de peso e^{-x^2}. Los mismos pueden ser definidos recursivamente:

$$H_0(x) = 1,$$
$$H_1(x) = 2x,$$
$$H_{k+1}(x) = 2xH_k(x) - 2kH_{k-1}(x),$$

o también se pueden expresar mediante la siguiente fórmula general:

$$H_k(x) = (-1)^k e^{x^2}\frac{d^k}{dx^k}e^{-x^2}.$$

Para calcular los pesos w_i en la cuadratura de Gauss-Hermite, Abramowitz y Stegun (1964) demuestran la siguiente fórmula:

$$w_i = \frac{m!2^{m-1}\sqrt{\pi}}{m^2 H_{m-1}(z_i)^2},$$

donde m es el número de nodos escogidos para la aproximación.

Para aproximar la función de verosimilitud en la práctica, el paquete *lme4* utilizará la cuadratura de Gauss-Hermite adaptativa, que es una técnica que permite mejorar la precisión de la cuadratura de Gauss-Hermite al dividir el intervalo de integración en subintervalos y aplicar la cuadratura de Gauss-Hermite a cada

subintervalo de manera independiente. La idea es que al dividir el intervalo en subintervalos se pueden identificar las áreas de la función donde se requiere una mayor precisión y aplicar la cuadratura de Gauss-Hermite con un mayor número de nodos en esas áreas. Otras formas de aproximación de la integral por métodos Montecarlo pueden encontrarse en el capítulo 4 de Jiang (2007).

Una vez aproximada la función de verosimilitud a partir de la estimación numérica de la integral, la cual dependerá de los valores de los efectos fijos, aleatorios y de las realizaciones de y, X y Z (siendo estas últimas datos conocidos), el siguiente paso consiste en maximizar las mismas con respecto a los efectos fijos y aleatorios. Si llamamos $\hat{\mathcal{L}}_\beta(y, u)$ a la aproximación de la función de verosimilitud que resulta de remplazar los valores de las integrales por sus respectivas aproximaciones en la Ecuación 2.4, podemos buscar, con algún método de optimización, los valores de β y u que la maximizan. Así obtendremos el **Estimador** de máxima verosimilitud de β (que denotaremos $\hat{\beta}$) y el **Predictor** de máxima verosimilitud de u (denotado $\hat{u}$). Esta diferencia terminológica es explicada en el texto de McCulloch y Searle (2001):

> "The assumptions about random effects differ from those for fixed effects and so treatment of the two kinds of effects is not the same. A fixed effect is considered to be a constant, which we wish to estimate. But a random effect is considered as just an effect coming from a population of effects. It is this population that is an extra assumption, compared to fixed effects, and we would hope it would lead to an estimation method for random effects being an improvement over that for fixed effects. To emphasize this distinction we use the term prediction of random effects rather than estimation"

> "Las suposiciones sobre los efectos aleatorios difieren de aquellas para los efectos fijos, por lo que el tratamiento de ambos tipos de efectos no es el mismo. Se considera que un efecto fijo es constante y que queremos estimar. Pero un efecto aleatorio se considera simplemente como un efecto que proviene de una población de efectos. Esta población es una suposición adicional en comparación con los efectos fijos, y esperaríamos que conduzca a un método de estimación de efectos aleatorios que sea mejor que el de los efectos fijos. Para enfatizar esta distinción, utilizamos el término predicción de efectos aleatorios en lugar de estimación."

Por estos motivos, usaremos el término **Estimación** cuando nos refiramos a efectos fijos y **Predicción** para efectos aleatorios.

2.2.3. Inferencia

Una vez obtenida una aproximación de la función de máxima verosimilitud, es posible replicar los tests tradicionales de modelos lineales generalizados que utilicen dicha función. Siempre teniendo en cuenta que el hecho de que se trate de una aproximación puede implicar problemas númericos, particularmente cuando se dispone de pocos datos, como señalan McCuloch y Searle (2001).

Test de Wald

Una primera aproximación es el test de significancia individual, el cual aplica únicamente a los efectos fijos. Supongamos que tenemos un candidato a valor del coeficiente del efecto fijo β_j, el cual típicamente será 0 pero que podría ser cualquier valor real. A dicho candidato a valor lo llamaremos b_j. En su texto de 1985, Fahrmeir y Kaufmann probaron la siguiente aproximación asintótica para modelos lineales generalizados de efectos fijos, bajo hipótesis nula de que $\beta_j = b_j$:

$$W = \frac{\hat{\beta}_j - b_j}{\widehat{\mathrm{SE}}(\hat{\beta}_j)} \underset{a}{\approx} \mathcal{N}(0,1),$$

donde $\hat{\beta}_j$ es el valor estimado por máxima verosimilitud del coeficiente de regresión de la variable explicativa j y $\mathrm{SE}(\hat{\beta}_j)$ es su desviación estándar estimada. McCuloch y Searle (2001) proponen extender su uso a los coeficientes de efectos fijos de los modelos con efectos mixtos estimados por máxima verosimilitud aproximada, que es la utilizada en la subsección anterior.

El hecho de tener una distribución aproximadamente normal nos permite realizar el test de hipótesis:

$$H_0\text{: } \beta_j = b_j$$
$$H_1\text{: } \beta_j \neq b_j.$$

Nuestro criterio de decisión será rechazar la hipótesis si $|W| > z_{\alpha/2}$ y no hacerlo en caso contrario, donde α es nuestro nivel de significancia y $z_{\alpha/2}$ es el valor tal que la distrbución acumulada de la normal estándar vale $1 - \alpha/2$.

Test de Cocientes de Verosimilitud

El enfoque anterior es de **significancia individual,** en el cual determinamos si el estimador de un coeficiente de una única variable es estadísticamente diferente del valor b_j. El segundo enfoque es el de la **significancia conjunta**: supongamos que tenemos dos modelos logísticos con efectos mixtos $\mathcal{M}_1$ y $\mathcal{M}_2$ anidados, es decir, que todas las variables incluidas en $\mathcal{M}_1$ lo están en $\mathcal{M}_2$. En este caso, lo

que queremos someter a un test es si las variables estrictamente propias de $\mathcal{M}_2$ (es decir, aquellas no incluidas en $\mathcal{M}_1$) añaden información de forma conjunta. Nuevamente, dado que nosotros disponemos de aproximaciones numéricas de las funciones de verosimilitud, un test posible es el de cocientes de verosimilitud. El test de ratio de verosimilitud calcula el cociente entre la verosimilitud del modelo más complejo ($\mathcal{M}_2$) y la verosimilitud del modelo más simple ($\mathcal{M}_1$), el cual, bajo la hipótesis de que $\mathcal{M}_2$ no agrega información relevante debería ser cercano a 1. Por lo tanto, definimos:

$$\Lambda = \frac{\mathcal{L}_2}{\mathcal{L}_1},$$

donde $\mathcal{L}_k$ (para $k = 1, 2$) es la verosimilitud de $\mathcal{M}_k$ con sus parámetros estimados en dicho modelo dada por $\mathcal{L}_k = \mathcal{L}_{\hat{\beta}_{\mathcal{M}_k}}(y, u)$ (donde $\hat{\beta}_{\mathcal{M}_k}$ es el vector de parámetros estimados del modelo k). Esto nos permite definir la deviance:

$$D = -2\log(\Lambda).$$

Puede demostrarse que la deviance, bajo hipótesis nula de que $\mathcal{M}_2$ no agrega información al modelo, tiene asintóticamente una distribución chi cuadrado con grados de libertad igual a la cantidad de variables estrictamente propias de $\mathcal{M}_2$, que notaremos m_{gl}. Pueden verse los detalles en McCuloch y Searle (2001). Esto nos permite realizar un test de hipótesis para una significancia dada. En el caso particular de los modelos con efectos mixtos hemos de reemplazar las verosimilitudes por sus aproximaciones numéricas. Lo implementaremos con la función *anova* del paquete *stats*. En este contexto, nuestro test de hipótesis será el siguiente:

- H_0: "los coeficientes de los parámetros de las variables propias de $\mathcal{M}_2$ son 0"
- H_1: "al menos uno de los coeficientes de los parámetros de las variables propias de $\mathcal{M}_2$ es distinto de 0"

Mientras que nuestro estadístico será $D \sim \chi^2_{m_{gl}}$ y rechazaremos la hipótesis nula si $D > \chi^2_{m_{gl}\,1-\alpha}$ (siendo α nuestro nivel de significancia).

Finalmente, es necesario observar que la deviance tiene la propiedad de disminuir con la incorporación de nuevas variables al modelo, lo cual es un problema para comparar modelos con distinto número de covariables. Este es un fenómeno similar al que ocurre con en el R^2 en el contexto de las regresiones lineales. Para ello, se define el **criterio de selección de Akaike (AIC)** que penaliza la incorporación de nuevas variables:

$$AIC = D - 2K$$

donde K es la cantidad de covariables incorporadas al modelo. Detalles al respecto pueden encontrarse en McCullagh y Nelder (1989).

2.3. Métricas de Clasificación

Recordemos que el objetivo de este trabajo es detectar fraudes, lo cual supone un problema de clasificación binaria supervisada. Es un hecho bien conocido que no existe una forma unívoca para medir el rendimiento de un modelo por lo cual mostraremos cuáles serán las que tendremos en cuenta en esta tesis. La principal fuente de esta sección será James et. al. (2013) y el libro de Hosmer y Lemeshow (2000).

Tasa de Aciertos:

La tasa de aciertos es una métrica que mide la proporción de predicciones correctas realizadas por un modelo de clasificación sobre el total de predicciones.

La tasa de aciertos es una métrica simple y fácil de entender, lo que la hace muy popular en la práctica. Sin embargo, la tasa de aciertos tiene algunas desventajas. En primer lugar, puede ser engañosa cuando hay clases desbalanceadas, ya que el modelo puede obtener una tasa de aciertos alta simplemente prediciendo la clase mayoritaria en todos los casos. En segundo lugar, la tasa de aciertos no proporciona información sobre la capacidad del modelo para identificar cada clase individualmente, sino que simplemente mide el rendimiento global del modelo. Por ello, en algunos casos es conveniente utilizar otras métricas como la precisión, la sensibilidad o la especificidad para tener una evaluación más detallada del rendimiento del modelo.

Tablas de confusión, sensibilidad y especificidad:

Una tabla de confusión es una herramienta comúnmente utilizada en el análisis de resultados de una prueba o evaluación. Se utiliza para evaluar la precisión de una prueba o modelo y se basa en cuatro tipos de resultados: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). Estos resultados se presentan en una tabla de dos por dos, donde se comparan las predicciones del modelo con los resultados verdaderos. A continuación en el cuadro 2.2 se presenta un ejemplo de una tabla de confusión:

	Predicción		
Resultado real	Positivo	Negativo	Total
Positivo	TP	FN	TP+FN
Negativo	FP	TN	FP+TN
Total	TP+FP	FN+TN	n

Cuadro 2.2: Tabla de confusión

La **sensibilidad** (también conocida como precisión) es una medida de la capacidad de una prueba o modelo para detectar correctamente los casos positivos. Se calcula como el número de verdaderos positivos dividido entre el número total de casos positivos:

$$\text{sensibilidad} = \frac{TP}{TP + FN}.$$

Por otro lado, la **especificidad** es una medida de la capacidad de una prueba o modelo para detectar correctamente los casos negativos. Se calcula como el número de verdaderos negativos dividido entre el número total de casos negativos:

$$\text{especificidad} = \frac{TN}{TN + FP}.$$

Es importante tener en cuenta que la sensibilidad y la especificidad no son mutuamente excluyentes. Es decir, una prueba o modelo puede tener alta sensibilidad y baja especificidad, o viceversa. El equilibrio adecuado entre estas dos medidas depende del contexto y del propósito de la prueba o modelo. Este balance entre sensibilidad y especificidad sugiere la introducción de curvas ROC.

Curvas ROC

En primer lugar, observemos que el umbral de clasificación en un modelo logístico es el valor que se utiliza para determinar la clase de una observación. El modelo logístico genera una probabilidad de pertenecer a cada clase, y el umbral de clasificación se utiliza para asignar una observación a la clase cuya probabilidad es mayor que el umbral. Por ejemplo, si el umbral es 0.5, una observación se asignará a la clase 1 si la probabilidad de pertenecer a la clase 1 es mayor que 0.5, y se asignará a la clase 0 en caso contrario. El umbral de clasificación se puede ajustar para aumentar o disminuir la sensibilidad o la especificidad del modelo. Una curva ROC (del inglés Receiver Operating Characteristic) es un gráfico que se utiliza para evaluar el rendimiento de un modelo de clasificación binaria. La curva ROC muestra la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos (1 - especificidad) a medida que se varía el umbral de decisión para clasificar una observación como positiva o negativa.

La curva ROC se puede utilizar para comparar diferentes modelos de clasificación y elegir el que tenga el mejor rendimiento. Una curva ROC ideal tiene una forma de media luna y se encuentra en la esquina superior izquierda del gráfico, lo que significa que tiene alta sensibilidad y especificidad. Una línea diagonal desde la esquina inferior izquierda hasta la esquina superior derecha del gráfico representa un modelo que simplemente predice de manera aleatoria sin ningún poder discriminativo (un clasificador con un rendimiento superior debería estar por encima

de esta diagonal en la curva ROC). El **Área bajo la curva (AUC)** de la curva ROC sirve como un indicador cuantitativo del rendimiento del modelo. Entre más cerca se encuentre de uno, mejor será el mismo. Esta será la principal métrica a maximizar en nuestro trabajo. En la Figura 2.2 se muestra un ejemplo de curva ROC con su valor de AUC y el punto más cercano al $(0, 1)$, es decir, el clasificador más cercano a una clasificación perfecta con distancia euclídea.

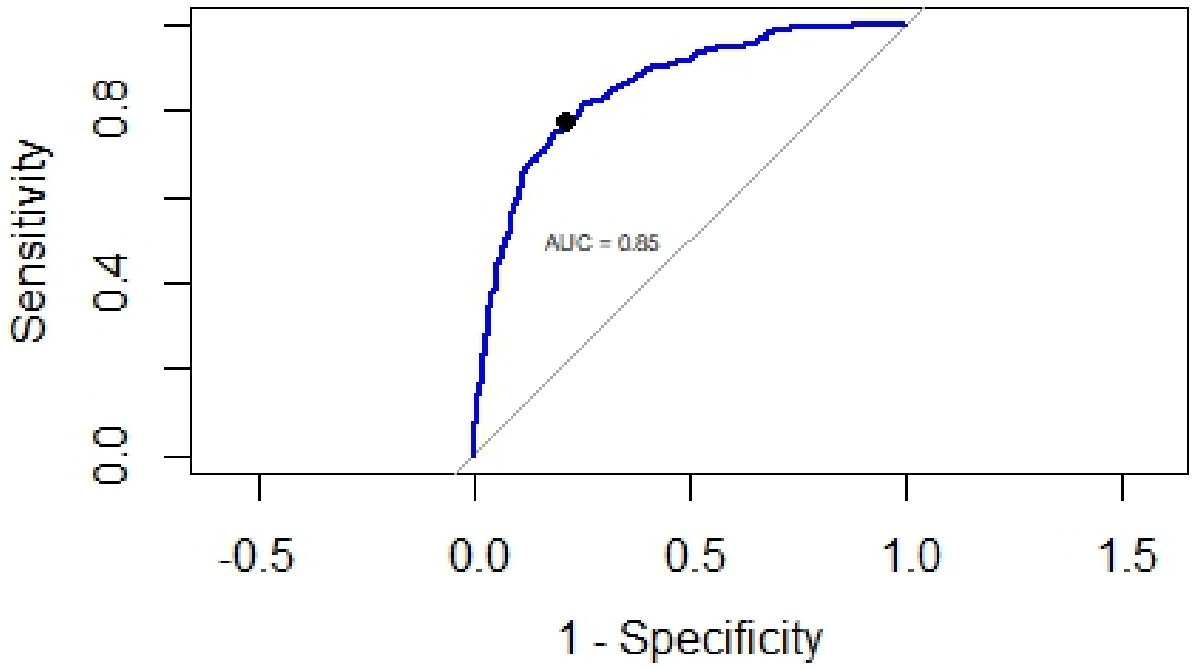


Figura 2.2: Ejemplo Curva ROC

2.4. Aplicación al Ejemplo inicial

Revisitemos el ejemplo motivador de la primera sección del presente capítulo. Allí modelamos la proporción de votos en función de cada estado como si estos fueran efectos fijos. La nueva aproximación del modelo sera considerarlos como efectos aleatorios. Consideremos el siguiente modelo:

$$\sum_{j=1}^{n_i} y_{ij} = y_i \sim Bin(n_i, \pi_i),$$
$$logit(\pi_i) = \beta_0 + u_i,$$
$$u_i = \mathcal{N}(0, \sigma_u^2),$$

con los efectos u_i independientes entre sí.

Para realizar las estimaciones de β_0 y las predicciones de los efectos aleatorios, utilizamos la función *glmer* del paquete *lme4* de R.

Aquí obtenemos predicciones de los resultados que no presentan el problema de la varianza que presentaban para el caso de efectos fijos. Por ejemplo, para este último la proporción muestral varía de 0.1 (Idaho) a 1.0 (DC), mientras que el de efectos aleatorios tenemos un rango de variabilidad mucho más acotado: desde 0.34 (Arkansas) a 0.64 (Maryland). Esto se puede apreciar en la Figura 2.3.

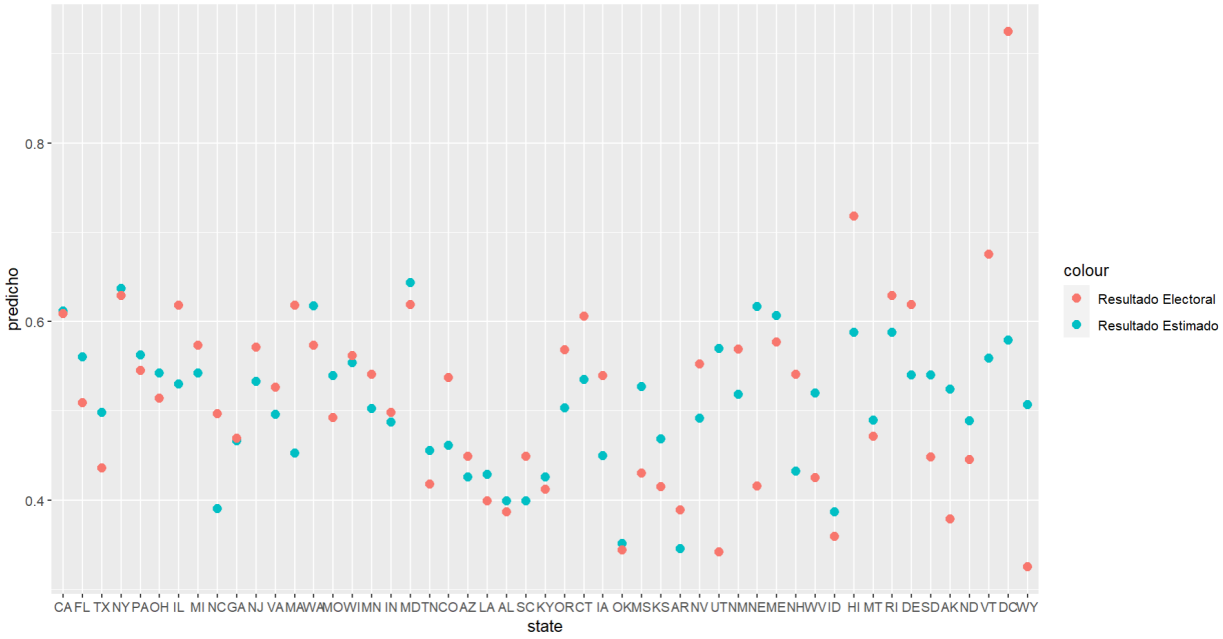


Figura 2.3: Proporciones observadas y estimadas por efectos aleatorios.

También podemos considerar el error cuadrático medio que se define como

$$ECM = \frac{\sum (y_{it} - \hat{y_{it}})^2}{n},$$

donde $\hat{y_{it}}$ es el valor predicho para la observacion y_{it} en cada modelo. En este caso, valores más pequeños del ECM indican un mejor ajuste al modelo.

Con respecto a esta métrica, también hay una mejora significativa al considerar efectos aleatorios, puesto que para este caso el error cuadrático medio es 0.00882 mientras que para la aproximación inicial de efectos fijos es de 0.02004.

2.5. ¿Efectos Fijos o Aleatorios?

Como hemos observado, la incorporación de efectos aleatorios puede ser una gran ventaja a la hora de controlar la variabilidad del modelo y de realizar predicciones en sujetos de los cuales no se posee observación. Sin embargo, esto abre una nueva disyuntiva que es cómo determinar si a una covariable se la debe considerar como un efecto fijo o aleatorio. No hay una respuesta única a esta problemática, pero existen múltiples discusiones al respecto, una de ellas puede ver en McCullagh (1997). Nosotros enumeraremos algunas consideraciones que pueden sugerir utilizar efectos aleatorios o fijos en circunstancias particulares.

1. Si una variable es controlada o manipulada por el investigador, entonces es un motivo para considerarla un efecto fijo. Por ejemplo, en un estudio sobre el efecto de una droga en la supervivencia de pacientes con cáncer, la dosis de la droga sería un efecto fijo.

2. Si la variable es un factor con un número limitado de niveles, entonces se considera un efecto fijo. Un posible ejemplo sería el sexo de un animal en un estudio de laboratorio. En cambio, si la variable es un factor con un número ilimitado de niveles, entonces se lo puede considerar un efecto aleatorio. Ejemplo: la marca de un automóvil.

3. Es preciso tener en cuenta la naturaleza de la variable: si esta es una característica que se asigna a un grupo o individuo, como la edad o el género, es probable que sea un efecto fijo. Por otro lado, si la variable es una medida tomada en diferentes individuos o grupos, como el peso o la altura, se podría incorporar un efecto aleatorio en el cual la pendiente que multiplica a la medida en cuestión dependerá del individuo o grupo en el cual se está efectuando la medición.

4. Cuando los datos están desbalanceados y se cuenta con pocas observaciones de un determinado grupo o individuo, puede ser recomendable utilizar efectos aleatorios porque, como se vio anteriormente, realiza un *trade off* entre la información dada por el grupo y la dada por la totalidad de las observaciones teniendo en cuenta la cantidad de ejemplares de cada uno. Sin embargo, cabe resaltar que esto, en algunas condiciones puede llevar a sesgos como los mencionados por George et al. (2016) donde muestran que un mal uso

de los efectos aleatorios implicó una subestimación sistemática en la tasa de mortalidad en hospitales con pocas observaciones.

5. McCuloch y Searle (2001) sugieren que " los efectos son fijos si son de interés en sí mismos o aleatorios si hay interés en la población subyacente".

Cabe aclarar que esta enumeración no pretende ser exhaustiva y que incluso puede ser contradictoria. Gelman y Hill (2007) argumentan:

> "A question that commonly arises is when to use fixed effects (in the sense of varying coefficients that are unmodeled) and when to use random effects. The statistical literature is full of confusing and contradictory advice. Some say that fixed effects are appropriate if group-level coefficients are of interest, and random effects are appropriate if interest lies in the underlying population. Others recommend fixed effects when the groups in the data represent all possible groups, and random effects when the population includes groups not in the data. These two recommendations (and others) can be unhelpful. For example, in the child support example, we are interested in these particular cities and also the country as a whole. The cities are only a sample of cities in the United States −but if we were suddenly given data from all the other cities, we would not want then to change our model. Our advice (elaborated upon in the rest of this book) is to always use multilevel modeling ("random effects"). Because of the conflicting definitions and advice, we avoid the terms "fixed" and "random" entirely, and focus on the description of the model itself (for example, varying intercepts and constant slopes), with the understanding that batches of coefficients will themselves be modeled."
>
> "Una pregunta que comúnmente surge es cuándo utilizar efectos fijos (en el sentido de coeficientes variables que no están modelados) y cuándo utilizar efectos aleatorios. La literatura estadística está llena de consejos confusos y contradictorios. Algunos afirman que los efectos fijos son apropiados si se están investigando los coeficientes a nivel de grupo, y los efectos aleatorios son apropiados si el interés se centra en la población subyacente. Otros recomiendan efectos fijos cuando los grupos en los datos representan todos los grupos posibles, y efectos aleatorios cuando la población incluye grupos que no están en los datos. Estas dos recomendaciones (y otras) pueden resultar poco útiles. Por ejemplo, en el caso del ejemplo de manutención infantil, nos interesa tanto estas ciudades en particular como el país en su conjunto. Las ciudades son solo una muestra de las ciudades en los Estados Unidos,

> pero si de repente tuviéramos datos de todas las demás ciudades, no querríamos cambiar nuestro modelo. Nuestro consejo (detallado en el resto de este libro) es utilizar siempre el modelado multinivel ("efectos aleatorio"). Debido a las definiciones y consejos conflictivos, evitamos completamente los términos "fijo" y "aleatorio" y nos enfocamos en la descripción del modelo en sí (por ejemplo, interceptos variables y pendientes constantes), con la comprensión de que los lotes de coeficientes serán modelados en sí mismos."

En conclusión, la decisión de si considerar un efecto fijo o aleatorio no es unívoca y ha de recaer en el investigador en cada situación, teniendo en cuenta las ventajas y desventajas en cada caso.

Capítulo 3

Modelo de Detección de Fraudes

En este capítulo aplicaremos el modelo logístico con efectos mixtos explicado en el capítulo anterior a la detección de fraudes en seguros automotores. En la primera sección se describirá el dataset disponible. Luego, se realizará un análisis exploratorio de los datos, con la utilización de herramientas gráficas que nos permitirán preseleccionar candidatos para una posterior selección de variables, para lo que tendremos en cuenta los criterios de maximización del área bajo la curva ROC (AUC) y minimización del Criterio de Información de Akaike. Una vez realizado esto, se procederá a un análisis de los resultados y una comparación de ambos modelos.

3.1. Datos

Disponemos de un dataset con 1000 observaciones de 40 variables relativas siniestros de automóviles ocurridos en Estados Unidos. El mismo está públicamente disponible y fue obtenido en Kaggle[1], una plataforma en línea donde los usuarios pueden participar en competiciones de inteligencia artificial y aprendizaje automático para resolver problemas utilizando datos. Las variables del dataset pueden ser clasificadas en:

- **Datos del asegurado:**
    - Edad
    - Género (M o F)
    - Nivel educativo, profesión, hobbies, estado civil

[1]El dataset puede encontrarse en https://www.kaggle.com/code/buntyshah/insurance-fraud-claims-detection

    - Meses como asegurado

- **Datos de la póliza:**
    - Fecha de inicio de vigencia de la póliza
    - Estado correspondiente (solo tenemos datos de Ohio, Illinois e Indiana)
    - Valor del deducible
    - Valor de la prima
- **Datos del auto:**
    - Marca
    - Modelo
    - Año
- **Datos del Siniestro:**
    - Fecha de ocurrencia
    - Monto del siniestro en conceptos de daños por lesiones a terceros y reparaciones patrimoniales (*property* y *vehicle*).
    - Tipo de incidente (robo, colisión, etc.)
    - Severidad del accidente (daños menores, mayores, totales, triviales)
    - Autoridades contactadas (policía, ambulancia, etc.)
    - Ubicación del incidente (estado, ciudad y dirección)
    - Hora del incidente.
    - Cantidad de vehículos involucrados
    - Cantidad de cuerpos humanos dañados
    - Cantidad de testigos
    - Variable dummy que indica si hubo daños a propiedades no automotores y otra que indica el monto.
    - Variable dummy que indica si hubo reporte policial
- **"Fraud reported"** Esta es nuestra variable dependiente a explicar. Toma el valor 1 si se detectó fraude y 0 en el caso contrario.

A primera vista, podría tener sentido incorporar cualquiera de estas variables al análisis, puesto que pueden estar relacionadas con la presencia de fraude de manera directa o indirecta. Por esto, resulta necesario hacer un adecuado análisis exploratorio de datos para visualizar el fenómeno en cuestión.

3.2. Análisis Exploratorio

El primer paso es ver qué proporción de la muestra es clasificada como fraudulenta. En la Figura 3.1 puede verse cómo el 24,75 % se trata de un fraude.

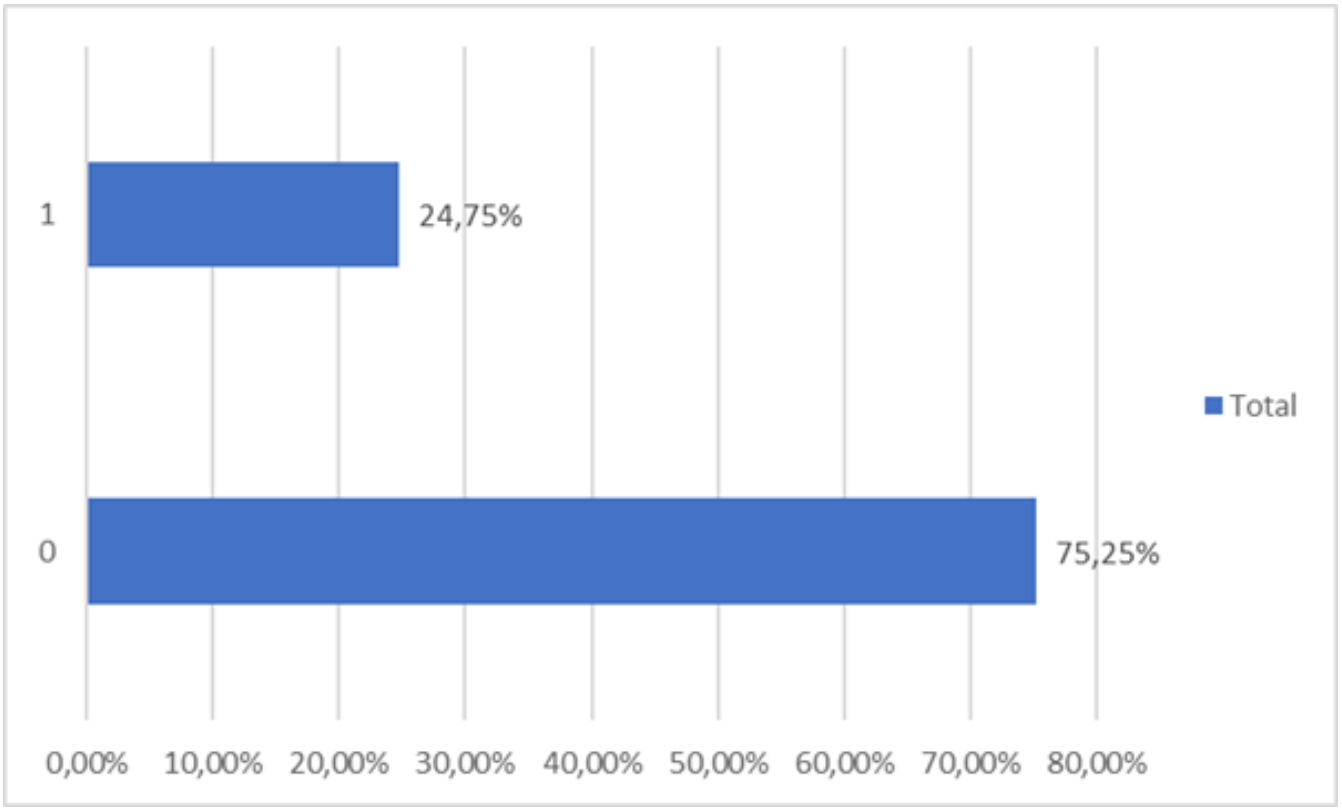


Figura 3.1: Proporción de fraudes en la muestra.

3.2.1. Variables Cualitativas

Antes de comenzar esta subsección, es necesario aclarar que a partir de este punto **el análisis exploratorio se realizará con un subconjunto de entrenamiento del** 80 % **de las observaciones del dataset para que no haya riesgo de sobreajuste.**

El próximo paso será analizar las variables cualitativas una por una (incluyendo algunas cuantitativas que solo pueden asumir un rango de datos pequeño, como es el de cantidad de vehículos involucrados) y ver si alguna segmenta los datos en una proporción significativamente distinta de ese valor, al menos en términos visuales

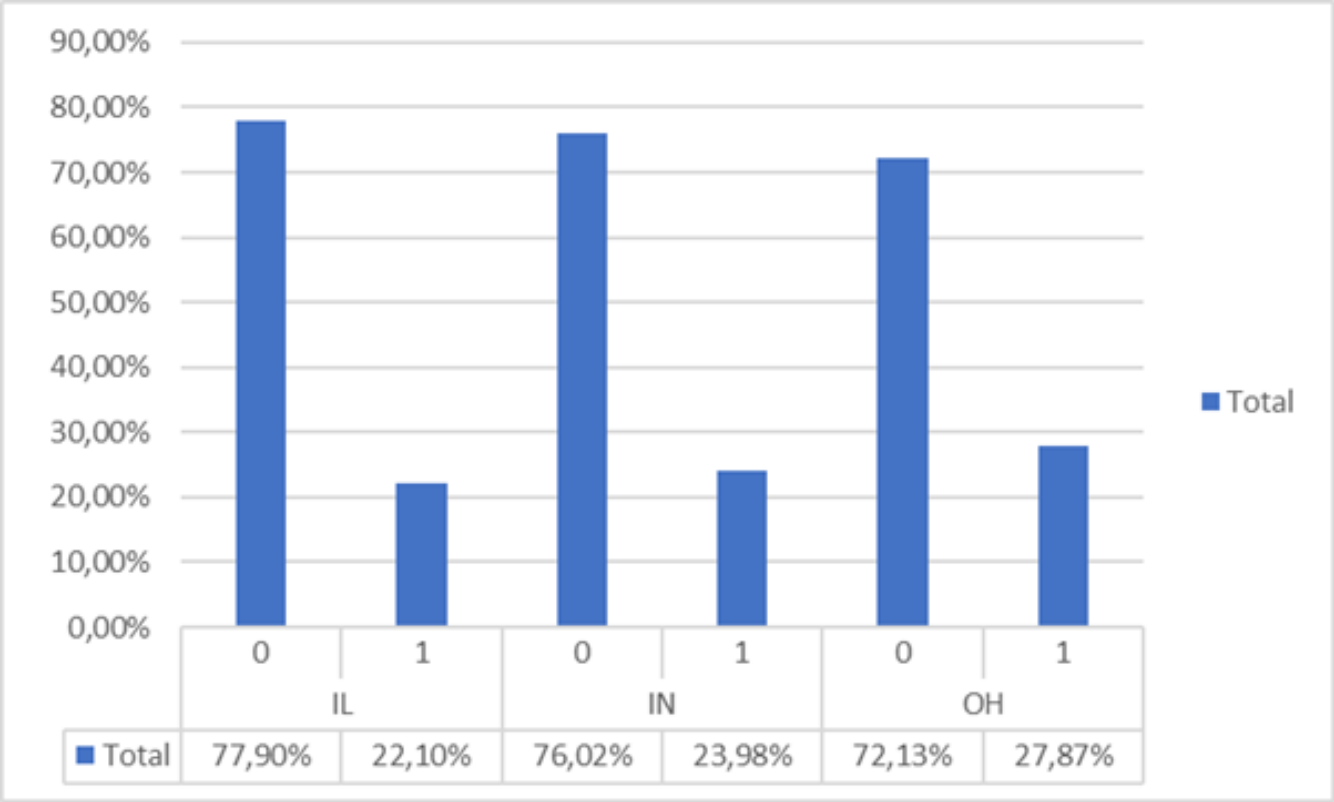


Figura 3.2: Proporción de fraudes por estado

en primera instancia. Por ejemplo, veamos el caso de la proporción de fraudes en función del Estado en el que se suscribió la póliza en la Figura 3.2.

En términos generales, podemos ver que no hay una clara diferencia de proporción por cada segmento. Este fenómeno se repite en la mayoría de las variables analizadas, mostraremos las excepciones:

- Severidad del incidente (ver Figura 3.3)
- Tipo de incidente
- Tipo de colisión
- Autoridades contactadas
- Estado en el que ocurrió el accidente (el cual no necesariamente es el mismo que el de la póliza)
- Cantidad de testigos

- Marca del automóvil
- Hobbies del asegurado

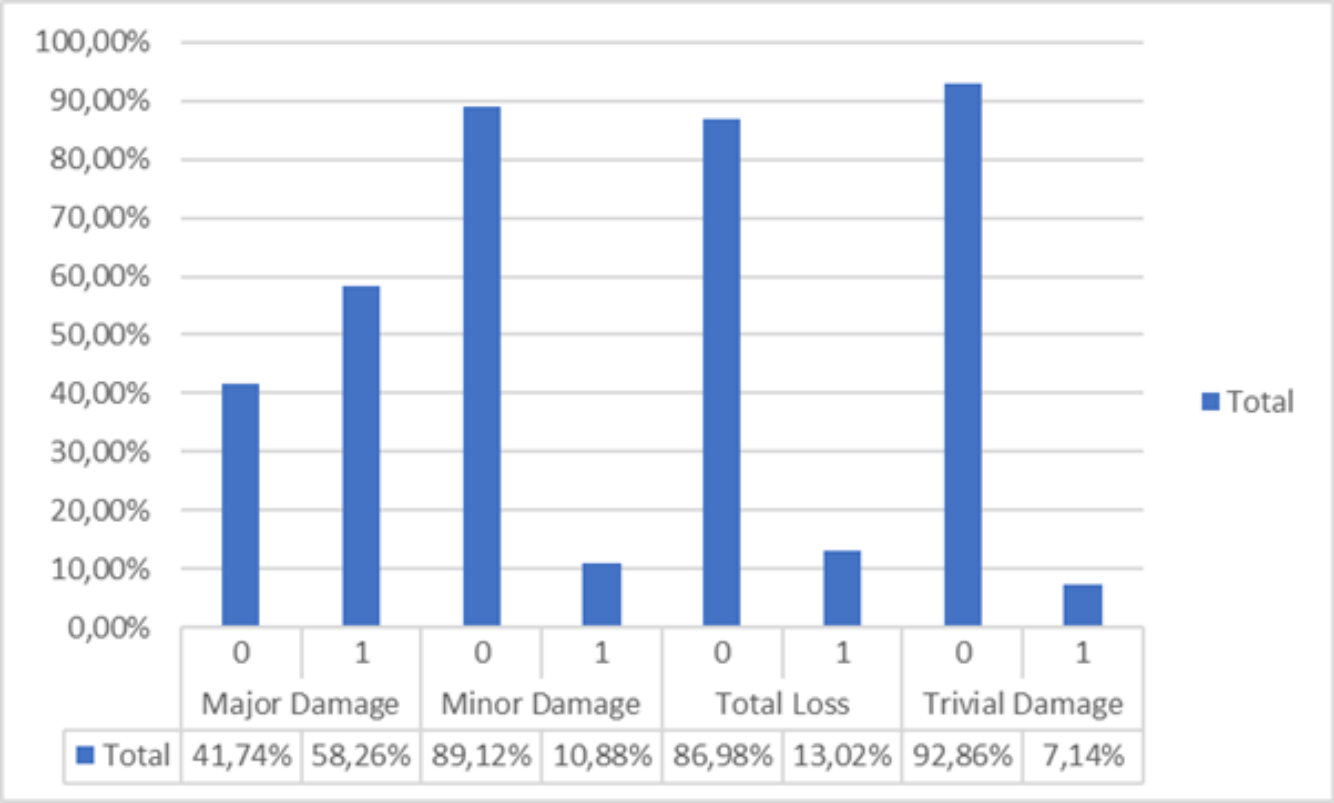
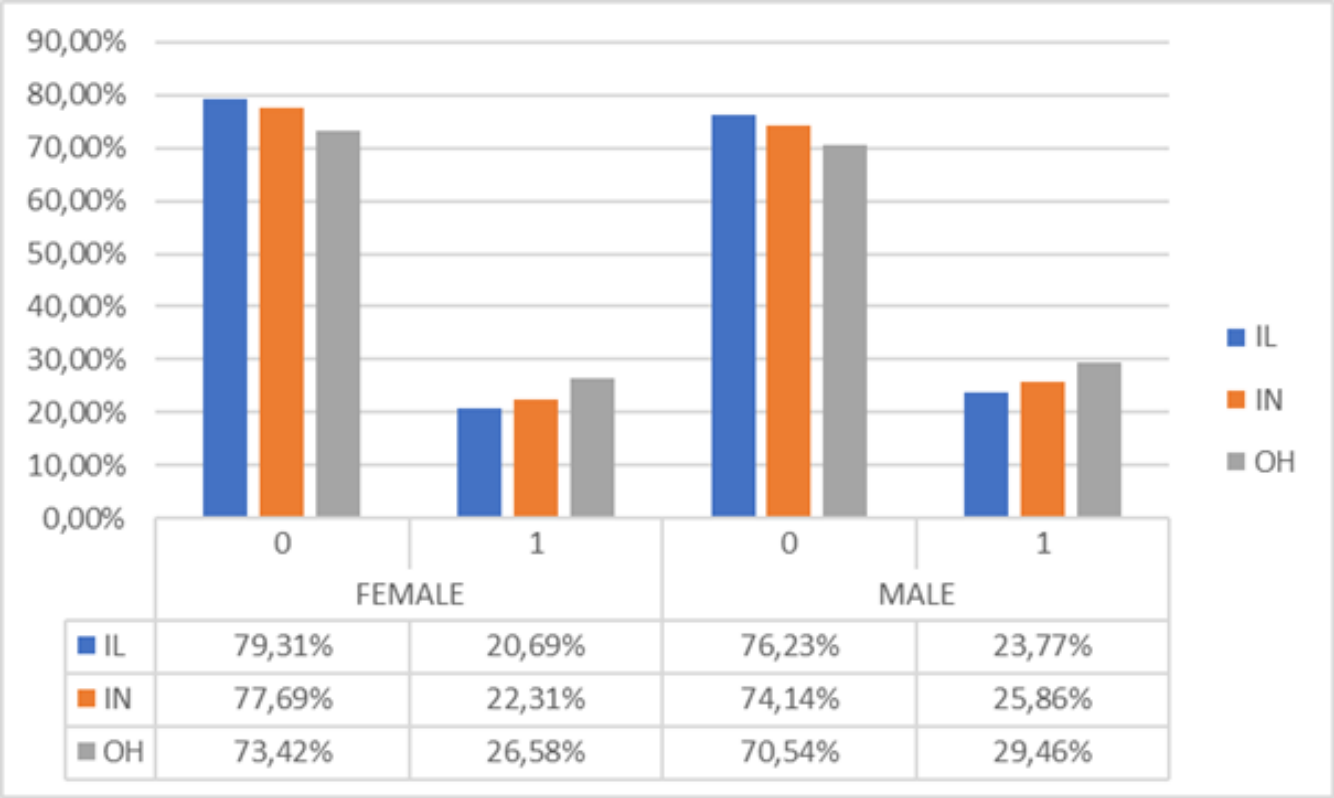


Figura 3.3: Proporción de fraudes por severidad de incidente

Por lo tanto, en primera instancia parece importante incorporar estas variables al análisis para luego hacer una **selección de variables**. Aquí es necesario hacer una aclaración: es sabido que pueden existir términos de interacción entre las variables. Para ello se procedió realizando gráficos como la Figura 3.4 que permitan advertir la presencia de interacciones relevantes. En ningún caso se encontraron términos de interacción claros, como en la Figura 3.4, en la cual puede verse la proporción de fraudes según género y estado de la póliza. Esto se debe a que la proporción de fraudes en cada estado parece ser independiente del género del titular, al menos en la muestra de entrenamiento. Por ejemplo, en Ohio, entre las mujeres, en el 73,42 % se encuentra fraude y entre los hombres, en el 70,54 %, lo cual a priori no parece ser una diferencia significativa y dada la cantidad de observaciones disponibles, es preciso evitar complejizar el modelo sin sólidos fundamentos. Por lo tanto, por estas observaciones gráficas se decidió no incorporar términos de interacción al análisis.



	FEMALE		MALE	
	0	1	0	1
IL	79,31%	20,69%	76,23%	23,77%
IN	77,69%	22,31%	74,14%	25,86%
OH	73,42%	26,58%	70,54%	29,46%

Figura 3.4: Proporción de fraudes según género y Estado de la póliza

3.2.2. Variables Cuantitativas

Para el análisis de las variables cuantitativas se ha utilizado la herramienta del diagrama de cajas o boxplot. En términos generales, no se han encontrado variables cuya influencia se perciba gráficamente, con la excepción de:

- Monto de daños físicos a terceros (*injury_claims*)
- Monto de daños patrimoniales no relacionados al vehículo (*property_claims*)
- Monto de daños patrimoniales relacionados con el vehículo (*vehicle_claims*)

Veamos el boxplot de *property_claims* en la Figura 3.5. Puede observarse que los casos fraudulentos tienden a tener mayores valores de indemnización por daños patrimoniales no relacionados con vehículos, sin embargo, no está completamente separado y la variable dista de ser determinante. El mismo fenómeno ocurre con los demás casos de montos de indemnización.

Para profundizar en el análisis, exploraremos en la Figura 3.6 la distribución con un histograma espejado. En la parte de arriba, tenemos el histograma de

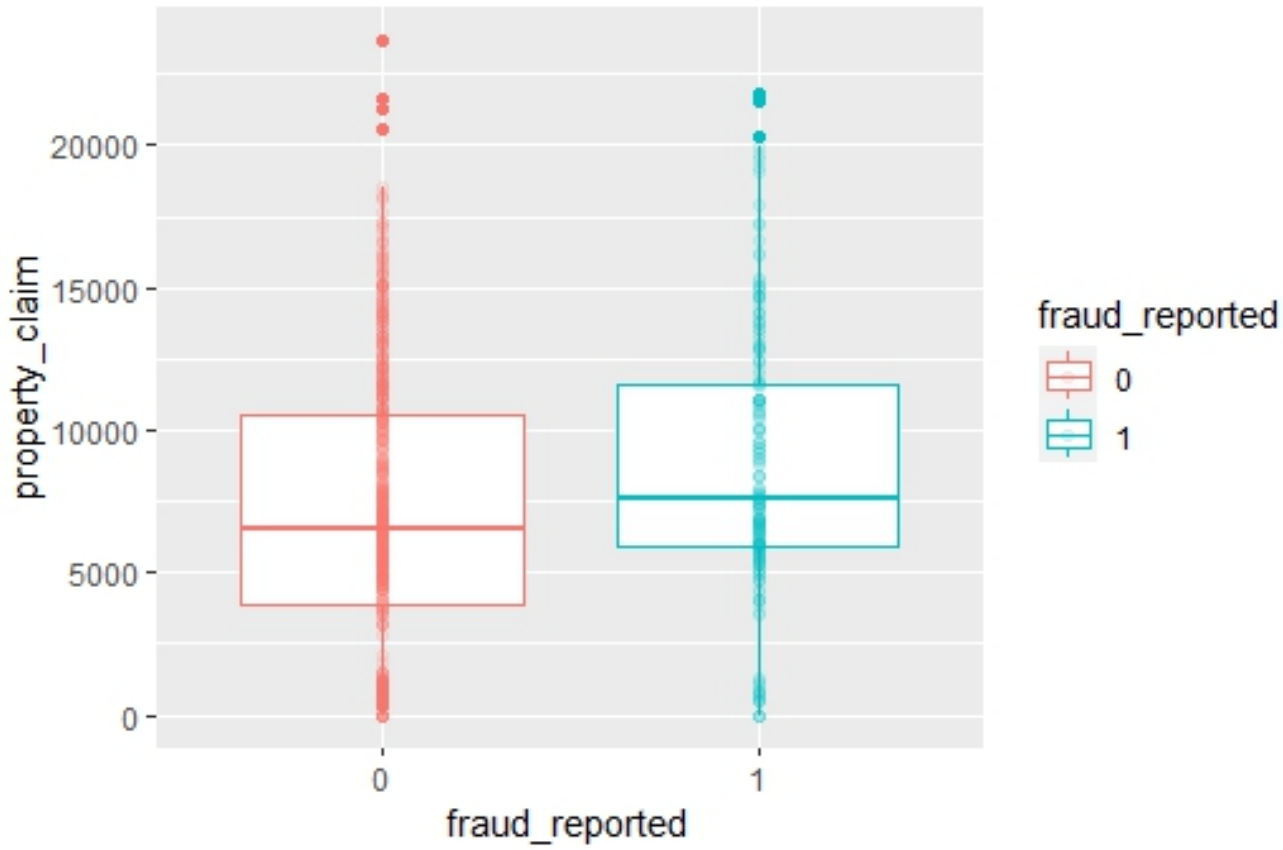


Figura 3.5: Boxplot Daños patrimoniales tipo property según presencia de Fraude.

property_claims de casos no fraudulentos; debajo, los fraudulentos. En las Figuras 3.7 y 3.8 vemos que el mismo fenómeno se repite en los demás tipos de siniestros (Injury y Vehicle).

Esta nueva visualización nos permite detectar una bimodalidad en la distribución de los montos de los siniestros que separa la muestra en dos grupos bien diferenciados que denominaremos siniestros "caros para property" y "baratos para property". Notemos que en términos relativos, la proporción caros/baratos difiere notablemente entre los casos fraudulentos y los no fraudulentos. Esto sugiere definir una variable *dummy* que valga **1** si el monto es mayor a 2000 y **0** si no (la elección del número es visual). Con este criterio, el Cuadro 3.1 que relaciona fraude y tipo de siniestro según monto property resulta ilustrativo.

En términos numéricos, entre los siniestros "baratos" solo el 11 % es fraudulento; mientras que entre los "caros", el 28 % lo es.

Como se repite el mismo fenómeno para las variables *injury_claim* y *vehicle_claim* repetimos el proceso. Mostramos los resultados en Tablas 3.2 y 3.3.

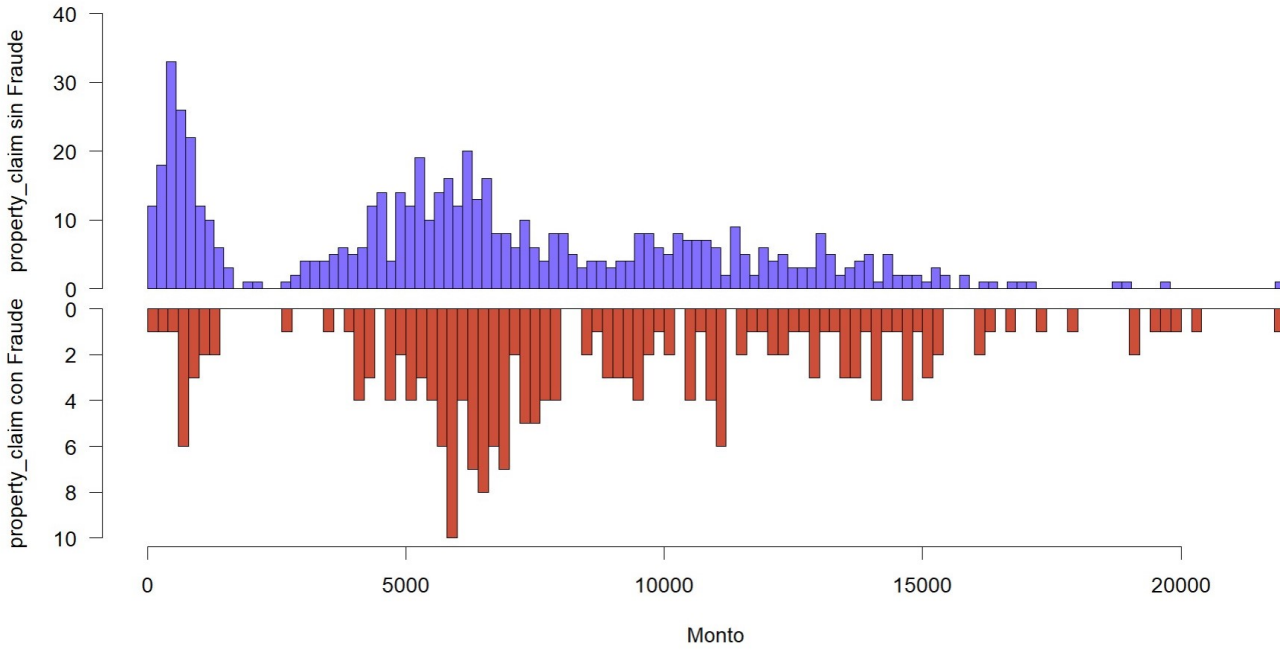


Figura 3.6: Histograma Daños patrimoniales tipo property según presencia de Fraude.

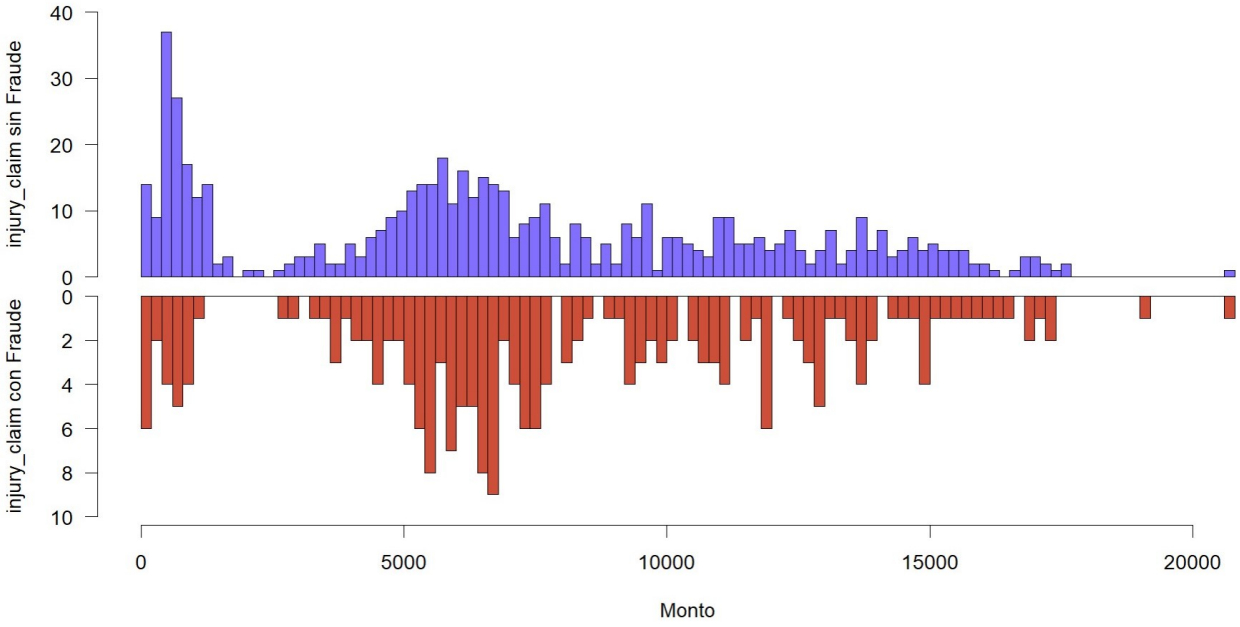


Figura 3.7: Histograma Daños tipo injury según presencia de Fraude.

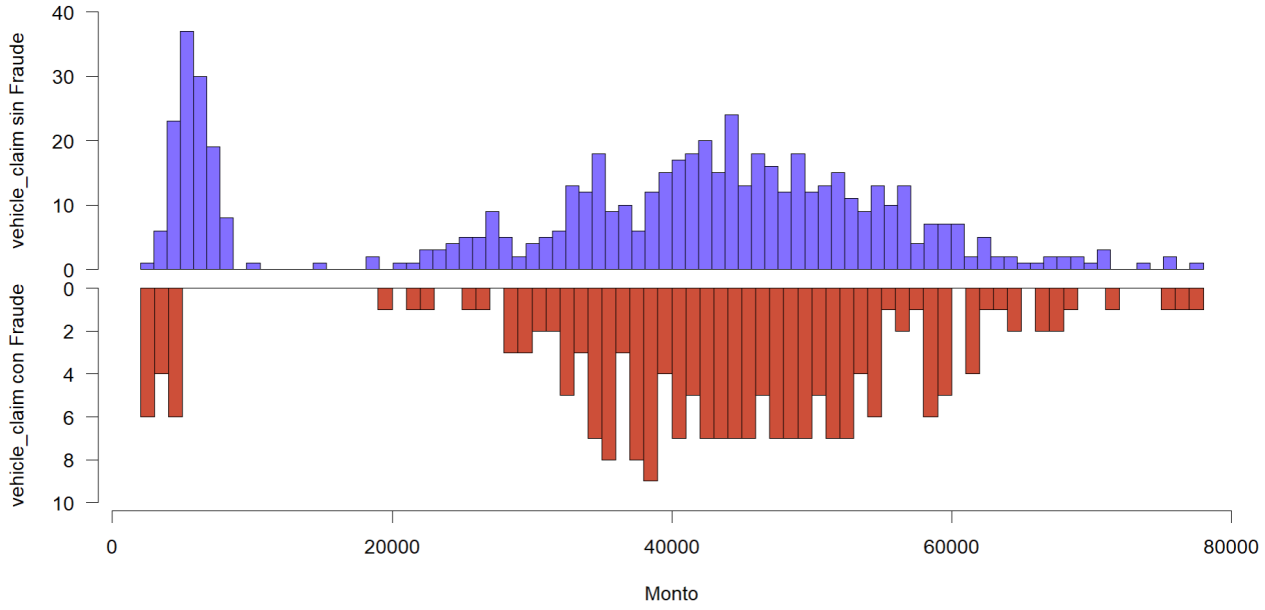


Figura 3.8: Histograma Daños patrimoniales tipo vehicle según presencia de Fraude.

	No Fraude	Fraude
Property Barato	135	17
Property Caro	467	181

Cuadro 3.1: Tabla de frecuencias absolutas Property-Fraude.

	No Fraude	Fraude
Injury Barato	136	19
Injury Caro	466	179

Cuadro 3.2: Tabla de frecuencias absolutas Injury-Fraude.

	No Fraude	Fraude
Vehicle Barato	127	13
Vehicle Caro	475	185

Cuadro 3.3: Tabla de frecuencias absolutas Vehicle-Fraude.

En conclusión, en una primera instancia nuestros candidatos a covariables cuan-

titativas son el monto de los siniestros según tipo de daño y las dicotomizaciones generadas a partir de ellos

3.3. Selección de Variables

Después de realizar un análisis exploratorio de los datos, hemos encontrado un conjunto de posibles covariables explicativas, que incluyen características generadas a partir de los datos, como los siniestros "baratos/caros" mencionados anteriormente. Se ha decidido que las variables Hobbies del Asegurado y Marca del Automóvil deben ser consideradas como efectos aleatorios, puesto que presentan muchos niveles, e incluso puede ocurrir que aparezcan nuevos para futuras predicciones, puesto que puede haber hobbies y marcas que no estén dentro de nuestro dataset de entrenamiento. A continuación, seguiremos un criterio *forward*[2] para elegir qué variables utilizar entre todas las candidatas encontradas, tanto para las que suponen efectos fijos como aleatorios. Esto significa que se irán añadiendo variables a nuestro modelo una a una, evaluando cómo afecta cada una de ellas al rendimiento del modelo, medido a través de la métrica AUC (área bajo la curva) en el conjunto de datos de prueba. Finalmente, utilizaremos solo aquellas variables que contribuyen significativamente al rendimiento del modelo, maximizando el AUC en el conjunto de testeo.

[2]En un método de selección *forward* se parte de un modelo sin covariables fuera del intercepto y van agregándose de a una las que mejoren la métrica a optimizar.

Selección de variables con el criterio de maximización de AUC con un método Forward

1. **Inicialización:** Separamos a nuestro conjunto en un subconjunto de entrenamiento y testeo. Le calculamos el AUC al modelo que no incluye ninguna de las covariables fuera del intercepto (con lo cual nuestro conjunto de covariables inicial es nulo). Este será nuestro AUC base.
2. **Cálculo de AUC:** Para cada uno de los candidatos a variables explicativas lo agregamos provisoriamente al conjunto de covariables y calculamos el AUC, considerándolo como efecto fijo o aleatorio según se haya predefinido.
3. **Actualización:** Si el máximo de los AUC calculados en el paso anterior resulta mayor que el AUC base con una diferencia mayor que 0,01, actualizamos el AUC base a dicho valor, agregamos la variable al conjunto de covariables y la quitamos del conjunto de candidatos a variables explicativas.
4. **Criterio de Cierre:** Si el AUC base no se modificó en el paso anterior, el proceso termina; si no, se vuelve al paso 2.

Por ejemplo, veamos el primer paso, de selección de variables. En la Figura 3.9 puede apreciarse un gráfico simultáneo de la curva ROC que resulta de incorporar cada variable al modelo. En rojo, está la que incorpora la variable cualitativa "severidad del incidente"; en azul claro, las demás. Puede observarse que esta variable resulta ser la que mayor aporte hace al AUC, cuyo valor es de 0,81.

En el segundo paso, se incorpora la variable "hobbies del asegurado", que es considerada un efecto aleatorio y lleva el AUC a 0,91. Resulta ostensible en la Figura 3.10 la dominancia de la curva que la incluye como covariable respecto a las demás. El proceso termina en este paso porque la incorporación de ninguna variable mejora el AUC.

Por lo tanto, iterando el proceso con el método *forward* llegamos a la conclusión de que la selección de variables a utilizar resulta de incorporar la variable severidad del incidente como efecto fijo y hobbies del asegurado como efecto aleatorio. Finalmente, el modelo seleccionado queda:

$$Fraude_{it}|u_i \sim Be(\pi_{it}) \tag{3.1}$$

$$logit(\pi_{it}) = \beta_0 + \beta_1 severidad_{it} + u_i \tag{3.2}$$

$$u_i \sim \mathcal{N}(0, \sigma_u^2) \quad iid \tag{3.3}$$

donde u_i es el efecto aleatorio del hobby i del asegurado en cuestión sobre el *logit* de la probabilidad de realizar fraude.

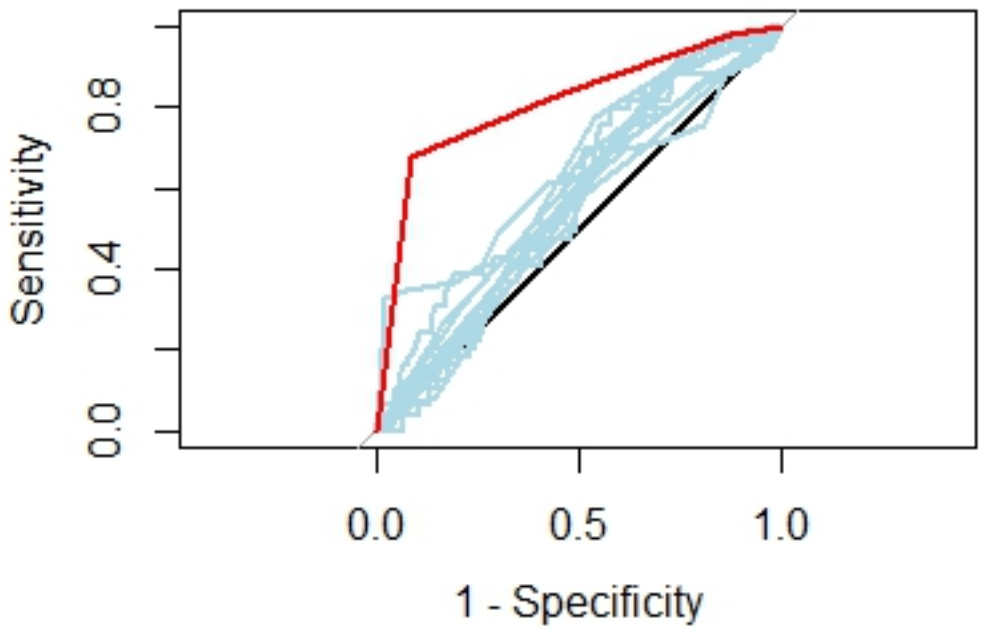


Figura 3.9: Curvas ROC según qué variable se incorpora al modelo en el **primer paso**. La curva roja es la variable seleccionada (*incident_severity*)

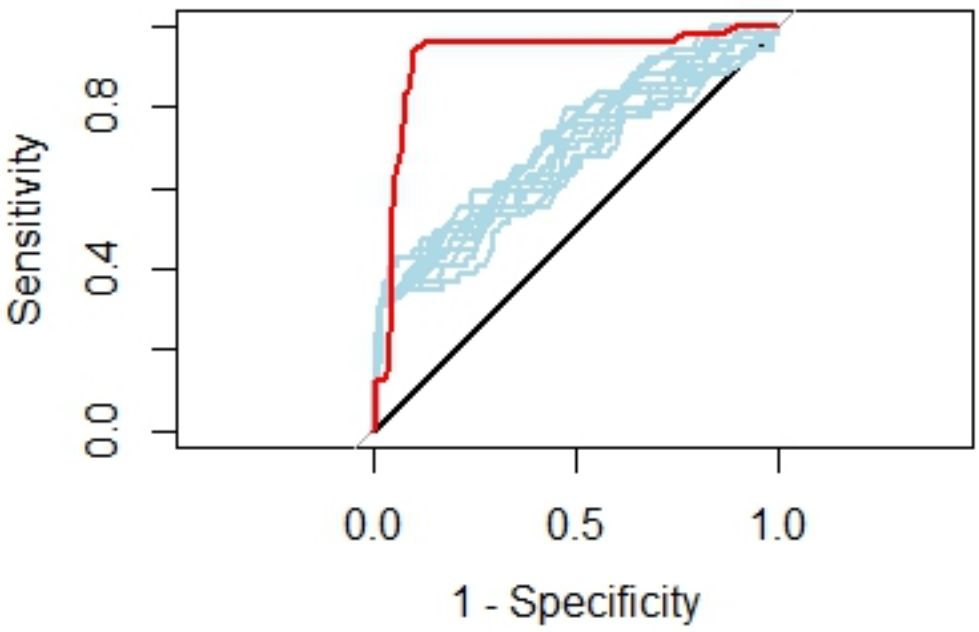


Figura 3.10: Curvas ROC según qué variable se incorpora al modelo en el **segundo paso**. La curva roja es la variable seleccionada (*insured_hobbies*)

Otro enfoque posible de selección de variables sería el de minimización de Criterio de Información de Akaike (AIC). Al igual que en AUC, la dirección del proceso será *forward*. Es decir, se partirá de un modelo que solo incluye al intercepto y se irá agregando la variable que más reduzca el valor de AIC (como efecto aleatorio o fijo, según corresponda). El proceso termina cuando ninguna lo reduce. El resultado puede verse en el Cuadro 3.4. Podemos observar cómo se sugieren las mismas variables en el mismo orden en los primeros dos pasos. Además, puede verse cómo las mismas contribuyen notablemente a disminuir el AIC. Sin embargo, la diferencia estriba en que se sugiere incorporar la variable "tipo de colisión". Pero si uno se detiene a observar el aporte que hace en términos de reducción de AIC, el mismo parece ser muy pequeño (0.18). Por lo tanto, le agregamos una columna al cuadro que incluye la deviance del modelo de cada paso y otra con el p-valor del test de deviance.

Paso	Variable incorporada	AIC	Var. AIC	Deviance	P-val test Chisq
0	(Intercept)	897.31	-	895.32	-
1	incident_severity	718.94	178.36	710.95	<2e-16 ***
2	insured_hobbies	625.75	93.19	615.75	<2e-16 ***
3	collision_type	625.57	0.18	609.58	0.10

Cuadro 3.4: Resultados del proceso de selección forward por minimización AIC

Tanto el enfoque de maximización de AUC como el de minimización de AIC parecen indicar que las variables "severidad del incidente" y "hobbies del asegurado" deben ser incorporadas al análisis. Con respecto a "tipo de colisión", por su bajo aporte al AIC, su no significancia del test de deviance y por evitar complejizar innecesariamente el modelo será excluida del análisis y el modelo propuesto será el ya mencionado, que maximiza el AUC.

Finalmente, en el Cuadro 3.5 encontramos la tabla resumen que sintetiza el proceso realizado hasta este punto.

3.4. Análisis de resultados

3.4.1. Estimaciones y Predicciones

En primer lugar, haremos un análisis de las estimaciones y predicciones del modelo.

Nombre de variable	Tipo	Preseleccionada	Efecto	Seleccionada
age	Numérica	-	-	-
authorities_contacted	Categórica binaria	Si	Fijo	-
auto_make	Cualitativa	-	-	-
auto_model	Cualitativa	Si	Aleatorio	-
auto_year	Numérica (Entero)	-	-	-
bodily_injuries	Numérica (Entero)	-	-	-
collision_type	Categórica	Si	Fijo	-
incident_city	Cualitativa	-	-	-
incident_date	Fecha/Hora	-	-	-
incident_hour_of_the_day	Fecha/Hora	-	-	-
incident_location	Cualitativa	-	-	-
incident_severity	Categórica	Si	Fijo	Si
incident_state	Cualitativa	Si	Fijo	-
incident_type	Categórica	Si	Fijo	-
injury_claim	Numérica continua	-	-	-
injury_dummy	Categórica binaria	Si	Fijo	-
insured_hobbies	Cualitativa	Si	Aleatorio	Si
insured_sex	Categórica binaria	-	-	-
months_as_customer	Numérica (Entero)	-	-	-
number_of_vehicles_involved	Numérica (Entero)	-	-	-
police_report_available	Categórica binaria	-	-	-
policy_annual_premium	Numérica continua	-	-	-
policy_bind_date	Fecha/Hora	-	-	-
policy_deductable	Numérica continua	-	-	-
policy_number	Numérica (Entero)	-	-	-
policy_state	Cualitativa	-	-	-
property_claim	Numérica continua	-	-	-
property_dummy	Categórica binaria	Si	Fijo	-
vehicle_claim	Numérica continua	-	-	-
vehicle_dummy	Categórica binaria	Si	Fijo	-
witnesses	Numérica (Entero)	Si	Fijo	-

Cuadro 3.5: Tabla resumen de las variables involucradas

Efectos Fijos Los coeficientes estimados de los efectos fijos junto con los p-valores del test de significancia individual de Wald se encuentran en el Cuadro 3.6.

Como puede apreciarse en dicha tabla, solo tenemos una variable cualitativa que es *incident_severity* y toma cuatro valores posibles: *Major Damage* (que es el

Efectos Fijos	Estimación	Std. Err.	P valor
(Intercept)	0.5612	0.3171	0.0767.
incident_severity Minor Damage	-3.047	0.2858	< 2e-16 ***
incident_severity Total Loss	-2.9827	0.3065	< 2e-16 ***
incident_severity Trivial Damage	-3.5058	0.5505	1.91e-10 ***

Cuadro 3.6: Estimación de Efectos Fijos para el Modelo

valor base y por eso no aparece en la tabla), *Minor Damage, Total Loss y Trivial Damage* que tienen un coeficiente estimado de $-3{,}04$, $-2{,}98$ y $-3{,}50$ respectivamente. El hecho de que mayores daños impliquen mayores probabilidades de fraude es consistente con la experiencia y con lo esperado a priori, además de lo que se vio en la Figura 3.3. Además, todos los niveles de la variable *incident_severity* son fuertemente significativos para tests de significancia individual de nivel $\alpha = 0{,}01$. Sin embargo, lo que esto nos dice es que son significativamente distintos de la base (es decir, de *Major Damage*), pero no nos garantiza la significancia conjunta de incorporar dicha variable.

Para testear la significancia conjunta tomamos como modelo base aquel que excluye a *incident_severity*. Es decir:

$$Fraude_{it}|u_i \sim Be(\pi_{it}) \tag{3.4}$$

$$logit(\pi_{it}) = \beta_0 + u_i \tag{3.5}$$

$$u_i \sim \mathcal{N}(0, \sigma_u^2) \;\; iid \tag{3.6}$$

Finalmente, con la función *anova* de R, hacemos un test de cociente de verosimilitud respecto al modelo que sí incorpora la severidad del incidente y obtenemos los resultados visibles en el Cuadro 3.7. Puede concluirse que la variable es *incident_severity* es estadísticamente significativa.

Modelo	Deviance	Dif Deviance	P-val test
Modelo sin incident_severity	832.98	-	-
Modelo Propuesto	615.75	217.23	<2e-16 ***

Cuadro 3.7: Test cociente de verosimilitud para *incident_severity*

Efectos Aleatorios En lo que respecta a la **predicción** de los efectos aleatorios, la única variable involucrada en él es el Hobby del asegurado. El Cuadro 3.8

nos muestra las predicciones según hobby para el modelo, decrecientemente por valor absoluto. Lo más notorio es el valor que se le asigna a aquellos que realizan ajedrez y crossfit, los cuales parecen estar relacionados con tendencias al fraude. La explicación de este fenómeno, a priori, no parece ser intuitiva. Si graficamos la proporción de fraudes según hobby, la influencia de estas variables se vuelve evidente. Esto puede observarse en la Figura 3.11.

Hobby	Pred. Modelo
chess	3.2679382
cross-fit	2.6841093
camping	-1.7190576
kayaking	-1.1210522
sleeping	-0.8395051
exercise	-0.6484961
golf	-0.6395591
paintball	-0.6056873
bungie-jumping	-0.5931328
yachting	0.5789181
dancing	-0.4969491
hiking	0.4918120
board-games	0.4870020
basketball	-0.4274640
movies	-0.4138440
base-jumping	0.2222892
reading	0.1378271
skydiving	-0.1184907
polo	0.1111808
video-games	0.0807666

Cuadro 3.8: Predicción de Efectos Aleatorios

3.4.2. Análisis de Residuos

En la regresión logística, al igual que en otros modelos GLM, es posible definir los residuos como la diferencia entre los valores observados y los valores esperados. Sin embargo, encontramos la diferencia de que en la regresión logística, cuando esta se usa para clasificación y no para estimación de proporciones subyacentes, los datos de nuestra variable explicada son discretos, por lo que los residuos también lo son. Por esta razón, los gráficos de residuos usuales no suelen ser de mucha utilidad. Gelman y Hill (2008) sugiere, en su lugar, utilizar un gráfico de residuos

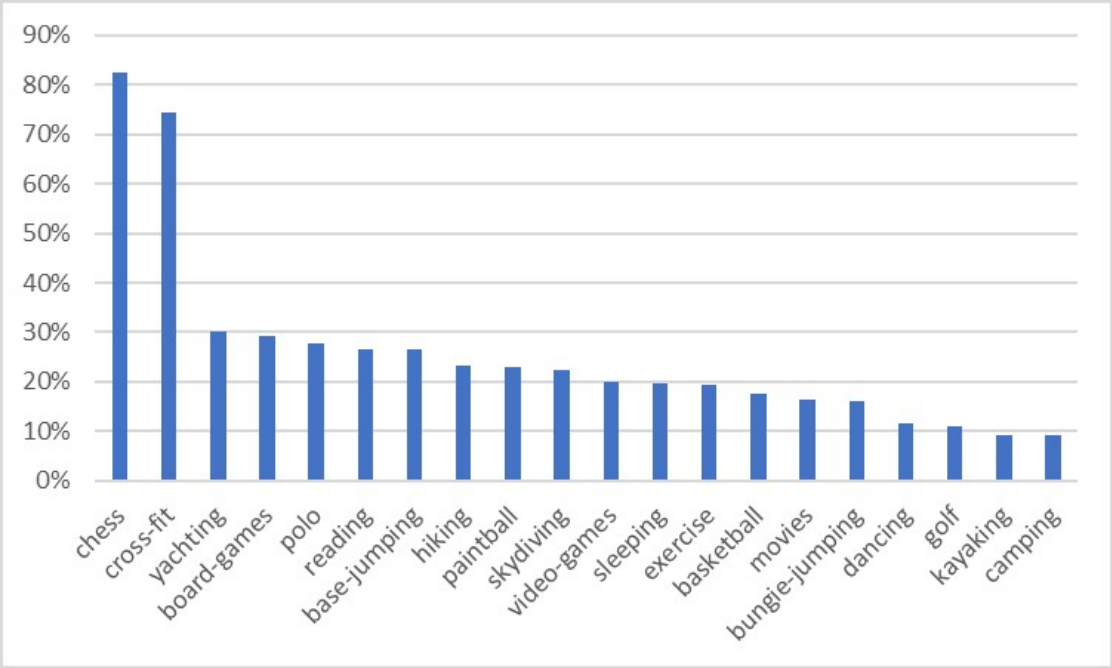


Figura 3.11: Proporción de Fraudes según Hobby

por categorías, en el que se dividen los datos en diferentes categorías o "bins" según sus valores ajustados, y se muestra la relación entre el residuo promedio y el valor ajustado promedio para cada categoría. De esta manera, es posible obtener una mejor comprensión de los patrones y tendencias en los residuos y cómo estos varían en función de los valores ajustados. En la Figura 3.12 se puede apreciar el gráfico resultante. En términos generales, en ninguno de los casos puede apreciarse un patrón inesperado o "patológico".

Otro enfoque alternativo es utilizar el paquete *DHARMa*[3], basado en los *Randomized quantile residuals* cuyo desarrollo teórico puede encontrarse en Dunn y Smyth (2018).

Este desarrollo permite hacer un QQ-plot para comparar la distribución de los residuos DHARMa contra los valores esperados. Esto puede verse en la Figura 3.13. En efecto, puede considerarse que no hay comportamientos indeseables y esto puede ser acompañado del test de Kolmogorov Smirnov que con un p-valor de 0,41 nos indica que no hay motivos para pensar que la distribución observada y la esperada difieran significativamente. Por lo tanto, podemos asumir que el modelo parece bien especificado.

[3]Puede encontrarse información al respecto en https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html

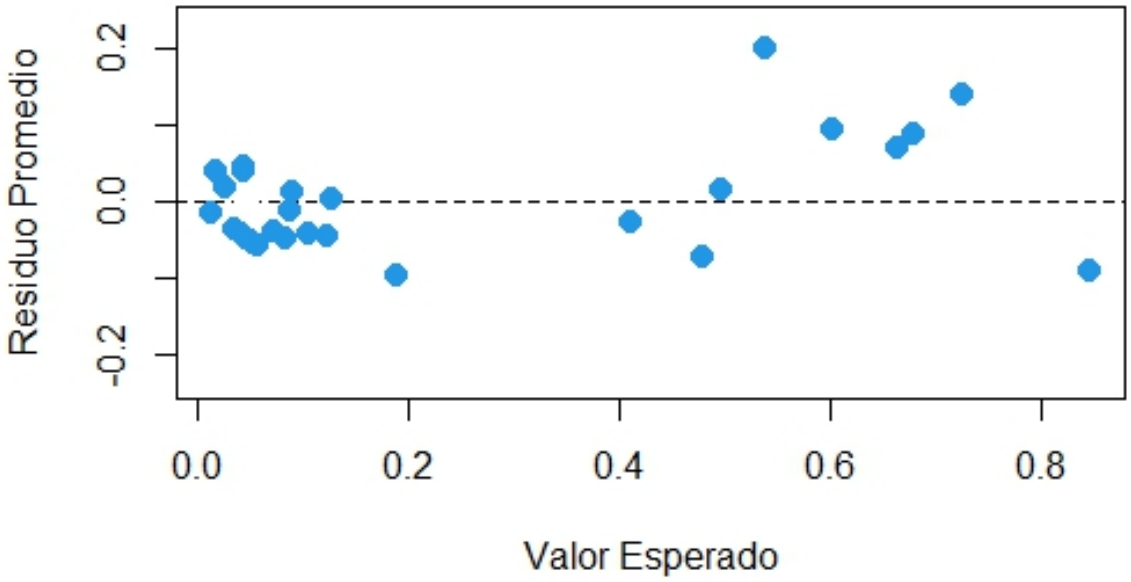


Figura 3.12: Binned plot de los residuos del modelo.

3.4.3. Bondad de Ajuste

En esta sección analizaremos los resultados del modelo seleccionado con diversas métricas, las cuales serán realizadas con el método *Leave one out cross validation* (LOOCV). Es decir, para predecir si una observación es fraudulenta o no se entrenará el modelo con todas las demás observaciones (excluyendo la que se quiere predecir) y se la clasificará de esta manera. Así, evitamos el sobreajuste y tenemos un conjunto de testeo del tamaño de la muestra, con lo cual es más representativo. Cabe preguntarse por qué no se lo utilizo en la sección anterior, esto se debe a que el costo computacional es alto para la cantidad de iteraciones que se requerían.

Tasa de Aciertos La primera métrica a evaluar será la **tasa de aciertos**. Si en cada modelo se clasificara una observación como fraudulenta, si la probabilidad estimada es mayor que el 50 %, la tasa de aciertos sería 84,8 %. Sin embargo, por muchos motivos podríamos optar por cambiar este umbral de probabilidad. Por ello, graficamos en cuál es la tasa de aciertos del modelo en función del umbral de

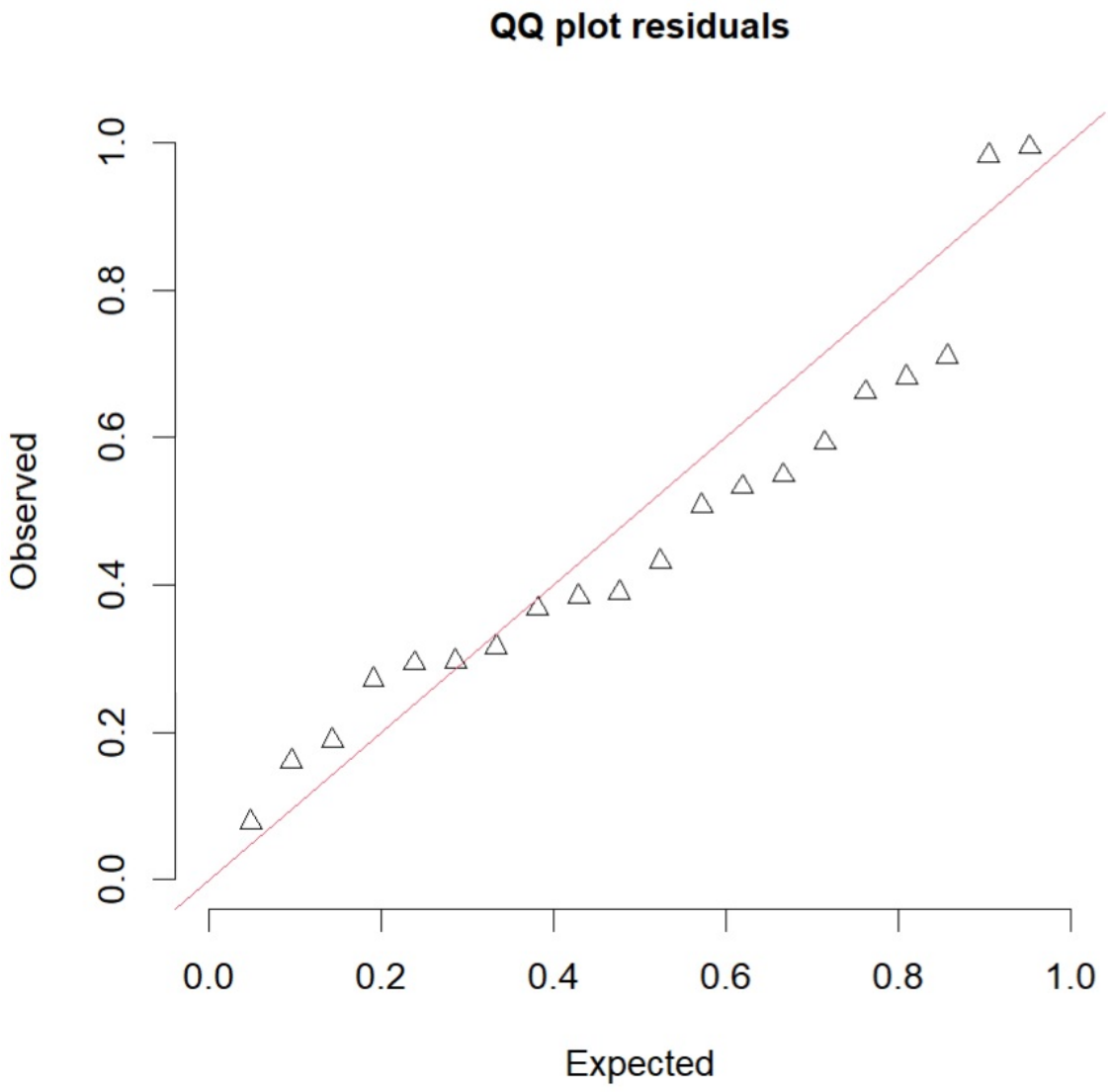


Figura 3.13: QQ-plot DHARMa

probabilidad. Véase la Figura 3.14 en la cual además se marca con un punto el umbral que maximizan la tasa de acierto del modelo. Este óptimo empírico se da cuando el umbral de probabilidad es 36 % y la tasa de aciertos resulta 86,1 %.

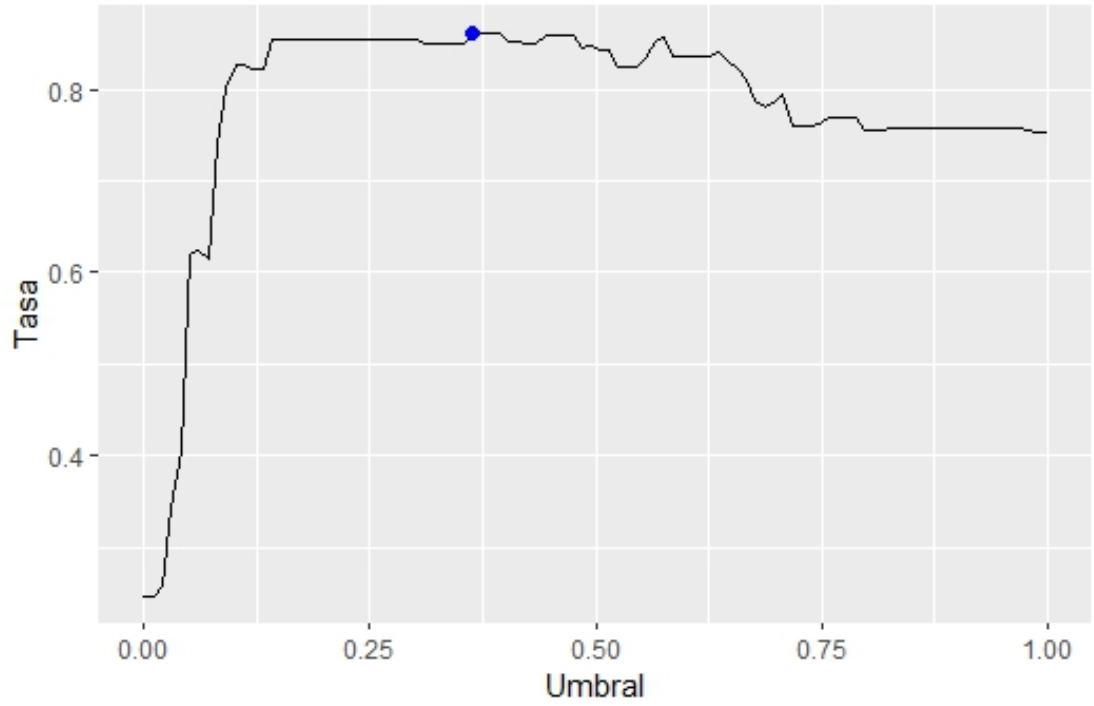


Figura 3.14: Tasa de aciertos en función de umbral de probabilidad. El punto azul indica el máximo valor alcanzado.

La tasa de acierto puede ser engañosa en conjuntos de datos desbalanceados: Si tenemos un conjunto de datos con una clase dominante, es posible que un modelo tenga una tasa de acierto alta simplemente prediciendo siempre la clase dominante. Esto puede dar una sensación de que el modelo es más preciso de lo que realmente es. Teniendo en cuenta que efectivamente estamos en datos desbalanceados por lo visto en el análisis exploratorio, es necesario considerar otras métricas en el análisis.

Tablas de Confusión Podemos observar en el Cuadro 3.9 la tabla de confusión del modelo con el umbral de probabilidad que maximiza su tasa de aciertos.

Esta tabla lleva implícita una sensibilidad del 87,0 % y una especificidad de 85,7 %

	No Fraude	Fraude
Predicho No Fraude	646	32
Predicho Fraude	107	215

Cuadro 3.9: Tabla de Confusión del Modelo con el umbral de probabilidad que maximiza la tasa de aciertos.

Curvas ROC Realizamos las curvas ROC con estimaciones efectuadas con LOOCV. Las graficamos en la Figura 3.15.

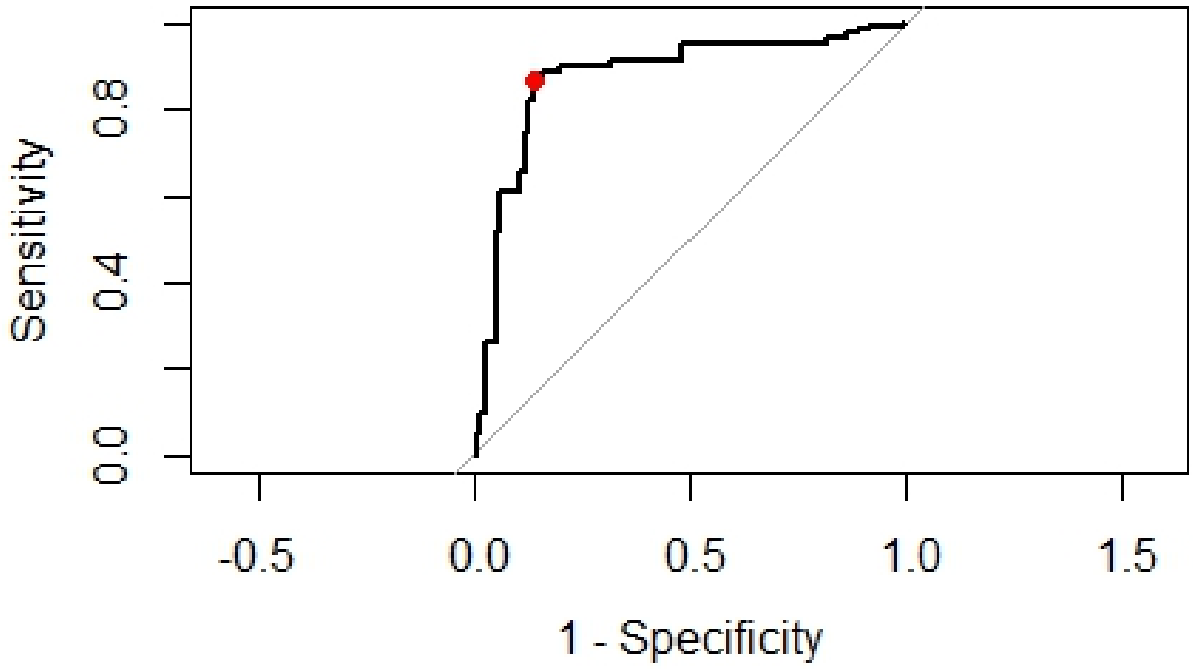


Figura 3.15: Curvas ROC con estimaciones LOOCV

Por último, analizaremos la curva ROC. En este caso, nos encontramos con que la curva construida con estimaciones LOOCV no difiere cualitativamente de la obtenida en la sección anterior, con la excepción de una mayor suavidad en la misma. Puede verse que en este caso, el AUC es 0,87 y se grafica el punto que maximiza la tasa de aciertos en rojo, el cual puede observarse que se encuentra en términos relativos cerca del punto $(1, 1)$. Con lo cual, parece razonable considerar un umbral de probabilidad de 36 %.

Capítulo 4

Conclusiones

En este trabajo hemos modelado los fraudes en seguros automotores por medio de un modelo logístico con efectos mixtos. En términos generales, este modelo tiene muchas ventajas como permitir una mejor estimación de las variables de interés y una mayor precisión en los resultados obtenidos. Además, el modelo logístico con efectos mixtos permite analizar la influencia de variables individuales y grupales en la probabilidad de fraude. A través de este análisis, se pueden identificar patrones y tendencias que pueden ser utilizados para desarrollar estrategias y medidas de prevención y detección de fraudes en seguros automotores. En conclusión, el modelo logístico con efectos mixtos se presenta como una herramienta útil y eficaz para el análisis de fraudes en seguros automotores, y su aplicación puede contribuir a la mejora de la eficacia en la lucha contra el fraude en esta industria, en la cual, como se ha mencionado, los modelos lineales generalizados tienden a realizarse con efectos fijos, por lo tanto, este trabajo puede funcionar como antecedente al respecto. Además, una reducción en la cantidad de Fraudes redundará en una prima pura más baja, lo cual implicará en beneficios para todo el ecosistema de seguros.

En particular, a través de nuestro análisis, se pudo observar que ciertas variables tienen un poder predictivo considerable, como es el caso de la severidad del incidente como efecto fijo y los hobbies del asegurado como efecto aleatorio.

Hemos desarrollado una selección de variables que tuvo varios enfoques. En primer lugar, uno exploratorio que sirvió como una poda inicial para descartar candidatos a covariables. Luego, se hizo un balance entre la maximización del AUC y la minimización del AIC, con una dirección *forward*, en el cual se parte de un modelo cuya única covariable explicativa es el intercepto y progresivamente se van agregando variables que optimicen la métrica de interés.

El modelo ha mostrado un poder predictivo no despreciable. En primer lugar, porque se logra una tasa de aciertos de 86,1 %, mientras que el balance de los datos es 75 % no fraudulentos y 25 % fraudulentos. En segundo lugar, a causa de

las observaciones relativas a la sensibilidad, especificidad y Curvas ROC, con su respectivo AUC de 0,87. A efectos de tener una comparación informal del poder predictivo con otros modelos, podemos observar la publicación en Kaggle de Nitesh Yadav *"Insurance Fraud Detection (Using 12 models)"*[1] en la cual se implementan 12 modelos de clasificación, principalmente basadas en árboles y se obtienen tasas de acierto de aproximadamente 80 %. Esto puede visualizarse en la Figura 4.1.

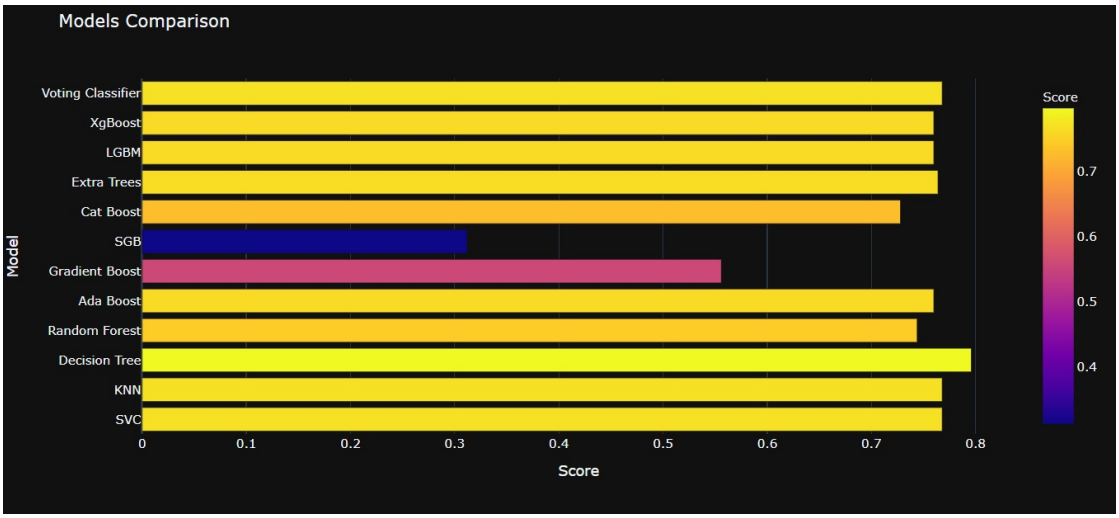


Figura 4.1: Comparación de 12 modelos de la publicación *"Insurance Fraud Detection (Using 12 models)"*

Además, el modelo podría seguir siendo utilizado sin dificultades ante la aparición de asegurados con hobbies de los cuales no se tiene experiencia previa, e incluso ir incorporándolos y "aprender" de ellos a medida que se acumulan observaciones. Esta es una gran diferencia respecto a lo que podría obtenerse con un modelo logístico de efectos fijos.

Por otro lado, cabe destacar la poca cantidad de covariables que tiene en cuenta el modelo, sin tener una pérdida de poder predictivo. Como hemos visto, solo considera la Severidad del incidente como efecto fijo y los hobbies del asegurado como efecto aleatorio. En conclusión, consideramos que el modelo puede ser útil para la detección de fraudes como primera instancia. Pero sería importante actualizar los datos de entrenamiento a medida que estos se vayan acumulando y analizar permanentemente si vale la pena incorporar nuevas variables.

[1]La misma puede encontrarse en https://www.kaggle.com/code/niteshyadav3103/insurance-fraud-detection-using-12-models/notebook

Bibliografía

[1] Abramowitz, M. & Stegun, I.A. *Handbook of Mathematical Functions.* National Bureau of Standards. Washington. 1964

[2] Agresti, A. *Categorical Data Analysis.* 3rd Edition. Wiley Series in probability and statistics. 2013.

[3] Antonio, K. Berlaint, J. *Actuarial statistics with generalized linear mixed models.* Vol. 40. Insurance: Mathematics and Economics. 2007.

[4] Dunn, P. Smyth, G. *Generalized Linear Models With Examples in R.* Springer Texts in Statistics. 2018.

[5] Fahrmeir, L. Kaufmann, H. *Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models.* Institute of Mathematical Statistics. 1985.

[6] Gelman, A. & Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press. 2007.

[7] George, Rosenbaum, et al. *Mortality Rate Estimation and Standardization for Public Reporting: Medicare's Hospital Compare.* University of Pennsylvania. 2016.

[8] Hosmer, D. Lemeshow, S. *Applied Logistic Regression* Wiley Series in probability and statistics. 2000.

[9] James, Gareth, et al. *An introduction to Statistical Learning.* Vol. 112. New York: springer, 2013.

[10] Jiang, J. *Linear and Generalized Linear Mixed Models and Their Applications.* Springer. 2007.

[11] McCullagh, P. Nelder, J.A. *Generalized Linear Models.* 2nd Edition. Chapman Hall, New York. 1989.

[12] McCulloch, C. *An Introduction to Generalized Linear Models.* Biometrics Unit and Statistics Center Cornell University. 1997.

[13] McCulloch, C. Searle, S. *Generalized, Linear and Mixed Models* Wiley Series in Probability and Statistics Cornell University. 2001.

[14] McNeil, A.J. Wendin, J. *Bayesian inference for generalized linear mixed models of portfolio credit risk.* 2005.

[15] Zeger, S. L.; K.Y. Liang Albert P.S. *Models for longitudinal data: a generalized estimating equation approach.* Biometrics. 1988.

Anexo: Script R

Aquí presentamos el script de R utilizado para el capítulo 4 del trabajo. El mismo se divide en seis secciones diferenciadas:

- Primera parte: incluye tratamiento y limpieza de los datos, separación de conjunto de entrenamiento y testeo. Además, incluye algunas cuestiones exploratorias.
- Segunda Parte: Se realiza la selección de variables con criterio AUC.
- Tercera Parte: Selección de variables, esta vez con criterio AIC.
- Cuarta Parte: Se evalúan los resultados del modelo final.
- Quinta parte: Se realiza análisis de residuos.
- Sexta parte: Se realiza el análisis de performance con *Leave one out cross-validation*

```
library(dplyr)
library(lme4)
library(pROC)
library(ggplot2)



######################################################################
########## PRIMERA PARTE: TRATAMIENTO PREVIO DE DATOS ############
######################################################################


insurance_claims <- read.csv("C:/Users/ian
    bounos/OneDrive/Escritorio/ESTADISTICA/Tesis
    Maestria/insurance_claims_original.csv", header=TRUE, sep=",")


insurance_claims$fraud_reported= as.numeric(insurance_claims$fraud_reported=="Y")
insurance_claims$fraud_reported=as.factor(insurance_claims$fraud_reported)

### Definimos Siniestros "Grandes" y "Pequeos" por las observaciones hechas:
insurance_claims$injury_dummy= insurance_claims$injury_claim > 2000
insurance_claims$property_dummy = insurance_claims$property_claim > 2000
insurance_claims$vehicle_dummy = insurance_claims$vehicle_claim > 10000


## Separamos conjunto de entrenamiento y testeo
set.seed(1996)
n= nrow(insurance_claims)
n
index = sample(1:n,size=0.8*n)
insurance.train = insurance_claims[index,]
insurance.test = insurance_claims[-index,]


### algunas cuestiones exploratorias

ggplot(insurance.train, aes(x=fraud_reported, y=property_claim,
    color=fraud_reported)) +
  geom_boxplot()+geom_point(alpha=0.1)

cor(insurance.train%>%
```

```
      select(vehicle_claim,
             injury_claim,
             property_claim,
             vehicle_dummy,
             property_dummy,
             injury_dummy))

table(insurance.train$property_dummy,insurance.train$fraud_reported)
table(insurance.train$vehicle_dummy,insurance.train$fraud_reported)
table(insurance.train$injury_dummy,insurance.train$fraud_reported)


#### Variables preseleccionadas

insurance.train = insurance.train%>%
  select(fraud_reported,
         incident_severity,
         insured_hobbies,
         incident_type,
         collision_type,
         authorities_contacted,
         incident_state,
         witnesses,
         auto_make,
         injury_claim,
         vehicle_claim,
         property_claim,
         injury_dummy,
         vehicle_dummy,
         property_dummy)
insurance.test = insurance.test%>%
  select(fraud_reported,
         incident_severity,
         insured_hobbies,
         incident_type,
         collision_type,
         authorities_contacted,
         incident_state,
         witnesses,
         auto_make,
         injury_claim,
         vehicle_claim,
         property_claim,
```

```
        injury_dummy,
        vehicle_dummy,
        property_dummy)


############################################################
###### SEGUNDA PARTE: SELECCION AUC ######################
############################################################


covariables_fijas_todas = names(insurance.train%>%
                                select(incident_severity,
                                       incident_type,
                                       collision_type,
                                       authorities_contacted,
                                       incident_state,
                                       witnesses,
                                       injury_claim,
                                       vehicle_claim,
                                       property_claim,
                                       injury_dummy,
                                       vehicle_dummy,
                                       property_dummy))
covariables_random_todas = names(insurance.train%>%
                                 select(insured_hobbies,auto_make))

insurance_seleccion.train = insurance.train
insurance_seleccion.test = insurance.test



#### definimos funcion auxiliar calculo_auc
#  calculo_auc toma los indices de las variables a utilizar y devuelve el auc
calculo_auc = function(indice_fijas,indice_random)
{

  variables_fijas_este_modelo =  covariables_fijas_todas[indice_fijas]
  if(length(indice_random) >0)
  {
    variables_random_este_modelo = paste("(1|",
        covariables_random_todas[indice_random],")",sep="")
```

```
    variables_todas_este_modelo =
        c(variables_fijas_este_modelo,variables_random_este_modelo )
  }else{variables_todas_este_modelo = variables_fijas_este_modelo}

  formula_este_modelo =
     paste("fraud_reported~",paste(variables_todas_este_modelo, collapse = "+"))

  if(length(indice_random) >0){f=glmer}else{f=glm}
  modelo_logistico <- f(as.formula(formula_este_modelo), data =
     insurance_seleccion.train, family =binomial(link="logit"))
  auc <- roc(response = insurance_seleccion.test$fraud_reported, predictor =
     predict(modelo_logistico,newdata= insurance_seleccion.test,type =
     "response"))
 return(auc)
}



##### A partir de ahora realizaremos las condiciones iniciales para el proceso
    de iteracion:


n_fijas = length(covariables_fijas_todas)
n_random = length(covariables_random_todas)
n_total = n_fijas+n_random


indice_fijas = c()
indice_random = c()

### generamos un nuevo indice que es el indice de todas las variables, para que
    sea mas comodo ir iterando sobre el
indice_total= c()

### definimos el auc inicial
modelo_logistico_base <- glm(fraud_reported~1, data = insurance_seleccion.train,
    family =binomial(link="logit"))
auc0 =roc(response = insurance_seleccion.test$fraud_reported, predictor =
    predict(modelo_logistico_base,newdata= insurance_seleccion.test,type =
    "response"))$auc


actualizar = -1 ### Mientras esta variable no sea 0, se seguiran agregando
```

```
    variables al indice
registro_auc<- data.frame(matrix(0, nrow = 14, ncol = 10))
paso = 1
#### Nos fijaremos quin es la que maximiza el valor
while(actualizar!=0){

  actualizar= 0
  auc_inicial = auc0

    #### iteramos sobrelas que no estan en el indice
    for( k in (1:n_total)) {

           if(! k %in% indice_total){
         indice_total_provisorio = c(indice_total,k)
         indice_fijas_provisorio =
            indice_total_provisorio[indice_total_provisorio<=n_fijas]
         indice_random_provisorio =
            indice_total_provisorio[indice_total_provisorio>n_fijas]-n_fijas

         auc_provisorio =
            calculo_auc(indice_fijas_provisorio,indice_random_provisorio)$auc
         registro_auc[k,paso]= auc_provisorio


         ### si supera el auc, cambiamos actualizar a 1
           if(auc0<auc_provisorio){
             actualizar = 1
             auc0 = auc_provisorio
             mejor_indice_total = indice_total_provisorio
             }

           }

    }
    if(actualizar == 1 ){
      if(auc0-auc_inicial>0.01){ ### pediremos tolerancia de 0.01
      print(auc0)
      indice_total =  mejor_indice_total
      print(indice_total)}else{actualizar=0}}
  paso = paso+1

}
```

```
rownames(registro_auc) = c(covariables_fijas_todas,covariables_random_todas)


#### Aqu guardamos los aportes al AUC de cada paso

registro_auc

#################### Graficamos AUC###################################
##### Aqu graficamos los pasos.
#No es importante para el desarrollo del codigo ms all del grafico en s:



#### Primer paso
indice_fijas = c()
indice_random = c()
indice_total= c()


modelo_logistico_base <- glm(fraud_reported~1, data = insurance_seleccion.train,
    family =binomial(link="logit"))
auc0 =roc(response = insurance_seleccion.test$fraud_reported, predictor =
    predict(modelo_logistico_base,newdata= insurance_seleccion.test,type =
    "response"))$auc
plot(roc(response = insurance_seleccion.test$fraud_reported, predictor =
    predict(modelo_logistico_base,newdata= insurance_seleccion.test,type =
    "response")))


for( k in (1:n_total)) {

  if(! k %in% indice_total){
    indice_total_provisorio = c(indice_total,k)
    indice_fijas_provisorio =
        indice_total_provisorio[indice_total_provisorio<=n_fijas]
    indice_random_provisorio =
        indice_total_provisorio[indice_total_provisorio>n_fijas]-n_fijas

    roc_provisorio =
        calculo_auc(indice_fijas_provisorio,indice_random_provisorio)
    auc_provisorio = roc_provisorio$auc
    plot(roc_provisorio,add=TRUE,col="light blue")
```

```
  }
}

plot(calculo_auc(c(1),c()),add=TRUE,col="red")




#### Segundo paso
indice_fijas = c()
indice_random = c(1)
indice_total= c(13)


modelo_logistico_base <- glm(fraud_reported~1, data = insurance_seleccion.train,
    family =binomial(link="logit"))
auc0 =roc(response = insurance_seleccion.test$fraud_reported, predictor =
    predict(modelo_logistico_base,newdata= insurance_seleccion.test,type =
    "response"))$auc
plot(roc(response = insurance_seleccion.test$fraud_reported, predictor =
    predict(modelo_logistico_base,newdata= insurance_seleccion.test,type =
    "response")))


for( k in (1:n_total)) {

  if(! k %in% indice_total){
    indice_total_provisorio = c(indice_total,k)
    indice_fijas_provisorio =
        indice_total_provisorio[indice_total_provisorio<=n_fijas]
    indice_random_provisorio =
        indice_total_provisorio[indice_total_provisorio>n_fijas]-n_fijas

    roc_provisorio =
        calculo_auc(indice_fijas_provisorio,indice_random_provisorio)
    auc_provisorio = roc_provisorio$auc
    plot(roc_provisorio,add=TRUE,col="light blue")

  }
}
```

```
plot(calculo_auc(c(1),c(1)),add=TRUE,col="red")



#####################################################################
############### TERCERA PARTE: SELECCION AIC ####################
#####################################################################


### Analogamente, definimos funcion calculo_aic

calculo_aic = function(indice_fijas,indice_random)
{

  variables_fijas_este_modelo =  covariables_fijas_todas[indice_fijas]
  if(length(indice_random) >0)
  {
    variables_random_este_modelo = paste("(1|",
        covariables_random_todas[indice_random],")",sep="")
    variables_todas_este_modelo =
        c(variables_fijas_este_modelo,variables_random_este_modelo )
  }else{variables_todas_este_modelo = variables_fijas_este_modelo}

  formula_este_modelo =
      paste("fraud_reported~",paste(variables_todas_este_modelo, collapse = "+"))

  if(length(indice_random) >0){f=glmer}else{f=glm}
  modelo_logistico <- f(as.formula(formula_este_modelo), data = insurance.train,
      family =binomial(link="logit"))
  aic <- AIC(modelo_logistico)
  return(aic)
}



indice_fijas = c()
indice_random = c()

### generamos un nuevo indice que es el indice de todas las variables, para que
    sea mas comodo ir iterando sobre el
indice_total= c()
```

```
### definimos el auc inicial
modelo_logistico_base <- glm(fraud_reported~1, data = insurance.train, family
    =binomial(link="logit"))
aic0 = AIC(modelo_logistico_base)
aic0
actualizar = -1 ### Mientras esta variable no sea 0, se seguiran agregando
    variables al indice


registro_aic<- data.frame(matrix(0, nrow = 14, ncol = 10))
paso = 1

while(actualizar!=0){

  actualizar= 0
  aic_inicial = aic0

  #### iteramos sobrelas que no estan en el indice
  for( k in (1:n_total)) {

    if(! k %in% indice_total){
      indice_total_provisorio = c(indice_total,k)
      indice_fijas_provisorio =
          indice_total_provisorio[indice_total_provisorio<=n_fijas]
      indice_random_provisorio =
          indice_total_provisorio[indice_total_provisorio>n_fijas]-n_fijas

      aic_provisorio =
          calculo_aic(indice_fijas_provisorio,indice_random_provisorio)

      registro_aic[k,paso] = aic_provisorio


      ### si supera el aic, cambiamos actualizar a 1
      if(aic0>aic_provisorio){
        actualizar = 1
        aic0 = aic_provisorio
        mejor_indice_total = indice_total_provisorio
      }

    }
```

```
  }
  if(actualizar == 1 ){
      ### pediremos tolerancia de 0.01
      print(aic0)
      indice_total =  mejor_indice_total
      print(indice_total)}else{actualizar=0}
  paso=paso+1

}


registro_aic


### Anova de cada etapa de la seleccion de variables

modelo_base = glm(fraud_reported~1,
                  family=binomial(link="logit"),
                  data=insurance.train )
modelo_paso1 = glm(fraud_reported~incident_severity,
                   family=binomial(link="logit"),
                   data=insurance.train )
modelo_paso2 = glmer(fraud_reported~incident_severity+(1|insured_hobbies),
                     family=binomial(link="logit"),
                     data=insurance.train )
modelo_paso3 =
    glmer(fraud_reported~incident_severity+(1|insured_hobbies)+collision_type,
                     family=binomial(link="logit"),
                     data=insurance.train )

anova(modelo_paso3,modelo_paso2,modelo_paso1, modelo_base)



################################################################################################
##############################CUARTA PARTE: RESULTADOS MODELO FINAL
    ############################################################
################################################################################



modelo_final <- glmer(fraud_reported ~ incident_severity
```

```
   +(1|insured_hobbies),data=insurance.train, family=binomial(link="logit") )
summary(modelo_final)


######## Comparamos el modelo final con la exclusion de incident severity

modelo_sin_severity = glmer(fraud_reported ~
   (1|insured_hobbies),data=insurance.train, family=binomial(link="logit") )
anova(modelo_final,modelo_sin_severity,test="LRT")

#### Curva ROC

roc_final <- roc(response = insurance.test$fraud_reported, predictor =
   predict(modelo_logistico,newdata= insurance.test,type = "response"))
plot(roc_final,col="blue")


summary(modelo_final)
anova(modelo_final)


### Efectos aleatorios
x = ranef(modelo_final) $insured_hobbies

efrandom = as.numeric(x[order(-abs(x[1:20,] )),])
efrandom
hobby = rownames(x)[order(-abs(x[1:20,] ))]
data.frame(hobby,efrandom)



##################################################
#### QUINTA PARTE: ANALISIS DE RESIDUOS ##########
##################################################



library(arm)
binnedplot(fitted(modelo_final),
          residuals(modelo_final, type = "response"),

          xlab = "Valor Esperado",
          ylab = "Residuo Promedio",
```

```
          main= "",
          cex.pts = 1.4,
          col.pts = 4,
          col.int = "white")


library(DHARMa)


simulationOutput <- simulateResiduals(fittedModel = modelo_final, n = 1000)


simulationOutput = recalculateResiduals(simulationOutput , group =
    insurance.train$insured_hobbies)

  plot(simulationOutput)




#####################################################################
############## SEXTA PARTE: ANALISIS LOOCV ######################
#####################################################################


#### Definimos probabilidad estimada con LOOCV
 probmodelo=c()
  for(k in 1:n){
    print(k)
    datos = insurance_claims[-k,]
    individuo = insurance_claims[k,]
   modelo_final_loocv = glmer(fraud_reported
       ~incident_severity+(1|insured_hobbies),data = datos, family = binomial)
    probmodelo = c(probmodelo,predict(
        modelo_final_loocv,newdata=individuo,type="response"))

  }

##### Funcion que dice, segun el umbral de probabilidad que clasificacion
    corresponde a dicha probabilidad
  pred = function(vector,p=0.5    ){
    return(as.numeric(vector>p))
    }
```

```
  #### tasas de acierto para p = 0.5
  mean((pred(probmodelo)==insurance_claims$fraud_reported))


#### Grafico tasa de acierto en funcin del umbral de probabilidad

  grilla =seq(0,1,length=100)
  acierto = c()
  for( p in grilla){  acierto =
      c(acierto,mean((pred(probmodelo,p)==insurance_claims$fraud_reported))) }

  ### Guardamos aca el maximo
  p_max = grilla[which.max(acierto) ]
  tasa_max = max(acierto)

  df = data.frame(Umbral = grilla,Tasa = acierto)
  ggplot(data=df,aes(x=Umbral,y=Tasa))+
    geom_line()+
    geom_point(data=data.frame(Umbral=c(p_max),Tasa =
        c(tasa_max)),color="blue",size=2)

 ### Tabla de confusin para p_max

  tabla_confusion_maximizatasa =
      table(insurance_claims$fraud_reported,pred(probmodelo,p_max))
  tabla_confusion_maximizatasa
  sens_maximizatasa =  tabla_confusion_maximizatasa[2,2]/(
      tabla_confusion_maximizatasa[2,2]+ tabla_confusion_maximizatasa[2,1])
  spec_maximizatasa = tabla_confusion_maximizatasa[1,1]/(
      tabla_confusion_maximizatasa[1,2]+ tabla_confusion_maximizatasa[1,1])
  #### Curva roc
  roc_loocv = roc(response = insurance_claims$fraud_reported, predictor
      =probmodelo)
roc_loocv$specificities
roc_loocv$auc

plot(roc_loocv)
points(spec_maximizatasa,sens_maximizatasa,pch=16,col="red")
```