UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques i Informàtica

GRADO DE MATEMÁTICAS

Trabajo de fin de grado

REGRESIÓN LOGÍSTICA: ESTUDIO DE LA PROBABILIDAD DE TENER UN SALARIO ELEVADO

Autor: Ana Victoria Galindo

Director: Dra. Esther Vayá y Dr. Josep Vives
Realizado en: Departamento de Matemáticas e Informática

Barcelona, 13 de junio de 2023

Abstract

Logistic models have a crucial role in data analysis and decision-making across diverse fields. They enable the modeling and prediction of binary events, making them valuable in areas such as medicine, psychology, marketing, and social sciences. This study aims to understand the mathematical theory behind logistic models and apply this knowledge to construct a practical model for predicting high income among university graduates. By combining theoretical exploration and practical implementation using R programming, we analyze relevant data from the National Institute of Statistics and evaluate the model's predictive capabilities. The study's methodology provides a comprehensive and insightful perspective, contributing to a better understanding of the factors influencing the likelihood of attaining a high salary.

Resumen

Los modelos logísticos son fundamentales en el análisis de datos y la toma de decisiones en diversas áreas. Este trabajo tiene como objetivo comprender la teoría matemática detrás de estos modelos y aplicarla en la construcción de un modelo práctico para analizar la probabilidad de tener un sueldo elevado en titulados universitarios. La metodología del trabajo se divide en dos partes: el estudio teórico y la aplicación práctica utilizando el programa R. Se utilizarán datos del Instituto Nacional de Estadística (INE) y se evaluará la capacidad predictiva del modelo. El trabajo combina rigurosidad teórica con aplicaciones prácticas, buscando obtener conclusiones sólidas y relevantes.

2020 Mathematics Subject Classification. 62J12, 62J20, 62P20

Agradecimientos

Me gustaría empezar dando las gracias a los dos tutores de este trabajo. Gracias por ayudarme a llevar a cabo la idea inicial de combinar las matemáticas con la economía. Pero gracias sobretodo por confiar en mí y en este trabajo.

También me gustaría agradecer a mi familia y a mi pareja que simplemente han estado a mi lado durante el proceso de este trabajo (y de toda la carrera) y han mostrado interés por ayudarme en todo momento aunque según ellos "no entienden nada de matemáticas". Gracias Eva por acompañarme durante las noches cuando nos quedábamos cada una haciendo su trabajo y gracias Buba por cuidarnos siempre.

Por último, me gustaría recordar una frase de un gran libro: *No caminamos en círculo, vamos hacia arriba, el círculo es una espiral, hemos subido ya muchos escalones*. Gracias a todas esas personas que empezaron siendo compañeros de clase y hoy son amigos. Ellos me han ayudado a no entrar en bucle y dejar de caminar en círculos. Me han dado la perspectiva necesaria para ver que realmente el círculo es una espiral hacia arriba.

Índice

1. Introducción

En la actualidad, los modelos logísticos desempeñan un papel fundamental en el análisis de datos y la toma de decisiones en diversas áreas. Su importancia radica en su capacidad para modelar y predecir eventos binarios, lo cual es de gran relevancia en campos como la medicina, la psicología, el marketing y la ciencia social. Estos modelos permiten comprender la relación entre variables independientes y una variable dependiente discreta, brindando información clave para la toma de decisiones y la identificación de factores de influencia.

El objetivo de este trabajo es adquirir un profundo entendimiento de la teoría matemática que respalda los modelos logísticos, con el propósito de aplicar este conocimiento en la construcción de un modelo práctico con para evidenciar cómo los modelos logit pueden emplearse para el análisis de situaciones reales. He seleccionado modelizar la probabilidad de que los titulados universitarios tengan un sueldo elevado como caso práctico. La elección ha sido basada en mi interés personal y su gran relevancia en el contexto actual. Este tema me resulta especialmente interesante, ya que me encuentro en el momento de finalizar mi carrera y adentrarme en el mercado laboral. A través de este estudio, espero obtener una visión rigurosa y fundamentada sobre los factores que influyen en la probabilidad de obtener un salario alto.

La metodología de este trabajo se divide en dos grandes bloques. En el primer bloque, me enfocaré en el estudio de teoremas, definiciones y explicaciones teóricas relacionadas con los modelos logísticos. Este conocimiento teórico será fundamental para comprender y aplicar de manera práctica los conceptos en la segunda parte de mi trabajo. En el segundo bloque, pondré en práctica los conocimientos adquiridos para estimar el modelo logístico concreto utilizando el programa R. Utilizaré datos relevantes al tema de estudio procedentes del Instituto Nacional de Estadística (INE), y evaluaré la capacidad predictiva del modelo. Esto permitirá obtener conclusiones concretas y verificar la aplicabilidad de la teoría en un contexto real.

A través de esta metodología dual, mi trabajo combinará la rigurosidad teórica con la aplicación práctica, aportando una visión completa y enriquecedora del tema estudiado. Siempre siguiendo el objetivo de obtener un modelo con conclusiones sólidas y relevantes.

Parte I
FUNDAMENTOS TEÓRICOS

2. Conceptos previos

En este trabajo se abordarán conceptos matemáticos que requieren de algunas definiciones previas para poder comprender las secciones posteriores.

2.1. Tipos de Variables de un Modelo

Definición 2.1. *En un modelo de regresión la* ***variable endógena*** *es la variable dependiente que se está tratando de predecir a partir de una o más variables independientes. También recibe el nombre de variable dependiente o respuesta.*

Definición 2.2. *La* ***variable explicativa*** *se refiere a la variable independiente o predictor que se utiliza para predecir o explicar la variable dependiente.*

2.2. Datos categóricos

Definición 2.3. *Los* ***datos categóricos*** *son un tipo de datos en los que la información se clasifica en categorías discretas y mutuamente excluyentes. Estas categorías pueden ser cualitativas o nominales y representan características que no pueden ser medidas en una escala numérica.*

Ejemplos de datos categóricos incluyen género, estado civil, nivel de educación, entre otros. Los datos categóricos son fundamentales en el análisis de datos, ya que permiten clasificar y agrupar información en diferentes categorías, lo que nos permite detectar patrones y tendencias que pueden ayudar en la toma de decisiones en una variedad de áreas, desde la investigación de mercado hasta la salud pública y la política.

2.2.1. Tablas de Contingencia

Sean X e Y dos variables categóricas, X con I categorías e Y con J categorías. La clasificación de los sujetos según estas dos variables tiene IJ posibles combinaciones. Cuando las dos variables son variables de respuesta usaremos su *distribución conjunta*, lo que determina sus distribuciones marginal y condicional. Sin embargo, cuando Y es una variable de respuesta pero X es una variable explicativa, estudiaremos la *distribución condicional* de Y y como varía cuando la categoría de X cambia.

Definición 2.4. *Una tabla con I filas para las categorías de X y J columnas para las categorías de Y muestra las IJ posibles combinaciones de resultados. Cuando las casillas de la tabla contienen la frecuencia de cada resultado para una muestra, se conoce como* ***tabla de contingencia****. Este término fue introducido por Karl Pearson (1904) y se denota como tabla* $I \times J$*.*

Las tablas de contingencia se pueden extender a más de dos variables. Veamos un ejemplo con una variable dependiente Y que tiene dos categorías: 0 y 1, y dos variables

predictoras X y Z, con I y K categorías respectivamente. La Tabla 1 muestra la tabla de contingencia correspondiente a este ejemplo. Nótese que n_{ijk} indica el número de sujetos (frecuencia) que cumplen que la categoría de la variable X es i, la categoría de Y es j y la de Z es k.

		Y=1	Y=0
Z=categoría 1	X=categoría 1	n_{111}	n_{101}
	X=categoría 2	n_{211}	n_{201}
	⋮	⋮	⋮
	X=categoría i	n_{i11}	n_{i01}
	⋮	⋮	⋮
	X=categoría I	n_{I11}	n_{I01}
⋮	⋮	⋮	⋮
Z=categoría k	X=categoría 1	n_{11k}	n_{10k}
	X=categoría 2	n_{21k}	n_{20k}
	⋮	⋮	⋮
	X=categoría i	n_{i1k}	n_{i0k}
	⋮	⋮	⋮
	X=categoría I	n_{I1k}	n_{I0k}
⋮	⋮	⋮	⋮
Z=categoría K	X=categoría 1	n_{11K}	n_{10K}
	X=categoría 2	n_{21K}	n_{20K}
	⋮	⋮	⋮
	X=categoría i	n_{i1K}	n_{i0K}
	⋮	⋮	⋮
	X=categoría I	n_{I1K}	n_{I0K}

Tabla 1: Tabla de contingencia $I \times 2 \times K$.

2.2.2. Distribuciones Conjuntas, Marginales y Condicionales para Tablas de Contingencia

Si X e Y son dos variables de respuesta denotamos π_{ij} como la probabilidad de que suceda (X, Y) en la celda que ocupa la fila i y la columna j. La *distribución conjunta* es la distribución de probabilidad π_{ij} de X e Y. Las *distribuciones marginales* son el total de cada fila y columna que resulta de sumar las probabilidades conjuntas. Se usa la notación π_{i+} para las filas y π_{+j} para las columnas, donde el subíndice "+"denota la suma de ese índice; esto es,

$$\pi_{i+} = \sum_j \pi_{ij} \qquad \text{y} \qquad \pi_{+j} = \sum_i \pi_{ij}.$$

En la práctica es común que una de las variables sea de respuesta, pongamos Y, y la otra sea explicativa, es decir X. Para una categoría fija de X, Y tiene una distribución de probabilidad. Por lo tanto, es relevante estudiar como cambia esta distribución según la categoría de X. Dado un sujeto clasificado en la fila i de X, usamos $\pi_{j|i}$ para denotar la

probabilidad de clasificación en la columna j de Y, $j = 1, \ldots, J$. Entonces, $\sum_j \pi_{j|i} = 1$. Las probabilidades $\{\pi_{1|i}, \ldots, \pi_{J|i}\}$ forman la *distribución condicional* de Y en la categoría i de X. La distribución condicional de Y dado X se relaciona con las otras distribuciones del siguiente modo:

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} \qquad \text{para cualquier i y j.}$$

La Tabla 2 muestra las diferentes distribuciones para el caso 2×2

	Columna		
Fila	1	2	Total
1	π_{11}	π_{12}	π_{1+}
	$(\pi_{1\|1})$	$(\pi_{2\|1})$	(1.0)
2	π_{21}	π_{22}	π_{2+}
	$(\pi_{1\|2})$	$(\pi_{2\|2})$	(1.0)
Total	π_{+1}	π_{+2}	1.0

Tabla 2: Notación para las Distribuciones Conjuntas, Marginales y Condicionales.

2.2.3. Independencia de Variables Categóricas

Definición 2.5. *Se dice que dos variables categóricas X e Y (con I y J categorías respectivamente) son **independientes** si todas las probabilidades conjuntas equivalen al producto de sus probabilidades marginales, esto es*

$$\pi_{ij} = \pi_{i+}\pi_{+j} \qquad \text{para} \qquad i = 1, \ldots, I \qquad \text{y} \qquad j = 1, \ldots, J \tag{2.1}$$

Cuando X e Y son independientes,

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} = (\pi_{i+}\pi_{+j})/\pi_{i+} = \pi_{+j} \qquad \text{para } i = 1, \ldots, I.$$

Es decir, cada distribución condicional de Y es idéntica a la distribución marginal de Y. De este modo, dos variables son independientes cuando $\{\pi_{j|1} = \cdots = \pi_{j|I}$, para $j = 1, \ldots, J\}$; esto significa que la probabilidad de cada columna de respuesta es la misma en cada fila. Cuando Y es una variable de respuesta y X es una variable explicativa, esto es una manera más natural de definir la independencia que usando (2.1). Por esto, la independencia a menudo se denota como *homogeneidad* de las distribuciones condicionales.

2.3. Odds

Definición 2.6. *Para una probabilidad π de éxito, los **odds** se definen como*

$$\text{odds } \Omega = \pi/(1 - \pi).$$

Los odds son no negativos, con odds $\Omega > 1$ cuando el éxito es más probable que el fracaso. Si por ejemplo, $\pi = 0{,}75$, entonces odds $\Omega = 0{,}75/0{,}25 = 3$; un éxito es tres veces más probable que un fracaso, y se esperan tres éxitos por cada fracaso. Cuando $\Omega = \frac{1}{3}$, un fallo es tres veces más probable que un éxito. De manera inversa,

$$\pi = (\text{odds } \Omega)/(\text{odds } \Omega + 1).$$

Por ejemplo, cuando odds $\Omega = \frac{1}{3}$, entonces la probabilidad $\pi = 0{,}25$.

Volvemos a una tabla 2×2. Dentro de la fila i, los odds de éxito en lugar de fracaso son $\Omega_i = \pi_i/(1-\pi_i)$.

Definición 2.7. *La ratio de los odds Ω_1 y Ω_2 en las dos filas,*

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)},$$

es conocida como ***odds ratio*** *y abreviado* ***OR****.*

Su interpretación es sencilla y muy común en los modelos logísticos. Cuando OR es igual 1, significa que la probabilidad del evento es la misma en ambos grupos. No hay diferencias significativas en la ocurrencia del evento entre los dos grupos. Si OR es mayor a 1, indica que la probabilidad del evento es mayor en el primer grupo en comparación con el segundo grupo. Esto sugiere que el primer grupo tiene una mayor predisposición o mayor probabilidad de experimentar el evento en cuestión. Por último, cuando OR es menor a 1, significa que la probabilidad del evento es menor en el primer grupo en comparación con el segundo grupo. Esto implica que el segundo grupo tiene una mayor probabilidad o mayor propensión al evento en comparación con el primer grupo.

Para distribuciones conjuntas con probabilidades $\{\pi_{ij}\}$ la definición equivalente para los odds en la fila i es $\Omega_i = \pi_{i1}/\pi_{i2}$, $i = 1, 2$. Entonces, el odds ratio es

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

2.3.1. Odds Ratio Marginal y Condicional

Los odds ratio pueden describir relaciones marginales y condicionales. Veamos el caso para una tabla $2 \times 2 \times K$, donde K denota el número de categorías de una variable explicativa Z. Sea $\{\mu_{ijk}\}$ las frecuencias esperadas en la casilla ijk de la tabla de contingencia para un modelo, por ejemplo binomial.

Para una categoría fija k de Z, el odds ratio

$$\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}$$

es el *odds ratio condicional* entre X e Y dado $Z = k$. Por otra parte, sea $\mu_{ij+} = \sum_k \mu_{ijk}$, definimos el *odds ratio marginal* de X e Y como

$$\theta_{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}.$$

3. Modelos Lineales Generalizados

El Modelo Lineal Generalizado es una generalización de la regresión lineal ordinaria. Para referirse a este modelo es común el uso del acrónimo en inglés GLM. Un GLM tiene tres componentes básicos:

1. Una componente aleatoria que identifica la distribución de probabilidad de la respuesta.

2. Una componente sistemática que especifica las variables explicativas utilizadas en la función predictora lineal.

3. Función link o de enlace que describe la relación funcional entre el componente sistemático y el valor esperado del componente aleatorio.

3.1. Componentes del Modelo Lineal Generalizado

La componente aleatoria de un GLM consiste en una variable respuesta Y con observaciones independientes $Y = (y_1, \dots, y_N)$ que siguen una distribución de la familia exponencial natural. Esto quiere decir que para cada observación y_i tiene una función de densidad de probabilidad de la forma

$$f(y_i; \theta_i) = a(\theta_i)b(y_i)\exp[y_i Q(\theta_i)]. \tag{3.1}$$

El parámetro $Q(\theta)$ recibe el nombre de *parámetro natural*. En esta familia se incluyen diversas distribuciones como la distribución de Poisson y la binomial.
La componente sistemática de un GLM relaciona un vector $\eta = (\eta_1, ..., \eta_N)$ con un conjunto de variables explicativas mediante un modelo lineal. Sea x_{ij} el valor de la variable explicativa j $(j = 0, 1, 2, \dots)$ para el sujeto i. Entonces

$$\eta_i = \sum_j \beta_j x_{ij}, \qquad i = 1, \dots, N.$$

Esta combinación lineal de variables explicativas recibe el nombre de *predictor lineal*. Es común que $x_{i0} = 1$ para cualquier i, haciendo que el coeficiente β_0 (o α, depende de la notación) se denomine término independiente.

Por último, la función de enlace conecta las dos componentes anteriores. Sea $\mu_i = E(Y_i)$ para i=1,...,N. La función μ_i está relacionado de la siguiente manera:

$$\eta_i = g(\mu_i)$$

donde g es una función monótona diferenciable. Por lo tanto, el modelo enlaza los valores esperados de las observaciones y las variables explicativas mediante la fórmula

$$g(\mu_i) = \sum_j \beta_j x_{ij}, \qquad i = 1, ..., N.$$

La función de enlace cumple dos própositos en los MLG:

- Primero, garantiza que la combinación lineal η tome valores en el rango apropiado para la variable de respuesta,
- y segundo, permite que la relación entre las variables predictoras y la variable de respuesta sea modelada de manera adecuada.

La función de enlace $g(\mu) = \mu$ es concocida como *función de enlace identidad* y establece una relación lineal directa entre las variables predictoras y la variable de respuesta, donde la media de la variable de respuesta se modela como una combinación lineal de las variables predictoras. Otra función de enlace relevante es la *función de enlace canónica*, que transforma la media en el parámetro natural. Para ello, $g(\mu_i) = Q(\theta_i)$.

4. Función Logit

Es común que una variable categórica esté compuesta por dos categorías: "éxito" y "fracaso". Estos posibles resultados se representan mediante un 1 o un 0, respectivamente. La distribución de Bernoulli se usa para variables aleatorias discretas que solo pueden resultar en dos sucesos mutuamente excluyentes. Estos tienen probabilidad $P(Y = 1) = \pi$ y $P(Y = 0) = 1 - \pi$ y esperanza $E(Y) = 0 \cdot P(Y = 0) + 1 \cdot P(Y = 1) = \pi$. Cuando Y_i tiene una distribución de Bernoulli con parámetro π_i la función de probabilidad (pdf) es:

$$f(y_i; \pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = (1-\pi_i)[\pi_i/(1-\pi_i)]^{y_i} = (1-\pi_i)\exp\left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right)\right] \quad (4.1)$$

para $y_i = 0$ y 1. Esta distribución pertenece a la familia exponencial natural. En este caso $Q(\pi) = \log[\pi/(1-\pi)]$. Obsérvese que dentro del logaritmo aparece la función *odds* que se ha visto en la sección 2.3. Por lo tanto $Q(\pi)$ es el logaritmo de la función odds de obtener como respuesta el resultado 1 y para abreviar esto se define

$$\text{logit}(\pi) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \log[\text{odds}(\pi)]. \quad (4.2)$$

Para hacer referencia al logaritmo de la función odds en el resto del trabajo se usará la abreviatura *log odds* o directamente *logit*.

5. Modelo de Regresión Logística

Sea Y una variable de respuesta binaria (con dos categorías: éxito y fracaso que como se ha visto antes toman valores 1 y 0 respectivamente). Tratamos cada observación como una variable binomial para un solo ensayo de Bernoulli. La media es $E(Y) = P(Y = 1)$ y denotamos $P(Y = 1)$ usando $\pi(\mathbf{x})$, reflejando así su dependencia de los valores $\mathbf{x} = (x_1, \ldots, x_p)$.

Para una variable de respuesta binaria, el modelo de regresión

$$\pi(\mathbf{x}) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p \quad (5.1)$$

es conocido como *modelo de probabilidad lineal*. Si tenemos observaciones independientes estamos hablando de un GLM con componente aleatorio binomial y función de enlace identidad. Este modelo tiene un gran defecto estructural: las probabilidades toman valores entre 0 y 1, pero las funciones lineales toman valores en toda la recta real. El modelo (5.1) puede tener $\pi(x) < 0$ para algunos valores de $\mathbf{x}$. No obstante, este modelo presenta una gran ventaja en el momento de interpretar los coeficiente.

Es común que la relación entre $\pi(x)$ y x no sea lineal, por ejemplo puede suceder que un cambio fijo en x tenga menos impacto cuando $\pi(x)$ está cerca de 0 o 1 que cuando se

aproxima a 0,5. Pongamos un ejemplo más concreto. Sea $\pi(x)$ la probabilidad de aprobar un examen dependiendo de $x =$ "número de horas de estudio". Un aumento de 3 horas de estudio seguramente tendrá menos impacto cuando $x = 20$ (es decir, cuando $\pi(x)$ está cerca de 1) que cuando $x = 1$.

En la práctica, las relaciones no lineales entre $\pi(x)$ y x son monótonas, con $\pi(x)$ incrementando continuamente o con $\pi(x)$ decreciendo continuamente a medida que x aumenta. Las curvas con forma de "S" son usuales, pero la más importante es la que tiene la siguiente expresión:

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}. \tag{5.2}$$

Esto es un *modelo de regresión logística*. A medida que x aumenta, $\pi(x)$ incrementa cuando $\beta > 0$ y decrece cuando $\beta < 0$. Esto se puede ver representado en las Figuras 1 y 2 a continuación.

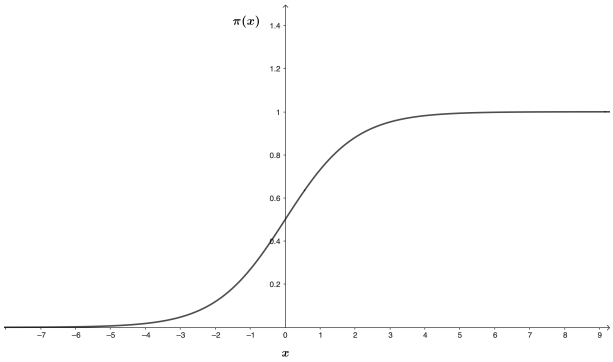

Figura 1: $\pi(\mathbf{x})$ con $\beta > 0$

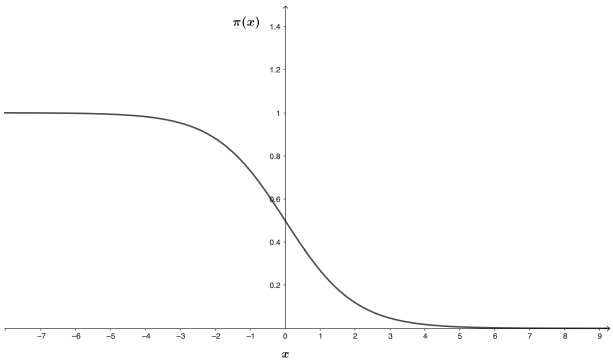

Figura 2: $\pi(\mathbf{x})$ con $\beta > 0$

Buscamos ahora la función link para la cual la regresión logística es un GLM. Para facilitar la búsqueda primero calculamos

$$1 - \pi(x) = 1 - \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} = \frac{1}{1 + \exp(\alpha + \beta x)}$$

y ahora ya es sencillo ver que los odds son

$$\frac{\pi(x)}{1-\pi(x)} = \frac{\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}}{\frac{1}{1+\exp(\alpha+\beta x)}} = \exp(\alpha+\beta x).$$

El logaritmo de odds presenta una relación lineal

$$\log \frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x. \tag{5.3}$$

Por lo tanto, la función link apropiada es el log odds, es decir el logit. En resumen, los modelos de regresión logística son GLMs con componente aleatoria binomial y con función de enlace logit.

El logit es el parámetro natural para la distribución binormal ya que la función de probabilidad de la distribución binomial es

$$\begin{aligned} f(y;n) = \binom{n}{y}\pi^y(1-\pi)^{n-y} &= \binom{n}{y}(1-\pi)^n\left(\frac{\pi}{1-\pi}\right)^y \\ &= \binom{n}{y}(1-\pi)^n \exp\left[y\log\left(\frac{\pi}{1-\pi}\right)\right] \end{aligned} \tag{5.4}$$

donde π es la probabilidad de éxito. Si recordamos (3.1) es fácil ver que el parámetro natural para la distribución binormal es el logit, por lo tanto la función de enlace logit es la función de enlace canónica. Mientras que $\pi(x)$ debe pertenecer al rango [0,1], el logit puede ser cualquier número real. Los números reales también están en el rango de los predictores lineales que forman la componente sistemática de un GLM. Por lo tanto, este modelo no tiene el problema estructural que presentaba el modelo de probabilidad lineal.

5.1. Regresión Logística Múltiple

La regresión logística se puede extender a modelos con varias variables explicativas, que pueden ser una mezcla de cuantitativas y cualitativas. El modelo para $\pi(\mathbf{x}) = P(Y=1)$ donde $\mathbf{x} = (x_1, ..., x_p)$ es el vector de p predictores es

$$\text{logit}[\pi(\mathbf{x})] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p. \tag{5.5}$$

O de manera alternativa, si directamente especificamos $\pi(\mathbf{x})$:

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}. \tag{5.6}$$

5.2. Regresión Logística con Predictores Categóricos

Como en las regresiones ordinarias, la regresión logística puede admitir variables explicativas cualitativas, que se conocen como *factores*. Para hacer esto se usan variables ficticias, como veremos en la sección 9.2.

5.3. Distribución logística

La distribución logística es una distribución de probabilidad continua que se utiliza en estadística para modelar variables aleatorias con comportamientos de crecimiento y decrecimiento logísticos. Cuando $\beta > 0$, la curva de regresión logística (5.2) es la función de distribución de probabilidad (cdf) de la distribución logística. La cdf de la distribución logística con media μ y parámetro de escala $\tau > 0$ es

$$F(x) = \frac{\exp[(x-\mu)/\tau]}{1+\exp[(x-\mu)/\tau}, \qquad -\infty < x < \infty.$$

La función de densidad correspondiente es simétrica y tiene forma de campana. Recuerda a la pdf de la distribución normal pero con colas más pesadas (tiene mayor curtosis). Para la distribución logística estándar $\mu = 0$ y $\tau = 1$ la cdf es $\Phi(x) = e^x/(1+e^x)$. Entonces podemos reescribir (5.2) como $\pi(x) = \Phi(\alpha+\beta x)$. Por lo tanto el logit es la función inversa de la cdf de la distribución logística enstándar:

$$\Phi(x) = \pi(x) = e^x/(1+e^x) \Longrightarrow x = \Phi^{-1}[\pi(x)] = \log[\pi(x)/(1-\pi(x)] = \text{logit}(\pi(x)).$$

5.4. Interpretación de parámetros de la Regresión logística

Veamos el caso más simple. Para una variable de respuesta binaria Y y una variable explicativa X, definimos $\pi(x) = P(Y=1|X=x) = 1 - P(Y=0|X=x)$. El modelo de regresión logística es

$$\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}. \tag{5.7}$$

Equivalentemente, el logit tiene la relación lineal

$$\text{logit}[\pi(x)] = \log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x. \tag{5.8}$$

5.4.1. Interpretación de β: Odds, Probabilidades y Aproximaciones Lineales

¿Cómo se interpreta β en (5.8)? Ya hemos visto que su signo determina si $\pi(x)$ crece o decrece a medida que x aumenta. La tasa de aumento o decrecimiento aumenta a medida que $|\beta|$ crece; cuando $\beta \to 0$ la curva se aplana hacia una línea horizontal recta. Cuando $\beta = 0$, Y es independiente de X. Para x cuantitativa con $\beta > 0$, la curva de $\pi(x)$ tiene la forma de la función de distribución (cdf) de la distribución logística (ver apartado 5.3). Como la pdf de la distribución logística es simétrica, $\pi(x)$ se acerca a 1 con la misma tasa de variación con la que se acerca a 0.

Exponenciando ambos lados de la igualdad (5.8) obtenemos odds $= e^\alpha e^{\beta x}$. Esto proporciona una interpretación básica de la magnitud de β: los odds se multiplican por e^β cada vez que x aumenta una unidad. En otras palabras, e^β es la ratio de los odds en $X = x+1$ y $X = x$:

$$\theta = \frac{\text{odds en X=x+1}}{\text{odds en X=x}} = \frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}} = \frac{e^{\alpha+\beta(x+1)}}{e^{\alpha+\beta x}} = \frac{e^\alpha e^{\beta x} e^\beta}{e^\alpha e^{\beta x}} = e^\beta.$$

Sin embargo, si no se está familiarizado con los odds o logits, la interpretación de los coeficientes es compleja. Podemos usar un argumento de linearización para buscar

una interpretación más simple. Como (5.8) no tiene una apariencia lineal, sino curva, la velocidad de cambio en $\pi(x)$ por unidad de cambio en x varía. Si dibujamos una línea tangente a la curva en un valor particular de x, como se muestra en la Figura 3, esta describe la tasa de cambio instantánea en ese punto concreto. Calculamos

$$\begin{aligned}
\frac{d}{dx}[\pi(x)] &= \frac{\frac{d}{dx}[\exp(\alpha+\beta x)](1+\exp(\alpha+\beta x)) - \exp(\alpha+\beta x)\frac{d}{dx}[1+\exp(\alpha+\beta x)]}{[1+\exp(\alpha+\beta x)]^2} \\
&= \frac{\beta\exp(\alpha+\beta x)(1+\exp(\alpha+\beta x)) - \beta\exp(2(\alpha+\beta x))}{[1+\exp(\alpha+\beta x)]^2} \\
&= \frac{\beta\exp(\alpha+\beta x) + \beta\exp(2(\alpha+\beta x)) - \beta\exp(2(\alpha+\beta x))}{[1+\exp(\alpha+\beta x)]^2} \\
&= \beta\cdot\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}\cdot\frac{1}{1+\exp(\alpha+\beta x)} \\
&= \beta\pi(x)[1-\pi(x))]
\end{aligned}$$

y vemos que por ejemplo, la tangente de la curva en el punto x tal que $\pi(x) = \frac{1}{2}$ tiene pendiente $\beta\cdot\frac{1}{2}\cdot\frac{1}{2} = \beta\cdot\frac{1}{4}$; cuando $\pi(x) = 0,9$ o $0,1$ entonces la pendiente vale $0,09\beta$. La pendiente se aproxima a cero cuando $\pi(x)$ se aproxima o bien a 1,0 o bien 0. Por lo contrario, la pendiente más pronunciada sucede cuando $\pi(x) = \frac{1}{2}$, es decir, cuando $x = -\frac{\alpha}{\beta}$. Esto se obtiene si aislamos x en la igualdad siguiente:

$$\pi(x) = \frac{1}{2} \iff \exp(\alpha+\beta x) = 1 \iff \alpha+\beta x = 0 \iff x = -\frac{\alpha}{\beta}.$$

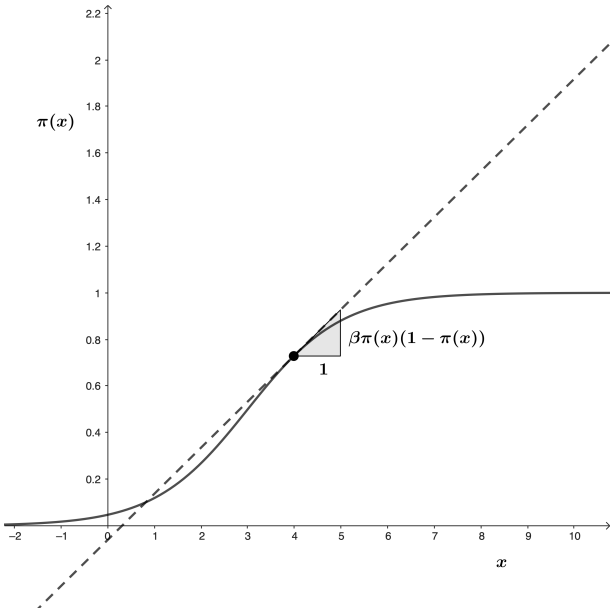

Figura 3: Aproximación lineal a la curva de la regresión logística

Por último veamos una última interpretación en este caso de los modelos de regresión logística múltiples. Recordamos (5.5) y vemos la siguiente observación:

Observación 5.1. En un modelo de regresión logística múltiple, el coeficiente β_i $(i = 1, \ldots, p)$ representa el cambio en el logit resultante al aumentar una unidad en la i-ésima variable x_i $(i = 1, \ldots, p)$.

Demostración: Sean $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_p)$ y $\mathbf{x}' = (x'_1, \ldots, x'_i, \ldots, x'_p)$ dos observaciones de variables explicativas que cumplen $x_j = x'_j \quad \forall j \neq i$ y $x'_i = x_i + 1$. Calculamos el cambio en el logit obteniendo

$$\text{logit}(\mathbf{x}') - \text{logit}(\mathbf{x}) = \alpha + \sum_{i=1}^{p} \beta_i x'_i - \left(\alpha + \sum_{i=1}^{p} \beta_i x_i \right) = \beta_i (x'_i - x_i) = \beta_i (x_i + 1 - x_i) = \beta_i.$$

tal y como queríamos ver. □

5.4.2. Análisis de los datos

Antes de ajustar el modelo y hacer las interpretaciones es conveniente analizar los datos para comprobar que el uso de la regresión logística es adecuado. Como y solo toma los valores 0 y 1, es complicado verificar esto mediante un gráfico de dispersión regular de los valores (x, y) observados.

Puede ser útil graficar las proporciones o *logits empíricos* en función de x. Sea n_i el número de observaciones con la configuración i de la variable x y sea y_i el número de respuestas "1", con $p_i = \frac{y_i}{n_i}$. El *logit empírico* i es $\log[p_i/(1-p_i)] = \log[y_i/(n_i - y_i)]$. El gráfico de dispersión de los logits empíricos debería ser prácticamente lineal. Obsérvese que el logit empírico no es finito cuando $y_i = 0$ o $n_i = 0$. Podemos hacer una pequeña modificación y sumar una constante al número de respuestas de cada resultado obteniendo

$$\log \frac{y_i + \frac{1}{2}}{n_i - y_i + \frac{1}{2}}.$$

Este ajuste es el estimador menos sesgado para el logit verdadero.

Cuando x es continua y $n_i = 1 \quad \forall i$, o cuando x es esencialmente continua y todos los n_i son pequeños, esto no se cumple. Podríamos agrupar los datos con valores cercanos a x en categorías antes de calcular los logits empíricos. Una mejor aproximación que no requiere la elección de categorías arbitrarias utiliza un mecanismo de suavizado para revelar tendencias. Un enfoque de suavizado de este tipo ajusta un modelo aditivo generalizado, que reemplaza el predictor lineal de un modelo lineal generalizado (GLM) por una función infinitamente diferenciable. Un gráfico de este ajuste revela si se producen discrepancias graves respecto a la tendencia en forma de "S" predicha por la regresión logística.

5.5. La Regresión Logística se deduce a partir de Variables Explicativas Normales

Independientemente del mecanismo de muestreo, la regresión logística puede describir bien o no la relación. Dado $Y \in \{0, 1\}$, supongamos que X tiene una distribución $N(\mu_i, \sigma^2)$. Entonces, tenemos el siguiente teorema que demostró Cornfield (1962):

Teorema 5.2. *$\pi(x) = P(Y = 1|X = x)$ satisface el modelo logístico con $\beta = (\mu_1 - \mu_0)/\sigma^2$.*

Demostración: Recordemos que $P(Y = 0|X = x) = 1 - P(Y = 1|X = x) = 1 - \pi(x)$. Entonces, usando el teorema de Bayes obtenemos:

$$\begin{aligned} P(Y = 1 \mid X = x) &= \frac{\pi(x) \exp\left[-(x-\mu_1)^2/2\sigma^2\right]}{\pi(x) \exp\left[-(x-\mu_1)^2/2\sigma^2\right] + (1-\pi(x)) \exp\left[-(x-\mu_0)^2/2\sigma^2\right]} \\ &= 1/\left\{1 + [(1-\pi(x))/\pi(x)] \exp\left\{-\left[\mu_0^2 - \mu_1^2 + 2x(\mu_1 - \mu_0)\right]/2\sigma^2\right\}\right. \\ &= 1/\{1 + \exp[-(\alpha + \beta x)]\} = \exp(\alpha + \beta x)/[1 + \exp(\alpha + \beta x)], \end{aligned}$$

donde $\beta = (\mu_1 - \mu_0)/\sigma^2$ y $\alpha = -\log[(1-\pi(x))/\pi(x)] + \left[\mu_0^2 - \mu_1^2\right]/2\sigma^2$. □

Así, cuando la población es una mezcla de dos tipos de sujetos, un tipo con $y = 1$ que se distribuye aproximadamente siguiendo una normal en X y el otro tipo con $y = 0$ que se distribuye aproximadamente de manera normal en X con una varianza igual, la función de regresión logística aproxima bien la curva para $\pi(x)$.

Los resultados se extienden a un vector de variables explicativas que tienen distribuciones normales multivariadas en cada caso. Si las distribuciones son normales pero con diferentes varianzas, el modelo se aplica pero con un término cuadrático. En ese caso, la relación no es monótona, con $\pi(x)$ aumentando y luego disminuyendo, o al revés.

6. Estimación Máximo Verosímil de los Parámetros

La estimación de los parámetros $\hat{\alpha}, \hat{\beta}_1, ..., \hat{\beta}_p$ de un modelo de regresión logística se realiza mediante el método de estimación por máxima verosimiltud (MV). Usando este método obtenemos los parámetros que maximizan la función de verosimilitud ℓ, que es una función que representa la probabilidad de obtener los datos observados, dadas las estimaciones de los parámetros del modelo.
Sea N el tamaño de la muestra, Y una variable dicotómica y $\alpha, \beta_1, ..., \beta_p$ los parámetros de la regresión logística, denotamos $\ell = \ell\left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(N)}, y^{(N)}\right), \beta_0, \beta_1, \ldots, \beta_p\right)$ como la función de verosimilitud asociada. Entonces, se tiene que:

$$\begin{aligned} \ell =& \ell\left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(N)}, y^{(N)}\right), \beta_0, \beta_1, \ldots, \beta_p\right) \\ =& \prod_{j=1}^{N} P\left(Y = 1 \mid \mathbf{x}^{(j)}\right)^{y^{(j)}} \left(1 - P\left(Y = 1 \mid \mathbf{x}^{(j)}\right)\right)^{1-y^{(j)}}. \end{aligned}$$

Además, considerando que la función $\ln(z)$ es estrictamente creciente, los valores de los parámetros que maximizan ℓ son los mismos que maximizan $L := \ln(\ell)$.
Desarrollamos el logaritmo neperiano de la función de verosimilitud:

$$
\begin{aligned}
L &= \ln \ell \left(\left(\mathbf{x}^{(1)}, y^{(1)} \right), \ldots, \left(\mathbf{x}^{(N)}, y^{(N)} \right), \beta_0, \beta_1, \ldots, \beta_n \right) \\
&= \sum_{j=1}^{N} y^{(j)} \ln P\left(Y = 1 \mid \mathbf{x}^{(j)} \right) + \sum_{j=1}^{N} \left(1 - y^{(j)} \right) \ln \left(1 - P\left(Y = 1 \mid \mathbf{x}^{(j)} \right) \right) \\
&= \sum_{j=1}^{N} y^{(j)} \left[\ln P\left(Y = 1 \mid \mathbf{x}^{(j)} \right) - \ln \left(1 - P\left(Y = 1 \mid \mathbf{x}^{(j)} \right) \right) \right] + \sum_{j=1}^{N} \ln \left(1 - P\left(Y = 1 \mid \mathbf{x}^{(j)} \right) \right) \\
&= \sum_{j=1}^{N} y^{(j)} \ln \frac{P\left(Y = 1 \mid \mathbf{x}^{(j)} \right)}{1 - P\left(Y = 1 \mid \mathbf{x}^{(j)} \right)} + \sum_{j=1}^{N} \ln \left(1 - P\left(Y = 1 \mid \mathbf{x}^{(j)} \right) \right) \\
&= \sum_{j=1}^{N} y^{(j)} \ln \frac{\pi(\mathbf{x}^{(j)})}{1 - \pi(\mathbf{x}^{(j)})} + \sum_{j=1}^{N} \ln(1 - \pi(\mathbf{x}^{(j)})) \\
&= \sum_{j=1}^{N} y^{(j)} \left(\alpha + \sum_{i=1}^{p} \beta_i x_i^{(j)} \right) - \sum_{j=1}^{N} \ln \left(1 + e^{\left(\alpha + \sum_{i=1}^{p} \beta_i x_i^{(j)} \right)} \right)
\end{aligned}
$$

donde hemos usado las igualdades (5.6) y

$$
1 - \pi(x) = 1 - \frac{\exp(\alpha + \sum_{i=1}^{p} \beta_i x_i)}{1 + \exp(\alpha + \sum_{i=1}^{p} \beta_i x_i)} = \frac{1}{1 + \exp(\alpha + \sum_{i=1}^{p} \beta_i x_i)}.
$$

Los estimadores máximo verosímiles $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_p$ para los parámetros $\alpha, \beta_1, \ldots, \beta_p$ se obtienen al resolver el siguiente sistema de ecuaciones:

$$
\begin{aligned}
\frac{\partial \ln \ell}{\partial \alpha} &= \sum_{j=1}^{N} y^{(j)} - \sum_{j=1}^{N} \frac{e^{\left(\alpha + \sum_{i=1}^{p} \beta_i x_i^{(j)} \right)}}{1 + e^{\left(\alpha + \sum_{i=1}^{p} \beta_i x_i^{(j)} \right)}} = 0 \\
\frac{\partial \ln \ell}{\partial \beta_1} &= \sum_{j=1}^{N} y^{(j)} x_1^{(j)} - \sum_{j=1}^{N} x_1^{(j)} \frac{e^{\left(\alpha + \sum_{i=1}^{p} \beta_i x_i^{(j)} \right)}}{1 + e^{\left(\alpha + \sum_{i=1}^{p} \beta_i x_i^{(j)} \right)}} = 0 \\
&\vdots \\
\frac{\partial \ln \ell}{\partial \beta_p} &= \sum_{j=1}^{N} y^{(j)} x_p^{(j)} - \sum_{j=1}^{N} x_p^{(j)} \frac{e^{\left(\alpha + \sum_{i=1}^{p} \beta_i x_i^{(j)} \right)}}{1 + e^{\left(\alpha + \sum_{i=1}^{p} \beta_i x_i^{(j)} \right)}} = 0.
\end{aligned}
$$

Estas $p + 1$ ecuaciones forman un sistema de ecuaciones con $p + 1$ incógnitas. A pesar de esto, no es factible encontrar una expresión cerrada para estimar los valores de los parámetros $\alpha, \beta_1, \ldots, \beta_p$. Por lo tanto, es común utilizar técnicas iterativas para realizar estas estimaciones. Una de estas técnicas es el método de Newton-Raphson.

6.1. Método de Newton-Raphson

Este es un método iterativo que se utiliza para resolver ecuaciones no lineales, como aquellas cuya solución determina el punto donde una función toma su valor máximo. Se empieza con una suposición inicial de la solución. Se obtiene una segunda suposición aproximando la función a maximizar en un entorno de la suposición inicial mediante un

polinomio de segundo grado y luego se encuentra la ubicación del valor máximo de ese polinomio. Luego aproxima la función en un entorno de la segunda suposición mediante otro polinomio de segundo grado, y la tercera suposición es su máximo. De esta manera, el método genera una secuencia de suposiciones que convergen al máximo cuando la función es adecuada y/o la suposición inicial es buena.

Con más detalle, a continuación se expone cómo el método de Newton-Raphson determina el valor de $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots \hat{\beta}_p)$ que maximiza L. Sea

$$u^T = (\partial L(\beta)/\partial\alpha, \partial L(\beta)/\partial\beta_1, \dots, \partial L(\beta)/\partial\beta_p),$$

y sea $\mathbf{H}$ la matriz Hessiana, con entradas $h_{ab} = \partial^2 L(\beta)/\partial\beta_a\partial\beta_b$. Denotamos $\mathbf{u}^{(t)}$ y $\mathbf{H}^{(t)}$ a $\mathbf{u}$ y $\mathbf{H}$ evaluados en $\beta^{(t)}$, suposición t-ésima para $\hat{\beta}$. Calculamos primero:

$$u_j^{(t)} = \left.\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\beta^{(t)}} = \sum_i \left(y_i - n_i\pi_i^{(t)}\right) x_{ij},$$
$$h_{ab}^{(t)} = \left.\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_a \partial \beta_b}\right|_{\beta^{(t)}} = -\sum_i x_{ia}x_{ib}n_i\pi_i^{(t)}\left(1 - \pi_i^{(t)}\right),$$

donde $\pi^{(t)}$, que es la aproximación t para $\hat{\pi}$, se obtiene a través de $\boldsymbol{\beta}^{(t)}$ del siguiente modo:

$$\pi_i^{(t)} = \frac{\exp\left(\sum_{j=1}^p \beta_j^{(t)} x_{ij}\right)}{1 + \exp\left(\sum_{j=1}^p \beta_j^{(t)} x_{ij}\right)}.$$

El paso t del proceso iterativo ($t = 0, 1, 2, \dots$) aproxima $L(\boldsymbol{\beta})$ en un entorno de $\boldsymbol{\beta}^{(t)}$ usando la expansión de Taylor de segundo orden,

$$L(\boldsymbol{\beta}) \approx L\left(\boldsymbol{\beta}^{(t)}\right) + \boldsymbol{u}^{(t)T}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\right) + \left(\frac{1}{2}\right)\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\right)^T \boldsymbol{H}^{(t)}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\right).$$

Aislando $\boldsymbol{\beta}$ de $\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} \approx \boldsymbol{u}^{(t)} + \boldsymbol{H}^{(t)}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\right) = \mathbf{0}$ encontramos la siguiente suposición de $\boldsymbol{\beta}$. Arreglando la expresión

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\boldsymbol{H}^{(t)}\right)^{-1}\boldsymbol{u}^{(t)}, \tag{6.1}$$

asumiendo que $\boldsymbol{H}^{(t)}$ es no singular (Los procedimientos informáticos utilizan métodos estándar para resolver ecuaciones lineales en lugar de calcular explícitamente la inversa).

Usamos $\boldsymbol{u}^{(t)}$ y $\boldsymbol{H}^{(t)}$ con la fórmula (6.1) para obtener el siguiente valor $\boldsymbol{\beta}^{(t+1)}$, que en el caso concreto de la regresión logística es

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left\{\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right\}^{-1}\boldsymbol{X}^T\left(\boldsymbol{y} - \boldsymbol{\mu}^{(t)}\right), \tag{6.2}$$

donde $\mu_i^{(t)} = n_i\pi_i^{(t)}$, $\mathbf{X} \in M(N, p)$ es la matriz cuyas filas son $\mathbf{x}^{(j)}, j = 1, \dots, N$ y $\mathbf{W} \in M(N, N)$ es la matriz diagonal con elementos $\pi^{(j)}\left(1 - \pi^{(j)}\right), j = 1, \dots, N$

$$\mathbf{W} = \begin{pmatrix} \pi^{(1)}\left(1-\pi^{(1)}\right) & \cdots & \cdots & 0 \\ 0 & \pi^{(2)}\left(1-\pi^{(2)}\right) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \pi^{(N)}\left(1-\pi^{(N)}\right) \end{pmatrix}.$$

A partir de (6.2) se obtiene $\pi^{(t+1)}$, y así sucesivamente.

Los criterios de convergencia del método iterativo utilizado para estimar los parámetros pueden ser varios, pero en todos ellos la idea subyacente es que bien $\widehat{\beta}^{\text{nuevo}} \simeq \widehat{\beta}^{\text{viejo}}$ o $\ln L\left(\widehat{\beta}^{\text{nuevo}}\right) \simeq \ln L\left(\widehat{\beta}^{\text{viejo}}\right)$ o $\widehat{\pi}^{\text{nuevo}} \simeq \widehat{\pi}^{\text{viejo}}$.

7. Tests de Hipótesis

Los tests de hipótesis son herramientas estadísticas utilizadas para evaluar afirmaciones o suposiciones sobre una población o un fenómeno en particular. En general, los tests de hipótesis se emplean para tomar decisiones basadas en evidencia empírica y determinar si los resultados observados son consistentes con una hipótesis nula o si proporcionan suficiente evidencia para respaldar una hipótesis alternativa. Veamos dos tests de hipótesis relevantes para la regresión logística.

7.1. Test de Razón de Verosimilitud

Suponemos que queremos contrastar la hipótesis $H_0 : \theta \in \Theta_0$ frente $H_a : \theta \in \Theta_a$, donde Θ_0 y Θ_a son una partición de Θ. El test de la razón de verosimilitud propone como la región crítica:

$$A_1 = \{\Lambda_N(x_1, ..., x_N) \leq K\}$$

donde $K \in (0, 1)$ y

$$\Lambda(x_1, ..., x_N) = \frac{\sup_{\theta\in\Theta_0} \ell(x_1, ..., x_N, \theta)}{\sup_{\theta\in\Theta} \ell(x_1, ..., x_N, \theta)}$$

Observamos que $\Lambda_N(x_1, \ldots, _N) \in [0, 1]$. Además, si verdaderamente el parámetro se encuentra en Θ_0 tenemos que $\Lambda_N \simeq 1$ y por lo tanto la región crítica es razonable.

Teorema 7.1. *(Wilks) Sea $p(x, \theta)$ un modelo estadístico regular. Sea $\Theta \in \mathbb{R}^d$. Suponemos que queremos contrastar la hipótesis $\theta \in \Theta_0 \subset \Theta$ donde*

$$\Theta_0 = \{\theta = (\theta_1, ..., \theta_d) \in \Theta : \theta_1 = \theta_1^0, ..., \theta_r = \theta_r^0\}$$

con $1 \leq r \leq d$. Entonces, el estadístico de Wilks cumple que

$$W_n(X_1, ..., X_n) = -2 \log \Lambda_n(X_1, ..., X_n) \xrightarrow{\mathcal{L}} \chi^2_r.$$

Los grados de libertad equivalen a la diferencia de las dimensiones de los espacios de los parámetros en $H_0 \cup H_a$ menos en H_0.

El test de la razón de verosimilitud tiene como objetivo comparar dos modelos de regresión logística; el denominado modelo completo frente al que se conoce como modelo reducido. Este segundo modelo puede verse como un submodelo del modelo completo. La hipótesis nula establece que los parámetros correspondientes a las variables que forman parte del modelo completo, pero no del modelo reducido, valen cero.

7.2. Demostración del Teorema de Wilks

Para evitar prolongar en exceso el alcance de este trabajo, nos centraremos únicamente en la demostración del caso en que $d = r = 1$. La demostración para cualquier valor r y d

se encuentra en los apuntes de la asignatura de Estadística de la Universidad de Barcelona escritos por Corcuera, J.M.

Antes de demostrar el teorema debemos demostrar y definir algunos conceptos que necesitaremos posteriormente.

Definición 7.2. *Sea $\{p(x,\theta), x \in \mathcal{X}, \theta \in \Theta\}$, se define la **información de Fisher** de un modelo como $I(\theta) = (\partial_\theta L(X_1, \ldots, X_n, \theta))$, donde L es el logaritmo de la función de verosimilitud.*

Cuando hablamos de modelo estadístico nos referimos a un modelo que cumple las 4 condiciones siguientes:

C1. El soporte de P_θ no depende de θ.

C2. Para cada x en el soporte, $f_x(\theta) := \log p_\theta(x)$ es tres veces diferenciable con respecto a θ en un intervalo $(\theta - \delta, \theta + \delta)$; además, $E_\theta \left| f'_X(\theta) \right|$ y $E_\theta \left| f''_X(\theta) \right|$ son finitos, y existe una función $M(x)$ tal que

$$\sup_{\theta \in (\theta^* - \delta, \theta^* + \delta)} \left| f'''_x(\theta) \right| \leq M(x), \quad E_{\theta^*}[M(X)] < \infty. \tag{7.1}$$

C3. La esperanza con respecto a P_θ y la diferenciación en θ se pueden intercambiar, lo cual implica que la función de score tiene media cero y que la información de Fisher existe y puede ser evaluada utilizando cualquiera de las dos fórmulas familiares.

C4. La información de Fisher en θ^* es positiva.

Definición 7.3. *Se dice que un estimador $\hat{\theta}$ es **consistente** si para cualquier valor verdadero del parámetro θ, a medida que el tamaño de la muestra n tiende a infinito, la probabilidad de que el estimador $\hat{\theta}$ difiera del valor verdadero θ, en más de un valor pequeño ϵ tiende a cero. Matemáticamente, esto se puede expresar como:*

$$\lim_{n \to \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0.$$

Teorema 7.4. *(Teorema de Slutsky). Suponemos que $X_n \xrightarrow{\mathcal{L}} X$ y que $Y_n \xrightarrow{\mathcal{L}} c \in \mathbb{R}$. Entonces,*
1. $X_n + Y_n \xrightarrow{\mathcal{L}} X + c$,
2. $X_n \cdot Y_n \xrightarrow{\mathcal{L}} c \cdot X$.

Teorema 7.5. *Supongamos que $X_1, X_2, \ldots, X_n$ son independientes e idénticamente distribuidas con distribución P_θ, donde $\theta \in \Theta \subset \mathbb{R}$.Supongamos C1-C4. Sea $\hat{\theta}_n$ una secuencia de estimadores consistentes de máxima verosimilitud. Entonces para cualquier punto interior θ^*,*

$$n^{1/2}\left(\hat{\theta}_n - \theta^*\right) \to N\left(0, I\left(\theta^*\right)^{-1}\right), \text{ en distribución bajo } P_{\theta^*}.$$

Demostración: Sea $\bar{L}_n(\theta) = n^{-1} \log \ell_n(\theta)$ el logaritmo de la función de verosimilitud escalado. Como θ^* es un punto interior, entonces existe un entorno abierto A de θ^* contenido en Θ. Por la consistencia de $\hat{\theta}_n$, tenemos en cuenta el comportamiento de $\hat{\theta}_n$ únicamente cuando se encuentra en A, donde la log-verosimilitud presenta un buen comportamiento, en particular, $\bar{L}'_n(\hat{\theta}_n) = 0$. A continuación, tomamos la aproximación de Taylor de segundo orden de $\bar{L}'_n(\hat{\theta}_n)$ alrededor de θ^*:

$$0 = \bar{L}'_n\left(\theta^*\right) + \bar{L}''_n\left(\theta^*\right)\left(\hat{\theta}_n - \theta^*\right) + \frac{1}{2}\bar{L}'''_n\left(\tilde{\theta}_n\right)\left(\hat{\theta}_n - \theta^*\right)^2,$$

donde $\tilde{\theta}_n$ se encuentra entre $\hat{\theta}_n$ y θ^*. Operando obtenemos

$$n^{1/2}\left(\hat{\theta}_n - \theta^*\right) = -\frac{n^{1/2}\bar{L}'_n(\theta^*)}{\bar{L}''_n(\theta^*) + 0{,}5\bar{L}'''_n\left(\tilde{\theta}_n\right)\left(\hat{\theta}_n - \theta^*\right)}, \text{ para } \hat{\theta}_n \text{ cerca de } \theta^*.$$

Por lo tanto, solo nos queda ver que el lado derecho de la igualdad de arriba tiene una distribución asintótica $N\left(0, I(\theta^*)^{-1}\right)$. Estudiaremos por partes separadas el numerador y el denominador.

Numerador. El numerador se puede reescribir como

$$n^{1/2}\bar{L}'_n(\theta^*) = n^{1/2}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log p_\theta(X_i)\Bigg|_{\theta=\theta^*}.$$

Los sumandos son iid con media 0 y varianza $I(\theta^*)$, por la suposición de que se pueden intercambiar derivadas e integrales. Por lo tanto, por el Teorema Central del Límite tenemos que $n^{1/2}\bar{L}'_n(\theta^*)$ se comporta asintóticamente como una distribución $\mathrm{N}(0, I(\theta^*))$.

Denominador. El primer término del denominador converge en P^*_θ-probabilidad a $-I(\theta^*)$ por la ley de los números grandes. Finalmente solo queda demostrar que el segundo término del denominador es negligible. Para eso, tengamos en cuenta que por (7.1) en la condición C2

$$\left|\bar{L}'''_n\left(\tilde{\theta}_n\right)\right| \leq \frac{1}{n}\sum_{i=1}^{n} M(X_i), \text{ para } \quad \hat{\theta}_n \quad \text{ cerca de } \theta^*.$$

Otra vez usando la ley de los números grandes, tenemos que la cota superior converge a $E^*_\theta[M(X_1)]$, que es finito. Consecuentemente, $\bar{L}'''\left(\tilde{\theta}_n\right)$ está acotado en probabilidad y como $\hat{\theta}_n - \theta^* \to 0$ en $P^*_\theta -$ probabilidad por suposición, podemos concluir que

$$\left(\hat{\theta}_n - \theta^*\right)\bar{L}'''_n\left(\tilde{\theta}_n\right) \to 0, \quad \text{en} \quad P_{\theta^*} - \text{ probabilidad.}$$

Entonces, se deduce del Teorema de Slutsky que,

$$-\frac{n^{1/2}\bar{L}'_n(\theta^*)}{\bar{L}''_n(\theta^*) + \frac{1}{2}\left(\hat{\theta}_n - \theta^*\right)\bar{L}'''_n\left(\tilde{\theta}_n\right)} \to \frac{N(0, I(\theta^*))}{-I(\theta^*)} = N\left(0, I(\theta^*)^{-1}\right),$$

en distribución, que es lo que queríamos ver. □

Ahora ya podemos demostrar demostrar el Teorema de Wilks.

Demostración del Teorema de Wilks. Como estamos en el caso $d = r = 1$ entonces $\Theta_0 = \{\theta_0\}$ es un singleton, y queremos conocer la distribución asintótica de W_n bajo P_{θ_0}. Claramente,

$$W_n = -2L_n(\theta_0) + 2L_n\left(\hat{\theta}_n\right)$$

donde $\hat{\theta}_n$ es el estimador máximo verosímil y L_n es la función log-verosimilitud. Debido a la continuidad supuesta de la log-verosimilitud, podemos calcular una aproximación de Taylor de segundo orden de $L_n(\theta_0)$ alrededor de $\hat{\theta}_n$:

$$L_n(\theta_0) = L_n\left(\hat{\theta}_n\right) + L'_n\left(\hat{\theta}_n\right)\left(\theta_0 - \hat{\theta}_n\right) + \frac{L''_n\left(\tilde{\theta}_n\right)}{2}\left(\theta_0 - \hat{\theta}_n\right)^2$$

donde $\tilde{\theta}_n$ está entre θ_0 y $\hat{\theta}_n$. Como $L_n'\left(\hat{\theta}_n\right) = 0$, obtenemos

$$W_n = -L_n''\left(\tilde{\theta}_n\right)\left(\theta_0 - \hat{\theta}_n\right)^2 = -\frac{L_n''\left(\tilde{\theta}_n\right)}{n}\left\{n^{1/2}\left(\hat{\theta} - \theta_0\right)\right\}^2 d$$

A partir del Teorema 7.4, $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \to \mathrm{N}\left(0, I\left(\theta_0\right)^{-1}\right)$ en distribución, cuando $n \to \infty$. También,

$$L_n''\left(\tilde{\theta}_n\right) = L_n''\left(\theta_0\right) + L_n''\left(\tilde{\theta}_n\right) - L_n''\left(\theta_0\right)$$

y tenemos que

$$\left|L_n''\left(\tilde{\theta}_n\right) - L_n''\left(\theta_0\right)\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial^2}{\partial\theta^2}\log p_\theta\left(X_i\right)|_{\theta=\tilde{\theta}_n} - \frac{\partial^2}{\partial\theta^2}\log p_\theta\left(X_i\right)|_{\theta=\theta_0}\right|$$

Usando la condición C2, la cota superior está acotada por $n^{-1}\sum_{i=1}^{n} M\left(X_i\right)\left|\tilde{\theta}_n - \theta_0\right|$, que tiende a 0 en probabilidad cuando bajo P_{θ_0} ya que $\tilde{\theta}_n$ es consistente. Por este motivo, $L_n''\left(\tilde{\theta}_n\right)$ tiene el mismo comportamiento asintótico que $L_n''\left(\theta_0\right)$. Finalmente, usando Slutsky, tenemos

$$W_n \to I\left(\theta_0\right)\mathrm{N}\left(0, I\left(\theta_0\right)^{-1}\right)^2 \equiv \mathrm{N}(0,1)^2 \equiv \chi_1^2,$$

tal y como queríamos ver □

7.3. Test de Wald

El objetivo del test de Wald es valorar si un parámetro en un modelo de regresión es significativamente diferente de cero. Sea $H_0 : \beta = \beta_0$ la hipótesis nula que queremos contrastar. Para llevar a cabo este test en el caso univariado se debe calcular el estadístico de Wald (z), que se obtiene de la siguiente manera:

$$z = \frac{(\hat{\beta} - \beta_0)}{\mathrm{SE}}, \tag{7.2}$$

donde $SE = 1/\sqrt{\iota(\hat{\beta})}$ y $\iota(\beta) = -E[\partial^2 L(\beta)/\partial\beta^2]$.

Este estadístico sigue aproximadamente una distribución normal estándar cuando $\beta = \beta_0$. Equivalentemente, z^2 tiene aproximadamente una distribución chi-cuadrado con un grado de libertad.

La extensión multivariada del test de Wald para contrastar la hipótesis nula $H_o : \beta = \beta_0$ usa el estadístico

$$\boldsymbol{W = (\hat{\beta} - \beta_0)^T[\mathrm{cov}(\hat{\beta})]^{-1}(\hat{\beta} - \beta_0)} \tag{7.3}$$

que sigue una distribución chi-cuadrado y los grados de libertad equivalen al rango de la matriz $\mathrm{cov}(\hat{\beta})$.

8. Inferencia para la Regresión Logística

Para el modelo logístico con una sola variable predictora,

$$\text{logit}[\pi(x)] = \alpha + \beta x, \tag{8.1}$$

acabamos de ver que los tests de significación contrastan la hipótesis nula $H_0 : \beta = 0$. Como hemos visto anteriormente el test de máxima verosimilitud usa el estadístico de Wilks, que sigue una distribución χ^2_1. Además, coincide que el estadístico del test de Wald sigue la misma distribución χ^2_1.

8.1. Intervalos de Confianza para Parámetros

En la práctica, es más informativo construir intervalos de confianza para los parámetros que realizar tests de hipótesis sobre sus valores. El intervalo de confianza de β se puede calcular invirtiendo el test. Por ejemplo, un intervalo de confianza del 95 % es el conjunto de β_0 tales que el test $H_0 : \beta = \beta_0$ tiene un p-valor mayor que 0,05.

Sea z_α el z-score de una distribución normal estándar con probabilidad en la cola derecha a; esto es el percentil $100(1-\alpha)$ de esa distribución. Un intervalo de confianza del $100(1-\alpha)$ % usa un valor crítico $z_{\alpha/2}$. Éste es el valor de la abcisa en una distribución normal que deja a su derecha un área igual $\alpha/2$ siendo $1-\alpha$ el nivel de confianza. Por ejemplo, $z_{0,025} = 1{,}96$ para 95 % de confianza.

8.1.1. Intervalo de Confianza de Wald

El intervalo de confianza de Wald es el conjunto de β_0 tales que $|\hat{\beta} - \beta_0|/SE < z_{\alpha/2}$. El intervalo resultante es $\hat{\beta} \pm z_{\alpha/2}(SE)$.

8.1.2. Intervalo de Confianza Basado en la Razón de Verosimilitud

Sea χ^2_{gl} el percentil $100(1-\alpha)$ de la distribución chi-cuadrado con gl grados de libertad. El intervalo de confianza basado en la razón de verosimilitud es el conjunto de β_0 tales que $-2[L(\beta_0) - L(\hat{\beta})] < \chi^2_1(\alpha)$

8.2. Intervalos de Confianza para Otros Elementos

También es interesante calcular los intervalos de confianza para otros elementos, como es el caso de $\pi(x)$ evaluado en diferentes valores de x. Para $x = x_0$ fijo, $\text{logit}[\hat{\pi}(x_0)] = \hat{\alpha} + \hat{\beta}x_0$ tiene un SE dado por la estimación de la raíz cuadrada de

$$\text{var}(\hat{\alpha} + \hat{\beta}x_0) = \text{var}(\hat{\alpha}) + x_0^2\text{var}(\hat{\beta}) + 2x_0\text{cov}(\hat{\alpha}, \hat{\beta}).$$

Un intervalo de confianza del 95 % para el $\text{logit}[\pi(x_0)]$ es $(\hat{\alpha}+\hat{\beta}x_0) \pm 1{,}96(SE)$. Si aplicamos a cada extremo del intervalo la transformación inversa $\pi(x_0) = \exp(\text{logit})/[1+\exp(\text{logit})]$ obtenemos el intervalo que corresponde a $\pi(x_0)$

8.3. Bondad del Ajuste

El estadístico de la razón de verosimilitudes $G^2(M_0|M_1) = -2(L_0 - L_1)$ prueba si ciertos parámetros del modelo son cero, suponiendo que M_1 es válido, mediante la comparación del logaritmo de la función de verosimilitud L_1 del modelo ajustado M_1 con L_0 de un modelo más simple M_0. El estadístico de bondad del ajuste $G^2(M)$ es un caso particular en el cual $M_0 = M$ y M_1 es un modelo saturado. Al probar si M se ajusta estamos probando si todos los parámetros que aparecen en el modelo saturado pero no en M son todos igual a cero. Los grados de libertad son la diferencia en el número de parámetros de los dos modelos.

8.4. Comparación de Modelos mediante comparación de Desviaciones

Sea L_S el logaritmo de la verosimilitud maximizado de un modelo saturado. Como hemos visto en la sección anterior

$$\begin{aligned} G^2(M_0|M_1) =& -2(L_0 - L_1) = -2L_0 + 2L_1 + 2L_S - 2L_S \\ &= -2(L_0 - L_S) - [-2(L_1 - L_S)] = G^2(M_0) - G^2(M_1). \end{aligned}$$

El estadístico de prueba que compara los dos modelos es idéntico a la diferencia de los estadísticos de bondad del ajuste G^2 (desviaciones) de los dos modelos.

El estadístico de comparación de modelos normalmente tiene una distribución chi-cuadrado incluso cuando por separado $G^2(M_i)$ no la tiene. Por ejemplo, cuando al menos un predictor es continuo o cuando la tabla de contingencia tiene valores ajustados muy pequeños, la distribución de la muestra de $G^2(M_i)$ puede alejarse de una distribución chi-cuadrado. No obstante, si gl (grados de libertad) del estadístico de comparación es pequeño (como cuando se comparan dos modelos que difieren en pocos parámetros), la distribución de $G^2(M_0|M_1)$ es aproximadamente chi-cuadrado.

9. Representación de Variables Categóricas

Como hemos visto en la sección (5.2) también es posible usar variables explicativas cualitativas (*factores*) en la regresión logística.

9.1. Representación de Factores: Tipo ANOVA

Con el propósito de simplificar, consideramos un único factor X, con I categorías. En la fila i de la tabla $I \times 2$, denotamos y_i como el número de resultados en la primera columna (éxitos) de un total de n_i experimentos. Tratamos y_i como binomial con parámetro π_i. El modelo de regresión logística con un único factor como predictor es

$$\log \frac{\pi_i}{1 - \pi_i} = \alpha + \beta_i. \tag{9.1}$$

Cuanto mayor sea β_i, mayor será el valor de π_i. La parte de la derecha de (9.1) se asemeja a la media del tratamiento i en el análisis de varianza (ANOVA).
Como en ANOVA, el factor tiene tantos parámetros β_i como categorías. A menos que

eliminemos α, una β_i es redundante. Una β_i puede fijarse como 0, pongamos β_I para la última categoría. Si los valores no satisfacen esta condición podemos hacer una recodificación para que se cumpla. Por ejemplo, pongamos $\tilde{\beta}_i = \beta_i - \beta_I$ y $\tilde{\alpha} = \alpha + \beta_I$, obsérvese que esta condición cumple $\tilde{\beta}_I = 0$. Entonces

$$\text{logit}(\pi_i) = \alpha + \beta_i = (\tilde{\alpha} - \beta_I) + (\tilde{\beta}_i + \beta_I) = \tilde{\alpha} + \tilde{\beta}_i$$

donde los nuevos parámetros satisfacen la condición. Cuando $\beta_I = 0$, α es igual al logit en la fila I y β_i es la diferencia de los logits en las filas i e I. De este modo, β_i equivale al logaritmo de odds ratio para ese par de filas.

Un odds ratio igual a 1 indica que no hay asociación entre la variable predictora y la variable de respuesta. Esto significa que las probabilidades de que ocurra el evento de interés son iguales en ambos grupos comparados. Un OR mayor a 1 para una categoría específica indica una mayor probabilidad de que ocurra el evento de interés en esa categoría en comparación con la categoría base. Y si por el contrario, el OR para una categoría específica es menor que 1, hay una menor probabilidad de que ocurra el evento de interés en esa categoría en comparación con la categoría de referencia. Para cualquier $\pi_i > 0$, existe β_i tal que el modelo (9.1) se sostiene. El modelo tiene tantos parámetros (I) como observaciones binomiales y es saturado. Se dice que un modelo es saturado si tiene tantos parámetros como observaciones, este tipo de modelos se ajustan perfectamente a los datos observados pero no tienen capacidad predictiva. Cuando un factor no tiene efecto entonces $\beta_1 = \beta_2 = \cdots = \beta_I$. Como esto es equivalente a $\pi_1 = \cdots = \pi_I$, este caso corresponde a la independencia estadística de X e Y.

9.2. Representación de Factores: Uso de Variables Ficticias

Una alternativa para representar el modelo (9.1) es usar variables ficticias. Para $i = 1, \ldots, I-1$ definimos $x_i = 1$ para las observaciones en la fila i y $x_i = 0$ en caso contrario. El modelo es

$$\text{logit}(\pi_i) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{I-1} x_{I-1}.$$

Al no crear una variable ficticia para la categoría I el modelo ya tiene en cuenta la redundancia de los parámetros. La categoría que se excluye del modelo es arbitraria y se conoce como *categoría base*. Algunos programas fijan $\beta_1 = 0$; esto corresponde a un modelo con variables ficticias para las categorías 2 hasta la I, pero no para la categoría 1.

Otro modo de imponer las condiciones es imponer $\sum_i \beta_i = 0$. Cuando X tiene $I = 2$ categorías, entonces $\beta_1 = -\beta_2$. Esto resulta de los efectos de la codificación de una variable ficticia, $x = 1$ en la categoría 1 y $x = -1$ en la categoría 2.

Independientemente del esquema de codificación que se use, los resultados sustanciales sobre los efectos que se están estimando son los mismos. Para el modelo (9.1), sin tener en cuenta la restricción de $\{\beta_i\}$, los valores de los predictores lineales $\{\hat{\alpha} + \hat{\beta}_i\}$, y por tanto $\{\hat{\pi}_i\}$, son los mismos. Las diferencias $\hat{\beta}_a - \hat{\beta}_b$ para pares de categorías (a, b) de X son idénticas y representan el log odds ratio estimado. Por lo tanto $\exp(\hat{\beta}_a - \hat{\beta}_b)$ es la estimación del odds de éxito en la categoría a de X dividido por la estimación del odds de éxito en la categoría b de X. Reparametrizar un modelo puede cambiar las estimaciones de los parámetros pero no afecta al ajuste ni a los efectos de interés.

El valor β_i o $\hat{\beta}_i$ para una única categoría es irrelevante. Dependiendo del sistema de restricciones utilizado, los valores asignados a las variables ficticias pueden variar. Para un

predictor binario, por ejemplo usando variables ficticias que impongan $\beta_2 = 0$, el log odds ratio equivale a $\beta_1 - \beta_2 = \beta_1$; por otro lado, si usamos la restricción $\sum_i \beta_i = \beta_1 + \beta_2 = 0$, el log odds ratio es $\beta_1 - \beta_2 = \beta_1 - (-\beta_1) = 2\beta_1$. Con esto se refleja que un parámetro o su estimación solo tiene sentido cuando se compara con otra categoría.

10. Modelos Lineales Logit para Tablas de Contingencia

Cuando todas las variables son categóricas las tablas de contingencia multidimensionales nos pueden ayudar a visualizar los datos. Veamos el caso en que tenemos dos variables categóricas binarias X y Z. Tratamos la muestra como combinaciones de X y Z fijas y cada resultado de Y como una variable binomial, con todos los resultados de la variable Y tratados como binomiales independientes. Usamos las variables ficticias x y z que toman el valor 1 en la primera categoría y 0 en la segunda. El modelo

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x + \beta_2 z \tag{10.1}$$

indica que cuando se cambia el valor de X o Z hay un efecto en la variable de respuesta Y pero las variables predictoras no interactúan entre ellas. El efecto de X es el mismo para los diferentes valores de Z y sucede lo mismo con el efecto de Z aunque X varíe.

Para un valor de Z fijado, pongamos z, el efecto sobre el logit de cambiar las categorías de X es

$$[\alpha + \beta_1(1) + \beta_2 z] - [\alpha + \beta_1(0) + \beta_2 z] = \beta_1. \tag{10.2}$$

Esta diferencia de logits equivale a la diferencia de los logaritmos de los odds, que coincide con el logaritmo del odds ratio entre X e Y, fijando únicamente Z

$$\ln \theta = \ln\left(\frac{\Omega_{X=1,Z=z}}{\Omega_{X=0,Z=z}}\right) = \ln\left(\frac{\exp(\alpha + \beta_1 + \beta_2 z)}{\exp(\alpha + \beta_2 z)}\right) = \ln\left(\frac{e^{\alpha} e^{\beta_1} e^{\beta_2 z}}{e^{\alpha} e^{\beta_2 z}}\right) = \beta_1.$$

Una variable categórica con I categorías necesita $I - 1$ variables ficticias. Con I categorías para una variable predictora X y K categorías para otra variable Z, el modelo (10.1) se extiende a

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1^X x_1 + \cdots + \beta_{I-1}^X x_{I-1} + \beta_1^Z z_1 + \cdots + \beta_{K-1}^Z z_{K-1} \tag{10.3}$$

donde, por ejemplo, $z_k = 1$ para las observaciones en la categoría k de Z y $z_k = 0$ en cualquier otro caso, para $k = 1, \ldots, K - 1$. Esta ecuación representa los efectos de X con parámetros $\{\beta_k^X\}$ y los efectos de Z con parámetros $\{\beta_k^Z\}$. Los superíndices X y Z son simplemente etiquetas para hacer referencia a las variables, no son potencias. Este modelo se puede extender a cualquier número de categorías. El parámetro, por ejemplo, β_k^Z, denota el efecto en el logit de clasificación en la categorías k de Z.

11. Ajuste de Modelos de Regresión Logística

12. Estrategias para la Selección del Modelo

El modelo de regresión logística se enfrenta a los mismos problemas que la regresión ordinaria. El proceso de selección se hace más difícil a medida que el número de variables

explicativas aumenta, esto se debe a que también crecen los posibles efectos e interacciones. A la hora de seleccionar un modelo hay dos objetivos en conflicto: el modelo debe ser lo suficientemente complejo para ajustar correctamente todos los datos, pero por otra parte, debería ser relativamente fácil de interpretar intentando alisar los datos (eliminar volatilidad o cualquier otro ruido). Pueden surgir complicaciones debido a la naturaleza binaria de la variable de respuesta, como estimaciones infinitas de parámetros ML para algunos modelos cuando un resultado de respuesta es mucho más común que el otro.

La mayoría de los estudios de investigación están diseñados para responder ciertas preguntas. Estas preguntas guían la elección de los términos del modelo. Los análisis confirmatorios utilizan entonces un conjunto restringido de modelos. Por ejemplo, una hipótesis de estudio acerca de un efecto puede ser probada comparando modelos con y sin ese efecto. Para los estudios que son exploratorios en lugar de confirmatorios, una búsqueda entre posibles modelos puede proporcionar pistas sobre la estructura de dependencia y plantear preguntas para investigaciones futuras. En cualquier caso, es útil primero estudiar el efecto de cada predictor o las distribuciones condicionales dentro de una tabla de contingencia para un predictor discreto.

12.1. Estrategias Secuenciales

Cuando la enumeración completa de todos los modelos posibles resulta computacionalmente costosa, se utilizan estrategias de modelización destinadas a encontrar el mejor subconjunto de variables predictoras. Las estrategias secuenciales más extendidas son: selección hacia adelante y eliminación hacia atrás.

12.1.1. Selección hacia delante

La selección hacia adelante agrega términos secuencialmente. En cada etapa, selecciona el término que proporciona la mayor mejora en el ajuste. El valor p mínimo para probar el término en el modelo es un criterio sensato, ya que las reducciones en la desviación para diferentes términos pueden tener diferentes valores de gl. Al añadir predictores se llega a un punto de rendimiento decreciente, esto sucede cuando los nuevos predictores están tan correlacionados con los ya utilizados y no mejoran el poder predictivo. El proceso se detiene cuando las nuevas adiciones no mejoran significativamente el ajuste. Una variación de este procedimiento vuelve a probar, en cada etapa, los términos agregados en etapas anteriores para ver si todavía son significativos.

12.1.2. Eliminación hacia atrás

La eliminación hacia atrás comienza con un modelo complejo y elimina términos secuencialmente. En cada etapa, selecciona el término cuya eliminación tiene el efecto menos perjudicial en el modelo (por ejemplo, el valor p más grande). El proceso se detiene cuando cualquier eliminación adicional conduce a un ajuste significativamente peor. En el caso de los predictores cualitativos con más de dos categorías, el proceso debe considerar toda la variable en cualquier etapa en lugar de solo variables ficticias individuales. Debemos agregar o eliminar toda la variable en lugar de solo una de sus ficticias. De lo contrario, el resultado depende de la elección de la categoría base para la codificación de ficticias. La misma observación se aplica a las interacciones que contienen esa variable.

13. Selección del Modelo y el Modelo Correcto

A la hora de seleccionar un modelo entre un conjunto de candidatos estamos equivocados si creemos que existe uno completamente "correcto", pues todos son simplificaciones de la realidad. ¿Cual es el objetivo de hacer pruebas en el ajuste del modelo cuando sabemos que ninguno se ajusta realmente? Un modelo simple que ajusta de manera adecuada los datos tiene la ventaja de la parsimonia. Este concepto se refiere a la idea de que se debe preferir la simplicidad, siempre que sea posible, en lugar de agregar elementos innecesarios a un modelo. Si un modelo tiene relativamente poco sesgo y describe bien la realidad, tiende a proporcionar estimaciones más precisas. Además de las pruebas de significación, existen otros criterios que pueden ayudar a seleccionar un buen modelo en términos de la estimación de los coeficientes. A continuación, presentamos el más conocido de tales criterios.

14. Criterio de Información Akaike

El criterio de información Akaike (AIC, por sus siglas en inglés) es una medida utilizada en estadística para comparar diferentes modelos y determinar cuál de ellos es el más adecuado para los datos disponibles. El AIC se basa en la idea de que un buen modelo debe tener un buen ajuste de datos, pero también debe ser lo más simple posible. El modelo óptimo es aquel que tiende a tener un ajuste más cercano a los valores reales.

La medida de distancia de Kullback-Leibler se utiliza para medir la cantidad de información que se pierde al aproximar una distribución de probabilidad con otra. Akaike definió la cercanía en términos de una medida de distancia de Kullback-Leibler. Sea $p(y)$ la probabilidad de los datos bajo el modelo verdadero y $p_M(y)$ la probabilidad bajo el modelo elegido. La medida de distancia es $E\{\log[p(y)/p_M(y)]\}$, donde el valor esperado se calcula en relación a la distribución verdadera. Para los datos categóricos esta medida se asemeja a G^2 en la forma. Con una muestra, este criterio selecciona el modelo que minimiza

$$\text{AIC} = -2(\text{log verosimilitud maximizado} - \text{número de parámetros en el modelo}).$$

Esto penaliza a los modelos por tener muchos parámetros.

Con muchos posibles predictores, podemos utilizar el AIC para ayudar en la selección de variables. De un conjunto de modelos candidatos, identificamos aquel con el AIC más pequeño. Sin embargo, también son relevantes los modelos que tienen valores de AIC similares. Por ejemplo, también consideraríamos modelos más simples que tengan valores de AIC relativamente cercanos al mínimo.

15. Diagnósticos de Regresión Logística

En la sección 8, presentamos estadísticas para verificar el ajuste del modelo en un sentido global. Después de seleccionar un modelo preliminar, obtenemos una mayor comprensión al cambiar a un modo de análisis microscópico.

15.1. Errores residuales: Pearson, Desviación y Estandarizado

Con predictores categóricos es útil calcular los errores residuales para comparar los datos observados con los ajustados. Sea y_i el resultado binomial para n_i ensayos en el nivel i de las variables explicativas, $i = 1, \ldots, N$. Sea $\hat{\pi}_i$ la estimación de $P(Y = 1)$. Entonces $\hat{\mu}_i = n_i\hat{\pi}_i$ es el valor ajustado de éxitos y podemos definir el residuo de Pearson como

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\text{var}}(Y_i)}} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{[n_i\hat{\pi}_i(1-\hat{\pi}_i)]}}. \tag{15.1}$$

El estadístico de Pearson para probar el ajuste del modelo satisface

$$X^2 = \sum_{i=1}^{N} e_i^2$$

Un residuo alternativo es la desviación residual, que utiliza componentes del estadístico de ajuste G^2. Para la regresión logística esto es

$$\sqrt{d_i} \times \text{signo}(y_i - n_i\hat{\pi}_i), \tag{15.2}$$

donde la función signo(x) indica el signo de x y donde

$$d_i = 2\left(y_i \log\frac{y_i}{n_i\hat{\pi}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right).$$

Una versión estandarizada del residuo de Pearson es la siguiente

$$r_i = \frac{e_i}{\sqrt{1-\hat{h}_i}} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{[n_i\hat{\pi}_i(1-\hat{\pi}_i)(1-\hat{h}_i)]}}, \tag{15.3}$$

donde usa el *leverage* $(\hat{h}_i)$. El leverage mide la influencia de una observación individual en la estimación de los parámetros de un modelo de regresión. Más específicamente, se relaciona con qué tan extremos o diferentes son los valores de las variables independientes de una observación en comparación con las demás observaciones. El valor $\hat{h}_i$ se define como el elemento i-ésimo de la diagonal de la matriz $\boldsymbol{H_{at}} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X^T W X})^{-1}\boldsymbol{X^T W}^{1/2}$.

La principal ventaja que presenta este residuo si lo comparamos con el de Pearson y con la desviación residual es que sigue aproximadamente una distribución $N(0,1)$ cuando el modelo es válido. Los valores absolutos mayores que aproximadamente 2 o 3 proporcionan evidencia de falta de ajuste.

16. Resumiendo la Capacidad Predictiva del Modelo de Regresión Logística

En una regresión ordinaria, R^2 es una medida estadística utilizada para evaluar qué tan bien se ajusta un modelo de regresión lineal a los datos observados. Mide la proporción de la variabilidad total de la variable dependiente que puede ser explicada por el modelo de regresión. El valor de R^2 varía de 0 a 1, siendo 0 indicativo de que el modelo no explica ninguna variabilidad en los datos y 1 indicativo de que el modelo explica toda la variabilidad en los datos.

A pesar de varios intentos de definir medidas análogas a R^2 para datos categóricos, no hay ninguna medida tan útil como R^2. En esta sección presentamos algunos métodos para resumir la capacidad predictiva del modelo.

16.1. Resumiendo la Capacidad Predictiva: Medidas R y R^2

Para cualquier GLM, la correlación $R(\mathbf{y}, \hat{\mu})$ entre las respuestas observadas $\{y_i\}$ y los valores pronosticados por el modelo $\{\hat{\mu}\}$ mide la capacidad predictiva. En la regresión de mínimos cuadrados, R es la correlación múltiple entre Y y los predictores.

En una regresión logística, $\hat{\mu}_i$ es la estimación de la probabilidad $\hat{\pi}_i$ para la observación binaria i. Entonces, $R(\mathbf{y}, \hat{\mu})$ es la correlación entre las n observaciones binarias $\{y_i\}$ (1 o 0 cada una) y sus probabilidades estimadas. La naturaleza altamente discreta de la variable Y puede limitar el rango de valores posibles de R. Sin embargo, R es útil para comparar el ajuste de diferentes modelos para los mismos datos. El hecho de que un modelo de regresión tenga muchos predictores puede conducir a una sobreestimación del coeficiente de correlación $R(\mathbf{Y}, E(\mathbf{Y}|\mathbf{X}))$. Esto se debe a que, cuando hay muchos predictores, la variabilidad de los datos puede ser "explicada"de manera artificial por los predictores, lo que lleva a una sobreestimación del coeficiente de correlación R. Por lo tanto, es importante tener cuidado al comparar los valores de R de diferentes modelos con diferentes grados de libertad.

Otra manera de medir la relación entre las respuestas binarias $\{y_i\}$ y sus valores ajustados $\{\hat{\pi}_i\}$ es usando la reducción proporcional del error cuadrático

$$1 - \frac{\sum_i (y_i - \hat{\pi}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{16.1}$$

que se obtiene usando $\hat{\pi}_i$ en lugar de $\bar{y} = \sum_j y_j/n$ como predictor de y_i. Anemiya (1981) sugirió una medida relacionada que pondera las desviaciones al cuadrado por las varianzas predichas inversas. A medida que se agregan más variables explicativas a un GLM, estas medidas y $R(\mathbf{y}, \hat{\mu})$ deben aumentar o mantenerse iguales. Sin embargo, en la regresión logística, no es necesario que estas medidas sigan este patrón. Estas medidas pueden depender fuertemente del rango de valores observados de las variables explicativas. Además, cuando se calculan para datos de muestra, estas medidas están sesgadas hacia arriba, lo que significa que tienden a sobreestimar las medidas poblacionales correspondientes. Es posible corregir el sesgo pero no entraremos en detalles acerca de esto en este trabajo (Liao y McGee 2003).

16.2. Resumiendo la Capacidad Predictiva: Medidas de Verosimilitud y Desviación

Otras medidas de la capacidad predictiva usan directamente la función de verosimilitud. Sea L_M el logaritmo de la verosimilitud maximizado de un modelo dado, sea L_S el mismo concepto pero para el modelo saturado, y L_0 para el modelo nulo, que es aquel que solo tiene término independiente. Como las probabilidades no son mayores que 1, el logaritmo neperiano de la verosimilitud es negativa. A medida que la complejidad del modelo incrementa, el espacio de parámetros aumenta y por lo tanto el log de la verosimilitud maximizado puede ser mayor. Esto es $L_0 \leq L_M \leq L_S \leq 0$. La medida

$$\frac{L_M - L_0}{L_S - L_0} \tag{16.2}$$

está entre 0 y 1. Es igual a 0 cuando el modelo no presenta ninguna mejora en el ajuste en comparación con el modelo nulo y es igual a 1 cuando el modelo ajusta los datos tan bien como lo hace el modelo saturado. Sin embargo, esta medida tiene la desventaja de que la interpretación del log de la verosimilitud es compleja. Interpretar el valor numérico es difícil, excepto en un sentido comparativo para diferentes modelos.

Para N observaciones de Bernoulli independientes el logaritmo de la verosimilitud maximizado es

$$\log \prod_{i=1}^{N} [\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{1-y_i}] = \sum_{i=1}^{N} [y_i \log \hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i)]. \tag{16.3}$$

En el modelo nulo $\hat{\pi}_i = (\sum_i y_i)/N = \bar{y}$, esto implica

$$L_0 = N[\bar{y}(\log \bar{y}) + (1-\bar{y})\log(1-\bar{y})]. \tag{16.4}$$

El modelo saturado tiene un parámetro para cada sujeto, esto supone que $\hat{\pi}_i = y_i$ para cualquier i. De este modo, como $y_i \in \{0,1\}$ o bien $\hat{\pi}_i = y_i = 0$ o $(1-\hat{\pi}_i) = (1-y_i) = 0$ entonces tenemos que $L_S = 0$. Por lo tanto (16.2) se puede simplificar:

$$D = \frac{L_0 - L_M}{L_0}. \tag{16.5}$$

Cabe remarcar que esta medida fue propuesta por McFadden (1974).

16.3. Resumiendo la Capacidad Predictiva: Tablas de Clasificación

Una tabla de clasificación es una matriz que muestra la cantidad de casos clasificados correctamente e incorrectamente por un modelo. En el caso de la regresión logística muestra si el modelo ha predicho correctamente el valor $y = 0$ o $y = 1$. La predicción para la observación i es $\hat{y} = 1$ cuando $\hat{\pi}_i > \pi_0$ e $\hat{y} = 0$ cuando $\hat{\pi}_i \leq \pi_0$, para algún punto de corte π_0. Una posibilidad es $\pi_0 = 0{,}50$. Otra es usar la proporción de la muestra de 1, que es $\hat{\pi}_i$ cuando el modelo solo contiene el término independiente. En lugar de utilizar $\hat{\pi}_i$ del modelo ajustado a todo el conjunto de datos (incluyendo y_i), es preferible utilizar $\hat{\pi}_i$ basado en el modelo ajustado a las otras $n-1$ observaciones.

Usando tablas de clasificación podemos resumir la capacidad predictiva:

$$\text{sensibilidad} = P(\hat{y} = 1 | y = 1) \qquad y \qquad \text{especificidad} = P(\hat{y} = 0 | y = 0). \tag{16.6}$$

Un resumen general de la capacidad predictiva es la proporción de clasificaciones correctas. Esto estima

$$\begin{aligned} P(\text{clasificación correcta}) &= P(y = 1 \cap \hat{y} = 1) + P(y = 0 \cap \hat{y} = 0) \\ &= P(\hat{y} = 1 | y = 1)P(y = 1) + P(\hat{y} = 0 | y = 0)P(y = 0), \end{aligned}$$

que es un promedio ponderador de la sensibilidad y la especificidad.

Las tablas de clasificación también tienen limitaciones: convierte los valores continuos $\hat{\pi}$ en binarios. La elección de π_0 es arbitraria. Los resultados son sensibles a las proporciones relativas de las veces que $y = 1$ y $y = 0$. Por ejemplo, si una pequeña proporción de observaciones cumple $y = 1$, el ajuste del modelo nunca debe tener $\hat{\pi}_i > 0{,}50$, en este caso nunca se predice $\hat{y} = 1$. Una vez más, su uso es principalmente para comparar diferentes modelos con los mismos datos.

16.4. Resumiendo la Capacidad Predictiva: Curva ROC

Las tablas de clasificación dependen del punto de corte π_0. Una curva de Característica Operativa del Receptor, ROC por sus siglas en inglés, es un gráfico de la sensibilidad en función de (1-especificidad) para los posibles valores de π_0. Una curva ROC es más informativa que una tabla de clasificación, ya que resume el poder predictivo para todos los posibles valores de π_0. Cuando π_0 es aproximadamente 0, casi todas las predicciones $\hat{y}_i = 1$, entonces la sensibilidad está cerca de 1 y la especificidad es prácticamente 0, y el punto (1−especificidad, sensibilidad)$\approx (1, 1)$. Cuando π_0 se aproxima a 1, la mayoría de predicciones son $\hat{y}_i = 0$, por lo tanto la sensibilidad es casi 0 y la especificidad es aproximadamente 1, y (1−especificidad, sensibilidad)$\approx (0, 0)$. La curva ROC normalmente tiene una forma cóncava que conecta los puntos $(0, 0)$ y $(1, 1)$.

Para una especificidad dada, una mejor capacidad predictiva corresponde a una mayor sensibilidad. Por lo tanto, cuanto mayor es la capacidad predictiva, mayor es la curva ROC. A modo de resumen, cuanto mayor es el área por debajo de la curva ROC, mejor son las predicciones. De hecho, el área bajo la curva ROC llamada AUC es la medida que se utiliza para interpretar las curvas ROC. Estea área tiene un valor que pertenece al intervalo [0.5, 1], dónde 1 representa un valor diagnóstico perfecto y 0,5 corresponde a un modelo sin capacidad predictiva. En la interpretación de las curvas ROC, se han establecido los siguientes intervalos para los valores de AUC. Cuando el valor de AUC es igual a 0.5, indica una capacidad predictiva muy limitada. En el rango de valores entre 0.5 y 0.6, se considera que el test es malo, con una precisión muy baja y poca confiabilidad para realizar predicciones precisas. Si el valor de AUC se encuentra entre 0.6 y 0.75, el test se clasifica como regular, lo que indica una capacidad predictiva mejorada, aunque modesta en términos de precisión. Si pertenece al intervalo [0.75, 0.9] se considera que el test es bueno, con una precisión razonable y una capacidad para realizar predicciones con un grado aceptable de confianza. Cuando el valor de AUC se encuentra entre 0.9 y 0.97, se considera que el test es muy bueno, con una alta precisión y una confiabilidad alta para realizar predicciones precisas. Finalmente, en el rango de valores entre 0.97 y 1, se clasifica como un test excelente, con una capacidad predictiva casi perfecta.

Parte II
ENFOQUE PRÁCTICO

17. Elección del Tema de Estudio

El objetivo de este estudio se centra en la aplicación del modelo logístico para calcular la probabilidad de obtener un sueldo alto entre los titulados universitarios. El tema elegido es de suma relevancia, ya que el nivel de ingresos constituye una medida significativa del éxito profesional y del retorno de la inversión en educación superior.

La elección de este enfoque se basa en el interés por comprender los factores que influyen en la probabilidad de alcanzar un sueldo alto después de obtener un título universitario. Para ello, se utilizará el modelo logístico, que permite modelar una variable binaria (en este caso, la categoría de "sueldo alto" o "no sueldo alto") en función de diversas variables independientes, como la especialización, la experiencia laboral, el tamaño de la empresa, entre otras.

18. Datos utilizados

Los datos para realizar el estudio han sido extraídos del Instituto Nacional de Estadística (INE). Concretamente de la *Encuesta de Inserción Laboral de Titulados Universitarios*. El propósito fundamental de la encuesta radica en comprender la situación laboral de los graduados universitarios, así como los múltiples aspectos relacionados con su transición al empleo, es decir, su entrada al mercado laboral. El tamaño muestral es de 31651 graduados universitarios.

Los ámbitos de la investigación fueron los que se comentan a continuación. La *población objetivo* de la encuesta abarca a los individuos que han completado su educación universitaria. El *ámbito geográfico* de estudio se extiende a todo el territorio español, es decir, abarca a los graduados de las universidades españolas. Por último, para determinar el *marco temporal*, se ha tomado en consideración el tiempo necesario para que los individuos establezcan una conexión sólida con el mercado laboral después de concluir sus estudios. Con esto en mente, se ha decidido que los encuestados sean los del curso académico 2013-2014, realizando la encuesta a finales de 2019.

Como el objetivo de nuestro modelo logístico será calcular la probabilidad de tener un sueldo elevado, debemos eliminar de la muestra aquellos individuos que se encuentran en situación de desempleo o de inactividad. Para ello, observamos la variable $TRBPRN1$, que indica la situación laboral actual del encuestado asignando un 1 a las personas que están trabajando, un 2 a los que se encuentran en desempleo y un 3 a los inactivos. En otras palabras, debemos eliminar de la muestra todas las observaciones tales que $TRBPRN1_i \in \{2, 3\}\ \forall i = 1, 2, \ldots, 31651$. Esto nos deja con una muestra de $31651 - 4527 = 27124$ individuos.

19. Especificación del Modelo

En esta sección seleccionaremos las variables que intervendrán en nuestro modelo y en caso de que sea necesario recodificaremos estas variables.

19.1. Variable Endógena

La variable endógena que utilizaremos es la variable de la encuesta del INE llamada TR_SUELDO que indica el sueldo mensual actual neto. La codificación original asigna un número entero a cada respuesta como muestra la Tabla 3 a continuación.

código	descripción
1	Menos de 700 euros
2	De 700 a 999 euros
3	De 1000 a 1499 euros
4	De 1500 a 1999 euros
5	De 2000 a 2499 euros
6	De 2500 a 2999 euros
7	De 3000 euros en adelante
9	NS/NC
333	No aplicable

Tabla 3: Codificación original de la variable TR_SUELDO propuesta por el INE

Observamos que las respuestas que tienen un 333 asignado son aquellas que corresponden a personas inactivas o en desempleo que ya han sido eliminadas de la muestra. Además debemos eliminar aquellas que su respuesta ha sido "No sabe o no contesta", abreviado NS/NC, y a la cual se le ha asignado el número 9. Esto nos deja con una muestra de tamaño $N = 27124 - 739 = 26385$.

Sin embargo, esta codificación no es útil para nuestro estudio, pues nuestra variable endógena TR_SUELDO, no es una variable binaria. Con el objetivo de estudiar la probabilidad de tener un sueldo elevado recodificaremos la variable TR_SUELDO usando la variable ficticia Y de la siguiente manera:

$$Y_i = \begin{cases} 0 & si \quad TR\_SUELDO_i \in \{1,2,3,4,5\} \\ 1 & si \quad TR\_SUELDO_i \in \{6,7\} \end{cases} \qquad \forall i = 1,2,\ldots,N.$$

Considerando de esta manera que un sueldo mensual neto elevado es aquel que iguala o supera los 2500 euros.

Una vez que hemos codificado la variable endógena, procedemos a seleccionar y codificar las variables explicativas.

19.2. Variables explicativas

En el Anexo 1 se encuentran las tablas de contingencia de las variables del modelo. Las variables predictivas que añadiremos al modelo son las siguientes:

- $SEXO$: si la persona es hombre o mujer.

- $EDAD$: menor de 30 años, entre 30 y 34 años y 35 o mayor.
- $NACIO$: si la nacionalidad es únicamente española, únicamente otra o ambas.
- $RAMA$: la rama de la titulación puede ser Artes, Ciencias, Ciencias sociales y jurídicas, Ingeniería y arquitectura y Ciencias de la salud.
- T_UNIV: el tipo de universidad puede ser pública o privada y presencial o a distancia.
- $DISCA$: si la persona presenta una discapacidad superior al 33 %.
- EST_B2_2: si disfrutó de alguna beca, en concreto un premio o beca de excelencia.
- EST_M2: si realizó un programa o beca de movilidad
- EST_B11_2: si la persona tiene un máster.
- $IDIOMAS$: número de idiomas que habla sin contar las lenguas maternas (0, 1, 2, 3, 4 o 5 o más).
- TIC: capacidad de usar el ordenador que puede ser: básica, intermedia o avanzada.
- SIT_PRO: la situación profesional puede ser: trabajador en prácticas, asalariado con trabajo permanente, asalariado con trabajo temporal, empresario con asalariados, trabajador independiente o ayuda en la empresa familiar.
- $JORNADA$: parcial o completa.
- TR_TAM: la empresa puede no tener trabajadores, ser una micro empresa, pequeña, mediana o grande.
- HL_E1: si se han realizado prácticas.

La codificación explícita de las variables y cómo se han creado las variables ficticias necesarias se encuentra en el Anexo 2.

19.3. Eliminación de Observaciones sin Información Relevante

Antes de estimar el modelo, eliminaremos de la muestra aquellas observaciones que cumplen

$$EST\_B2\_2_i = 9 \vee EST\_M2_i = 9 \vee IDIOMAS_i = 9 \vee TIC_i = 9 \vee TR\_TAM_i = 9$$

El INE utiliza el número 9 para hacer referencia a aquellas respuestas que el encuestado no sabe o no contestan. Como no son relevantes para la estimación de nuestro modelo, eliminaremos estas observaciones de la muestra.

El número de observaciones que cumplen la condición anterior son 554, entonces, haciendo un abuso de notación, $N = 26385 - 554 = 25831$.

20. Estimación del Modelo

Una vez que hemos seleccionado y codificado las variables explicativas relevantes para nuestro modelo, procederemos a estimar los parámetros utilizando el programa R. R es un lenguaje de programación ampliamente utilizado en análisis estadístico que proporciona una variedad de funciones y paquetes para ajustar modelos estadísticos. A continuación se muestra la salida del programa al estimar el logit de la probabilidad de tener un sueldo elevado de los titulados universitarios usando las variables ficticias de la sección anterior. Al modelo resultante le denominamos LOGIT.1. Llama la atención un mensaje en la última línea que indica que un coeficiente no ha podido ser estimado debido a singularidades.

```
LOGIT.1
Call:
glm(formula = Y ~ SEXO_M + EDAD_30_34 + EDAD_35 + NACIO_ESP_OTR + NACIO_OTR +
    RAMA_CIENCIAS + RAMA_CCSS + RAMA_ING + RAMA_SALUD + T_UNIV_PUB_DIST +
    T_UNIV_PRIV_PRE + T_UNIV_PRIV_DIST + DISCA + EST_B2_2_OTRA + EST_B2_2_NO +
    EST_M2_NO + EST_B11_2_NO + IDIOMAS_1 + IDIOMAS_2 + IDIOMAS_3 + IDIOMAS_4 +
    IDIOMAS_5 + TIC_INTER + TIC_AVAN + SIT_PRO_AS_IND + SIT_PRO_AS_TEMP +
    SIT_PRO_EMPR + SIT_PRO_IND + SIT_PRO_FAM + JORNADA_COMPL +  TR_TAM_MICRO +
    TR_TAM_PEQUE + TR_TAM_MEDI + TR_TAM_GRAN + HL_E1_NO,
    family = binomial(logit), data = DATADEF)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8932  -0.4665  -0.3032  -0.1559   3.8914

Coefficients: (1 not defined because of singularities)
```

Esto es debido a que dos de las variables de nuestro modelo están perfectamente correlacionadas. Es lógico pensar que la variable "situación profesional" (SIT_PRO) y la variable "Número de personas trabajando en la empresa donde trabaja actualmente" (TR_TAM) están correlacionadas cuando $SIT\_PRO = 5$ ="Trabajador independiente o empresario sin asalariados" y $TR\_TAM = 0$ ="Trabajadores independientes o autónomos sin asalariados". Por este motivo debemos eliminar una de las variables ficticias que toma el valor 1 cuando $SIT\_PRO = 5$ o $TR\_TAM = 0$; elegiremos eliminar la variable ficticia SIT_PRO_IND y cambiar la categoría base de la variable TR_TAM (añadiendo la variable ficticia TR_TAM_IND y eliminando TR_TAM_MICRO). Volvemos a estimar el modelo (ahora llamado LOGIT.2) obteniendo los resultados que se muestran a continuación.

```
LOGIT.2
Call:
glm(formula = Y ~ SEXO_M + EDAD_30_34 + EDAD_35 + NACIO_ESP_OTR +
    NACIO_OTR + RAMA_CIENCIAS + RAMA_CCSS + RAMA_ING + RAMA_SALUD +
    T_UNIV_PUB_DIST + T_UNIV_PRIV_PRE + T_UNIV_PRIV_DIST + DISCA +
    EST_B2_2_OTRA + EST_B2_2_NO + EST_M2_NO + EST_B11_2_NO +
    IDIOMAS_1 + IDIOMAS_2 + IDIOMAS_3 + IDIOMAS_4 + IDIOMAS_5 +
    TIC_INTER + TIC_AVAN + SIT_PRO_AS_IND + SIT_PRO_AS_TEMP +
    SIT_PRO_EMPR + SIT_PRO_FAM + JORNADA_COMPL + TR_TAM_IND +
    TR_TAM_PEQUE + TR_TAM_MEDI + TR_TAM_GRAN + HL_E1_NO, family = binomial(logit),
    data = DATADEF)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8932  -0.4665  -0.3032  -0.1559   3.8914

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -6.48952    0.53636 -12.099  < 2e-16 ***
SEXO_M[T.1]           -0.71754    0.05068 -14.159  < 2e-16 ***
EDAD_30_34[T.1]       -0.15781    0.06044  -2.611 0.009024 **
EDAD_35[T.1]           0.13989    0.06678   2.095 0.036204 *
NACIO_ESP_OTR[T.1]     0.62670    0.19026   3.294 0.000988 ***
NACIO_OTR[T.1]        -0.05645    0.23071  -0.245 0.806715
RAMA_CIENCIAS[T.1]     0.14299    0.15691   0.911 0.362154
RAMA_CCSS[T.1]         0.63999    0.12223   5.236 1.64e-07 ***
RAMA_ING[T.1]          1.15258    0.12412   9.286  < 2e-16 ***
RAMA_SALUD[T.1]        1.60615    0.12838  12.511  < 2e-16 ***
T_UNIV_PUB_DIST[T.1]   0.60832    0.10870   5.596 2.19e-08 ***
T_UNIV_PRIV_PRE[T.1]   0.23930    0.06552   3.652 0.000260 ***
T_UNIV_PRIV_DIST[T.1]  0.52706    0.12081   4.363 1.29e-05 ***
DISCA                 -0.29393    0.20486  -1.435 0.151352
EST_B2_2_OTRA[T.1]    -1.11360    0.09216 -12.083  < 2e-16 ***
EST_B2_2_NO[T.1]      -0.67862    0.09240  -7.344 2.07e-13 ***
EST_M2_NO[T.1]        -0.49884    0.08151  -6.120 9.36e-10 ***
EST_B11_2_NO[T.1]     -0.05403    0.04837  -1.117 0.263991
IDIOMAS_1[T.1]        -0.05822    0.12417  -0.469 0.639161
IDIOMAS_2[T.1]         0.32670    0.12632   2.586 0.009701 **
IDIOMAS_3[T.1]         0.55937    0.13986   4.000 6.35e-05 ***
IDIOMAS_4[T.1]         1.02840    0.17995   5.715 1.10e-08 ***
IDIOMAS_5[T.1]         1.00603    0.30934   3.252 0.001145 **
TIC_INTER[T.1]         0.23172    0.08994   2.577 0.009979 **
TIC_AVAN[T.1]          0.38993    0.10003   3.898 9.70e-05 ***
SIT_PRO_AS_IND[T.1]    1.64420    0.16433  10.006  < 2e-16 ***
SIT_PRO_AS_TEMP[T.1]   1.16292    0.16856   6.899 5.23e-12 ***
SIT_PRO_EMPR[T.1]      3.47432    0.21862  15.892  < 2e-16 ***
SIT_PRO_FAM[T.1]       1.18218    0.62383   1.895 0.058087 .
JORNADA_COMPL[T.1]     2.34517    0.21781  10.767  < 2e-16 ***
TR_TAM_IND[T.1]        2.79403    0.21317  13.107  < 2e-16 ***
TR_TAM_PEQUE[T.1]      0.63617    0.12408   5.127 2.94e-07 ***
TR_TAM_MEDI[T.1]       0.97034    0.12273   7.906 2.65e-15 ***
TR_TAM_GRAN[T.1]       1.56777    0.11461  13.679  < 2e-16 ***
HL_E1_NO[T.4]          0.12180    0.05490   2.219 0.026518 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15547  on 25830  degrees of freedom
Residual deviance: 13127  on 25796  degrees of freedom
AIC: 13197

Number of Fisher Scoring iterations: 7


> exp(coef(LOGIT.2))  # Exponentiated coefficients ("odds ratios")
        (Intercept)          SEXO_M[T.1]       EDAD_30_34[T.1]          EDAD_35[T.1]    NACIO_ESP_OTR[T.1]
         0.00151928           0.48795187           0.85401226           1.15014477           1.87141959
```

NACIO_OTR[T.1]	RAMA_CIENCIAS[T.1]	RAMA_CCSS[T.1]	RAMA_ING[T.1]	RAMA_SALUD[T.1]
0.94511559	1.15371705	1.89646552	3.16636434	4.98357016
T_UNIV_PUB_DIST[T.1]	T_UNIV_PRIV_PRE[T.1]	T_UNIV_PRIV_DIST[T.1]	DISCA	EST_B2_2_OTRA[T.1]
1.83733749	1.27035471	1.69394061	0.74532943	0.32837441
EST_B2_2_NO[T.1]	EST_M2_NO[T.1]	EST_B11_2_NO[T.1]	IDIOMAS_1[T.1]	IDIOMAS_2[T.1]
0.50731684	0.60723650	0.94740140	0.94344236	1.38638809
IDIOMAS_3[T.1]	IDIOMAS_4[T.1]	IDIOMAS_5[T.1]	TIC_INTER[T.1]	TIC_AVAN[T.1]
1.74956897	2.79658982	2.73473315	1.26077182	1.47688144
SIT_PRO_AS_IND[T.1]	SIT_PRO_AS_TEMP[T.1]	SIT_PRO_EMPR[T.1]	SIT_PRO_FAM[T.1]	JORNADA_COMPL[T.1]
5.17688175	3.19925437	32.27596931	3.26148617	10.43505017
TR_TAM_IND[T.1]	TR_TAM_PEQUE[T.1]	TR_TAM_MEDI[T.1]	TR_TAM_GRAN[T.1]	HL_E1_NO[T.4]
16.34679190	1.88923266	2.63884255	4.79594758	1.12953211

La interpretación de los signos de los parámetros y de los odds ratio (OR) en un modelo logístico con variables categóricas es fundamental para comprender la relación entre las variables independientes y la variable dependiente. A continuación, se presentan las interpretaciones sobre los signos y los OR obtenidos en la estimación de la Figura 5.

El signo del parámetro que acompaña a la variable ficticia $SEXO\_M$ es negativo, esto implica que la probabilidad de tener un sueldo elevado para las mujeres es menor que para los hombres. El OR de esta variable tiene un valor de $0,488$, lo que también indica una probabilidad mucho menor de que ocurra el evento de interés en comparación con la categoría base, que en este caso es ser hombre.

En relación a las variables de edad, el OR de 0.854 para $EDAD\_30\_34$ sugiere que la categoría correspondiente tiene una probabilidad ligeramente más baja de que se cobre un sueldo elevado en comparación con la categoría base que es ser menor de 30 años. Por otro lado, el OR de 1.150 para $EDAD\_35$ indica una probabilidad ligeramente más alta de que ocurra el evento de interés en comparación con la categoría base.

En cuanto a la variable $NACIO\_ESP\_OTR$, se obtuvo un odds ratio de 1.8714, lo que indica un aumento en las probabilidades de tener un suledo alto en comparación con la categoría base que es tener únicamente la nacionalidad española. Por otro lado, la variable $NACIO\_OTR$ mostró un odds ratio de 0.9451, lo que implica una ligera disminución en las probabilidades. Como es de esperar, el coeficiente que acompaña a la variable $NACIO\_ESP\_OTR$ es positivo mientras que el que acompaña a $NACIO\_OTR$ es negativo.

Al considerar las variables relacionadas con las ramas de estudio, encontramos que la variable $RAMA\_CIENCIAS$ presenta un odds ratio de 0.9451, lo que implica una ligera disminución en las probabilidades del suceso estudiado en comparación con los titulados en una carrera de artes (que es la categoría base). Por otro lado, la variable $RAMA\_CCSS$ muestra un odds ratio de 1.1537, lo que representa un aumento en las probabilidades. Además, la variable $RAMA\_ING$ tiene un odds ratio de 1.8965, indicando un aumento significativo en las probabilidades, mientras que la variable $RAMA\_SALUD$ presenta un odds ratio de 3.1664, lo que implica un incremento sustancial de las probabilidades.

En cuanto al tipo de universidad, los resultados revelan que la variable $T\_UNIV\_PUB\_DIST$ tiene un odds ratio de 1.8373, mostrando un aumento significativo en las probabilidades en comparación con aquellas personas que estudiaron en una universidad pública presencial. Por otro lado, la variable $T\_UNIV\_PRIV\_PRE$ presenta un odds ratio de 1.2704, lo que implica un incremento en las probabilidades. Además, la variable $T\_UNIV\_PRIV\_DIST$ muestra un odds ratio de 1.6939, indicando un aumento en las probabilidades de tener un sueldo elevado.

La variable $DISCA$ presenta un odds ratio de 0.7453, lo que implica una disminución en las probabilidades de tener un sueldo elevado en comparación con las personas que no tienen una discapacidad. Esto sugiere que la presencia de una discapacidad puede influir

en las probabilidades del evento considerado.

Asimismo, se encontraron odds ratio diversos para el resto de variables. Estos odds ratio nos indican aumentos o disminuciones en las probabilidades en comparación con las categorías base correspondientes.

21. Validación del Modelo

Después de estimar un modelo logístico concreto, es fundamental evaluar la significación estadística de las variables predictoras incluidas en el modelo. Esto nos permite determinar si dichas variables tienen un efecto significativo en la probabilidad que tienen los titulados universitarios de cobrar un sueldo elevado. Seguiremos una estrategia mixta: primero realizaremos una eliminación hacia atrás y luego una pequeña selección hace delante.

Comenzaremos realizando un test de razón de verosimilitudes para comprobar si la variable $DISCA$ es realmente significativa. Sea β_0 el término independiente, sea β_1 el coeficiente que acompaña a la primera variable ficticia ($SEXO\_M$), β_2 el parámetro que acompaña a la segunda variable ficticia y así hasta β_{35}. Podría parecer que $\beta_{13}(= -0,29393)$, que es el coeficiente que acompaña a la variable que toma un 1 cuando la persona titulada es discapacitada, es poco significativa ya que se aproxima mucho a cero. Sin embargo, a simple vista no podemos confirmarlo, para hacerlo debemos realizar un test de hipótesis como el de la razón de verosimilitud.

Concretamente queremos contrastar

$$\begin{cases} H_0: & \beta_{13} = 0 \\ H_1: & \beta_{13} \neq 0 \end{cases}.$$

Para ello necesitaremos el logaritmo de las funciones de máxima verosimilitud del modelo que cumple la restricción de la hipótesis nula y la que cumple la restricción de la hipótesis alternativa. El modelo que cumple $\beta_{13} \neq 0$ ya lo hemos estimado, es el modelo LOGIT.2.

Existen dos vías para encontrar el logaritmo de la función de verosimilitud maximizada. La primera es usar el valor Akaike. Anteriormente vimos que

$$\text{AIC} = -2(\text{log verosimilitud maximizado} - \text{número de parámetros en el modelo}).$$

Por lo tanto,

$$\text{log verosimilitud maximizado} = \text{número de parámetros en el modelo} - \frac{1}{2}\text{AIC}$$

Sea $\Theta_0 = \{\beta = (\beta_1, \ldots, \beta_{35}) \in \mathbb{R}^{35} : \beta_{13} = 0\}$, y si calculamos usando R el AIC con tres decimales obtenemos

```
> aic <- AIC(LOGIT.2)
> aic_decimal <- sprintf("%.3f", aic)
> print(aic_decimal)
[1] "13197.476"
```

Entonces, ya podemos calcular el logaritmo de la función de verosimilitud maximizado:

$$\begin{aligned}\sup_{\beta\in\mathbb{R}^{35}} L &= \text{número de parámetros en el modelo LOGIT.2} - \frac{1}{2}\cdot \text{AIC}_{\text{LOGIT.2}}\\ &= 35 - \frac{1}{2}\cdot 13197,476\\ &= -6563,738.\end{aligned}$$

El segundo método para encontrar este valor es usar directamente R de la siguiente manera

```
> logLik(LOGIT.2.)
'log Lik.' -6563.738 (df=35)
```

Para calcular $\sup_{\beta\in\Theta_0} L$ debemos calcular un modelo exactamente igual que LOGIT.2 pero omitiendo la variable predictora $DISCA$. Este modelo recibirá el nombre de LOGIT.3. A continuación insertamos su estimación en R:

```
LOGIT.3
Call:
glm(formula = Y ~ SEXO_M + EDAD_30_34 + EDAD_35 + NACIO_ESP_OTR +
    NACIO_OTR + RAMA_CIENCIAS + RAMA_CCSS + RAMA_ING + RAMA_SALUD +
    T_UNIV_PUB_DIST + T_UNIV_PRIV_PRE + T_UNIV_PRIV_DIST + EST_B2_2_OTRA +
    EST_B2_2_NO + EST_M2_NO + EST_B11_2_NO + IDIOMAS_1 + IDIOMAS_2 +
    IDIOMAS_3 + IDIOMAS_4 + IDIOMAS_5 + TIC_INTER + TIC_AVAN +
    SIT_PRO_AS_IND + SIT_PRO_AS_TEMP + SIT_PRO_EMPR + SIT_PRO_FAM +
    JORNADA_COMPL + TR_TAM_IND + TR_TAM_PEQUE + TR_TAM_MEDI +
    TR_TAM_GRAN + HL_E1_NO, family = binomial(logit), data = DATADEF)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8934  -0.4663  -0.3035  -0.1561   3.8901

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -7.06902    0.35399 -19.969  < 2e-16 ***
SEXO_M[T.1]           -0.71694    0.05067 -14.149  < 2e-16 ***
EDAD_30_34[T.1]       -0.15707    0.06043  -2.599 0.009347 **
EDAD_35[T.1]           0.14532    0.06666   2.180 0.029272 *
NACIO_ESP_OTR[T.1]     0.62436    0.19027   3.281 0.001033 **
NACIO_OTR[T.1]        -0.05820    0.23070  -0.252 0.800831
RAMA_CIENCIAS[T.1]     0.14168    0.15690   0.903 0.366548
RAMA_CCSS[T.1]         0.64045    0.12223   5.240 1.61e-07 ***
RAMA_ING[T.1]          1.15125    0.12411   9.276  < 2e-16 ***
RAMA_SALUD[T.1]        1.60675    0.12838  12.516  < 2e-16 ***
T_UNIV_PUB_DIST[T.1]   0.60709    0.10874   5.583 2.36e-08 ***
T_UNIV_PRIV_PRE[T.1]   0.23693    0.06549   3.618 0.000297 ***
T_UNIV_PRIV_DIST[T.1]  0.52519    0.12080   4.348 1.38e-05 ***
EST_B2_2_OTRA[T.1]    -1.11255    0.09215 -12.073  < 2e-16 ***
EST_B2_2_NO[T.1]      -0.67903    0.09240  -7.349 1.99e-13 ***
EST_M2_NO[T.1]        -0.49802    0.08149  -6.111 9.88e-10 ***
```

```
EST_B11_2_NO[T.1]     -0.05434    0.04837  -1.123 0.261295
IDIOMAS_1[T.1]        -0.06334    0.12407  -0.511 0.609671
IDIOMAS_2[T.1]         0.32145    0.12622   2.547 0.010873 *
IDIOMAS_3[T.1]         0.55347    0.13974   3.961 7.47e-05 ***
IDIOMAS_4[T.1]         1.02149    0.17980   5.681 1.34e-08 ***
IDIOMAS_5[T.1]         1.00015    0.30936   3.233 0.001225 **
TIC_INTER[T.1]         0.23188    0.08994   2.578 0.009934 **
TIC_AVAN[T.1]          0.38989    0.10004   3.897 9.73e-05 ***
SIT_PRO_AS_IND[T.1]    1.64247    0.16427   9.998  < 2e-16 ***
SIT_PRO_AS_TEMP[T.1]   1.16000    0.16850   6.884 5.81e-12 ***
SIT_PRO_EMPR[T.1]      3.47209    0.21859  15.884  < 2e-16 ***
SIT_PRO_FAM[T.1]       1.17880    0.62381   1.890 0.058800 .
JORNADA_COMPL[T.1]     2.34350    0.21780  10.760  < 2e-16 ***
TR_TAM_IND[T.1]        2.79199    0.21315  13.098  < 2e-16 ***
TR_TAM_PEQUE[T.1]      0.63777    0.12411   5.139 2.77e-07 ***
TR_TAM_MEDI[T.1]       0.97305    0.12276   7.927 2.25e-15 ***
TR_TAM_GRAN[T.1]       1.57016    0.11464  13.697  < 2e-16 ***
HL_E1_NO[T.4]          0.12156    0.05491   2.214 0.026839 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15547  on 25830  degrees of freedom
Residual deviance: 13129  on 25797  degrees of freedom
AIC: 13197

Number of Fisher Scoring iterations: 7
```

Calculamos el logaritmo de la función de verosimilitud maximizado

```
> logLik(LOGIT.3)
'log Lik.' -6564.709 (df=34)
```

obteniendo $\sup_{\beta\in\Theta_0} L = -6564{,}709$, entonces ya podemos obtener la razón de verosimilitudes

$$\Lambda = \frac{\exp(-6564,709)}{\exp(-6563,738)} = e^{6563{,}738-6564{,}709}$$

y el estadístico de Wilks

$$W = -2\log\Lambda = 2\cdot 6564,709 - 2\cdot 6563,738 = 1,942 \sim \chi^2_{35-34} = \chi^2_1$$

El valor en tablas de χ^2_1 es $3,8415$. Como $W = 1,942 < 3,8415$ no rechazamos la hipótesis nula y por lo tanto β_{13} del modelo LOGIT.2 no es significativo y debemos aceptar el modelo LOGIT.3. Con esto llegamos a la conclusión de que tener una discapacidad no influye en la probabilidad de que los graduados universitarios tengan un sueldo elevado.

A continuación vamos a estudiar la significación de la variable EST_B11_2. Recordamos que esta variable hace referencia a si el graduado estudió un máster. La variable ficticia que hemos añadido al modelo asigna un 1 a aquellas personas que no estudiaron

un máster ($EST\_B11\_2\_NO$). Si usamos la notación anterior adaptada al modelo LOGIT.3, ahora el coeficiente que acompaña a la variable $EST\_B11\_2\_NO$ es β_{16}. Queremos contrastar

$$\begin{cases} H_0: & \beta_{16} = 0 \\ H_1: & \beta_{16} \neq 0 \end{cases}$$

pero esta vez usaremos el test de Wald. El estadístico de prueba es

$$z^2 = \left(\frac{(\hat{\beta}_{16} - 0)}{\text{SE}}\right)^2 = \left(\frac{-0{,}05434}{0{,}04837}\right)^2 = (-1{,}123424)^2 = 1{,}262...$$

Previamente hemos visto que este estadístico sigue una distribución chi-cuadrado con un grado de libertad. En tablas $\chi^2_1 = 3,8415$. Como 1,262... $<$ 3,8415 no rechazamos la hipótesis nula y concluimos que el hecho de haber cursado un máster no influye en la probabilidad de tener un sueldo elevado. Por lo tanto, debemos eliminar del modelo LOGIT.3 la variable ficticia $EST\_B11\_2\_NO$ y volver a estimar el modelo. Este nuevo modelo recibirá el nombre LOGIT.4.

El output de R nos proporciona de manera directa este test. SE, que es la desviación estándar, se puede encontrar en la tercera columna llamada "Std. Error"y el valor de z (no de z^2) está en la cuarta columna. En la sección 7.2 ya hemos visto que el estadístico z sigue una distribución normal estándar. En la quinta columna tenemos el p-valor asociado a cada estadístico z. El criterio de decisión es el siguiente: si el p-valor es mayor que el nivel de significación no rechazamos la hipótesis nula, y si es menor entonces rechazamos la hipótesis nula.

Pasamos a estimar el modelo LOGIT.4. A continuación, se proporciona el resultado de R.

```
LOGIT.4
Call:
glm(formula = Y ~ SEXO_M + EDAD_30_34 + EDAD_35 + NACIO_ESP_OTR +
    NACIO_OTR + RAMA_CIENCIAS + RAMA_CCSS + RAMA_ING + RAMA_SALUD +
    T_UNIV_PUB_DIST + T_UNIV_PRIV_PRE + T_UNIV_PRIV_DIST + EST_B2_2_OTRA +
    EST_B2_2_NO + EST_M2_NO + IDIOMAS_1 + IDIOMAS_2 + IDIOMAS_3 +
    IDIOMAS_4 + IDIOMAS_5 + TIC_INTER + TIC_AVAN + SIT_PRO_AS_IND +
    SIT_PRO_AS_TEMP + SIT_PRO_EMPR + SIT_PRO_FAM + JORNADA_COMPL +
    TR_TAM_IND + TR_TAM_PEQUE + TR_TAM_MEDI + TR_TAM_GRAN + HL_E1_NO,
    family = binomial(logit), data = DATADEF)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8890  -0.4662  -0.3030  -0.1559   3.8929

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -7.09197    0.35345 -20.065  < 2e-16 ***
SEXO_M[T.1]          -0.71721    0.05067 -14.153  < 2e-16 ***
EDAD_30_34[T.1]      -0.16706    0.05979  -2.794 0.005203 **
EDAD_35[T.1]          0.13300    0.06575   2.023 0.043109 *
NACIO_ESP_OTR[T.1]    0.62911    0.19024   3.307 0.000943 ***
```

```
NACIO_OTR[T.1]         -0.05995    0.23071  -0.260 0.794981
RAMA_CIENCIAS[T.1]      0.14473    0.15690   0.922 0.356314
RAMA_CCSS[T.1]          0.63271    0.12206   5.183 2.18e-07 ***
RAMA_ING[T.1]           1.14194    0.12385   9.221  < 2e-16 ***
RAMA_SALUD[T.1]         1.60146    0.12834  12.479  < 2e-16 ***
T_UNIV_PUB_DIST[T.1]    0.60522    0.10874   5.566 2.61e-08 ***
T_UNIV_PRIV_PRE[T.1]    0.23765    0.06548   3.629 0.000284 ***
T_UNIV_PRIV_DIST[T.1]   0.52608    0.12079   4.355 1.33e-05 ***
EST_B2_2_OTRA[T.1]     -1.11391    0.09215 -12.088  < 2e-16 ***
EST_B2_2_NO[T.1]       -0.67777    0.09240  -7.335 2.21e-13 ***
EST_M2_NO[T.1]         -0.50017    0.08147  -6.139 8.30e-10 ***
IDIOMAS_1[T.1]         -0.05944    0.12404  -0.479 0.631828
IDIOMAS_2[T.1]          0.32711    0.12614   2.593 0.009506 **
IDIOMAS_3[T.1]          0.56086    0.13961   4.017 5.88e-05 ***
IDIOMAS_4[T.1]          1.02947    0.17970   5.729 1.01e-08 ***
IDIOMAS_5[T.1]          1.01105    0.30903   3.272 0.001069 **
TIC_INTER[T.1]          0.23811    0.08976   2.653 0.007982 **
TIC_AVAN[T.1]           0.39698    0.09984   3.976 7.00e-05 ***
SIT_PRO_AS_IND[T.1]     1.63994    0.16430   9.982  < 2e-16 ***
SIT_PRO_AS_TEMP[T.1]    1.16058    0.16852   6.887 5.70e-12 ***
SIT_PRO_EMPR[T.1]       3.47015    0.21858  15.876  < 2e-16 ***
SIT_PRO_FAM[T.1]        1.17071    0.62385   1.877 0.060575 .
JORNADA_COMPL[T.1]      2.34296    0.21779  10.758  < 2e-16 ***
TR_TAM_IND[T.1]         2.79199    0.21316  13.098  < 2e-16 ***
TR_TAM_PEQUE[T.1]       0.63651    0.12406   5.131 2.89e-07 ***
TR_TAM_MEDI[T.1]        0.97369    0.12271   7.935 2.11e-15 ***
TR_TAM_GRAN[T.1]        1.57107    0.11459  13.711  < 2e-16 ***
HL_E1_NO[T.4]           0.12038    0.05490   2.193 0.028343 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15547  on 25830  degrees of freedom
Residual deviance: 13131  on 25798  degrees of freedom
AIC: 13197

Number of Fisher Scoring iterations: 7
```

Para ejemplificar el test de Wald multivariado podemos contrastar

$$\begin{cases} H_0: & (\beta_2 \quad \beta_3)^T = (0,0)^T \\ H_1: & (\beta_2 \quad \beta_3)^T \neq (0,0)^T, \end{cases}$$

donde β_2 es el coeficiente que acompaña a la variable ficticia $EDAD\_30\_34$ y β_3 el que acompaña a $EDAD\_35$. Con este test realmente estamos contrastando si la edad influye en la probabilidad de tener un sueldo elevado. Calculamos el estadístico de prueba:

$$\boldsymbol{W = (\hat{\beta} - \beta_0)^T [\mathrm{cov}(\hat{\beta})]^{-1} (\hat{\beta} - \beta_0)}$$

En este caso $\hat{\beta}^T = (-0{,}16706 \quad 0{,}013300)^T$ y ${\beta_0}^T = (0 \quad 0)^T$. Para calcular el estadístico W usaremos un pequeño código en R.

```
> beta_hat <- matrix(coef(LOGIT.4.), ncol=1,
+                    dimnames=list(names(coef(LOGIT.4)), NULL))
> L <-matrix(c(0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
+              0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
+        ncol=33, byrow=TRUE)
> C <- matrix(c(0, 0), ncol=1, byrow=TRUE)
> aux <- L %*% beta_hat - C
> W <- t(aux) %*% solve(L %*% vcov(LOGIT.4.) %*% t(L)) %*% aux
> W <- as.numeric(W)
> W
[1] 20.03092
```

obteniendo $\mathbf{W} = 20{,}03092$. Este estadístico ha de ser comparado con el valor en tablas de χ^2_2 (con un nivel de significación de 0,05) que vale 5,9915. Como $20{,}03092 > 5{,}9915$ rechazamos la hipótesis nula y concluimos que la variable $EDAD$ es significativa para calcular la probabilidad de tener un sueldo elevado.

Si observamos detenidamente el output del modelo LOGIT.4 observamos que todas las variables restantes son significativas. Los asteriscos al lado de cada estimación nos indican con qué significación es relevante cada variable. Cabe destacar que no podemos eliminar del modelo una única variable ficticia si la variable está codificada en más de una ficticia sin que afecte a la interpretación de los parámetros. Ahora podemos ver si alguna otra codificación proporciona un modelo mejor.

La variable $NACIO\_OTR$ tiene un estadístico $z = -0,260$ con p-valor $0,794981 > 0,05$ por lo tanto podemos eliminar la variable ficticia. Esto afecta a la interpretación de los parámetros ya que ahora la categoría base de la variable $NACIO$ es tener una única nacionalidad (ya sea española u otra).

Para el resto de variables ficticias no significativas (sin asteriscos), que son $RAMA\_CIENCIAS$ e $IDIOMAS\_1$ no repetiremos el proceso ya que la interpretación de los parámetros dejaría de ser tan intuitiva. Nótese que para ser conservadores con los resultados conservaremos las variables que son relevantes con una significación de 0,1, como es el caso de la variable ficticia SIT_PRO_FAM.

Adjuntamos a continuación el output de R al estimar el modelo LOGIT.5, que es el mismo modelo que LOGIT.4 extrayendo la variable ficticia $NACIO\_OTR$

```
LOGIT.5
Call:
glm(formula = Y ~ SEXO_M + EDAD_30_34 + EDAD_35 + NACIO_ESP_OTR +
    RAMA_CIENCIAS + RAMA_CCSS + RAMA_ING + RAMA_SALUD + T_UNIV_PUB_DIST +
    T_UNIV_PRIV_PRE + T_UNIV_PRIV_DIST + EST_B2_2_OTRA + EST_B2_2_NO +
    EST_M2_NO + IDIOMAS_1 + IDIOMAS_2 + IDIOMAS_3 + IDIOMAS_4 +
    IDIOMAS_5 + TIC_INTER + TIC_AVAN + SIT_PRO_AS_IND + SIT_PRO_AS_TEMP +
    SIT_PRO_EMPR + SIT_PRO_FAM + JORNADA_COMPL + TR_TAM_IND +
    TR_TAM_PEQUE + TR_TAM_MEDI + TR_TAM_GRAN + HL_E1_NO, family = binomial(logit),
    data = DATADEF)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8871  -0.4661  -0.3029  -0.1561   3.8930

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -7.09300    0.35344 -20.069  < 2e-16 ***
SEXO_M[T.1]          -0.71733    0.05067 -14.156  < 2e-16 ***
EDAD_30_34[T.1]      -0.16733    0.05978  -2.799 0.005124 **
EDAD_35[T.1]          0.13297    0.06575   2.022 0.043145 *
NACIO_ESP_OTR[T.1]    0.63018    0.19019   3.313 0.000922 ***
RAMA_CIENCIAS[T.1]    0.14542    0.15688   0.927 0.353947
RAMA_CCSS[T.1]        0.63306    0.12206   5.186 2.14e-07 ***
RAMA_ING[T.1]         1.14268    0.12382   9.229  < 2e-16 ***
RAMA_SALUD[T.1]       1.60185    0.12833  12.482  < 2e-16 ***
T_UNIV_PUB_DIST[T.1]  0.60552    0.10873   5.569 2.56e-08 ***
T_UNIV_PRIV_PRE[T.1]  0.23785    0.06548   3.633 0.000281 ***
T_UNIV_PRIV_DIST[T.1] 0.52621    0.12079   4.356 1.32e-05 ***
EST_B2_2_OTRA[T.1]   -1.11434    0.09213 -12.095  < 2e-16 ***
EST_B2_2_NO[T.1]     -0.67745    0.09239  -7.333 2.26e-13 ***
EST_M2_NO[T.1]       -0.50096    0.08142  -6.153 7.60e-10 ***
IDIOMAS_1[T.1]       -0.05912    0.12404  -0.477 0.633650
IDIOMAS_2[T.1]        0.32705    0.12614   2.593 0.009523 **
IDIOMAS_3[T.1]        0.55984    0.13955   4.012 6.03e-05 ***
IDIOMAS_4[T.1]        1.02775    0.17958   5.723 1.05e-08 ***
IDIOMAS_5[T.1]        1.00645    0.30846   3.263 0.001103 **
TIC_INTER[T.1]        0.23803    0.08976   2.652 0.008001 **
TIC_AVAN[T.1]         0.39675    0.09983   3.974 7.06e-05 ***
SIT_PRO_AS_IND[T.1]   1.64017    0.16430   9.983  < 2e-16 ***
SIT_PRO_AS_TEMP[T.1]  1.16107    0.16851   6.890 5.57e-12 ***
SIT_PRO_EMPR[T.1]     3.47023    0.21857  15.877  < 2e-16 ***
SIT_PRO_FAM[T.1]      1.17112    0.62383   1.877 0.060477 .
JORNADA_COMPL[T.1]    2.34327    0.21778  10.760  < 2e-16 ***
TR_TAM_IND[T.1]       2.79196    0.21316  13.098  < 2e-16 ***
TR_TAM_PEQUE[T.1]     0.63684    0.12405   5.134 2.84e-07 ***
TR_TAM_MEDI[T.1]      0.97390    0.12271   7.937 2.07e-15 ***
TR_TAM_GRAN[T.1]      1.57137    0.11458  13.714  < 2e-16 ***
HL_E1_NO[T.4]         0.12027    0.05490   2.191 0.028479 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15547  on 25830  degrees of freedom
Residual deviance: 13131  on 25799  degrees of freedom
AIC: 13195

Number of Fisher Scoring iterations: 7
```

Hasta ahora hemos estado comparando modelos anidados, pero a continuación estudiaremos diferentes modelos usando el criterio de información Akaike.

Comenzaremos por ver qué sucede si eliminamos las variables ficticias $IDIOMAS\_4$ e $IDIOMAS\_5$ y añadimos una variable ficticia $IDIOMAS\_4OMAS$ que tome el valor 1 cuando se hablan 4 idiomas o más (sin tener en cuenta la lengua materna) y 0 en cualquier otro caso. La codificación concreta de esta variable fictia se encuentra en el anexo. Para mantener la fluidez de este trabajo, se ha tomado la decisión de no incluir directamente los outputs generados por el programa R. Sin embargo, añadiremos la información necesaria para poder seleccionar qué modelo es mejor y si se selecciona un modelo sí que se añadirán los outputs.

Creamos un modelo que recibe el nombre de LOGIT.6 que es idéntico a LOGIT.5 eliminando $IDIOMAS\_4$ e $IDIOMAS\_5$ y añadiendo la nueva variable ficticia $IDIOMAS\_4OMAS$. Comparamos los AIC de los dos modelos

```
> AIC(LOGIT.5)
[1] 13194.75
> AIC(LOGIT.6)
[1] 13192.75
```

y seleccionamos el modelo LOGIT.6 ya que AIC(LOGIT.6) $<$ AIC(LOGIT.5). Con esta selección concluimos que con hablar 4 lenguas o más es suficiente para que afecte a la probabilidad de tener un sueldo elevado.

A continuación vamos a ver si es necesario diferenciar el tipo de universidad según si es privada o pública y presencial o a distancia. El modelo LOGIT.6 clasifica el tipo de universidad según los 4 criterios comentados. Creamos un modelo LOGIT.7 que elimine todas las variables ficticias que hacen referencia a la variable T_UNIV y añadimos una variable ficticia llamada T_UNIV_PRIV que tome el valor 1 cuando el graduado ha asistido a una universidad privada independientemente de si es presencial o a distancia y 0 en el caso de haber asistido a una universidad pública. De manera similar creamos un modelo LOGIT.8 que excluya las variables $T\_UNIV\_PUB\_DIST$, $T\_UNIV\_PRIV\_PRE$ y $T\_UNIV\_PRIV\_DIST$ y añada la variable ficiticia T_UNIV_DIST que toma el valor 1 para las universidades a distancia y 0 para las presenciales. Calculamos el AIC de cada uno de estos modelos

```
> AIC(LOGIT.6)
[1] 13192.75
> AIC(LOGIT.7)
[1] 13220.8
> AIC(LOGIT.8)
[1] 13201.92
```

y como AIC(LOGIT.6) $<$ AIC(LOGIT.8) $<$ AIC(LOGIT.7) seleccionamos el modelo LOGIT.6, que es el que diferencia el tipo de universidad según los 4 criterios.

Una vez que se ha seleccionado el modelo LOGIT.6, se ha detenido el proceso de eliminación de variables adicionales. Considerando la posible influencia del sector de empleo en la probabilidad de tener un sueldo alto, se ha decidido recodificar la variable TR_CNAE. Según los datos obtenidos de la Encuesta de Estructura Salarial del Instituto Nacional de

Estadística (INE) correspondientes a los años 2019-2020, se ha identificado que los sectores con salarios más altos son: Suministro de energía eléctrica, gas, vapor y aire acondicionado, Información y comunicaciones (actividades de edición, cinematográficas, de vídeo, de sonido, programas de televisión, telecomunicaciones, programación informática y servicios de información), Actividades financieras y de seguros, y Administración Pública y defensa; Seguridad Social obligatoria.

Para incorporar esta información al análisis, se procederá a crear una variable ficticia llamada TR_CNAE_SEL (ver anexo). Esta variable asignará el valor 1 a las observaciones correspondientes a los sectores mencionados anteriormente, y un valor de 0 al resto de las categorías. Luego, esta variable ficticia será añadida al modelo LOGIT.9, el cual será ajustado para evaluar su efecto en la probabilidad de tener un sueldo alto.

El output del modelo LOGIT.9 es

```
LOGIT.9
Call:
glm(formula = Y ~ SEXO_M + EDAD_30_34 + EDAD_35 + NACIO_ESP_OTR +
    RAMA_CIENCIAS + RAMA_CCSS + RAMA_ING + RAMA_SALUD + T_UNIV_PUB_DIST +
    T_UNIV_PRIV_PRE + T_UNIV_PRIV_DIST + EST_B2_2_OTRA + EST_B2_2_NO +
    EST_M2_NO + IDIOMAS_1 + IDIOMAS_2 + IDIOMAS_3 + IDIOMAS_4OMAS +
    TIC_INTER + TIC_AVAN + SIT_PRO_AS_IND + SIT_PRO_AS_TEMP +
    SIT_PRO_EMPR + SIT_PRO_FAM + JORNADA_COMPL + TR_TAM_IND +
    TR_TAM_PEQUE + TR_TAM_MEDI + TR_TAM_GRAN + HL_E1_NO + TR_CNAE_SEL,
    family = binomial(logit), data = DATADEF)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0219  -0.4654  -0.3009  -0.1558   3.8829

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -7.10058    0.35360 -20.081  < 2e-16 ***
SEXO_M[T.1]          -0.71106    0.05071 -14.021  < 2e-16 ***
EDAD_30_34[T.1]      -0.16394    0.05983  -2.740 0.006142 **
EDAD_35[T.1]          0.12392    0.06586   1.882 0.059902 .
NACIO_ESP_OTR[T.1]    0.64383    0.19022   3.385 0.000713 ***
RAMA_CIENCIAS[T.1]    0.15487    0.15695   0.987 0.323762
RAMA_CCSS[T.1]        0.60040    0.12225   4.911 9.05e-07 ***
RAMA_ING[T.1]         1.13252    0.12388   9.142  < 2e-16 ***
RAMA_SALUD[T.1]       1.62863    0.12848  12.676  < 2e-16 ***
T_UNIV_PUB_DIST[T.1]  0.57704    0.10905   5.291 1.21e-07 ***
T_UNIV_PRIV_PRE[T.1]  0.24294    0.06554   3.706 0.000210 ***
T_UNIV_PRIV_DIST[T.1] 0.51987    0.12101   4.296 1.74e-05 ***
EST_B2_2_OTRA[T.1]   -1.11528    0.09215 -12.103  < 2e-16 ***
EST_B2_2_NO[T.1]     -0.67640    0.09243  -7.318 2.51e-13 ***
EST_M2_NO[T.1]       -0.50021    0.08147  -6.140 8.27e-10 ***
IDIOMAS_1[T.1]       -0.04477    0.12436  -0.360 0.718846
IDIOMAS_2[T.1]        0.34771    0.12648   2.749 0.005978 **
IDIOMAS_3[T.1]        0.58525    0.13992   4.183 2.88e-05 ***
```

```
IDIOMAS_4OMAS[T.1]     1.05666    0.17128   6.169 6.86e-10 ***
TIC_INTER[T.1]         0.23954    0.08986   2.666 0.007685 **
TIC_AVAN[T.1]          0.37875    0.10003   3.786 0.000153 ***
SIT_PRO_AS_IND[T.1]    1.60213    0.16439   9.746  < 2e-16 ***
SIT_PRO_AS_TEMP[T.1]   1.15207    0.16849   6.838 8.04e-12 ***
SIT_PRO_EMPR[T.1]      3.47687    0.21857  15.907  < 2e-16 ***
SIT_PRO_FAM[T.1]       1.17919    0.62398   1.890 0.058786 .
JORNADA_COMPL[T.1]     2.32461    0.21787  10.670  < 2e-16 ***
TR_TAM_IND[T.1]        2.79185    0.21320  13.095  < 2e-16 ***
TR_TAM_PEQUE[T.1]      0.64680    0.12422   5.207 1.92e-07 ***
TR_TAM_MEDI[T.1]       0.96840    0.12289   7.880 3.27e-15 ***
TR_TAM_GRAN[T.1]       1.53698    0.11491  13.376  < 2e-16 ***
HL_E1_NO[T.4]          0.11653    0.05497   2.120 0.034031 *
TR_CNAE_SEL[T.1]       0.33460    0.05859   5.711 1.12e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15547  on 25830  degrees of freedom
Residual deviance: 13099  on 25799  degrees of freedom
AIC: 13163

Number of Fisher Scoring iterations: 7
```

Podemos ver que efectivamente la variable TR_CNAE_SEL es relevante, por lo tanto el hecho de trabajar en uno de los 4 sectores mencionados aumenta las probabilidades de tener un sueldo elevado. Más adelante se comentarán los resultados de la estimación de este modelo (ver sección 23).

22. Capacidad Predictiva del Modelo

Una vez que hemos seleccionado las variables relevantes y validado nuestro modelo, es importante analizar su capacidad predictiva. Para ello usaremos tres mecanismos vistos en los fundamentos teóricos.

22.1. Pseudo R^2 de McFadden

Para calcular el pseudo R^2 necesitamos el logaritmo de la función de verosimilitud del modelo LOGIT.9 y el logaritmo de la función de verosimilitud del modelo nulo que llamaremos NULO. Usando el programa R rápidamente obtenemos

```
> logLik(NULO)
'log Lik.' -7773.62 (df=1)
> logLik(LOGIT.9)
'log Lik.' -6549.571 (df=32)
```

y usando la fórmula (16.5) tenemos

$$D = \frac{L_0 - L_M}{L_0} = \frac{-7773,62 - (-6549,541)}{-7773,62} \approx 0,16.$$

Sin embargo, McFadden contribuyó con el capítulo 15 "Quantitative Methods for Analyzing Travel Behaviour on Individuals: Some Recent Developments" del libro Behavioural Travel Modelling (1979) donde explica que los valores de D suelen ser considerablemente más bajos que los del índice R^2 y que los valores entre 0,2 y 0,4 representan un excelente ajuste. De este modo, el modelo LOGIT.9 no es excelente pero tampoco es totalmente deficiente. Para seguir analizando la capacidad predictiva del modelo usaremos otro método.

22.2. Tablas de Clasificación

Usaremos el programa R para generar la tabla de clasificación. Como hemos visto anteriormente al tener pocas observaciones que cumplen Y=1 debemos seleccionar correctamente el punto de corte π_0. Es común usar la proporción de la muestra de éxitos. En nuestro caso, la muestra contiene 25831 observaciones y 2307 cumplen Y=1, por lo tanto podemos usar $\pi_0 = 2307/25831 = 0{,}08931129 \approx 0,1$. Con un pequeño código en R que adjuntamos a continuación obtenemos la tabla de clasificación de nuestro modelo.

```
> predicciones <- ifelse(test = LOGIT.9$fitted.values > 0.1, yes = 1,
+                        no = 0)
> matriz_confusion <- table(DATADEF$Y, predicciones,
+                           dnn = c("observaciones", "predicciones"))
> matriz_confusion
                      predicciones
    observaciones         0     1
                  0 16933  6591
                  1   612  1695
```

Este modelo tiene una capacidad de acierto de $CA = \frac{16933+1695}{25831} \approx 72\,\%$. Sin embargo, es importante ver como se distribuye el error. El número de falsos éxitos es muy elevado (6591). Veamos como se comportan las tablas de clasificación cuando aumenta el punto de corte π_0 mediante las Figuras 5, 6 y 7.

Observamos que a medida que aumenta π_0 también aumentan los éxitos que predice el modelo y disminuyen los falsos éxitos. Sin embargo, aumentan los falsos fracasos y disminuyen los éxitos reales. Aunque pueda parecer que la tabla de clasificación con $\pi_0 = 0,2$ tiene una capacidad de acierto mayor, esto es a costa de que el modelo no detecte correctamento los éxitos. Éste solo detecta el $\frac{832}{2307} \approx 36\,\%$ de los éxitos. Por otro lado, el modelo con menor CA detecta el $\frac{1695}{2307} \approx 73\,\%$ de los éxitos reales.

Con esta comparativa se ve perfectamente que las tablas de clasificación no son una herramienta perfecta para valorar la capacidad predictiva del modelo pues cuando parece que la capacidad de acierto del modelo aumenta disminuye la capacidad de predecir correctamente los éxitos. Además, todo depende de una selección subjetiva del punto de corte π_0.

Por último, y con intención de encontrar una medida que pueda valorar objetivamente la capacidad predictiva del modelo estudiaremos la curva ROC.

```
      0     1
0 16933  6591
1   612  1695
```

$$CA = \frac{16933 + 1695}{25831} \approx 72\,\%$$

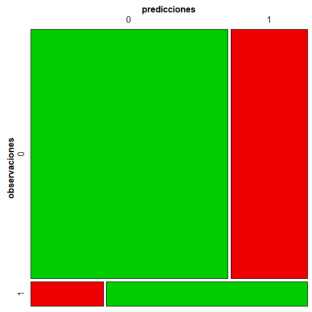

Figura 4: Tabla de Clasificación con $\pi_0 = 0,1$

```
      0     1
0 19703  3821
1  1065  1242
```

$$CA = \frac{19703 + 1242}{25831} \approx 81\,\%$$

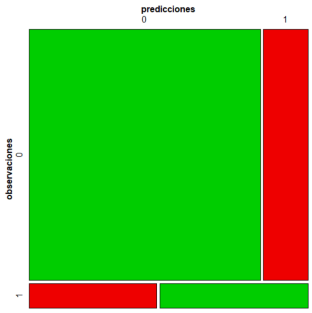

Figura 5: Tabla de Clasificación con $\pi_0 = 0,15$

```
      0     1
0 21292  2232
1  1475   832
```

$$CA = \frac{21292 + 832}{25831} \approx 86\,\%$$

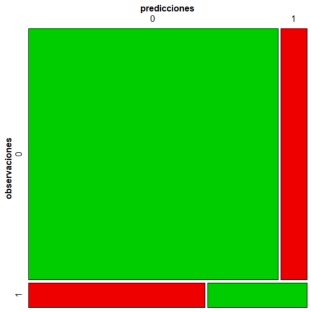

Figura 6: Tabla de Clasificación con $\pi_0 = 0,2$

22.3. Curva Roc

Con un código simple en el programa R es fácil obtener una representación de la curva ROC y el valor AUC.

```
> probabilidades <- predict(LOGIT.9, type = "response")
> roc_obj <- roc(DATADEF$Y, probabilidades)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> auc(roc_obj)
Area under the curve: 0.7962
> plot(roc_obj, main = "Curva ROC")
```

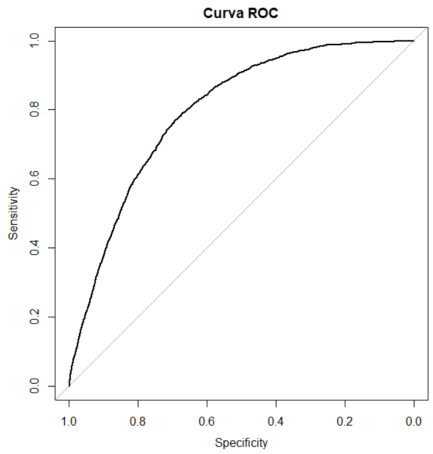

Figura 7: Curva ROC del modelo LOGIT.9

En el output de R vemos que $AUC = 0,7962 \in [0,75, 0,9)$ por lo tanto el modelo es

bueno según el área bajo la curva ROC.

23. Análisis del Modelo Final

En este último apartado del Enfoque Práctico se analizan y discuten los resultados y hallazgos obtenidos a partir del modelo LOGIT.9 que es el que finalmente hemos seleccionado.

Las variables que han resultado relevantes para estimar la probabilidad de tener un sueldo elevado han sido: el sexo, la edad, la nacionalidad, la rama de la titulación, el tipo de universidad, la posesión de un premio de excelencia, el hecho de haber participado en un programa de movilidad, el número de idiomas hablado, la capacidad para usar el ordenador, la situación profesional, el tipo de jornada, el tamaño de la empresa, el desarrollo de prácticas y el sector donde se trabaja. Es interesante ver como el modelo predice hechos que son lógicos, como que cuántos más idiomas se hablen mayor es el odds de tener un sueldo elevado, o que la jornada completa tiene mayor odds que la jornada parcial. La rama con mayor probabilidad es la rama de la salud y el tipo de universidad con mayor odds es la universidad pública a distancia. Observamos también que tener capacidades TIC o un premio de excelencia o haber realizado un programa de movilidad aumenta la probabilidad de éxito en el suceso estudiado. Es coherente esperar que las personas que tienen su propia empresa tengan mayor probabilidad de tener un sueldo elevado. Observamos que la situación profesional con más probabilidad de tener un sueldo elevado son los trabajadores independientes o autónomos sin asalariados. Sin embargo, no es únicamente importante analizar lo que muestra el modelo final. Resulta curioso analizar el proceso y ver qué variables no son significativas. Sorprendentemente, el hecho de haber estudiado un máster o tener una discapacidad no afecta a la probabilidad de tener un sueldo elevado. Por último, me gustaría remarcar la diferencia en la probabilidad del suceso estudiado entre hombres y mujeres. Al tener la variable $SEXO\_M$ un coeficiente negativo, ésta tiene un odds ratio menor que 1. Esto quiere decir que manteniendo todas las otras variables iguales, las mujeres tienen menos probabilidad de tener un sueldo elevado que los hombres.

La capacidad predictiva del modelo es imprescindible para valorar nuestro modelo. Dependiendo del método utilizado las conclusiones son más optimistas o más objetivas. Podríamos resumir los resultados de los tres métodos vistos diciendo que el modelo ajusta bien los datos. No es un modelo ineficiente pero tampoco es perfecto. Al tener una curva ROC por encima de la diagonal, queda demostrado que nuestro modelo predice mejor la probabilidad de tener un sueldo elevado que si lo analizáramos aleatoriamente, por ejemplo lanzando una moneda al aire.

También es importante comentar la aplicabilidad del modelo y sus limitaciones. No hay que olvidar el origen del modelo: los datos. Al usar unos datos de 2019 es posible que actualmente las estimaciones cambien. Pero no solo es importante el marco temporal, sino también el geográfico. Seguramente lo que en España hemos considerado un sueldo elevado en otro país no lo sea, o también es posible que las variables predictoras cambien. Además, no creo que deban tomarse decisiones únicamente en base a este modelo. Por ejemplo, la gran diversidad de titulaciones hace que en el modelo se tengan que agrupar según la rama. Es posible que alguna rama con menor probabilidad de tener un sueldo elevado contenga una titulación con sueldos elevados. Con la finalidad de mantener un modelo simple no se ha detallado tanto en algunas variables, como por ejemplo el sector

en el que se trabaja ya que deberíamos añadir 30 variables ficticias más. Por último, cabe destacar que la inclusión de alguna variable más podría mejorar la capacidad predictiva del modelo. Ejemplos de posibles variables a incluir son: la productividad, la capacidad de pago de la empresa, el coste de la vida, etc.

En resumen, hemos visto qué variables afectan a la probabilidad de tener un sueldo elevado y que el modelo ajusta bien los datos y muestra capacidad predictiva, aunque presenta ciertas limitaciones.

24. Conclusiones

El objetivo de este trabajo era abordar de manera exhaustiva los modelos logísticos como herramientas estadísticas, en concreto para calcular la probabilidad de tener un salario elevado. Como acabamos de ver en el análisis del modelo final este objetivo ha sido logrado.

Mediante los tests de hipótesis, la inferencia, el ajuste, los diagnósticos de la regresión logística, etc., hemos sido capaces de crear un modelo para estudiar la probabilidad de tener un sueldo elevado. Hemos seleccionado qué variables explican este hecho y hemos calculado cuál es la capacidad predictiva del modelo. En concreto se ha visto que el perfil de titulado universitario con mayor probabilidad de tener un sueldo elevado es un hombre de 35 años o más con doble nacionalidad (española y otra). Es relevante que tenga un premio de excelencia, haya estudiado en una universidad pública a distancia una carrera de la rama de salud y haya realizado algún programa de movilidad, pero no prácticas. Hablar 4 idiomas o más y tener una capacidad para usar el ordenador avanzada también ayuda a aumentar la probabilidad. Además, también aumenta las posibilidades que tenga una empresa sin asalariados trabajando a jornada completa en uno de los 4 sectores siguientes: Suministro de energía eléctrica, gas, vapor y aire acondicionado, Información y comunicaciones, Actividades financieras y de seguros y Administración Pública y defensa; Seguridad Social obligatoria.

La división entre los Fundamentos Teóricos y el Enfoque Práctico ha enriquecido el trabajo. La teoría ha ayudado a llevar a cabo el caso práctico y la práctica ha necesitado de más teoría para analizar lo que deseábamos.

En última instancia, este trabajo ha contribuido al campo de los modelos logísticos y su aplicación en el estudio de la probabilidad de tener un salario elevado. La combinación de enfoques teóricos y prácticos ha proporcionado una perspectiva integral y sólida sobre el tema. No obstante, se reconoce que existen áreas de mejora y posibles límites en el estudio, como la necesidad de considerar variables adicionales y la actualización de los datos utilizados.

Referencias

[1] Agresti, A.: Categorical Data Analysis, *Wiley series in probability and statistics; 729*, 2013.

[2] Corcuera, J.M..: Statistics. Apuntes de la asignatura de Estadística. Universidad de Barcelona, 2022.

[3] Frank, E; Harrell, Jr.: Regression Modelling Strategies. With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, *Springer*, 2015.

[4] Instituto Nacional de Estadística: Encuesta de inserción laboral de titulados universitarios. Resultados, `https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_Ccid=1254736176991menu=resultadosidp=1254735976597`, 2022.

[5] Jobson, J.D.: Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods, *Springer*, 1992.

[6] Larrañaga, P; Inza, I; Moujahid, A.: Tema 7. Regresión Logística. Apuntes del Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad del País Vasco–Euskal Herriko Unibertsitatea.

[7] Lloyd, C.: Statistical Analysis of Categorical Data, *Wiley series in probability and statistics; 729*, 1999.

[8] McFadden, D.: Quantitative Methods for analyzing Travel Behaviour of individuals: Some recent Developments. Univeristy of California, Berkeley, 1977.

[9] Pati, D.: Likelihood Maximum Likelihood Estimators. University of Texas. Department of Statistics, 2022.

[10] Vives, J.: Curs Elemental d'Estadística Matemàtica. Apuntes de la asignatura de Estadística. Universidad de Barcelona, 2011.

[11] Varios autores: Curva ROC, `https://es.wikipedia.org/wiki/Curva_ROC`, 2022.

A. Anexo 1: Tablas de Contingencia

Y=0	Y=1	TOTAL
23524	2307	25831

Tabla 4: Tabla de contingencia total

Categorías	Total	Y=0	%	Y=1	%
Hombre	11226	9760	86.94	1466	13.06
Mujer	14605	13764	94,24	841	5,76

Tabla 5: Tabla de contingencia según variable $SEXO$

Categorías	Total	Y=0	%	Y=1	%
Menores de 30 años	12768	11771	92,19	997	7,81
De 30 a 34 años	7243	6680	92,23	563	7,77
De 35 y más años	5820	5073	87,16	747	12,84

Tabla 6: Tabla de contingencia según variable $EDAD$

Categorías	Total	Y=0	%	Y=1	%
Española	25265	23020	91,11	2245	8,89
Española y otra	280	241	86,07	39	13,93
Otra nacionalidad	286	263	91,96	23	8,04

Tabla 7: Tabla de contingencia según variable $NACIO$

Categorías	Total	Y=0	%	Y=1	%
Artes y humanidades	2252	2163	96,05	89	3,95
Ciencias	2232	2135	95,65	97	4,35
Ciencias sociales y jurídicas	11533	10786	93,52	747	6,48
Ingeniería y arquitectura	5911	5018	84,89	893	15,11
Ciencias de la salud	3903	3422	87,68	481	12,32

Tabla 8: Tabla de contingencia según variable $RAMA$

Categorías	Total	Y=0	%	Y=1	%
Universidad Pública presencial	20970	19300	92,04	1670	7,96
Universidad Pública a distancia	876	726	82,88	150	17,12
Universidad Privada presencial	3226	2847	88,25	379	11,75
Universidad Privada a distancia	759	651	85,77	108	14,23

Tabla 9: Tabla de contingencia según variable T_UNIV

Categorías	Total	Y=0	%	Y=1	%
Sí	270	239	88,52	31	11,48
No	25561	23285	91,10	2276	8,90

Tabla 10: Tabla de contingencia según variable $DISCA$

Categorías	Total	Y=0	%	Y=1	%
Sí	1200	977	81,42	223	18,58
Otra beca	10614	9978	94,01	636	5,99
Ninguna beca	14017	12569	89,67	1448	10,33

Tabla 11: Tabla de contingencia según variable EST_B2_2

Categorías	Total	Y=0	%	Y=1	%
Sí	2781	2467	88,71	314	11,29
No	23050	21057	91,35	1993	8,65

Tabla 12: Tabla de contingencia según variable EST_M2

Categorías	Total	Y=0	%	Y=1	%
Sí	12370	11280	91,19	1090	8,81
No	13461	12244	90,96	1217	9,04

Tabla 13: Tabla de contingencia según variable EST_B11_2

Categorías	Total	Y=0	%	Y=1	%
Ninguno	1121	1036	92,42	85	7,58
1	13478	12484	92,63	994	7,37
2	8468	7617	89,95	851	10,05
3	2180	1906	87,43	274	12,57
4	462	377	81,60	85	18,40
5 o más	122	104	85,25	18	14,75

Tabla 14: Tabla de contingencia según variable $IDIOMAS$

Categorías	Total	Y=0	%	Y=1	%
Usuario de nivel básico	2799	2632	94,03	167	5,97
Usuario de nivel intermedio	17345	15987	92,17	1358	7,83
Usuario de nivel avanzado	5687	4905	86,25	782	13,75

Tabla 15: Tabla de contingencia según variable TIC

Categorías	Total	Y=0	%	Y=1	%
Trabajador en prácticas, formación o becario	1431	1388	97,00	43	3,00
Asalariado con trabajo permanente	14957	13374	89,42	1583	10,58
Asalariado con trabajo temporal	7183	6789	94,51	394	5,49
Empresario con asalariados	439	328	74,72	111	25,28
Trabajador independiente o empresario sin asalariados	1677	1504	89,68	173	10,32
Ayuda en la empresa o negocio familiar	144	141	97,92	3	2,08

Tabla 16: Tabla de contingencia según variable SIT_PRO

Categorías	Total	Y=0	%	Y=1	%
A tiempo parcial	3530	3508	99,38	22	0,62
A tiempo completo	22301	20016	89,75	2285	10,25

Tabla 17: Tabla de contingencia según variable $JORNADA$

Categorías	Total	Y=0	%	Y=1	%
Trabajadores independientes	1677	1504	89,68	173	10,32
Micro empresa	3574	3420	95,69	154	4,31
Empresa pequeña	5108	4842	94,79	266	5,21
Empresa mediana	5015	4639	92,50	376	7,50
Empresa grande	10457	9119	87,20	1338	12,80

Tabla 18: Tabla de contingencia según variable TR_TAM

Categorías	Total	Y=0	%	Y=1	%
Sí	19022	17477	91,88	1545	8,12
No	6809	6047	88,81	762	11,19

Tabla 19: Tabla de contingencia según variable HL_E1

B. Anexo 2: Variables Ficticias

Dividiremos en cuatro bloques las variables explicativas que proponemos para calcular la probabilidad de que los titulados universitarios cobren un sueldo mensual neto mayor o igual que 2500 euros.

Primer Bloque: Datos Personales y Sociodemográficos

La primera variable que introduciremos en el modelo como variable explicativa es el sexo. En la encuesta del INE, llamada $SEXO$, ésta toma dos valores: un 1 si el encuestado es hombre y un 2 si es una mujer. De ahora en adelante esta información la resumiremos en una tabla como la siguiente: Usaremos como categoría base la categoría "Hombre",

código	descripción
1	Hombre
2	Mujer

Tabla 20: Codificación original de la variable $SEXO$ propuesta por el INE

por lo tanto en el modelo añadiremos la variable ficticia

$$SEXO\_M_i = \begin{cases} 0 & si \quad SEXO_i = 1 = \text{“}Hombre\text{”} \\ 1 & si \quad SEXO_i = 2 = \text{“}Mujer\text{”} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

A continuación añadiremos la variable $EDAD$, que describe el grupo de edad al que pertenece el titulado universitario del siguiente modo

código	descripción
1	Menores de 30 años
2	De 30 a 34 años
3	De 35 y más años

Tabla 21: Codificación original de la variable $EDAD$ propuesta por el INE

Usaremos como categoría base durante todo el trabajo siempre la primera categoría que propone el INE. Por lo tanto, añadiremos a nuestro modelo las dos variables ficticias siguientes

$$EDAD\_30\_34_i = \begin{cases} 0 & si \quad EDAD_i = 1 = \text{“Menores de 30 años”} \\ 1 & si \quad EDAD_i = 2 = \text{“De 30 a 34 años”} \\ 0 & si \quad EDAD_i = 3 = \text{“De 35 y más años”} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$EDAD\_35_i = \begin{cases} 0 & si \quad EDAD_i = 1 = \text{“Menores de 30 años”} \\ 0 & si \quad EDAD_i = 2 = \text{“De 30 a 34 años”} \\ 1 & si \quad EDAD_i = 3 = \text{“De 35 y más años”} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Como acabamos de observar, la recodificaión de las variables es un proceso mecánico. Por este motivo a partir de ahora la información se presentará de una manera esquemática

y resumida.

Variable de la encuesta: $NACIO$
Descripción: Nacionalidad
Codificación original:

código	descripción
1	Española
2	Española y otra
3	Otra nacionalidad

Tabla 22: Codificación original de la variable $NACIO$ propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$NACIO\_ESP\_OTR_i = \begin{cases} 1 & \text{si } NACIO_i = 2 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$NACIO\_OTR_i = \begin{cases} 1 & \text{si } NACIO_i = 3 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Variable de la encuesta: $RAMA$
Descripción: Rama de conocimiento de la titulación
Codificación original:

código	descripción
1	Artes y humanidades
2	Ciencias
3	Ciencias sociales y jurídicas
4	Ingeniería y arquitectura
5	Ciencias de la salud

Tabla 23: Codificación original de la variable $RAMA$ propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$RAMA\_CIENCIAS_i = \begin{cases} 1 & \text{si } RAMA_i = 2 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$RAMA\_CCSS_i = \begin{cases} 1 & \text{si } RAMA_i = 3 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$RAMA\_ING_i = \begin{cases} 1 & \text{si } RAMA_i = 4 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$RAMA\_SALUD_i = \begin{cases} 1 & \text{si } RAMA_i = 5 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Variable de la encuesta: T_UNIV
Descripción: Tipo de Universidad
Codificación original:

código	descripción
1	Universidad Pública presencial
2	Universidad Pública a distancia
3	Universidad Privada presencial
4	Universidad Privada a distancia

Tabla 24: Codificación original de la variable T_UNIV propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$T\_UNIV\_PUB\_DIST_i = \begin{cases} 1 & \text{si } T\_UNIV_i = 2 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$T\_UNIV\_PRIV\_PRE_i = \begin{cases} 1 & \text{si } T\_UNIV_i = 3 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$T\_UNIV\_PRIV\_DIST_i = \begin{cases} 1 & \text{si } T\_UNIV_i = 4 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Variable de la encuesta: $DISCA$
Descripción: Tiene discapacidad reconocida superior al 33 %
Codificación original:

código	descripción
1	Sí
2	No

Tabla 25: Codificación original de la variable $DISCA$ propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$DISCA\_NO_i = \begin{cases} 1 & \text{si } DISCA_i = 2 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Segundo Bloque: Educación y Aprendizaje

Variable de la encuesta: EST_B2_2
Descripción: Disfrutó de alguna beca: Premio o beca de excelencia
Codificación original:[2]

código	descripción
1	Sí
2	Tuvo beca pero no de excelencia
9	NS/NC
333	No disfrutó de ninguna beca

Tabla 26: Codificación original de la variable EST_B2_2 propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$EST\_B2\_2\_OTRA_i = \begin{cases} 1 & \text{si } EST\_B2\_2_i = 2 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$EST\_B2\_2\_NO_i = \begin{cases} 1 & \text{si } EST\_B2\_2_i = \text{“333”} \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Variable de la encuesta: EST_M2
Descripción: Programa o beca de movilidad
Codificación original:

código	descripción
1	Programa Erasmus
2	Otros programas o becas dentro de la UE
3	Otros programas o becas fuera de la UE
9	NS/NC
333	No tuvo programa de movilidad

Tabla 27: Codificación original de la variable EST_M2 propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$EST\_M2\_NO_i = \begin{cases} 1 & \text{si } EST\_M2\_i = \text{“333”} \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Observemos que para esta categoría la categoría base es: “sí disfrutó de un programa o

[2](Al acabar de seleccionar y codificar todas las variables explicativas eliminaremos de la muestra las respuestas con código 9 que corresponden a la respuesta “no sabe o no contesta”)

beca de movilidad", independientemente de si fue un programa Erasmus, otro programa dentro de la UE, u otra beca fuera de la UE.

Variable de la encuesta: EST_B11_2
Descripción: Otros estudios: Máster universitario
Codificación original:

código	descripción
1	Sí
2	No

Tabla 28: Codificación original de la variable EST_B11_2 propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$EST\_B11\_2\_NO_i = \begin{cases} 1 & \text{si } EST\_B11\_2_i = 2 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Variable de la encuesta: $IDIOMAS$
Descripción: Nº de idiomas que habla (sin contar los maternos)
Codificación original:

código	descripción
0	Ninguno
1	1
2	2
3	3
4	4
5	5 o más
9	NS/NC

Tabla 29: Codificación original de la variable $IDIOMAS$ propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$IDIOMAS\_1_i = \begin{cases} 1 & \text{si } IDIOMAS_i = 1 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$IDIOMAS\_2_i = \begin{cases} 1 & \text{si } IDIOMAS_i = 2 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$IDIOMAS\_3_i = \begin{cases} 1 & \text{si } IDIOMAS_i = 3 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$IDIOMAS\_4_i = \begin{cases} 1 & \text{si } IDIOMAS_i = 4 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$IDIOMAS\_5_i = \begin{cases} 1 & \text{si } IDIOMAS_i = 5 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Variable de la encuesta: TIC
Descripción: Capacidad para usar el ordenador u otros dispositivos informáticos
Codificación original:

código	descripción
1	Usuario de nivel básico
2	Usuario de nivel intermedio
3	Usuario de nivel avanzado
9	NS/NC

Tabla 30: Codificación original de la variable TIC propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$TIC\_INTER_i = \begin{cases} 1 & \text{si } TIC_i = 2 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$TIC\_AVAN_i = \begin{cases} 1 & \text{si } TIC_i = 3 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Tercer Bloque: Situación Actual

Variable de la encuesta: SIT_PRO
Descripción: Situación profesional actual
Codificación original:

código	descripción
1	Trabajador en prácticas, formación (incluido MIR, EIR, FIR,...) o becario
2	Asalariado con trabajo permanente o contrato de trabajo de duración indefinida
3	Asalariado con trabajo temporal o contrato de trabajo de duración determinada
4	Empresario con asalariados
5	Trabajador independiente o empresario sin asalariados
6	Ayuda en la empresa o negocio familiar

Tabla 31: Codificación original de la variable SIT_PRO propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$SIT\_PRO\_AS\_IND_i = \begin{cases} 1 & \text{si } SIT\_PRO_i = 2 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$SIT\_PRO\_AS\_TEMP_i = \begin{cases} 1 & \text{si } SIT\_PRO_i = 3 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$SIT\_PRO\_EMPR_i = \begin{cases} 1 & \text{si } SIT\_PRO_i = 4 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$SIT\_PRO\_IND_i = \begin{cases} 1 & \text{si } SIT\_PRO_i = 5 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$SIT\_PRO\_FAM_i = \begin{cases} 1 & \text{si } SIT\_PRO_i = 6 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Variable de la encuesta: $JORNADA$
Descripción: Tipo de jornada de trabajo actual
Codificación original:

código	descripción
1	A tiempo parcial
2	A tiempo completo

Tabla 32: Codificación original de la variable $JORNADA$ propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$JORNADA\_COMPL_i = \begin{cases} 1 & \text{si } JORNADA_i = 2 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Variable de la encuesta: TR_TAM
Descripción: Número de personas trabajando en la empresa donde trabaja actualmente
Codificación original:

código	descripción
0	Trabajadores independientes o autónomos sin asalariados
1	De 1 a 9 personas
2	Entre 10 y 19 personas
3	Entre 20 y 49 personas
4	Entre 50 y 249 personas
5	250 o más personas
9	NS/NC

Tabla 33: Codificación original de la variable TR_TAM propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$TR\_TAM\_MICRO_i = \begin{cases} 1 & \text{si } TR\_TAM_i = 1 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$TR\_TAM\_PEQUE_i = \begin{cases} 1 & \text{si } TR\_TAM_i \in \{2, 3\} \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$TR\_TAM\_MEDI_i = \begin{cases} 1 & \text{si } TR\_TAM_i = 4 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$TR\_TAM\_GRAN_i = \begin{cases} 1 & \text{si } TR\_TAM_i = 5 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Cuarto Bloque: Historial Laboral

Variable de la encuesta: HL_E1
Descripción: Ha realizado prácticas en empresas, instituciones o similares
Codificación original:

código	descripción
1	Sí, como parte del plan de estudios
2	Sí, fuera del plan de estudios
3	Sí, ambos tipos de prácticas
333	No

Tabla 34: Codificación original de la variable HL_E1 propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$HL\_E1\_NO_i = \begin{cases} 1 & \text{si } HL\_E1_i = \text{“333”} \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Observemos que para esta categoría la categoría base es: “sí realizó prácticas en empresa”, sin importar si fueron parte del plan de estudios o no.

Otras Variables Ficticias que se añaden posteriormente

Variable de la encuesta: TR_TAM
Descripción: Número de personas trabajando en la empresa donde trabaja actualmente
Codificación original:

código	descripción
0	Trabajadores independientes o autónomos sin asalariados
1	De 1 a 9 personas
2	Entre 10 y 19 personas
3	Entre 20 y 49 personas
4	Entre 50 y 249 personas
5	250 o más personas
9	NS/NC

Tabla 35: Codificación original de la variable TR_TAM propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$TR\_TAM\_IND_i = \begin{cases} 1 & \text{si } TR\_TAM_i = 0 \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Variable de la encuesta: $IDIOMAS$
Descripción: Nº de idiomas que habla (sin contar los maternos)
Codificación original:

código	descripción
0	Ninguno
1	1
2	2
3	3
4	4
5	5 o más
9	NS/NC

Tabla 36: Codificación original de la variable $IDIOMAS$ propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$IDIOMAS\_4OMAS = \begin{cases} 1 & \text{si } IDIOMAS_i \in \{4, 5\} \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Variable de la encuesta: T_UNIV
Descripción: Tipo de Universidad
Codificación original:

código	descripción
1	Universidad Pública presencial
2	Universidad Pública a distancia
3	Universidad Privada presencial
4	Universidad Privada a distancia

Tabla 37: Codificación original de la variable T_UNIV propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$T\_UNIV\_PRI_i = \begin{cases} 1 & \text{si } T\_UNIV_i \in \{3,4\} \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

$$T\_UNIV\_DIST_i = \begin{cases} 1 & \text{si } T\_UNIV_i \in \{2,4\} \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$

Variable de la encuesta: TR_CNAE
Descripción: Actividad principal del establecimiento o local donde trabaja actual.
Codificación original:

código	descripción
AA	Agricultura, ganadería, silvicultura y pesca
BB	Industrias extractivas
CC	Industria manufacturera
DD	Suministro de energía eléctrica, gas, vapor y aire acondicionado
EE	Suministro de agua, actividades de saneamiento, gestión de residuos y descontaminación
FF	Construcción e ingeniería civil
GG	Comercio al por mayor y al por menor; reparación de vehículos de motor y motocicletas
HH	Transporte, almacenamiento, actividades postales y de correos
II	Hostelería
JJ	Información y comunicaciones
KK	Actividades financieras y de seguros
LL	Actividades inmobiliarias
M1	Actividades profesionales, científicas y técnicas
M2	Actividades de fotografía
M3	Actividades veterinarias
N1	Actividades de alquiler
N2	Actividades relacionadas con el empleo
N3	Actividades de agencias de viajes, operadores turísticos, servicios de reservas y actividades relacionadas con los mismos
N4	Actividades de seguridad e investigación
N5	Servicios a edificios y actividades de jardinería
N6	Actividades administrativas de oficina y otras actividades auxiliares a las empresas
OO	Administración Pública y defensa; Seguridad Social obligatoria
PP	Educación
QQ	Actividades sanitarias y de servicios sociales
RR	Actividades artísticas, culturales, recreativas, deportivas y de entrenimiento
S1	Actividades asociativas y sindicales
S2	Reparación de ordenadores
S3	Efectos personales y artículos de uso doméstico y otros servicios personales
S4	Actividades de mantenimiento físico
TT	Particulares como empleadores de personal doméstico; particulares como productores de bienes y servicios para uso propio
UU	Actividades de organizaciones y organismos extraterritoriales

Tabla 38: Codificación original de la variable TR_CNAE propuesta por el INE

Variables ficticias que añadiremos al modelo:

$$TR\_CNAE\_SEL = \begin{cases} 1 & \text{si } TR\_CNAE_i \in \{DD, JJ, KK, OO\} \\ 0 & \text{cualquier otro caso} \end{cases} \qquad \forall i = 1, 2, \ldots, N.$$