Curso Elemental de Regresión Logística y Análisis de Supervivencia

Carlos E Martínez-Rodríguez

Julio 2024

Índice general

Parte I

PRIMERA PARTE: Regresión Logística

CAPÍTULO 1

Día 1: Introducción

1.1. Conceptos Básicos

La regresión logística es una técnica de modelado estadístico utilizada para predecir la probabilidad de un evento binario (es decir, un evento que tiene dos posibles resultados) en función de una o más variables independientes. Es ampliamente utilizada en diversas disciplinas, como medicina, economía, biología, y ciencias sociales, para analizar y predecir resultados binarios. Un modelo de regresión logística describe cómo una variable dependiente binaria Y (que puede tomar los valores 0 o 1) está relacionada con una o más variables independientes $X_1, X_2, \ldots, X_n$. A diferencia de la regresión lineal, que predice un valor continuo, la regresión logística predice una probabilidad que puede ser interpretada como la probabilidad de que $Y = 1$ dado un conjunto de valores para $X_1, X_2, \ldots, X_n$.

1.2. Regresión Lineal

La regresión lineal es utilizada para predecir el valor de una variable dependiente continua en función de una o más variables independientes. El modelo de regresión lineal tiene la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \tag{1.1}$$

donde:

- Y es la variable dependiente.
- β_0 es la intersección con el eje Y o término constante.
- $\beta_1, \beta_2, \ldots, \beta_n$ son los coeficientes que representan la relación entre las variables independientes y la variable dependiente.
- $X_1, X_2, \ldots, X_n$ son las variables independientes.
- ϵ es el término de error, que representa la desviación de los datos observados de los valores predichos por el modelo.

El objetivo de la regresión lineal es encontrar los valores de los coeficientes $\beta_0, \beta_1, \ldots, \beta_n$ que minimicen la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos. Este método se conoce como mínimos cuadrados ordinarios (OLS, por sus siglas en inglés). La función de costo a minimizar es:

$$J(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{1.2}$$

donde:

- y_i es el valor observado de la variable dependiente para la i-ésima observación.

- $\hat{y}_i$ es el valor predicho por el modelo para la i-ésima observación, dado por:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in} \tag{1.3}$$

Para encontrar los valores óptimos de los coeficientes, se toman las derivadas parciales de la función de costo con respecto a cada coeficiente y se igualan a cero:

$$\frac{\partial J}{\partial \beta_j} = 0 \quad \text{para } j = 0, 1, \ldots, n \tag{1.4}$$

Resolviendo este sistema de ecuaciones, se obtienen los valores de los coeficientes que minimizan la función de costo.

1.3. Regresión Logística

La deducción de la fórmula de la regresión logística comienza con la necesidad de modelar la probabilidad de un evento binario. Queremos encontrar una función que relacione las variables independientes con la probabilidad de que la variable dependiente tome el valor 1. La probabilidad de que el evento ocurra, $P(Y = 1)$, se denota como p. La probabilidad de que el evento no ocurra, $P(Y = 0)$, es $1 - p$. Los *odds* (chances) de que ocurra el evento se definen como:

$$\text{odds} = \frac{p}{1-p} \tag{1.5}$$

Los *odds* nos indican cuántas veces más probable es que ocurra el evento frente a que no ocurra. Para simplificar el modelado de los *odds*, aplicamos el logaritmo natural, obteniendo la función logit:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \tag{1.6}$$

La transformación logit es útil porque convierte el rango de la probabilidad (0, 1) al rango de números reales $(-\infty, \infty)$. La idea clave de la regresión logística es modelar la transformación logit de la probabilidad como una combinación lineal de las variables independientes:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{1.7}$$

Aquí, β_0 es el término constante y $\beta_1, \beta_2, \ldots, \beta_n$ son los coeficientes asociados con las variables independientes $X_1, X_2, \ldots, X_n$. Para expresar p en función de una combinación lineal de las variables independientes, invertimos la transformación logit. Partimos de la ecuación:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Aplicamos la exponenciación a ambos lados:

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n}$$

Despejamos p:

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n}}$$

La expresión final que obtenemos es conocida como la función logística:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}} \tag{1.8}$$

Esta función describe cómo las variables independientes se relacionan con la probabilidad de que el evento de interés ocurra. Los coeficientes $\beta_0, \beta_1, \ldots, \beta_n$ se estiman a partir de los datos utilizando el método de máxima verosimilitud.

1.4. Método de Máxima Verosimilitud

Para estimar los coeficientes $\beta_0, \beta_1, \ldots, \beta_n$ en la regresión logística, utilizamos el método de máxima verosimilitud. La idea es encontrar los valores de los coeficientes que maximicen la probabilidad de observar los datos dados. Esta probabilidad se expresa mediante la función de verosimilitud L. La función de verosimilitud $L(\beta_0, \beta_1, \ldots, \beta_n)$ para un conjunto de n observaciones se define como el producto de las probabilidades de las observaciones dadas las variables independientes:

$$L(\beta_0, \beta_1, \ldots, \beta_n) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \tag{1.9}$$

donde:

- p_i es la probabilidad predicha de que $Y_i = 1$,
- y_i es el valor observado de la variable dependiente para la i-ésima observación.

Trabajar directamente con esta función de verosimilitud puede ser complicado debido al producto de muchas probabilidades, especialmente si n es grande. Para simplificar los cálculos, se utiliza el logaritmo de la función de verosimilitud, conocido como la función de log-verosimilitud. El uso del logaritmo simplifica significativamente la diferenciación y maximización de la función. La función de log-verosimilitud se define como:

$$\log L(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{1.10}$$

Aquí, log representa el logaritmo natural. Esta transformación es válida porque el logaritmo es una función monótona creciente, lo que significa que maximizar la log-verosimilitud es equivalente a maximizar la verosimilitud original. En la regresión logística, la probabilidad p_i está dada por la función logística:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})}} \tag{1.11}$$

Sustituyendo esta expresión en la función de log-verosimilitud, obtenemos:

$$\begin{aligned} \log L(\beta_0, \beta_1, \ldots, \beta_n) \quad &= \textstyle\sum_{i=1}^{n} \Big[y_i \log \Big(\frac{1}{1+e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})}} \Big) + \\ &\quad (1 - y_i) \log \Big(1 - \frac{1}{1+e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})}} \Big) \Big] \end{aligned}$$

Simplificando esta expresión, notamos que:

$$\log \left(\frac{1}{1 + e^{-z}} \right) = -\log(1 + e^{-z})$$

y

$$\log \left(1 - \frac{1}{1 + e^{-z}} \right) = \log \left(\frac{e^{-z}}{1 + e^{-z}} \right) = -z - \log(1 + e^{-z})$$

Aplicando estas identidades, la función de log-verosimilitud se convierte en:

$$\begin{aligned} \log L(\beta_0, \beta_1, \ldots, \beta_n) \quad = \quad & \sum_{i=1}^{n} \Big[y_i (-\log(1 + e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})})) + \\ & (1 - y_i) \left(-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}) - \log(1 + e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})}) \right) \Big] \end{aligned}$$

Simplificando aún más, obtenemos:

$$\begin{aligned}\log L(\beta_0, \beta_1, \ldots, \beta_n) &= \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}) \\ &- \log(1 + e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}})]\end{aligned}$$

Para simplificar aún más la notación, podemos utilizar notación matricial. Definimos la matriz $\mathbf{X}$ de tamaño $n \times (k+1)$ y el vector de coeficientes $\boldsymbol{\beta}$ de tamaño $(k+1) \times 1$ como sigue:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \tag{1.12}$$

Entonces, la expresión para la función de log-verosimilitud es:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i(\mathbf{X}_i \boldsymbol{\beta}) - \log(1 + e^{\mathbf{X}_i \boldsymbol{\beta}}) \right] \tag{1.13}$$

donde $\mathbf{X}_i$ es la i-ésima fila de la matriz $\mathbf{X}$. Esta notación matricial simplifica la implementación y la derivación de los estimadores de los coeficientes en la regresión logística. Utilizando métodos numéricos, como el algoritmo de Newton-Raphson, se pueden encontrar los coeficientes que maximizan la función de log-verosimilitud. Para maximizar la función de log-verosimilitud, derivamos esta función con respecto a cada uno de los coeficientes β_j y encontramos los puntos críticos. La derivada parcial de la función de log-verosimilitud con respecto a β_j es:

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_i X_{ij} - \frac{X_{ij} e^{\mathbf{X}_i \boldsymbol{\beta}}}{1 + e^{\mathbf{X}_i \boldsymbol{\beta}}} \right] \tag{1.14}$$

Simplificando, esta derivada se puede expresar como:

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} X_{ij}(y_i - p_i), \text{ donde } p_i = \frac{1}{1 + e^{-\mathbf{X}_i \boldsymbol{\beta}}} \tag{1.15}$$

Para encontrar los coeficientes que maximizan la log-verosimilitud, resolvemos el sistema de ecuaciones

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = 0 \text{ para todos los } j = 0, 1, \ldots, k.$$

Este sistema de ecuaciones no tiene una solución analítica cerrada, por lo que se resuelve numéricamente utilizando métodos iterativos como el algoritmo de Newton-Raphson.

1.5. Método de Newton-Raphson

El método de Newton-Raphson es un algoritmo iterativo que se utiliza para encontrar las raíces de una función. En el contexto de la regresión logística, se utiliza para maximizar la función de log-verosimilitud encontrando los valores de los coeficientes $\beta_0, \beta_1, \ldots, \beta_n$. Este método se basa en una aproximación de segundo orden de la función objetivo. Dado un valor inicial de los coeficientes $\boldsymbol{\beta}^{(0)}$, se actualiza iterativamente el valor de los coeficientes utilizando la fórmula:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\mathbf{H}(\boldsymbol{\beta}^{(t)}) \right]^{-1} \nabla \log L(\boldsymbol{\beta}^{(t)}) \tag{1.16}$$

donde:

- $\boldsymbol{\beta}^{(t)}$ es el vector de coeficientes en la t-ésima iteración.
- $\nabla \log L(\boldsymbol{\beta}^{(t)})$ es el gradiente de la función de log-verosimilitud con respecto a los coeficientes $\boldsymbol{\beta}$:

$$\nabla \log L(\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{y} - \mathbf{p}) \tag{1.17}$$

  donde $\mathbf{y}$ es el vector de valores observados y $\mathbf{p}$ es el vector de probabilidades.

- $\mathbf{H}(\boldsymbol{\beta}^{(t)})$ es la matriz Hessiana (matriz de segundas derivadas) evaluada en $\boldsymbol{\beta}^{(t)}$:

$$\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}^T\mathbf{W}\mathbf{X} \tag{1.18}$$

  donde $\mathbf{W}$ es una matriz diagonal de pesos con elementos $w_i = p_i(1 - p_i)$.

En resumen:

Algoritmo 1 *El algoritmo Newton-Raphson para la regresión logística se puede resumir en los siguientes pasos:*

1. *Inicializar el vector de coeficientes $\boldsymbol{\beta}^{(0)}$ (por ejemplo, con ceros o valores pequeños aleatorios).*
2. *Calcular el gradiente $\nabla \log L(\boldsymbol{\beta}^{(t)})$ y la matriz Hessiana $\mathbf{H}(\boldsymbol{\beta}^{(t)})$ en la iteración t.*
3. *Actualizar los coeficientes utilizando la fórmula:*

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[\mathbf{H}(\boldsymbol{\beta}^{(t)})\right]^{-1} \nabla \log L(\boldsymbol{\beta}^{(t)}) \tag{1.19}$$

4. *Repetir los pasos 2 y 3 hasta que la diferencia entre $\boldsymbol{\beta}^{(t+1)}$ y $\boldsymbol{\beta}^{(t)}$ sea menor que un umbral predefinido (criterio de convergencia).*

En resumen, el método de Newton-Raphson permite encontrar los coeficientes que maximizan la función de log-verosimilitud de manera eficiente.

1.6. Específicando

En específico para un conjunto de n observaciones, la función de verosimilitud L se define como el producto de las probabilidades individuales de observar cada dato:

$$L(\beta_0, \beta_1, \ldots, \beta_n) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \tag{1.20}$$

donde y_i es el valor observado de la variable dependiente para la i-ésima observación y p_i es la probabilidad predicha de que $Y_i = 1$. Aquí, p_i es dado por la función logística:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in})}} \tag{1.21}$$

Tomando el logaritmo:

$$\log L(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right] \tag{1.22}$$

Sustituyendo p_i:

$$\log L(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}) - \log(1 + e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}})\right] \tag{1.23}$$

Dado que el objetivo es encontrar los valores de $\beta_0, \beta_1, \ldots, \beta_n$ que maximicen la función de log-verosimilitud. Para β_j, la derivada parcial de la función de log-verosimilitud es:

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_i X_{ij} - \frac{X_{ij} e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\ldots+\beta_n X_{in}}}{1 + e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\ldots+\beta_n X_{in}}} \right] \tag{1.24}$$

Esto se simplifica a (comparar con la ecuación 1.14):

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{n} X_{ij}(y_i - p_i) \tag{1.25}$$

Para maximizar la log-verosimilitud, resolvemos el sistema de ecuaciones $\frac{\partial \log L}{\partial \beta_j} = 0$ para todos los j de 0 a n., mismo que se resuelve numéricamente utilizando métodos el algoritmo de Newton-Raphson. El método de Newton-Raphson se basa en una aproximación de segundo orden de la función objetivo. Dado un valor inicial de los coeficientes $\beta^{(0)}$, se iterativamente actualiza el valor de los coeficientes utilizando la fórmula:

$$\beta^{(k+1)} = \beta^{(k)} - \left[\mathbf{H}(\beta^{(k)}) \right]^{-1} \mathbf{g}(\beta^{(k)}) \tag{1.26}$$

donde:

- $\beta^{(k)}$ es el vector de coeficientes en la k-ésima iteración.
- $\mathbf{g}(\beta^{(k)})$ es el gradiente (vector de primeras derivadas) evaluado en $\beta^{(k)}$:

  $$\mathbf{g}(\beta) = \frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} \mathbf{X}_i (y_i - p_i) \tag{1.27}$$

  donde $\mathbf{X}_i$ es el vector de valores de las variables independientes para la i-ésima observación (comparar con ecuación 1.17).
- $\mathbf{H}(\beta^{(k)})$ es la matriz Hessiana (matriz de segundas derivadas) evaluada en $\beta^{(k)}$:

  $$\mathbf{H}(\beta) = \frac{\partial^2 \log L}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{n} p_i(1 - p_i)\mathbf{X}_i \mathbf{X}_i^T, \tag{1.28}$$

  comparar con ecuación 1.18

Algoritmo 2 *Los pasos del algoritmo Newton-Raphson para la regresión logística son:*

1. *Inicializar el vector de coeficientes $\beta^{(0)}$ (por ejemplo, con ceros o valores pequeños aleatorios).*
2. *Calcular el gradiente $\mathbf{g}(\beta^{(k)})$ y la matriz Hessiana $\mathbf{H}(\beta^{(k)})$ en la iteración k.*
3. *Actualizar los coeficientes utilizando la fórmula:*

   $$\beta^{(k+1)} = \beta^{(k)} - \left[\mathbf{H}(\beta^{(k)}) \right]^{-1} \mathbf{g}(\beta^{(k)}) \tag{1.29}$$

4. *Repetir los pasos 2 y 3 hasta que la diferencia entre $\beta^{(k+1)}$ y $\beta^{(k)}$ sea menor que un umbral predefinido (criterio de convergencia).*

Como se puede observar la diferencia entre el Algoritmo 1 y el Algoritmo 2 son mínimas

1.7. Notas finales

En el contexto de la regresión logística, los vectores $X_1, X_2, \ldots, X_n$ representan las variables independientes. Cada X_j es un vector columna que contiene los valores de la variable independiente j para cada una de las n observaciones. Es decir,

$$X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \tag{1.30}$$

Para simplificar la notación y los cálculos, a menudo combinamos todos los vectores de variables independientes en una única matriz de diseño $\mathbf{X}$ de tamaño $n \times (k+1)$, donde n es el número de observaciones y $k+1$ es el número de variables independientes más el término de intercepto. La primera columna de $\mathbf{X}$ corresponde a un vector de unos para el término de intercepto, y las demás columnas corresponden a los valores de las variables independientes:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \tag{1.31}$$

revisar la ecuación 1.12. De esta forma, el modelo logit puede ser escrito de manera compacta utilizando la notación matricial:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \mathbf{X}\boldsymbol{\beta} \tag{1.32}$$

donde $\boldsymbol{\beta}$ es el vector de coeficientes:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \tag{1.33}$$

Así, la probabilidad p se puede expresar como:

$$p = \frac{1}{1 + e^{-\mathbf{X}\boldsymbol{\beta}}} \tag{1.34}$$

Comparar la ecuación anterior con la ecuación 1.8. Esta notación matricial simplifica la implementación y la derivación de los estimadores de los coeficientes en la regresión logística. Para estimar los coeficientes $\boldsymbol{\beta}$ en la regresión logística, se utiliza el método de máxima verosimilitud. La función de verosimilitud $L(\boldsymbol{\beta})$ se define como el producto de las probabilidades de las observaciones dadas las variables independientes, recordemos la ecuación 1.9:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \tag{1.35}$$

donde y_i es el valor observado de la variable dependiente para la i-ésima observación, y p_i es la probabilidad predicha de que $Y_i = 1$. La función de log-verosimilitud, que es más fácil de maximizar, se obtiene tomando el logaritmo natural de la función de verosimilitud (1.13):

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{1.36}$$

Sustituyendo $p_i = \frac{1}{1+e^{-\mathbf{X}_i\boldsymbol{\beta}}}$, donde $\mathbf{X}_i$ es la i-ésima fila de la matriz de diseño $\mathbf{X}$, obtenemos:

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i(\mathbf{X}_i\boldsymbol{\beta}) - \log(1 + e^{\mathbf{X}_i\boldsymbol{\beta}}) \right] \tag{1.37}$$

Para encontrar los valores de $\boldsymbol{\beta}$ que maximizan la función de log-verosimilitud, se utiliza un algoritmo iterativo como el método de Newton-Raphson. Este método requiere calcular el gradiente y la matriz Hessiana de la función de log-verosimilitud.

El gradiente de la función de log-verosimilitud con respecto a $\boldsymbol{\beta}$ es (1.17 y 1.27):

$$\nabla \log L(\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{y} - \mathbf{p}) \tag{1.38}$$

donde $\mathbf{y}$ es el vector de valores observados y $\mathbf{p}$ es el vector de probabilidades predichas.

La matriz Hessiana de la función de log-verosimilitud es (1.18 y 1.28):

$$\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}^T\mathbf{W}\mathbf{X} \tag{1.39}$$

donde $\mathbf{W}$ es una matriz diagonal de pesos con elementos $w_i = p_i(1 - p_i)$.

El método de Newton-Raphson actualiza los coeficientes $\boldsymbol{\beta}$ de la siguiente manera:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - [\mathbf{H}(\boldsymbol{\beta}^{(t)})]^{-1} \nabla \log L(\boldsymbol{\beta}^{(t)}) \tag{1.40}$$

Iterando este proceso hasta que la diferencia entre $\boldsymbol{\beta}^{(t+1)}$ y $\boldsymbol{\beta}^{(t)}$ sea menor que un umbral predefinido (1.19 y 1.29), se obtienen los estimadores de máxima verosimilitud para los coeficientes de la regresión logística.

CAPÍTULO 2

Elementos de Probabilidad

2.1. Introducción

Los fundamentos de probabilidad y estadística son esenciales para comprender y aplicar técnicas de análisis de datos y modelado estadístico, incluyendo la regresión lineal y logística. Este capítulo proporciona una revisión de los conceptos clave en probabilidad y estadística que son relevantes para estos métodos.

2.2. Probabilidad

La probabilidad es una medida de la incertidumbre o el grado de creencia en la ocurrencia de un evento. Los conceptos fundamentales incluyen:

2.2.1. Espacio Muestral y Eventos

El espacio muestral, denotado como S, es el conjunto de todos los posibles resultados de un experimento aleatorio. Un evento es un subconjunto del espacio muestral. Por ejemplo, si lanzamos un dado, el espacio muestral es:

$$S = \{1, 2, 3, 4, 5, 6\}$$

Un evento podría ser obtener un número par:

$$E = \{2, 4, 6\}$$

2.2.2. Definiciones de Probabilidad

Existen varias definiciones de probabilidad, incluyendo la probabilidad clásica, la probabilidad frecuentista y la probabilidad bayesiana.

Probabilidad Clásica

La probabilidad clásica se define como el número de resultados favorables dividido por el número total de resultados posibles:

$$P(E) = \frac{|E|}{|S|}$$

donde $|E|$ es el número de elementos en el evento E y $|S|$ es el número de elementos en el espacio muestral S.

Probabilidad Frecuentista

La probabilidad frecuentista se basa en la frecuencia relativa de ocurrencia de un evento en un gran número de repeticiones del experimento:

$$P(E) = \lim_{n \to \infty} \frac{n_E}{n}$$

donde n_E es el número de veces que ocurre el evento E y n es el número total de repeticiones del experimento.

Probabilidad Bayesiana

La probabilidad bayesiana se interpreta como un grado de creencia actualizado a medida que se dispone de nueva información. Se basa en el teorema de Bayes:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

donde $P(A|B)$ es la probabilidad de A dado B, $P(B|A)$ es la probabilidad de B dado A, $P(A)$ y $P(B)$ son las probabilidades de A y B respectivamente.

2.3. Estadística Bayesiana

La estadística bayesiana proporciona un enfoque coherente para el análisis de datos basado en el teorema de Bayes. Los conceptos fundamentales incluyen:

2.3.1. Prior y Posterior

Distribución Prior

La distribución prior (apriori) representa nuestra creencia sobre los parámetros antes de observar los datos. Es una distribución de probabilidad que refleja nuestra incertidumbre inicial sobre los parámetros. Por ejemplo, si creemos que un parámetro θ sigue una distribución normal con media μ_0 y varianza σ_0^2, nuestra prior sería:

$$P(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}}$$

Verosimilitud

La verosimilitud (likelihood) es la probabilidad de observar los datos dados los parámetros. Es una función de los parámetros θ dada una muestra de datos X:

$$L(\theta; X) = P(X|\theta)$$

donde X son los datos observados y θ son los parámetros del modelo.

Distribución Posterior

La distribución posterior (a posteriori) combina la información de la prior y la verosimilitud utilizando el teorema de Bayes. Representa nuestra creencia sobre los parámetros después de observar los datos:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

donde $P(\theta|X)$ es la distribución posterior, $P(X|\theta)$ es la verosimilitud, $P(\theta)$ es la prior y $P(X)$ es la probabilidad marginal de los datos.

La probabilidad marginal de los datos $P(X)$ se puede calcular como:

$$P(X) = \int_{\Theta} P(X|\theta)P(\theta)d\theta$$

donde Θ es el espacio de todos los posibles valores del parámetro θ.

2.4. Distribuciones de Probabilidad

Las distribuciones de probabilidad describen cómo se distribuyen los valores de una variable aleatoria. Existen distribuciones de probabilidad discretas y continuas.

2.4.1. Distribuciones Discretas

Una variable aleatoria discreta toma un número finito o contable de valores. Algunas distribuciones discretas comunes incluyen:

Distribución Binomial

La distribución binomial describe el número de éxitos en una serie de ensayos de Bernoulli independientes y con la misma probabilidad de éxito. La función de probabilidad es:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

donde X es el número de éxitos, n es el número de ensayos, p es la probabilidad de éxito en cada ensayo, y $\binom{n}{k}$ es el coeficiente binomial.

La función generadora de momentos (MGF) para la distribución binomial es:

$$M_X(t) = \left(1 - p + pe^t\right)^n$$

El valor esperado y la varianza de una variable aleatoria binomial son:

$$\begin{aligned} E(X) &= np \\ \text{Var}(X) &= np(1-p) \end{aligned}$$

Distribución de Poisson

La distribución de Poisson describe el número de eventos que ocurren en un intervalo de tiempo fijo o en un área fija. La función de probabilidad es:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

donde X es el número de eventos, λ es la tasa media de eventos por intervalo, y k es el número de eventos observados.

La función generadora de momentos (MGF) para la distribución de Poisson es:

$$M_X(t) = e^{\lambda(e^t - 1)}$$

El valor esperado y la varianza de una variable aleatoria de Poisson son:

$$\begin{aligned} E(X) &= \lambda \\ \text{Var}(X) &= \lambda \end{aligned}$$

2.4.2. Distribuciones Continuas

Una variable aleatoria continua toma un número infinito de valores en un intervalo continuo. Algunas distribuciones continuas comunes incluyen:

Distribución Normal

La distribución normal, también conocida como distribución gaussiana, es una de las distribuciones más importantes en estadística. La función de densidad de probabilidad es:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

donde x es un valor de la variable aleatoria, μ es la media, y σ es la desviación estándar.

La función generadora de momentos (MGF) para la distribución normal es:

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

El valor esperado y la varianza de una variable aleatoria normal son:

$$\begin{aligned} E(X) &= \mu \\ \text{Var}(X) &= \sigma^2 \end{aligned}$$

Distribución Exponencial

La distribución exponencial describe el tiempo entre eventos en un proceso de Poisson. La función de densidad de probabilidad es:

$$f(x) = \lambda e^{-\lambda x}$$

donde x es el tiempo entre eventos y λ es la tasa media de eventos.

La función generadora de momentos (MGF) para la distribución exponencial es:

$$M_X(t) = \frac{\lambda}{\lambda - t}, \quad \text{para } t < \lambda$$

El valor esperado y la varianza de una variable aleatoria exponencial son:

$$\begin{aligned} E(X) &= \frac{1}{\lambda} \\ \text{Var}(X) &= \frac{1}{\lambda^2} \end{aligned}$$

2.5. Estadística Descriptiva

La estadística descriptiva resume y describe las características de un conjunto de datos. Incluye medidas de tendencia central, medidas de dispersión y medidas de forma.

2.5.1. Medidas de Tendencia Central

Las medidas de tendencia central incluyen la media, la mediana y la moda.

Media

La media aritmética es la suma de los valores dividida por el número de valores:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

donde x_i son los valores de la muestra y n es el tamaño de la muestra.

Mediana

La mediana es el valor medio cuando los datos están ordenados. Si el número de valores es impar, la mediana es el valor central. Si es par, es el promedio de los dos valores centrales.

Moda

La moda es el valor que ocurre con mayor frecuencia en un conjunto de datos.

2.5.2. Medidas de Dispersión

Las medidas de dispersión incluyen el rango, la varianza y la desviación estándar.

Rango

El rango es la diferencia entre el valor máximo y el valor mínimo de los datos:

$$Rango = x_{\max} - x_{\min}$$

Varianza

La varianza es la media de los cuadrados de las diferencias entre los valores y la media:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Desviación Estándar

La desviación estándar es la raíz cuadrada de la varianza:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

2.6. Inferencia Estadística

La inferencia estadística es el proceso de sacar conclusiones sobre una población a partir de una muestra. Incluye la estimación de parámetros y la prueba de hipótesis.

2.6.1. Estimación de Parámetros

La estimación de parámetros implica el uso de datos muestrales para estimar los parámetros de una población.

Estimador Puntual

Un estimador puntual proporciona un único valor como estimación de un parámetro de la población. Por ejemplo, la media muestral $\bar{x}$ es un estimador puntual de la media poblacional μ. Otros ejemplos de estimadores puntuales son:

- **Mediana muestral ($\tilde{x}$)**: Estimador de la mediana poblacional.
- **Varianza muestral (s^2)**: Estimador de la varianza poblacional σ^2, definido como:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **Desviación estándar muestral (s)**: Estimador de la desviación estándar poblacional σ, definido como:

$$s = \sqrt{s^2}$$

Propiedades de los Estimadores Puntuales

Los estimadores puntuales deben cumplir ciertas propiedades deseables, como:

- **Insesgadez**: Un estimador es insesgado si su valor esperado es igual al valor del parámetro que estima.

$$E(\hat{\theta}) = \theta$$

- **Consistencia**: Un estimador es consistente si converge en probabilidad al valor del parámetro a medida que el tamaño de la muestra tiende a infinito.
- **Eficiencia**: Un estimador es eficiente si tiene la varianza más baja entre todos los estimadores insesgados.

Estimador por Intervalo

Un estimador por intervalo proporciona un rango de valores dentro del cual se espera que se encuentre el parámetro poblacional con un cierto nivel de confianza. Por ejemplo, un intervalo de confianza para la media es:

$$\left(\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}}\right)$$

donde z es el valor crítico correspondiente al nivel de confianza deseado, σ es la desviación estándar poblacional y n es el tamaño de la muestra.

2.6.2. Prueba de Hipótesis

La prueba de hipótesis es un procedimiento para decidir si una afirmación sobre un parámetro poblacional es consistente con los datos muestrales.

Hipótesis Nula y Alternativa

La hipótesis nula (H_0) es la afirmación que se somete a prueba, y la hipótesis alternativa (H_a) es la afirmación que se acepta si se rechaza la hipótesis nula.

Nivel de Significancia

El nivel de significancia (α) es la probabilidad de rechazar la hipótesis nula cuando es verdadera. Un valor comúnmente utilizado es $\alpha = 0{,}05$.

Estadístico de Prueba

El estadístico de prueba es una medida calculada a partir de los datos muestrales que se utiliza para decidir si se rechaza la hipótesis nula. Por ejemplo, en una prueba t para la media:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

donde $\bar{x}$ es la media muestral, μ_0 es la media poblacional bajo la hipótesis nula, s es la desviación estándar muestral y n es el tamaño de la muestra.

P-valor

El p-valor es la probabilidad de obtener un valor del estadístico de prueba al menos tan extremo como el observado, bajo la suposición de que la hipótesis nula es verdadera. Si el p-valor es menor que el nivel de significancia α, se rechaza la hipótesis nula. El p-valor se interpreta de la siguiente manera:

- **P-valor bajo (p ¡0.05)**: Evidencia suficiente para rechazar la hipótesis nula.
- **P-valor alto (p ¿0.05)**: No hay suficiente evidencia para rechazar la hipótesis nula.

Tipos de Errores

En la prueba de hipótesis, se pueden cometer dos tipos de errores:

- **Error Tipo I (α)**: Rechazar la hipótesis nula cuando es verdadera.
- **Error Tipo II (β)**: No rechazar la hipótesis nula cuando es falsa.

Tabla de Errores en la Prueba de Hipótesis

A continuación se presenta una tabla que muestra los posibles resultados en una prueba de hipótesis, incluyendo los falsos positivos (error tipo I) y los falsos negativos (error tipo II):

	Hipótesis Nula Verdadera	**Hipótesis Nula Falsa**
Rechazar H_0	Error Tipo I (α)	Aceptar H_a
No Rechazar H_0	Aceptar H_0	Error Tipo II (β)

Cuadro 2.1: Resultados de la Prueba de Hipótesis

CAPÍTULO 3

Matemáticas Detrás de la Regresión Logística

3.1. Introducción

La regresión logística es una técnica de modelado estadístico utilizada para predecir la probabilidad de un evento binario en función de una o más variables independientes. Este capítulo profundiza en las matemáticas subyacentes a la regresión logística, incluyendo la función logística, la función de verosimilitud, y los métodos para estimar los coeficientes del modelo.

3.2. Función Logística

La función logística es la base de la regresión logística. Esta función transforma una combinación lineal de variables independientes en una probabilidad.

3.2.1. Definición

La función logística se define como:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}}$$

donde p es la probabilidad de que el evento ocurra, $\beta_0, \beta_1, \ldots, \beta_n$ son los coeficientes del modelo, y $X_1, X_2, \ldots, X_n$ son las variables independientes.

3.2.2. Propiedades

La función logística tiene varias propiedades importantes:

- **Rango**: La función logística siempre produce un valor entre 0 y 1, lo que la hace adecuada para modelar probabilidades.
- **Monotonía**: La función es monótona creciente, lo que significa que a medida que la combinación lineal de variables independientes aumenta, la probabilidad también aumenta.
- **Simetría**: La función logística es simétrica en torno a $p = 0{,}5$.

3.3. Función de Verosimilitud

La función de verosimilitud se utiliza para estimar los coeficientes del modelo de regresión logística. Esta función mide la probabilidad de observar los datos dados los coeficientes del modelo.

3.3.1. Definición

Para un conjunto de n observaciones, la función de verosimilitud L se define como el producto de las probabilidades individuales de observar cada dato:

$$L(\beta_0, \beta_1, \ldots, \beta_n) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

donde y_i es el valor observado de la variable dependiente para la i-ésima observación y p_i es la probabilidad predicha de que $Y_i = 1$.

3.3.2. Función de Log-Verosimilitud

Para simplificar los cálculos, trabajamos con el logaritmo de la función de verosimilitud, conocido como la función de log-verosimilitud. Tomar el logaritmo convierte el producto en una suma:

$$\log L(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} [y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

Sustituyendo p_i:

$$\log L(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}) - \log(1 + e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\ldots+\beta_n X_{in}})\right]$$

3.4. Estimación de Coeficientes

Los coeficientes del modelo de regresión logística se estiman maximizando la función de log-verosimilitud. Este proceso generalmente se realiza mediante métodos iterativos como el algoritmo de Newton-Raphson.

3.4.1. Gradiente y Hessiana

Para maximizar la función de log-verosimilitud, necesitamos calcular su gradiente y su matriz Hessiana.

Gradiente

El gradiente de la función de log-verosimilitud con respecto a los coeficientes β es:

$$\mathbf{g}(\beta) = \frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} \mathbf{X}_i(y_i - p_i)$$

donde $\mathbf{X}_i$ es el vector de valores de las variables independientes para la i-ésima observación.

Hessiana

La matriz Hessiana de la función de log-verosimilitud con respecto a los coeficientes β es:

$$\mathbf{H}(\beta) = \frac{\partial^2 \log L}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{n} p_i(1-p_i)\mathbf{X}_i\mathbf{X}_i^T$$

3.4.2. Algoritmo Newton-Raphson

El algoritmo Newton-Raphson se utiliza para encontrar los valores de los coeficientes que maximizan la función de log-verosimilitud. El algoritmo se puede resumir en los siguientes pasos:

1. Inicializar el vector de coeficientes $\beta^{(0)}$ (por ejemplo, con ceros o valores pequeños aleatorios).
2. Calcular el gradiente $\mathbf{g}(\beta^{(k)})$ y la matriz Hessiana $\mathbf{H}(\beta^{(k)})$ en la iteración k.

3. Actualizar los coeficientes utilizando la fórmula:

$$\beta^{(k+1)} = \beta^{(k)} - \left[\mathbf{H}(\beta^{(k)})\right]^{-1} \mathbf{g}(\beta^{(k)})$$

4. Repetir los pasos 2 y 3 hasta que la diferencia entre $\beta^{(k+1)}$ y $\beta^{(k)}$ sea menor que un umbral predefinido (criterio de convergencia).

3.5. Validación del Modelo

Una vez que se han estimado los coeficientes del modelo de regresión logística, es importante validar el modelo para asegurarse de que proporciona predicciones precisas.

3.5.1. Curva ROC y AUC

La curva ROC (Receiver Operating Characteristic) es una herramienta gráfica utilizada para evaluar el rendimiento de un modelo de clasificación binaria. El área bajo la curva (AUC) mide la capacidad del modelo para distinguir entre las clases.

3.5.2. Matriz de Confusión

La matriz de confusión es una tabla que resume el rendimiento de un modelo de clasificación al comparar las predicciones del modelo con los valores reales. Los términos en la matriz de confusión incluyen verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.

CAPÍTULO 4

Preparación de Datos y Selección de Variables

4.1. Introducción

La preparación de datos y la selección de variables son pasos cruciales en el proceso de modelado estadístico. Un modelo bien preparado y con las variables adecuadas puede mejorar significativamente la precisión y la interpretabilidad del modelo. Este capítulo proporciona una revisión detallada de las técnicas de limpieza de datos, tratamiento de datos faltantes, codificación de variables categóricas y selección de variables.

4.2. Importancia de la Preparación de Datos

La calidad de los datos es fundamental para el éxito de cualquier análisis estadístico. Los datos sin limpiar pueden llevar a modelos inexactos y conclusiones erróneas. La preparación de datos incluye varias etapas:

- Limpieza de datos
- Tratamiento de datos faltantes
- Codificación de variables categóricas
- Selección y transformación de variables

4.3. Limpieza de Datos

La limpieza de datos es el proceso de detectar y corregir (o eliminar) los datos incorrectos, incompletos o irrelevantes. Este proceso incluye:

- Eliminación de duplicados
- Corrección de errores tipográficos
- Consistencia de formato
- Tratamiento de valores extremos (outliers)

4.4. Tratamiento de Datos Faltantes

Los datos faltantes son un problema común en los conjuntos de datos y pueden afectar la calidad de los modelos. Hay varias estrategias para manejar los datos faltantes:

- **Eliminación de Datos Faltantes**: Se eliminan las filas o columnas con datos faltantes.
- **Imputación**: Se reemplazan los valores faltantes con estimaciones, como la media, la mediana o la moda.
- **Modelos Predictivos**: Se utilizan modelos predictivos para estimar los valores faltantes.

4.4.1. Imputación de la Media

Una técnica común es reemplazar los valores faltantes con la media de la variable. Esto se puede hacer de la siguiente manera:

$$x_i = \begin{cases} x_i & \text{si } x_i \text{ no es faltante} \\ \bar{x} & \text{si } x_i \text{ es faltante} \end{cases}$$

donde $\bar{x}$ es la media de la variable.

4.5. Codificación de Variables Categóricas

Las variables categóricas deben ser convertidas a un formato numérico antes de ser usadas en un modelo de regresión logística. Hay varias técnicas para codificar variables categóricas:

4.5.1. Codificación One-Hot

La codificación one-hot crea una columna binaria para cada categoría. Por ejemplo, si tenemos una variable categórica con tres categorías (A, B, C), se crean tres columnas:

$$\begin{aligned} \text{A} &= [1, 0, 0] \\ \text{B} &= [0, 1, 0] \\ \text{C} &= [0, 0, 1] \end{aligned}$$

4.5.2. Codificación Ordinal

La codificación ordinal asigna un valor entero único a cada categoría, preservando el orden natural de las categorías. Por ejemplo:

$$\begin{aligned} \text{Bajo} &= 1 \\ \text{Medio} &= 2 \\ \text{Alto} &= 3 \end{aligned}$$

4.6. Selección de Variables

La selección de variables es el proceso de elegir las variables más relevantes para el modelo. Existen varias técnicas para la selección de variables:

4.6.1. Métodos de Filtrado

Los métodos de filtrado seleccionan variables basadas en criterios estadísticos, como la correlación o la chi-cuadrado. Algunas técnicas comunes incluyen:

- **Análisis de Correlación**: Se seleccionan variables con alta correlación con la variable dependiente y baja correlación entre ellas.
- **Pruebas de Chi-cuadrado**: Se utilizan para variables categóricas para determinar la asociación entre la variable independiente y la variable dependiente.

4.6.2. Métodos de Wrapper

Los métodos de wrapper evalúan múltiples combinaciones de variables y seleccionan la combinación que optimiza el rendimiento del modelo. Ejemplos incluyen:

- **Selección hacia Adelante**: Comienza con un modelo vacío y agrega variables una por una, seleccionando la variable que mejora más el modelo en cada paso.
- **Selección hacia Atrás**: Comienza con todas las variables y elimina una por una, removiendo la variable que tiene el menor impacto en el modelo en cada paso.
- **Selección Paso a Paso**: Combina la selección hacia adelante y hacia atrás, agregando y eliminando variables según sea necesario.

4.6.3. Métodos Basados en Modelos

Los métodos basados en modelos utilizan técnicas de regularización como Lasso y Ridge para seleccionar variables. Estas técnicas añaden un término de penalización a la función de costo para evitar el sobreajuste.

Regresión Lasso

La regresión Lasso (Least Absolute Shrinkage and Selection Operator) añade una penalización L_1 a la función de costo:

$$J(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

donde λ es el parámetro de regularización que controla la cantidad de penalización.

Regresión Ridge

La regresión Ridge añade una penalización L_2 a la función de costo:

$$J(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

donde λ es el parámetro de regularización.

4.7. Implementación en R

4.7.1. Limpieza de Datos

Para ilustrar la limpieza de datos en R, considere el siguiente conjunto de datos:

```
data <- data.frame(
  var1 = c(1, 2, 3, NA, 5),
  var2 = c("A", "B", "A", "B", "A"),
  var3 = c(10, 15, 10, 20, 25)
)

# Eliminaci\'on de filas con datos faltantes
```

```
data_clean <- na.omit(data)

# Imputaci\'on de la media
data$var1[is.na(data$var1)] <- mean(data$var1, na.rm = TRUE)
```

4.7.2. Codificación de Variables Categóricas

Para codificar variables categóricas, utilice la función 'model.matrix':

```
data <- data.frame(
  var1 = c(1, 2, 3, 4, 5),
  var2 = c("A", "B", "A", "B", "A")
)

# Codificaci\'on one-hot
data_onehot <- model.matrix(~ var2 - 1, data = data)
```

4.7.3. Selección de Variables

Para la selección de variables, utilice el paquete 'caret':

```
library(caret)

# Dividir los datos en conjuntos de entrenamiento y prueba
set.seed(123)
trainIndex <- createDataPartition(data$var1, p = .8,
                                  list = FALSE,
                                  times = 1)
dataTrain <- data[trainIndex,]
dataTest <- data[-trainIndex,]

# Modelo de regresi\'on log\'istica
model <- train(var1 ~ ., data = dataTrain, method = "glm", family = "binomial")

# Selecci\'on de variables
model <- step(model, direction = "both")
summary(model)
```

CAPÍTULO 5

Evaluación del Modelo y Validación Cruzada

5.1. Introducción

Evaluar la calidad y el rendimiento de un modelo de regresión logística es crucial para asegurar que las predicciones sean precisas y útiles. Este capítulo se centra en las técnicas y métricas utilizadas para evaluar modelos de clasificación binaria, así como en la validación cruzada, una técnica para evaluar la generalización del modelo.

5.2. Métricas de Evaluación del Modelo

Las métricas de evaluación permiten cuantificar la precisión y el rendimiento de un modelo. Algunas de las métricas más comunes incluyen:

5.2.1. Curva ROC y AUC

La curva ROC (Receiver Operating Characteristic) es una representación gráfica de la sensibilidad (verdaderos positivos) frente a 1 - especificidad (falsos positivos). El área bajo la curva (AUC) mide la capacidad del modelo para distinguir entre las clases.

$$\begin{aligned} \text{Sensibilidad} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{Especificidad} &= \frac{\text{TN}}{\text{TN} + \text{FP}} \end{aligned}$$

5.2.2. Matriz de Confusión

La matriz de confusión es una tabla que muestra el rendimiento del modelo comparando las predicciones con los valores reales. Los términos incluyen:

- **Verdaderos Positivos (TP)**: Predicciones correctas de la clase positiva.
- **Falsos Positivos (FP)**: Predicciones incorrectas de la clase positiva.
- **Verdaderos Negativos (TN)**: Predicciones correctas de la clase negativa.
- **Falsos Negativos (FN)**: Predicciones incorrectas de la clase negativa.

	Predicción Positiva	**Predicción Negativa**
Real Positiva	TP	FN
Real Negativa	FP	TN

Cuadro 5.1: Matriz de Confusión

5.2.3. Precisión, Recall y F1-Score

$$\begin{aligned} \text{Precisión} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{F1-Score} &= 2 \cdot \frac{\text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}} \end{aligned}$$

5.2.4. Log-Loss

La pérdida logarítmica (Log-Loss) mide la precisión de las probabilidades predichas. La fórmula es:

$$\text{Log-Loss} = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

donde y_i son los valores reales y p_i son las probabilidades predichas.

5.3. Validación Cruzada

La validación cruzada es una técnica para evaluar la capacidad de generalización de un modelo. Existen varios tipos de validación cruzada:

5.3.1. K-Fold Cross-Validation

En K-Fold Cross-Validation, los datos se dividen en K subconjuntos. El modelo se entrena K veces, cada vez utilizando K-1 subconjuntos para el entrenamiento y el subconjunto restante para la validación.

$$\text{Error Medio} = \frac{1}{K} \sum_{k=1}^{K} \text{Error}_k$$

5.3.2. Leave-One-Out Cross-Validation (LOOCV)

En LOOCV, cada observación se usa una vez como conjunto de validación y las restantes como conjunto de entrenamiento. Este método es computacionalmente costoso pero útil para conjuntos de datos pequeños.

5.4. Ajuste y Sobreajuste del Modelo

El ajuste adecuado del modelo es crucial para evitar el sobreajuste (overfitting) y el subajuste (underfitting).

5.4.1. Sobreajuste

El sobreajuste ocurre cuando un modelo se ajusta demasiado bien a los datos de entrenamiento, capturando ruido y patrones irrelevantes. Los síntomas incluyen una alta precisión en el entrenamiento y baja precisión en la validación.

5.4.2. Subajuste

El subajuste ocurre cuando un modelo no captura los patrones subyacentes de los datos. Los síntomas incluyen baja precisión tanto en el entrenamiento como en la validación.

5.4.3. Regularización

La regularización es una técnica para prevenir el sobreajuste añadiendo un término de penalización a la función de costo. Las técnicas comunes incluyen:

- **Regresión Lasso (L1)**
- **Regresión Ridge (L2)**

5.5. Implementación en R

5.5.1. Evaluación del Modelo

```
# Cargar el paquete necesario
library(caret)

# Dividir los datos en conjuntos de entrenamiento y prueba
set.seed(123)
trainIndex <- createDataPartition(data$var1, p = .8,
                                  list = FALSE,
                                  times = 1)
dataTrain <- data[trainIndex,]
dataTest <- data[-trainIndex,]

# Entrenar el modelo de regresión logística
model <- train(var1 ~ ., data = dataTrain, method = "glm", family = "binomial")

# Predicciones en el conjunto de prueba
predictions <- predict(model, dataTest)

# Matriz de confusión
confusionMatrix(predictions, dataTest$var1)
```

5.5.2. Validación Cruzada

```
# K-Fold Cross-Validation
control <- trainControl(method = "cv", number = 10)
model_cv <- train(var1 ~ ., data = dataTrain, method = "glm",
                  family = "binomial", trControl = control)

# Evaluación del modelo con validación cruzada
print(model_cv)
```

CAPÍTULO 6

Diagnóstico del Modelo y Ajuste de Parámetros

6.1. Introducción

El diagnóstico del modelo y el ajuste de parámetros son pasos esenciales para mejorar la precisión y la robustez de los modelos de regresión logística. Este capítulo se enfoca en las técnicas para diagnosticar problemas en los modelos y en métodos para ajustar los parámetros de manera óptima.

6.2. Diagnóstico del Modelo

El diagnóstico del modelo implica evaluar el rendimiento del modelo y detectar posibles problemas, como el sobreajuste, la multicolinealidad y la influencia de puntos de datos individuales.

6.2.1. Residuos

Los residuos son las diferencias entre los valores observados y los valores predichos por el modelo. El análisis de residuos puede revelar patrones que indican problemas con el modelo.

$$\text{Residuo}_i = y_i - \hat{y}_i$$

Residuos Estudiantizados

Los residuos estudiantizados se ajustan por la variabilidad del residuo y se utilizan para detectar outliers.

$$r_i = \frac{\text{Residuo}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

donde h_i es el leverage del punto de datos.

6.2.2. Influencia

La influencia mide el impacto de un punto de datos en los coeficientes del modelo. Los puntos con alta influencia pueden distorsionar el modelo.

Distancia de Cook

La distancia de Cook es una medida de la influencia de un punto de datos en los coeficientes del modelo.

$$D_i = \frac{r_i^2}{p} \cdot \frac{h_i}{1 - h_i}$$

donde p es el número de parámetros en el modelo.

6.2.3. Multicolinealidad

La multicolinealidad ocurre cuando dos o más variables independientes están altamente correlacionadas. Esto puede inflar las varianzas de los coeficientes y hacer que el modelo sea inestable.

Factor de Inflación de la Varianza (VIF)

El VIF mide cuánto se inflan las varianzas de los coeficientes debido a la multicolinealidad.

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

donde R_j^2 es el coeficiente de determinación de la regresión de la variable j contra todas las demás variables.

6.3. Ajuste de Parámetros

El ajuste de parámetros implica seleccionar los valores óptimos para los hiperparámetros del modelo. Esto puede mejorar el rendimiento y prevenir el sobreajuste.

6.3.1. Grid Search

El grid search es un método exhaustivo para ajustar los parámetros. Se define una rejilla de posibles valores de parámetros y se evalúa el rendimiento del modelo para cada combinación.

6.3.2. Random Search

El random search selecciona aleatoriamente combinaciones de valores de parámetros dentro de un rango especificado. Es menos exhaustivo que el grid search, pero puede ser más eficiente.

6.3.3. Bayesian Optimization

La optimización bayesiana utiliza modelos probabilísticos para seleccionar iterativamente los valores de parámetros más prometedores.

6.4. Implementación en R

6.4.1. Diagnóstico del Modelo

```
# Cargar el paquete necesario
library(car)

# Residuos estudentizados
dataTrain$resid <- rstudent(model)
hist(dataTrain$resid, breaks = 20, main = "Residuos Estudentizados")
```

```
# Distancia de Cook
dataTrain$cook <- cooks.distance(model)
plot(dataTrain$cook, type = "h", main = "Distancia de Cook")

# Factor de Inflaci\'on de la Varianza
vif_values <- vif(model)
print(vif_values)
```

6.4.2. Ajuste de Parámetros

```
# Grid Search con caret
control <- trainControl(method = "cv", number = 10)
tune_grid <- expand.grid(.alpha = c(0, 0.5, 1), .lambda = seq(0.01, 0.1, by = 0.01))

model_tune <- train(var1 ~ ., data = dataTrain, method = "glmnet",
                    trControl = control, tuneGrid = tune_grid)

print(model_tune)
```

CAPÍTULO 7

Interpretación de los Resultados

7.1. Introducción

Interpretar correctamente los resultados de un modelo de regresión logística es esencial para tomar decisiones informadas. Este capítulo se centra en la interpretación de los coeficientes del modelo, las odds ratios, los intervalos de confianza y la significancia estadística.

7.2. Coeficientes de Regresión Logística

Los coeficientes de regresión logística representan la relación entre las variables independientes y la variable dependiente en términos de log-odds.

7.2.1. Interpretación de los Coeficientes

Cada coeficiente β_j en el modelo de regresión logística se interpreta como el cambio en el log-odds de la variable dependiente por unidad de cambio en la variable independiente X_j.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

7.2.2. Signo de los Coeficientes

- **Coeficiente Positivo**: Un coeficiente positivo indica que un aumento en la variable independiente está asociado con un aumento en el log-odds de la variable dependiente.
- **Coeficiente Negativo**: Un coeficiente negativo indica que un aumento en la variable independiente está asociado con una disminución en el log-odds de la variable dependiente.

7.3. Odds Ratios

Las odds ratios proporcionan una interpretación más intuitiva de los coeficientes de regresión logística. La odds ratio para una variable independiente X_j se calcula como e^{β_j}.

7.3.1. Cálculo de las Odds Ratios

$$\text{OR}_j = e^{\beta_j}$$

7.3.2. Interpretación de las Odds Ratios

- **OR ¿1**: Un OR mayor que 1 indica que un aumento en la variable independiente está asociado con un aumento en las odds de la variable dependiente.
- **OR ¡1**: Un OR menor que 1 indica que un aumento en la variable independiente está asociado con una disminución en las odds de la variable dependiente.
- **OR = 1**: Un OR igual a 1 indica que la variable independiente no tiene efecto sobre las odds de la variable dependiente.

7.4. Intervalos de Confianza

Los intervalos de confianza proporcionan una medida de la incertidumbre asociada con los estimadores de los coeficientes. Un intervalo de confianza del 95 % para un coeficiente β_j indica que, en el 95 % de las muestras, el intervalo contendrá el valor verdadero de β_j.

7.4.1. Cálculo de los Intervalos de Confianza

Para calcular un intervalo de confianza del 95 % para un coeficiente β_j, utilizamos la fórmula:

$$\beta_j \pm 1{,}96 \cdot \text{SE}(\beta_j)$$

donde $\text{SE}(\beta_j)$ es el error estándar de β_j.

7.5. Significancia Estadística

La significancia estadística se utiliza para determinar si los coeficientes del modelo son significativamente diferentes de cero. Esto se evalúa mediante pruebas de hipótesis.

7.5.1. Prueba de Hipótesis

Para cada coeficiente β_j, la hipótesis nula H_0 es que $\beta_j = 0$. La hipótesis alternativa H_a es que $\beta_j \neq 0$.

7.5.2. P-valor

El p-valor indica la probabilidad de obtener un coeficiente tan extremo como el observado, asumiendo que la hipótesis nula es verdadera. Un p-valor menor que el nivel de significancia α (típicamente 0.05) indica que podemos rechazar la hipótesis nula.

7.6. Implementación en R

7.6.1. Cálculo de Coeficientes y Odds Ratios

```
# Cargar el paquete necesario
library(broom)

# Entrenar el modelo de regresi\'on log\'istica
model <- glm(var1 ~ ., data = dataTrain, family = "binomial")

# Coeficientes del modelo
coef(model)

# Odds ratios
exp(coef(model))
```

7.6.2. Intervalos de Confianza

```
# Intervalos de confianza para los coeficientes
confint(model)

# Intervalos de confianza para las odds ratios
exp(confint(model))
```

7.6.3. P-valores y Significancia Estadística

```
# Resumen del modelo con p-valores
summary(model)
```

CAPÍTULO 8

Regresión Logística Multinomial y Análisis de Supervivencia

8.1. Introducción

La regresión logística multinomial y el análisis de supervivencia son extensiones de la regresión logística binaria. Este capítulo se enfoca en las técnicas y aplicaciones de estos métodos avanzados.

8.2. Regresión Logística Multinomial

La regresión logística multinomial se utiliza cuando la variable dependiente tiene más de dos categorías.

8.2.1. Modelo Multinomial

El modelo de regresión logística multinomial generaliza el modelo binario para manejar múltiples categorías. La probabilidad de que una observación pertenezca a la categoría k se expresa como:

$$P(Y = k) = \frac{e^{\beta_{0k} + \beta_{1k} X_1 + ... + \beta_{nk} X_n}}{\sum_{j=1}^{K} e^{\beta_{0j} + \beta_{1j} X_1 + ... + \beta_{nj} X_n}}$$

8.2.2. Estimación de Parámetros

Los coeficientes del modelo multinomial se estiman utilizando máxima verosimilitud, similar a la regresión logística binaria.

8.3. Análisis de Supervivencia

El análisis de supervivencia se utiliza para modelar el tiempo hasta que ocurre un evento de interés, como la muerte o la falla de un componente.

8.3.1. Función de Supervivencia

La función de supervivencia $S(t)$ describe la probabilidad de que una observación sobreviva más allá del tiempo t:

$$S(t) = P(T > t)$$

8.3.2. Modelo de Riesgos Proporcionales de Cox

El modelo de Cox es un modelo de regresión semiparamétrico utilizado para analizar datos de supervivencia:

$$h(t|X) = h_0(t)e^{\beta_1 X_1 + \ldots + \beta_p X_p}$$

donde $h(t|X)$ es la tasa de riesgo en el tiempo t dado el vector de covariables X y $h_0(t)$ es la tasa de riesgo basal.

8.4. Implementación en R

8.4.1. Regresión Logística Multinomial

```
# Cargar el paquete necesario
library(nnet)

# Entrenar el modelo de regresi\'on log\'istica multinomial
model_multinom <- multinom(var1 ~ ., data = dataTrain)

# Resumen del modelo
summary(model_multinom)
```

8.4.2. Análisis de Supervivencia

```
# Cargar el paquete necesario
library(survival)

# Crear el objeto de supervivencia
surv_object <- Surv(time = data$time, event = data$status)

# Ajustar el modelo de Cox
model_cox <- coxph(surv_object ~ var1 + var2, data = data)

# Resumen del modelo
summary(model_cox)
```

CAPÍTULO 9

Implementación de Regresión Logística en Datos Reales

9.1. Introducción

Implementar un modelo de regresión logística en datos reales implica varias etapas, desde la limpieza de datos hasta la evaluación y validación del modelo. Este capítulo presenta un ejemplo práctico de la implementación de un modelo de regresión logística utilizando un conjunto de datos real.

9.2. Conjunto de Datos

Para este ejemplo, utilizaremos un conjunto de datos disponible públicamente que contiene información sobre clientes bancarios. El objetivo es predecir si un cliente suscribirá un depósito a plazo fijo.

9.3. Preparación de Datos

9.3.1. Carga y Exploración de Datos

Primero, cargamos y exploramos el conjunto de datos para entender su estructura y contenido.

```
# Cargar el paquete necesario
library(dplyr)

# Cargar el conjunto de datos
data <- read.csv("bank.csv")

# Explorar los datos
str(data)
summary(data)
```

9.3.2. Limpieza de Datos

El siguiente paso es limpiar los datos, lo que incluye tratar los valores faltantes y eliminar las duplicidades.

```
# Eliminar duplicados
data <- data %>% distinct()

# Imputar valores faltantes (si existen)
data <- data %>% mutate_if(is.numeric, ~ifelse(is.na(.), mean(., na.rm = TRUE), .))
```

9.3.3. Codificación de Variables Categóricas

Convertimos las variables categóricas en variables numéricas utilizando la codificación one-hot.

```
# Codificaci\'on one-hot de variables categ\'oricas
data <- data %>% mutate(across(where(is.factor), ~ as.numeric(as.factor(.))))
```

9.4. División de Datos

Dividimos los datos en conjuntos de entrenamiento y prueba.

```
# Dividir los datos en conjuntos de entrenamiento y prueba
set.seed(123)
trainIndex <- createDataPartition(data$y, p = .8, list = FALSE, times = 1)
dataTrain <- data[trainIndex,]
dataTest <- data[-trainIndex,]
```

9.5. Entrenamiento del Modelo

Entrenamos un modelo de regresión logística utilizando el conjunto de entrenamiento.

```
# Entrenar el modelo de regresi\'on log\'istica
model <- glm(y ~ ., data = dataTrain, family = "binomial")

# Resumen del modelo
summary(model)
```

9.6. Evaluación del Modelo

Evaluamos el rendimiento del modelo utilizando el conjunto de prueba.

```
# Predicciones en el conjunto de prueba
predictions <- predict(model, dataTest, type = "response")

# Convertir probabilidades a etiquetas
predicted_labels <- ifelse(predictions > 0.5, 1, 0)

# Matriz de confusi\'on
confusionMatrix(predicted_labels, dataTest$y)
```

9.7. Interpretación de los Resultados

Interpretamos los coeficientes del modelo y las odds ratios.

```
# Coeficientes del modelo
coef(model)

# Odds ratios
exp(coef(model))
```

CAPÍTULO 10

Resumen y Proyecto Final

10.1. Resumen de Conceptos Clave

En este curso, hemos cubierto una variedad de conceptos y técnicas esenciales para la regresión logística. Los conceptos clave incluyen:

- **Fundamentos de Probabilidad y Estadística**: Comprensión de distribuciones de probabilidad, medidas de tendencia central y dispersión, inferencia estadística y pruebas de hipótesis.
- **Regresión Logística**: Modelo de regresión logística binaria y multinomial, interpretación de coeficientes y odds ratios, métodos de estimación y validación.
- **Preparación de Datos**: Limpieza de datos, tratamiento de valores faltantes, codificación de variables categóricas y selección de variables.
- **Evaluación del Modelo**: Curva ROC, AUC, matriz de confusión, precisión, recall, F1-score y validación cruzada.
- **Diagnóstico del Modelo**: Análisis de residuos, influencia, multicolinealidad y ajuste de parámetros.
- **Análisis de Supervivencia**: Modelos de supervivencia, función de supervivencia y modelos de riesgos proporcionales de Cox.

10.2. Buenas Prácticas

Al implementar modelos de regresión logística, es importante seguir buenas prácticas para garantizar la precisión y la robustez de los modelos. Algunas buenas prácticas incluyen:

- **Exploración y Preparación de Datos**: Realizar un análisis exploratorio exhaustivo y preparar los datos adecuadamente antes de construir el modelo.
- **Evaluación y Validación del Modelo**: Utilizar métricas adecuadas para evaluar el rendimiento del modelo y validar el modelo utilizando técnicas como la validación cruzada.
- **Interpretación de Resultados**: Interpretar correctamente los coeficientes del modelo y las odds ratios, y comunicar los resultados de manera clara y concisa.
- **Revisión y Ajuste del Modelo**: Diagnosticar problemas en el modelo y ajustar los parámetros para mejorar el rendimiento.

10.3. Proyecto Final

Para aplicar los conceptos y técnicas aprendidos en este curso, te proponemos realizar un proyecto final utilizando un conjunto de datos de tu elección. El proyecto debe incluir las siguientes etapas:

10.3.1. Selección del Conjunto de Datos

Elige un conjunto de datos relevante que contenga una variable dependiente binaria o multinomial y varias variables independientes.

10.3.2. Exploración y Preparación de Datos

Realiza un análisis exploratorio de los datos y prepara los datos para el modelado. Esto incluye la limpieza de datos, el tratamiento de valores faltantes y la codificación de variables categóricas.

10.3.3. Entrenamiento y Evaluación del Modelo

Entrena un modelo de regresión logística utilizando el conjunto de datos preparado y evalúa su rendimiento utilizando métricas apropiadas.

10.3.4. Interpretación de Resultados

Interpreta los coeficientes del modelo y las odds ratios, y proporciona una explicación clara de los resultados.

10.3.5. Presentación del Proyecto

Presenta tu proyecto en un informe detallado que incluya la descripción del conjunto de datos, los pasos de preparación y modelado, los resultados del modelo y las conclusiones.

Parte II

SEGUNDA PARTE

CAPÍTULO 11

Introducción al Análisis de Supervivencia

11.1. Conceptos Básicos

El análisis de supervivencia es una rama de la estadística que se ocupa del análisis del tiempo que transcurre hasta que ocurre un evento de interés, comúnmente referido como "tiempo de falla". Este campo es ampliamente utilizado en medicina, biología, ingeniería, ciencias sociales, y otros campos.

11.2. Definición de Eventos y Tiempos

En el análisis de supervivencia, un .ev"ento"se refiere a la ocurrencia de un evento específico, como la muerte, la falla de un componente, la recaída de una enfermedad, etc. El "tiempo de supervivencia.es el tiempo que transcurre desde un punto de inicio definido hasta la ocurrencia del evento.

11.3. Censura

La censura ocurre cuando la información completa sobre el tiempo hasta el evento no está disponible para todos los individuos en el estudio. Hay tres tipos principales de censura:

- **Censura a la derecha:** Ocurre cuando el evento de interés no se ha observado para algunos sujetos antes del final del estudio.
- **Censura a la izquierda:** Ocurre cuando el evento de interés ocurrió antes del inicio del periodo de observación.
- **Censura por intervalo:** Ocurre cuando el evento de interés se sabe que ocurrió en un intervalo de tiempo, pero no se conoce el momento exacto.

11.4. Función de Supervivencia

La función de supervivencia, $S(t)$, se define como la probabilidad de que un individuo sobreviva más allá de un tiempo t. Matemáticamente, se expresa como:

$$S(t) = P(T > t)$$

donde T es una variable aleatoria que representa el tiempo hasta el evento. La función de supervivencia tiene las siguientes propiedades:

- $S(0) = 1$: Esto indica que al inicio (tiempo $t = 0$), la probabilidad de haber experimentado el evento es cero, por lo tanto, la supervivencia es del 100

- $\lim_{t\to\infty} S(t) = 0$: A medida que el tiempo tiende al infinito, la probabilidad de que cualquier individuo aún no haya experimentado el evento tiende a cero.
- $S(t)$ es una función no creciente: Esto significa que a medida que el tiempo avanza, la probabilidad de supervivencia no aumenta.

11.5. Función de Densidad de Probabilidad

La función de densidad de probabilidad $f(t)$ describe la probabilidad de que el evento ocurra en un instante de tiempo específico. Se define como:

$$f(t) = \frac{dF(t)}{dt}$$

donde $F(t)$ es la función de distribución acumulada, $F(t) = P(T \leq t)$. La relación entre $S(t)$ y $f(t)$ es:

$$f(t) = -\frac{dS(t)}{dt}$$

11.6. Función de Riesgo

La función de riesgo, $\lambda(t)$, también conocida como función de tasa de fallas o hazard rate, se define como la tasa instantánea de ocurrencia del evento en el tiempo t, dado que el individuo ha sobrevivido hasta el tiempo t. Matemáticamente, se expresa como:

$$\lambda(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

Esto se puede reescribir usando $f(t)$ y $S(t)$ como:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

11.7. Relación entre Función de Supervivencia y Función de Riesgo

La función de supervivencia y la función de riesgo están relacionadas a través de la siguiente ecuación:

$$S(t) = \exp\left(-\int_0^t \lambda(u)\, du\right)$$

Esta fórmula se deriva del hecho de que la función de supervivencia es la probabilidad acumulativa de no haber experimentado el evento hasta el tiempo t, y $\lambda(t)$ es la tasa instantánea de ocurrencia del evento.

La función de riesgo también puede ser expresada como:

$$\lambda(t) = -\frac{d}{dt} \log S(t)$$

11.8. Deducción de la Función de Supervivencia

La relación entre la función de supervivencia y la función de riesgo se puede deducir integrando la función de riesgo:

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t \lambda(u)\,du\right) \\ \log S(t) &= -\int_0^t \lambda(u)\,du \\ \frac{d}{dt}\log S(t) &= -\lambda(t) \\ \lambda(t) &= -\frac{d}{dt}\log S(t) \end{aligned}$$

11.9. Ejemplo de Cálculo

Supongamos que tenemos una muestra de tiempos de supervivencia $T_1, T_2, \ldots, T_n$. Podemos estimar la función de supervivencia empírica como:

$$\hat{S}(t) = \frac{\text{Número de individuos que sobreviven más allá de } t}{\text{Número total de individuos en riesgo en } t}$$

y la función de riesgo empírica como:

$$\hat{\lambda}(t) = \frac{\text{Número de eventos en } t}{\text{Número de individuos en riesgo en } t}$$

11.10. Conclusión

El análisis de supervivencia es una herramienta poderosa para analizar datos de tiempo hasta evento. Entender los conceptos básicos como la función de supervivencia y la función de riesgo es fundamental para el análisis más avanzado.

CAPÍTULO 12

Función de Supervivencia y Función de Riesgo

12.1. Introducción

Este capítulo profundiza en la definición y propiedades de la función de supervivencia y la función de riesgo, dos conceptos fundamentales en el análisis de supervivencia. Entender estas funciones y su relación es crucial para modelar y analizar datos de tiempo hasta evento.

12.2. Función de Supervivencia

La función de supervivencia, $S(t)$, describe la probabilidad de que un individuo sobreviva más allá de un tiempo t. Formalmente, se define como:

$$S(t) = P(T > t)$$

donde T es una variable aleatoria que representa el tiempo hasta el evento.

12.2.1. Propiedades de la Función de Supervivencia

La función de supervivencia tiene varias propiedades importantes:

- $S(0) = 1$: Indica que la probabilidad de haber experimentado el evento en el tiempo 0 es cero.
- $\lim_{t\to\infty} S(t) = 0$: A medida que el tiempo tiende al infinito, la probabilidad de supervivencia tiende a cero.
- $S(t)$ es una función no creciente: A medida que el tiempo avanza, la probabilidad de supervivencia no aumenta.

12.2.2. Derivación de $S(t)$

Si la función de densidad de probabilidad $f(t)$ del tiempo de supervivencia T es conocida, la función de supervivencia puede derivarse como:

$$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - P(T \leq t) \\ &= 1 - F(t) \\ &= 1 - \int_0^t f(u)\, du \end{aligned}$$

donde $F(t)$ es la función de distribución acumulada.

12.2.3. Ejemplo de Cálculo de $S(t)$

Consideremos un ejemplo donde el tiempo de supervivencia T sigue una distribución exponencial con tasa λ. La función de densidad de probabilidad $f(t)$ es:

$$f(t) = \lambda e^{-\lambda t}, \quad t \geq 0$$

La función de distribución acumulada $F(t)$ es:

$$F(t) = \int_0^t \lambda e^{-\lambda u}\, du = 1 - e^{-\lambda t}$$

Por lo tanto, la función de supervivencia $S(t)$ es:

$$S(t) = 1 - F(t) = e^{-\lambda t}$$

12.3. Función de Riesgo

La función de riesgo, $\lambda(t)$, proporciona la tasa instantánea de ocurrencia del evento en el tiempo t, dado que el individuo ha sobrevivido hasta el tiempo t. Matemáticamente, se define como:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

12.3.1. Relación entre $\lambda(t)$ y $f(t)$

La función de riesgo se puede relacionar con la función de densidad de probabilidad $f(t)$ y la función de supervivencia $S(t)$ de la siguiente manera:

$$\lambda(t) \quad = \quad \frac{f(t)}{S(t)}$$

12.3.2. Derivación de $\lambda(t)$

La derivación de $\lambda(t)$ se basa en la definición condicional de la probabilidad:

$$\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\frac{P(t \leq T < t+\Delta t \text{ y } T \geq t)}{P(T \geq t)}}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\frac{P(t \leq T < t+\Delta t)}{P(T \geq t)}}{\Delta t} \\
&= \frac{f(t)}{S(t)}
\end{aligned}$$

12.4. Relación entre Función de Supervivencia y Función de Riesgo

La función de supervivencia y la función de riesgo están estrechamente relacionadas. La relación se expresa mediante la siguiente ecuación:

$$S(t) = \exp\left(-\int_0^t \lambda(u)\, du\right)$$

12.4.1. Deducción de la Relación

Para deducir esta relación, consideramos la derivada logarítmica de la función de supervivencia:

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t \lambda(u)\,du\right) \\ \log S(t) &= -\int_0^t \lambda(u)\,du \\ \frac{d}{dt}\log S(t) &= -\lambda(t) \\ \lambda(t) &= -\frac{d}{dt}\log S(t) \end{aligned}$$

12.5. Interpretación de la Función de Riesgo

La función de riesgo, $\lambda(t)$, se interpreta como la tasa instantánea de ocurrencia del evento por unidad de tiempo, dado que el individuo ha sobrevivido hasta el tiempo t. Es una medida local del riesgo de falla en un instante específico.

12.5.1. Ejemplo de Cálculo de $\lambda(t)$

Consideremos nuevamente el caso donde el tiempo de supervivencia T sigue una distribución exponencial con tasa λ. La función de densidad de probabilidad $f(t)$ es:

$$f(t) = \lambda e^{-\lambda t}$$

La función de supervivencia $S(t)$ es:

$$S(t) = e^{-\lambda t}$$

La función de riesgo $\lambda(t)$ se calcula como:

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{S(t)} \\ &= \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} \\ &= \lambda \end{aligned}$$

En este caso, $\lambda(t)$ es constante y igual a λ, lo que es una característica de la distribución exponencial.

12.6. Funciones de Riesgo Acumulada y Media Residual

La función de riesgo acumulada $H(t)$ se define como:

$$H(t) = \int_0^t \lambda(u)\,du$$

Esta función proporciona la suma acumulada de la tasa de riesgo hasta el tiempo t.

La función de vida media residual $e(t)$ se define como la esperanza del tiempo de vida restante dado que el individuo ha sobrevivido hasta el tiempo t:

$$e(t) = \mathbb{E}[T - t \mid T > t] = \int_t^\infty S(u)\,du$$

12.7. Ejemplo de Cálculo de Función de Riesgo Acumulada y Vida Media Residual

Consideremos nuevamente la distribución exponencial con tasa λ. La función de riesgo acumulada $H(t)$ es:

$$\begin{aligned} H(t) &= \int_0^t \lambda \, du \\ &= \lambda t \end{aligned}$$

La función de vida media residual $e(t)$ es:

$$\begin{aligned} e(t) &= \int_t^\infty e^{-\lambda u} \, du \\ &= \left[\frac{-1}{\lambda} e^{-\lambda u} \right]_t^\infty \\ &= \frac{1}{\lambda} e^{-\lambda t} \\ &= \frac{1}{\lambda} \end{aligned}$$

En este caso, la vida media residual es constante e igual a $\frac{1}{\lambda}$, otra característica de la distribución exponencial.

12.8. Conclusión

La función de supervivencia y la función de riesgo son herramientas fundamentales en el análisis de supervivencia. Entender su definición, propiedades, y la relación entre ellas es esencial para modelar y analizar correctamente los datos de tiempo hasta evento. Las funciones de riesgo acumulada y vida media residual proporcionan información adicional sobre la dinámica del riesgo a lo largo del tiempo.

CAPÍTULO 13

Estimador de Kaplan-Meier

13.1. Introducción

El estimador de Kaplan-Meier, también conocido como la función de supervivencia empírica, es una herramienta no paramétrica para estimar la función de supervivencia a partir de datos censurados. Este método es especialmente útil cuando los tiempos de evento están censurados a la derecha.

13.2. Definición del Estimador de Kaplan-Meier

El estimador de Kaplan-Meier se define como:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

donde:

- t_i es el tiempo del i-ésimo evento,
- d_i es el número de eventos que ocurren en t_i,
- n_i es el número de individuos en riesgo justo antes de t_i.

13.3. Propiedades del Estimador de Kaplan-Meier

El estimador de Kaplan-Meier tiene las siguientes propiedades:

- Es una función escalonada que disminuye en los tiempos de los eventos observados.
- Puede manejar datos censurados a la derecha.
- Proporciona una estimación no paramétrica de la función de supervivencia.

13.3.1. Función Escalonada

La función escalonada del estimador de Kaplan-Meier significa que $\hat{S}(t)$ permanece constante entre los tiempos de los eventos y disminuye en los tiempos de los eventos. Matemáticamente, si t_i es el tiempo del i-ésimo evento, entonces:

$$\hat{S}(t) = \hat{S}(t_i) \quad \text{para } t_i \leq t < t_{i+1}$$

13.3.2. Manejo de Datos Censurados

El estimador de Kaplan-Meier maneja datos censurados a la derecha al ajustar la estimación de la función de supervivencia sólo en los tiempos en que ocurren eventos. Si un individuo es censurado antes de experimentar el evento, no contribuye a la disminución de $\hat{S}(t)$ en el tiempo de censura. Esto asegura que la censura no sesga la estimación de la supervivencia.

13.3.3. Estimación No Paramétrica

El estimador de Kaplan-Meier es no paramétrico porque no asume ninguna forma específica para la distribución de los tiempos de supervivencia. En cambio, utiliza la información empírica disponible para estimar la función de supervivencia.

13.4. Deducción del Estimador de Kaplan-Meier

La deducción del estimador de Kaplan-Meier se basa en el principio de probabilidad condicional. Consideremos un conjunto de tiempos de supervivencia observados $t_1, t_2, \ldots, t_k$ con eventos en cada uno de estos tiempos. El estimador de la probabilidad de supervivencia más allá del tiempo t es el producto de las probabilidades de sobrevivir más allá de cada uno de los tiempos de evento observados hasta t.

13.4.1. Probabilidad Condicional

La probabilidad de sobrevivir más allá de t_i, dado que el individuo ha sobrevivido justo antes de t_i, es:

$$P(T > t_i \mid T \geq t_i) = 1 - \frac{d_i}{n_i}$$

donde d_i es el número de eventos en t_i y n_i es el número de individuos en riesgo justo antes de t_i.

13.4.2. Producto de Probabilidades Condicionales

La probabilidad de sobrevivir más allá de un tiempo t cualquiera, dada la secuencia de tiempos de evento, es el producto de las probabilidades condicionales de sobrevivir más allá de cada uno de los tiempos de evento observados hasta t. Así, el estimador de Kaplan-Meier se obtiene como:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

13.5. Ejemplo de Cálculo

Supongamos que tenemos los siguientes tiempos de supervivencia observados para cinco individuos: 2, 3, 5, 7, 8. Supongamos además que tenemos censura a la derecha en el tiempo 10. Los tiempos de evento y el número de individuos en riesgo justo antes de cada evento son:

Tiempo (t_i)	Eventos (d_i)	En Riesgo (n_i)
2	1	5
3	1	4
5	1	3
7	1	2
8	1	1

Cuadro 13.1: Ejemplo de cálculo del estimador de Kaplan-Meier

Usando estos datos, el estimador de Kaplan-Meier se calcula como:

$$\begin{aligned}
\hat{S}(2) &= 1 - \frac{1}{5} = 0{,}8 \\
\hat{S}(3) &= 0{,}8 \times \left(1 - \frac{1}{4}\right) = 0{,}8 \times 0{,}75 = 0{,}6 \\
\hat{S}(5) &= 0{,}6 \times \left(1 - \frac{1}{3}\right) = 0{,}6 \times 0{,}6667 = 0{,}4 \\
\hat{S}(7) &= 0{,}4 \times \left(1 - \frac{1}{2}\right) = 0{,}4 \times 0{,}5 = 0{,}2 \\
\hat{S}(8) &= 0{,}2 \times \left(1 - \frac{1}{1}\right) = 0{,}2 \times 0 = 0
\end{aligned}$$

13.6. Intervalos de Confianza para el Estimador de Kaplan-Meier

Para calcular intervalos de confianza para el estimador de Kaplan-Meier, se puede usar la transformación logarítmica y la aproximación normal. Un intervalo de confianza aproximado para $\log(-\log(\hat{S}(t)))$ se obtiene como:

$$\log(-\log(\hat{S}(t))) \pm z_{\alpha/2}\sqrt{\frac{1}{d_i(n_i - d_i)}}$$

donde $z_{\alpha/2}$ es el percentil correspondiente de la distribución normal estándar.

13.7. Transformación Logarítmica Inversa

La transformación logarítmica inversa se utiliza para obtener los límites del intervalo de confianza para $S(t)$:

$$\hat{S}(t) = \exp\left(-\exp\left(\log(-\log(\hat{S}(t))) \pm z_{\alpha/2}\sqrt{\frac{1}{d_i(n_i - d_i)}}\right)\right)$$

13.8. Cálculo Detallado de Intervalos de Confianza

Para un cálculo más detallado de los intervalos de confianza, consideremos un tiempo específico t_j. La varianza del estimador de Kaplan-Meier en t_j se puede estimar usando Greenwood's formula:

$$\text{Var}(\hat{S}(t_j)) = \hat{S}(t_j)^2 \sum_{t_i \leq t_j} \frac{d_i}{n_i(n_i - d_i)}$$

El intervalo de confianza aproximado para $\hat{S}(t_j)$ es entonces:

$$\hat{S}(t_j) \pm z_{\alpha/2}\sqrt{\text{Var}(\hat{S}(t_j))}$$

13.9. Ejemplo de Intervalo de Confianza

Supongamos que en el ejemplo anterior queremos calcular el intervalo de confianza para $\hat{S}(3)$. Primero, calculamos la varianza:

$$\begin{aligned}
\text{Var}(\hat{S}(3)) &= \hat{S}(3)^2 \left(\frac{1}{5 \times 4} + \frac{1}{4 \times 3} \right) \\
&= 0{,}6^2 \left(\frac{1}{20} + \frac{1}{12} \right) \\
&= 0{,}36 \, (0{,}05 + 0{,}0833) \\
&= 0{,}36 \times 0{,}1333 \\
&= 0{,}048
\end{aligned}$$

El intervalo de confianza es entonces:

$$0{,}6 \pm 1{,}96\sqrt{0{,}048} = 0{,}6 \pm 1{,}96 \times 0{,}219 = 0{,}6 \pm 0{,}429$$

Por lo tanto, el intervalo de confianza para $\hat{S}(3)$ es aproximadamente $(0{,}171, 1{,}029)$. Dado que una probabilidad no puede exceder 1, ajustamos el intervalo a $(0{,}171, 1{,}0)$.

13.10. Interpretación del Estimador de Kaplan-Meier

El estimador de Kaplan-Meier proporciona una estimación empírica de la función de supervivencia que es fácil de interpretar y calcular. Su capacidad para manejar datos censurados lo hace especialmente útil en estudios de supervivencia.

13.11. Conclusión

El estimador de Kaplan-Meier es una herramienta poderosa para estimar la función de supervivencia en presencia de datos censurados. Su cálculo es relativamente sencillo y proporciona una estimación no paramétrica robusta de la supervivencia a lo largo del tiempo. La interpretación adecuada de este estimador y su intervalo de confianza asociado es fundamental para el análisis de datos de supervivencia.

CAPÍTULO 14

Comparación de Curvas de Supervivencia

14.1. Introducción

Comparar curvas de supervivencia es crucial para determinar si existen diferencias significativas en las tasas de supervivencia entre diferentes grupos. Las pruebas de hipótesis, como el test de log-rank, son herramientas comunes para esta comparación.

14.2. Test de Log-rank

El test de log-rank se utiliza para comparar las curvas de supervivencia de dos o más grupos. La hipótesis nula es que no hay diferencia en las funciones de riesgo entre los grupos.

14.2.1. Fórmula del Test de Log-rank

El estadístico del test de log-rank se define como:

$$\chi^2 = \frac{\left(\sum_{i=1}^{k}(O_i - E_i)\right)^2}{\sum_{i=1}^{k} V_i}$$

donde:

- O_i es el número observado de eventos en el grupo i.
- E_i es el número esperado de eventos en el grupo i.
- V_i es la varianza del número de eventos en el grupo i.

14.2.2. Cálculo de E_i y V_i

El número esperado de eventos E_i y la varianza V_i se calculan como:

$$\begin{aligned} E_i &= \frac{d_i \cdot n_i}{n} \\ V_i &= \frac{d_i \cdot (n - d_i) \cdot n_i \cdot (n - n_i)}{n^2 \cdot (n - 1)} \end{aligned}$$

donde:

- d_i es el número total de eventos en el grupo i.
- n_i es el número de individuos en riesgo en el grupo i.
- n es el número total de individuos en todos los grupos.

14.3. Ejemplo de Cálculo del Test de Log-rank

Supongamos que tenemos dos grupos con los siguientes datos de eventos:

Grupo	Tiempo (t_i)	Eventos (O_i)	En Riesgo (n_i)
1	2	1	5
1	4	1	4
2	3	1	4
2	5	1	3

Cuadro 14.1: Ejemplo de datos para el test de log-rank

Calculemos E_i y V_i para cada grupo:

$$
\begin{aligned}
E_1 &= \frac{2 \cdot 5}{9} + \frac{2 \cdot 4}{8} = \frac{10}{9} + \frac{8}{8} = 1{,}11 + 1 = 2{,}11 \\
V_1 &= \frac{2 \cdot 7 \cdot 5 \cdot 4}{81 \cdot 8} = \frac{2 \cdot 7 \cdot 5 \cdot 4}{648} = \frac{280}{648} = 0{,}432 \\
E_2 &= \frac{2 \cdot 4}{9} + \frac{2 \cdot 3}{8} = \frac{8}{9} + \frac{6}{8} = 0{,}89 + 0{,}75 = 1{,}64 \\
V_2 &= \frac{2 \cdot 7 \cdot 4 \cdot 4}{81 \cdot 8} = \frac{2 \cdot 7 \cdot 4 \cdot 4}{648} = \frac{224}{648} = 0{,}346
\end{aligned}
$$

El estadístico de log-rank se calcula como:

$$
\begin{aligned}
\chi^2 &= \frac{((1 - 2{,}11) + (1 - 1{,}64))^2}{0{,}432 + 0{,}346} \\
&= \frac{(-1{,}11 - 0{,}64)^2}{0{,}778} \\
&= \frac{3{,}04}{0{,}778} \\
&= 3{,}91
\end{aligned}
$$

El valor p se puede obtener comparando χ^2 con una distribución χ^2 con un grado de libertad (dado que estamos comparando dos grupos).

14.4. Interpretación del Test de Log-rank

Un valor p pequeño (generalmente menos de 0.05) indica que hay una diferencia significativa en las curvas de supervivencia entre los grupos. Un valor p grande sugiere que no hay suficiente evidencia para rechazar la hipótesis nula de que las curvas de supervivencia son iguales.

14.5. Pruebas Alternativas

Además del test de log-rank, existen otras pruebas para comparar curvas de supervivencia, como el test de Wilcoxon (Breslow), que da más peso a los eventos en tiempos tempranos.

14.6. Conclusión

El test de log-rank es una herramienta esencial para comparar curvas de supervivencia entre diferentes grupos. Su cálculo se basa en la diferencia entre los eventos observados y esperados en cada grupo, y su interpretación puede ayudar a identificar diferencias significativas en la supervivencia.

CAPÍTULO 15

Modelos de Riesgos Proporcionales de Cox

15.1. Introducción

El modelo de riesgos proporcionales de Cox, propuesto por David Cox en 1972, es una de las herramientas más utilizadas en el análisis de supervivencia. Este modelo permite evaluar el efecto de varias covariables en el tiempo hasta el evento, sin asumir una forma específica para la distribución de los tiempos de supervivencia.

15.2. Definición del Modelo de Cox

El modelo de Cox se define como:

$$\lambda(t \mid X) = \lambda_0(t) \exp(\beta^T X)$$

donde:

- $\lambda(t \mid X)$ es la función de riesgo en el tiempo t dado el vector de covariables X.
- $\lambda_0(t)$ es la función de riesgo basal en el tiempo t.
- β es el vector de coeficientes del modelo.
- X es el vector de covariables.

15.3. Supuesto de Proporcionalidad de Riesgos

El modelo de Cox asume que las razones de riesgo entre dos individuos son constantes a lo largo del tiempo. Matemáticamente, si X_i y X_j son las covariables de dos individuos, la razón de riesgos se expresa como:

$$\frac{\lambda(t \mid X_i)}{\lambda(t \mid X_j)} = \frac{\lambda_0(t) \exp(\beta^T X_i)}{\lambda_0(t) \exp(\beta^T X_j)} = \exp(\beta^T (X_i - X_j))$$

15.4. Estimación de los Parámetros

Los parámetros β se estiman utilizando el método de máxima verosimilitud parcial. La función de verosimilitud parcial se define como:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\beta^T X_i)}{\sum_{j \in R(t_i)} \exp(\beta^T X_j)}$$

donde $R(t_i)$ es el conjunto de individuos en riesgo en el tiempo t_i.

15.4.1. Función de Log-Verosimilitud Parcial

La función de log-verosimilitud parcial es:

$$\log L(\beta) = \sum_{i=1}^{k} \left(\beta^T X_i - \log \sum_{j \in R(t_i)} \exp(\beta^T X_j) \right)$$

15.4.2. Derivadas Parciales y Maximización

Para encontrar los estimadores de máxima verosimilitud, resolvemos el sistema de ecuaciones obtenido al igualar a cero las derivadas parciales de $\log L(\beta)$ con respecto a β:

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{k} \left(X_i - \frac{\sum_{j \in R(t_i)} X_j \exp(\beta^T X_j)}{\sum_{j \in R(t_i)} \exp(\beta^T X_j)} \right) = 0$$

15.5. Interpretación de los Coeficientes

Cada coeficiente β_i representa el logaritmo de la razón de riesgos asociado con un incremento unitario en la covariable X_i. Un valor positivo de β_i indica que un aumento en X_i incrementa el riesgo del evento, mientras que un valor negativo indica una reducción del riesgo.

15.6. Evaluación del Modelo

El modelo de Cox se evalúa utilizando varias técnicas, como el análisis de residuos de Schoenfeld para verificar el supuesto de proporcionalidad de riesgos, y el uso de curvas de supervivencia estimadas para evaluar la bondad de ajuste.

15.6.1. Residuos de Schoenfeld

Los residuos de Schoenfeld se utilizan para evaluar la proporcionalidad de riesgos. Para cada evento en el tiempo t_i, el residuo de Schoenfeld para la covariable X_j se define como:

$$r_{ij} = X_{ij} - \hat{X}_{ij}$$

donde $\hat{X}_{ij}$ es la covariable ajustada.

15.6.2. Curvas de Supervivencia Ajustadas

Las curvas de supervivencia ajustadas se obtienen utilizando la función de riesgo basal estimada y los coeficientes del modelo. La función de supervivencia ajustada se define como:

$$\hat{S}(t \mid X) = \hat{S}_0(t)^{\exp(\beta^T X)}$$

donde $\hat{S}_0(t)$ es la función de supervivencia basal estimada.

15.7. Ejemplo de Aplicación del Modelo de Cox

Consideremos un ejemplo con tres covariables: edad, sexo y tratamiento. Supongamos que los datos se ajustan a un modelo de Cox y obtenemos los siguientes coeficientes:

$$\hat{\beta}_{edad} = 0{,}02, \quad \hat{\beta}_{sexo} = -0{,}5, \quad \hat{\beta}_{tratamiento} = 1{,}2$$

La función de riesgo ajustada se expresa como:

$$\lambda(t \mid X) = \lambda_0(t) \exp(0{,}02 \cdot \text{edad} - 0{,}5 \cdot \text{sexo} + 1{,}2 \cdot \text{tratamiento})$$

15.8. Conclusión

El modelo de riesgos proporcionales de Cox es una herramienta poderosa para analizar datos de supervivencia con múltiples covariables. Su flexibilidad y la falta de suposiciones fuertes sobre la distribución de los tiempos de supervivencia lo hacen ampliamente aplicable en diversas disciplinas.

CAPÍTULO 16

Diagnóstico y Validación de Modelos de Cox

16.1. Introducción

Una vez ajustado un modelo de Cox, es crucial realizar diagnósticos y validaciones para asegurar que el modelo es apropiado y que los supuestos subyacentes son válidos. Esto incluye la verificación del supuesto de proporcionalidad de riesgos y la evaluación del ajuste del modelo.

16.2. Supuesto de Proporcionalidad de Riesgos

El supuesto de proporcionalidad de riesgos implica que la razón de riesgos entre dos individuos es constante a lo largo del tiempo. Si este supuesto no se cumple, las inferencias hechas a partir del modelo pueden ser incorrectas.

16.2.1. Residuos de Schoenfeld

Los residuos de Schoenfeld se utilizan para evaluar la proporcionalidad de riesgos. Para cada evento en el tiempo t_i, el residuo de Schoenfeld para la covariable X_j se define como:

$$r_{ij} = X_{ij} - \hat{X}_{ij}$$

donde $\hat{X}_{ij}$ es la covariable ajustada. Si los residuos de Schoenfeld no muestran una tendencia sistemática cuando se trazan contra el tiempo, el supuesto de proporcionalidad de riesgos es razonable.

16.3. Bondad de Ajuste

La bondad de ajuste del modelo de Cox se evalúa comparando las curvas de supervivencia observadas y ajustadas, y utilizando estadísticas de ajuste global.

16.3.1. Curvas de Supervivencia Ajustadas

Las curvas de supervivencia aaustadas se obtienen utilizando la función de riesgo basal estimada y los coeficientes del modelo. La función de supervivencia ajustada se define como:

$$\hat{S}(t \mid X) = \hat{S}_0(t)^{\exp(\beta^T X)}$$

donde $\hat{S}_0(t)$ es la función de supervivencia basal estimada. Comparar estas curvas con las curvas de Kaplan-Meier para diferentes niveles de las covariables puede proporcionar una validación visual del ajuste del modelo.

16.3.2. Estadísticas de Ajuste Global

Las estadísticas de ajuste global, como el test de la desviación y el test de la bondad de ajuste de Grambsch y Therneau, se utilizan para evaluar el ajuste global del modelo de Cox.

16.4. Diagnóstico de Influencia

El diagnóstico de influencia identifica observaciones individuales que tienen un gran impacto en los estimados del modelo. Los residuos de devianza y los residuos de martingala se utilizan comúnmente para este propósito.

16.4.1. Residuos de Deviance

Los residuos de deviance se definen como:

$$D_i = \text{sign}(O_i - E_i)\sqrt{-2\left(O_i \log \frac{O_i}{E_i} - (O_i - E_i)\right)}$$

donde O_i es el número observado de eventos y E_i es el número esperado de eventos. Observaciones con residuos de deviance grandes en valor absoluto pueden ser influyentes.

16.4.2. Residuos de Martingala

Los residuos de martingala se definen como:

$$M_i = O_i - E_i$$

donde O_i es el número observado de eventos y E_i es el número esperado de eventos. Los residuos de martingala se utilizan para detectar observaciones que no se ajustan bien al modelo.

16.5. Ejemplo de Diagnóstico

Consideremos un modelo de Cox ajustado con las covariables edad, sexo y tratamiento. Para diagnosticar la influencia de observaciones individuales, calculamos los residuos de deviance y martingala para cada observación.

Observación	Edad	Sexo	Tratamiento	Residuo de Deviance
1	50	0	1	1.2
2	60	1	0	-0.5
3	45	0	1	-1.8
4	70	1	0	0.3

Cuadro 16.1: Residuos de deviance para observaciones individuales

Observaciones con residuos de deviance grandes en valor absoluto (como la observación 3) pueden ser influyentes y requieren una revisión adicional.

16.6. Conclusión

El diagnóstico y la validación son pasos críticos en el anślisis de modelos de Cox. Evaluar el supuesto de proporcionalidad de riesgos, la bondad de ajuste y la influencia de observaciones individuales asegura que las inferencias y conclusiones derivadas del modelo sean vślidas y fiables.

CAPÍTULO 17

Modelos Acelerados de Fallos

17.1. Introducción

Los modelos acelerados de fallos (AFT) son una alternativa a los modelos de riesgos proporcionales de Cox. En lugar de asumir que las covariables afectan la tasa de riesgo, los modelos AFT asumen que las covariables multiplican el tiempo de supervivencia por una constante.

17.2. Definición del Modelo AFT

Un modelo AFT se expresa como:

$$T = T_0 \exp(\beta^T X)$$

donde:

- T es el tiempo de supervivencia observado.
- T_0 es el tiempo de supervivencia bajo condiciones basales.
- β es el vector de coeficientes del modelo.
- X es el vector de covariables.

17.2.1. Transformación Logarítmica

El modelo AFT se puede transformar logarítmicamente para obtener una forma lineal:

$$\log(T) = \log(T_0) + \beta^T X$$

17.3. Estimación de los Parámetros

Los parámetros del modelo AFT se estiman utilizando el método de máxima verosimilitud. La función de verosimilitud se define como:

$$L(\beta) = \prod_{i=1}^{n} f(t_i \mid X_i; \beta)$$

donde $f(t_i \mid X_i; \beta)$ es la función de densidad de probabilidad del tiempo de supervivencia t_i dado el vector de covariables X_i y los parámetros β.

17.3.1. Función de Log-Verosimilitud

La función de log-verosimilitud es:

$$\log L(\beta) = \sum_{i=1}^{n} \log f(t_i \mid X_i; \beta)$$

17.3.2. Maximización de la Verosimilitud

Los estimadores de máxima verosimilitud se obtienen resolviendo el sistema de ecuaciones obtenido al igualar a cero las derivadas parciales de $\log L(\beta)$ con respecto a β:

$$\frac{\partial \log L(\beta)}{\partial \beta} = 0$$

17.4. Distribuciones Comunes en Modelos AFT

En los modelos AFT, el tiempo de supervivencia T puede seguir varias distribuciones comunes, como la exponencial, Weibull, log-normal y log-logística. Cada una de estas distribuciones tiene diferentes propiedades y aplicaciones.

17.4.1. Modelo Exponencial AFT

En un modelo exponencial AFT, el tiempo de supervivencia T sigue una distribución exponencial con parámetro λ:

$$f(t) = \lambda \exp(-\lambda t)$$

La función de supervivencia es:

$$S(t) = \exp(-\lambda t)$$

La transformación logarítmica del tiempo de supervivencia es:

$$\log(T) = \log\left(\frac{1}{\lambda}\right) + \beta^T X$$

17.4.2. Modelo Weibull AFT

En un modelo Weibull AFT, el tiempo de supervivencia T sigue una distribución Weibull con parámetros λ y k:

$$f(t) = \lambda k t^{k-1} \exp(-\lambda t^k)$$

La función de supervivencia es:

$$S(t) = \exp(-\lambda t^k)$$

La transformación logarítmica del tiempo de supervivencia es:

$$\log(T) = \log\left(\left(\frac{1}{\lambda}\right)^{1/k}\right) + \frac{\beta^T X}{k}$$

17.5. Interpretación de los Coeficientes

En los modelos AFT, los coeficientes β_i se interpretan como factores multiplicativos del tiempo de supervivencia. Un valor positivo de β_i indica que un aumento en la covariable X_i incrementa el tiempo de supervivencia, mientras que un valor negativo indica una reducción del tiempo de supervivencia.

17.6. Ejemplo de Aplicación del Modelo AFT

Consideremos un ejemplo con tres covariables: edad, sexo y tratamiento. Supongamos que los datos se ajustan a un modelo Weibull AFT y obtenemos los siguientes coeficientes:

$$\hat{\beta}_{edad} = -0{,}02, \quad \hat{\beta}_{sexo} = 0{,}5, \quad \hat{\beta}_{tratamiento} = -1{,}2$$

La función de supervivencia ajustada se expresa como:

$$S(t \mid X) = \exp\left(- \left(\frac{t \exp(-0{,}02 \cdot \text{edad} + 0{,}5 \cdot \text{sexo} - 1{,}2 \cdot \text{tratamiento})}{\lambda} \right)^{k} \right)$$

17.7. Conclusión

Los modelos AFT proporcionan una alternativa flexible a los modelos de riesgos proporcionales de Cox. Su enfoque en la multiplicación del tiempo de supervivencia por una constante permite una interpretación intuitiva y aplicaciones en diversas áreas.

CAPÍTULO 18

Análisis Multivariado de Supervivencia

18.1. Introducción

El análisis multivariado de supervivencia extiende los modelos de supervivencia para incluir múltiples covariables, permitiendo evaluar su efecto simultáneo sobre el tiempo hasta el evento. Los modelos de Cox y AFT son comúnmente utilizados en este contexto.

18.2. Modelo de Cox Multivariado

El modelo de Cox multivariado se define como:

$$\lambda(t \mid X) = \lambda_0(t) \exp(\beta^T X)$$

donde X es un vector de covariables.

18.2.1. Estimación de los Parámetros

Los parámetros β se estiman utilizando el método de máxima verosimilitud parcial, como se discutió anteriormente. La función de verosimilitud parcial se maximiza para obtener los estimadores de los coeficientes.

18.3. Modelo AFT Multivariado

El modelo AFT multivariado se expresa como:

$$T = T_0 \exp(\beta^T X)$$

18.3.1. Estimación de los Parámetros

Los parámetros β se estiman utilizando el método de máxima verosimilitud, similar al caso univariado. La función de verosimilitud se maximiza para obtener los estimadores de los coeficientes.

18.4. Interacción y Efectos No Lineales

En el análisis multivariado, es importante considerar la posibilidad de interacciones entre covariables y efectos no lineales. Estos se pueden incluir en los modelos extendiendo las funciones de riesgo o supervivencia.

18.4.1. Interacciones

Las interacciones entre covariables se pueden modelar añadiendo términos de interacción en el modelo:

$$\lambda(t \mid X) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2)$$

donde $X_1 X_2$ es el término de interacción.

18.4.2. Efectos No Lineales

Los efectos no lineales se pueden modelar utilizando funciones no lineales de las covariables, como polinomios o splines:

$$\lambda(t \mid X) = \lambda_0(t) \exp(\beta_1 X + \beta_2 X^2)$$

18.5. Selección de Variables

La selección de variables es crucial en el análisis multivariado para evitar el sobreajuste y mejorar la interpretabilidad del modelo. Métodos como la regresión hacia atrás, la regresión hacia adelante y la selección por criterios de información (AIC, BIC) son comúnmente utilizados.

18.5.1. Regresión Hacia Atrás

La regresión hacia atrás comienza con todas las covariables en el modelo y elimina iterativamente la covariable menos significativa hasta que todas las covariables restantes sean significativas.

18.5.2. Regresión Hacia Adelante

La regresión hacia adelante comienza con un modelo vacío y añade iterativamente la covariable más significativa hasta que no se pueda añadir ninguna covariable adicional significativa.

18.5.3. Criterios de Información

Los criterios de información, como el AIC (Akaike Information Criterion) y el BIC (Bayesian Information Criterion), se utilizan para seleccionar el modelo que mejor se ajusta a los datos con la menor complejidad posible:

$$\begin{aligned} AIC &= -2 \log L + 2k \\ BIC &= -2 \log L + k \log n \end{aligned}$$

donde L es la función de verosimilitud del modelo, k es el número de parámetros en el modelo y n es el tamaño de la muestra.

18.6. Ejemplo de Análisis Multivariado

Consideremos un ejemplo con tres covariables: edad, sexo y tratamiento. Ajustamos un modelo de Cox multivariado y obtenemos los siguientes coeficientes:

$$\hat{\beta}_{edad} = 0{,}03, \quad \hat{\beta}_{sexo} = -0{,}6, \quad \hat{\beta}_{tratamiento} = 1{,}5$$

La función de riesgo ajustada se expresa como:

$$\lambda(t \mid X) = \lambda_0(t) \exp(0{,}03 \cdot \text{edad} - 0{,}6 \cdot \text{sexo} + 1{,}5 \cdot \text{tratamiento})$$

18.7. Conclusión

El análisis multivariado de supervivencia permite evaluar el efecto conjunto de múltiples covariables sobre el tiempo hasta el evento. La inclusión de interacciones y efectos no lineales, junto con la selección adecuada de variables, mejora la precisión y la interpretabilidad de los modelos de supervivencia.

CAPÍTULO 19

Supervivencia en Datos Complicados

19.1. Introducción

El análisis de supervivencia en datos complicados se refiere a la evaluación de datos de supervivencia que presentan desafíos adicionales, como la censura por intervalo, datos truncados y datos con múltiples tipos de eventos. Estos escenarios requieren métodos avanzados para un análisis adecuado.

19.2. Censura por Intervalo

La censura por intervalo ocurre cuando el evento de interés se sabe que ocurrió dentro de un intervalo de tiempo, pero no se conoce el momento exacto. Esto es común en estudios donde las observaciones se realizan en puntos de tiempo discretos.

19.2.1. Modelo para Datos Censurados por Intervalo

Para datos censurados por intervalo, la función de verosimilitud se modifica para incluir la probabilidad de que el evento ocurra dentro de un intervalo:

$$L(\beta) = \prod_{i=1}^{n} P(T_i \in [L_i, U_i] \mid X_i; \beta)$$

donde $[L_i, U_i]$ es el intervalo de tiempo durante el cual se sabe que ocurrió el evento para el individuo i.

19.3. Datos Truncados

Los datos truncados ocurren cuando los tiempos de supervivencia están sujetos a un umbral, y solo se observan los individuos cuyos tiempos de supervivencia superan (o están por debajo de) ese umbral. Existen dos tipos principales de truncamiento: truncamiento a la izquierda y truncamiento a la derecha.

19.3.1. Modelo para Datos Truncados

Para datos truncados a la izquierda, la función de verosimilitud se ajusta para considerar solo los individuos que superan el umbral de truncamiento:

$$L(\beta) = \prod_{i=1}^{n} \frac{f(t_i \mid X_i; \beta)}{1 - F(L_i \mid X_i; \beta)}$$

donde L_i es el umbral de truncamiento para el individuo i.

19.4. Análisis de Competing Risks

En estudios donde pueden ocurrir múltiples tipos de eventos (competing risks), es crucial modelar adecuadamente el riesgo asociado con cada tipo de evento. La probabilidad de ocurrencia de cada evento compite con las probabilidades de ocurrencia de otros eventos.

19.4.1. Modelo de Competing Risks

Para un análisis de competing risks, la función de riesgo se descompone en funciones de riesgo específicas para cada tipo de evento:

$$\lambda(t) = \sum_{j=1}^{m} \lambda_j(t)$$

donde $\lambda_j(t)$ es la función de riesgo para el evento j.

19.5. Métodos de Imputación

Los métodos de imputación se utilizan para manejar datos faltantes o censurados en estudios de supervivencia. La imputación múltiple es un enfoque común que crea múltiples conjuntos de datos completos imputando valores faltantes varias veces y luego combina los resultados.

19.5.1. Imputación Múltiple

La imputación múltiple para datos de supervivencia se realiza en tres pasos:

1. Imputar los valores faltantes múltiples veces para crear varios conjuntos de datos completos.
2. Analizar cada conjunto de datos completo por separado utilizando métodos de supervivencia estándar.
3. Combinar los resultados de los análisis separados para obtener estimaciones y varianzas combinadas.

19.6. Ejemplo de Análisis con Datos Complicados

Consideremos un estudio con datos censurados por intervalo y competing risks. Ajustamos un modelo para los datos censurados por intervalo y obtenemos los siguientes coeficientes para las covariables edad y tratamiento:

$$\hat{\beta}_{edad} = 0{,}04, \quad \hat{\beta}_{tratamiento} = -0{,}8$$

La función de supervivencia ajustada se expresa como:

$$S(t \mid X) = \exp\left(-\left(\frac{t \exp(0{,}04 \cdot \text{edad} - 0{,}8 \cdot \text{tratamiento})}{\lambda}\right)^k\right)$$

19.7. Conclusión

El análisis de supervivencia en datos complicados requiere métodos avanzados para manejar censura por intervalo, datos truncados y competing risks. La aplicación de modelos adecuados y métodos de imputación asegura un análisis preciso y completo de estos datos complejos.

CAPÍTULO 20

Proyecto Final y Revisión

20.1. Introducción

El proyecto final proporciona una oportunidad para aplicar los conceptos y técnicas aprendidas en el curso de análisis de supervivencia. Este capítulo incluye una guía para desarrollar un proyecto de análisis de supervivencia y una revisión de los conceptos clave.

20.2. Desarrollo del Proyecto

El proyecto final debe incluir los siguientes componentes:

1. Definición del problema: Identificar la pregunta de investigación y los objetivos del análisis de supervivencia.
2. Descripción de los datos: Presentar los datos utilizados, incluyendo las covariables y la estructura de los datos.
3. Análisis exploratorio: Realizar un análisis descriptivo de los datos, incluyendo la censura y la distribución de los tiempos de supervivencia.
4. Ajuste del modelo: Ajustar modelos de supervivencia adecuados (Kaplan-Meier, Cox, AFT) y evaluar su bondad de ajuste.
5. Diagnóstico del modelo: Realizar diagnósticos para evaluar los supuestos del modelo y la influencia de observaciones individuales.
6. Interpretación de resultados: Interpretar los coeficientes del modelo y las curvas de supervivencia ajustadas.
7. Conclusiones: Resumir los hallazgos del análisis y proporcionar recomendaciones basadas en los resultados.

20.3. Revisión de Conceptos Clave

Una revisión de los conceptos clave del análisis de supervivencia incluye:

- **Función de Supervivencia:** Define la probabilidad de sobrevivir más allá de un tiempo específico.
- **Función de Riesgo:** Define la tasa instantánea de ocurrencia del evento.
- **Estimador de Kaplan-Meier:** Proporciona una estimación no paramétrica de la función de supervivencia.
- **Test de Log-rank:** Compara curvas de supervivencia entre diferentes grupos.

- **Modelo de Cox:** Evalúa el efecto de múltiples covariables sobre el tiempo hasta el evento, asumiendo proporcionalidad de riesgos.
- **Modelos AFT:** Modelan el efecto de las covariables multiplicando el tiempo de supervivencia por una constante.
- **Análisis Multivariado:** Considera interacciones y efectos no lineales entre múltiples covariables.
- **Supervivencia en Datos Complicados:** Maneja censura por intervalo, datos truncados y competing risks.

20.4. Ejemplo de Proyecto Final

A continuación se presenta un ejemplo de estructura de un proyecto final de análisis de supervivencia:

20.4.1. Definición del Problema

Analizar el efecto del tratamiento y la edad sobre la supervivencia de pacientes con una enfermedad específica.

20.4.2. Descripción de los Datos

Datos de supervivencia de 100 pacientes, con covariables: edad, sexo y tipo de tratamiento. Los tiempos de supervivencia están censurados a la derecha.

20.4.3. Análisis Exploratorio

Realizar histogramas y curvas de Kaplan-Meier para explorar la distribución de los tiempos de supervivencia y la censura.

20.4.4. Ajuste del Modelo

Ajustar un modelo de Cox y un modelo AFT con las covariables edad y tratamiento.

20.4.5. Diagnóstico del Modelo

Evaluar la proporcionalidad de riesgos y realizar análisis de residuos para identificar observaciones influyentes.

20.4.6. Interpretación de Resultados

Interpretar los coeficientes del modelo y las curvas de supervivencia ajustadas para diferentes niveles de las covariables.

20.4.7. Conclusiones

Resumir los hallazgos y proporcionar recomendaciones sobre el efecto del tratamiento y la edad en la supervivencia de los pacientes.

20.5. Conclusión

El proyecto final es una oportunidad para aplicar los conocimientos adquiridos en un contexto práctico. La revisión de los conceptos clave y la aplicación de técnicas adecuadas de análisis de supervivencia aseguran un análisis riguroso y significativo.

Parte III

APÉNDICES

CAPÍTULO 21

IMPLEMENTACIONES NUMÉRICAS

21.1. Día 1: Regresión Logística

Implementación Básica en R

Para implementar una regresión logística en R, primero es necesario instalar y cargar los paquetes necesarios.

Instalación y Configuración de R y RStudio

- Descargue e instale R desde `https://cran.r-project.org/`. Siga las instrucciones para su sistema operativo (Windows, MacOS, Linux).
- Descargue e instale RStudio desde `https://rstudio.com/products/rstudio/download/`.

21.1.1. Ejemplo de Regresión Logística en R

A continuación, se muestra un ejemplo de cómo ajustar un modelo de regresión logística en R utilizando un conjunto de datos simulado. El ejemplo incluye la instalación del paquete necesario, la carga de datos, el ajuste del modelo, y la interpretación de los resultados.

```
# Instalación del paquete necesario
install.packages("stats")

# Carga del paquete
library(stats)

# Ejemplo de conjunto de datos
data <- data.frame(
  outcome = c(1, 0, 1, 0, 1, 1, 0, 1, 0, 0),
  predictor = c(2.3, 1.9, 3.1, 2.8, 3.6, 2.4, 2.1, 3.3, 2.2, 1.7)
)

# Ajuste del modelo de regresión logística
model <- glm(outcome ~ predictor, data = data, family = binomial)

# Resumen del modelo
summary(model)
```

En este ejemplo, se utiliza el conjunto de datos *data* que contiene una variable de resultado binaria *outcome* y una variable predictora continua *predictor*. El modelo de regresión logística se ajusta utilizando la función `glm` con la familia binomial. La función `summary(model)` proporciona un resumen del modelo ajustado, incluyendo los coeficientes estimados, sus errores estándar, valores z, y p-valores.

- **Coeficientes**: Los coeficientes estimados β_0 y β_1 indican la dirección y magnitud de la relación entre las variables predictoras y la probabilidad del resultado.
- **Errores Estándar**: Los errores estándar proporcionan una medida de la precisión de los coeficientes estimados.
- **Valores z y p-valores**: Los valores z y p-valores se utilizan para evaluar la significancia estadística de los coeficientes. Un p-valor pequeño (generalmente ¡0.05) indica que el coeficiente es significativamente diferente de cero.

Este es solo un ejemplo básico, en aplicaciones reales, es posible que necesites realizar más análisis y validaciones, como la evaluación de la bondad de ajuste del modelo, el diagnóstico de posibles problemas de multicolinealidad, y la validación cruzada del modelo.

```
# Archivo: regresionlogistica.R

# Instalación del paquete necesario
#install.packages("stats")

# Carga del paquete
library(stats)

# Fijar la semilla para reproducibilidad
set.seed(123)

# Número de observaciones
n <- 100

# Generar las variables independientes X1, X2, ..., X15
# Creamos una matriz de tamaño n x 15 con valores generados aleatoriamente de una
 distribución normal
X <- as.data.frame(matrix(rnorm(n * 15), nrow = n, ncol = 15))
colnames(X) <- paste0("X", 1:15)  # Nombramos las columnas como X1, X2, ..., X15

# Coeficientes verdaderos para las variables independientes
# Generamos un vector de 16 coeficientes (incluyendo el intercepto) aleatorios entre -1 y 1
beta <- runif(16, -1, 1)  # 15 coeficientes más el intercepto

# Generar el término lineal
# Calculamos el término lineal utilizando los coeficientes y las variables independientes
linear_term <- beta[1] + as.matrix(X) %*% beta[-1]

# Generar la probabilidad utilizando la función logística
# Calculamos las probabilidades utilizando la función logística
p <- 1 / (1 + exp(-linear_term))

# Generar la variable dependiente binaria Y
# Generamos valores binarios (0 o 1) utilizando las probabilidades calculadas
Y <- rbinom(n, 1, p)

# Combinar las variables independientes y la variable dependiente en un data frame
data <- cbind(Y, X)

# Dividir el conjunto de datos en entrenamiento y prueba
set.seed(123)  # Fijar la semilla para reproducibilidad
train_indices <- sample(1:n, size = 0.7 * n)  # 70% de los datos para entrenamiento
train_set <- data[train_indices, ]  # Conjunto de entrenamiento
```

```
test_set <- data[-train_indices, ]  # Conjunto de prueba

# Ajuste del modelo de regresión logística en el conjunto de entrenamiento
# Ajustamos un modelo de regresión logística utilizando las variables independientes
para predecir Y
model <- glm(Y ~ ., data = train_set, family = binomial)

# Resumen del modelo
# Mostramos un resumen del modelo ajustado
summary(model)

# Guardar el modelo y los resultados en un archivo
# Guardamos el modelo ajustado en un archivo .RData
save(model, file = "regresion_logistica_modelo.RData")

# Guardar los datos simulados en archivos CSV
# Guardamos los conjuntos de datos de entrenamiento y prueba en archivos CSV
write.csv(train_set, "datos_entrenamiento_regresion_logistica.csv", row.names = FALSE)
write.csv(test_set, "datos_prueba_regresion_logistica.csv", row.names = FALSE)

# Hacer predicciones en el conjunto de prueba
# Utilizamos el modelo ajustado para hacer predicciones en el conjunto de prueba
test_set$prob_pred <- predict(model, newdata = test_set, type = "response")
test_set$Y_pred <- ifelse(test_set$prob_pred > 0.5, 1, 0)
# Convertimos probabilidades a clases binarias

# Calcular la precisión de las predicciones
# Calculamos la precisión de las predicciones comparando con los valores reales de Y
accuracy <- mean(test_set$Y_pred == test_set$Y)
cat("La precisión del modelo en el conjunto de prueba es:", accuracy, "\n")

# Guardar las predicciones en un archivo CSV
# Guardamos las predicciones y las probabilidades predichas en un archivo CSV
write.csv(test_set, "predicciones_regresion_logistica.csv", row.names = FALSE)

# Graficar los coeficientes estimados
# Graficamos los coeficientes estimados del modelo ajustado
plot(coef(model), main = "Coeficientes Estimados del Modelo de Regresión Logística",
     xlab = "Variables", ylab = "Coeficientes", type = "h", col = "blue")
abline(h = 0, col = "red", lwd = 2)

# Mostrar un mensaje indicando que el proceso ha finalizado
cat("El modelo de regresión logística se ha ajustado, se han hecho predicciones y los resultados se
```

21.1.2. Aplicación a Datos de Cáncer - Parte I

A continuación, se muestra un ejemplo de cómo ajustar un modelo de regresión logística en R utilizando el conjunto de datos del cáncer de mama de Wisconsin.

```
# Archivo: regresionlogistica_cancer.R

# Instalación del paquete necesario
install.packages("mlbench")
install.packages("dplyr")

# Carga de los paquetes
```

```
library(mlbench)
library(dplyr)

# Cargar el conjunto de datos BreastCancer
data("BreastCancer")

# Ver las primeras filas del conjunto de datos
head(BreastCancer)

# Preprocesamiento de los datos
# Eliminar la columna de identificación y filas con valores faltantes
breast_cancer_clean <- BreastCancer %>%
  select(-Id) %>%
  na.omit()

# Convertir la variable 'Class' a factor binario
breast_cancer_clean$Class <- ifelse(breast_cancer_clean$Class == "malignant", 1, 0)
breast_cancer_clean$Class <- as.factor(breast_cancer_clean$Class)

# Convertir las demás columnas a numéricas
breast_cancer_clean[, 1:9] <- lapply(breast_cancer_clean[, 1:9], as.numeric)

# Dividir el conjunto de datos en entrenamiento (70%) y prueba (30%)
set.seed(123)
train_indices <- sample(1:nrow(breast_cancer_clean), size = 0.7 * nrow(breast_cancer_clean))
train_set <- breast_cancer_clean[train_indices, ]
test_set <- breast_cancer_clean[-train_indices, ]

# Ajuste del modelo de regresión logística en el conjunto de entrenamiento
model <- glm(Class ~ ., data = train_set, family = binomial)

# Resumen del modelo
summary(model)

# Guardar el modelo y los resultados en un archivo
save(model, file = "regresion_logistica_cancer_modelo.RData")

# Guardar los datos simulados en archivos CSV
write.csv(train_set, "datos_entrenamiento_cancer.csv", row.names = FALSE)
write.csv(test_set, "datos_prueba_cancer.csv", row.names = FALSE)

# Hacer predicciones en el conjunto de prueba
test_set$prob_pred <- predict(model, newdata = test_set, type = "response")
test_set$Class_pred <- ifelse(test_set$prob_pred > 0.5, 1, 0)

# Calcular la precisión de las predicciones
accuracy <- mean(test_set$Class_pred == test_set$Class)
cat("La precisión del modelo en el conjunto de prueba es:", accuracy, "\n")

# Guardar las predicciones en un archivo CSV
write.csv(test_set, "predicciones_cancer.csv", row.names = FALSE)

# Graficar los coeficientes estimados
plot(coef(model), main = "Coeficientes Estimados del Modelo de Regresión Logística",
     xlab = "Variables", ylab = "Coeficientes", type = "h", col = "blue")
abline(h = 0, col = "red", lwd = 2)
```

```
# Mostrar un mensaje indicando que el proceso ha finalizado
cat("El modelo de regresión logística se ha ajustado, se han hecho predicciones y los resultados se h
```

Descripción del Código

Instalación y Carga de Paquetes:
Instalamos y cargamos el paquete `stats` necesario para la regresión logística.
Generación de Datos Simulados:

- Fijamos una semilla para la reproducibilidad.
- Generamos un conjunto de datos con 100 observaciones y 15 variables independientes (`X1, X2, ..., X15`) usando una distribución normal.
- Definimos los coeficientes verdaderos para las variables independientes y calculamos el término lineal.
- Calculamos las probabilidades usando la función logística y generamos una variable dependiente binaria `Y` basada en esas probabilidades.
- Combinamos las variables independientes y la variable dependiente en un `data frame`.

División de Datos en Conjuntos de Entrenamiento y Prueba:

- Dividimos los datos en un conjunto de entrenamiento (70 %) y un conjunto de prueba (30 %).

Ajuste del Modelo de Regresión Logística:

- Ajustamos un modelo de regresión logística en el conjunto de entrenamiento.
- Mostramos un resumen del modelo ajustado.

Guardado de Datos y Modelo:

- Guardamos el modelo ajustado en un archivo `.RData`.
- Guardamos los conjuntos de datos de entrenamiento y prueba en archivos CSV.

Predicciones y Evaluación del Modelo:

- Hacemos predicciones en el conjunto de prueba utilizando el modelo ajustado.
- Calculamos la precisión de las predicciones comparando con los valores reales de `Y`.
- Guardamos las predicciones y las probabilidades predichas en un archivo CSV.

Visualización de los Coeficientes del Modelo:

- Graficamos los coeficientes estimados del modelo ajustado.
- Mostramos un mensaje indicando que el proceso ha finalizado.

Para ejecutar este script, guarda el código en un archivo llamado *regresionlogistica.R*, abre R o RStudio, navega hasta el directorio donde guardaste el archivo y ejecuta el script usando *source(regresionlogistica.R")*.

Ejemplo Titanic

Cuando realizas una regresión logística, obtienes coeficientes para cada variable independiente en tu modelo. Estos coeficientes indican la dirección y la magnitud de la relación entre cada variable independiente y la variable dependiente (en este caso, *Survived*).

Interpretación de los Coeficientes

- **Intercepto** (*(Intercept)*): Este coeficiente representa el logaritmo de las probabilidades (log-odds) de que *Survived* sea 1 (supervivencia) cuando todas las variables independientes son cero.
- **Pclass**: El coeficiente asociado con *Pclass* indica cómo cambia el log-odds de supervivencia con cada incremento en la clase del pasajero. Si el coeficiente es negativo, sugiere que una clase más alta (por ejemplo, de primera clase a tercera clase) reduce las probabilidades de supervivencia.
- **Sex**: Este coeficiente muestra el efecto de ser hombre o mujer en las probabilidades de supervivencia. Generalmente, se espera que el coeficiente sea positivo para *female* indicando que las mujeres tenían mayores probabilidades de sobrevivir.
- **Age**: El coeficiente de *Age* indica cómo cambia el log-odds de supervivencia con cada año de incremento en la edad. Un coeficiente negativo sugiere que la probabilidad de supervivencia disminuye con la edad.
- **SibSp** y **Parch**: Estos coeficientes indican el efecto del número de hermanos/cónyuges a bordo y padres/hijos a bordo en las probabilidades de supervivencia.
- **Fare**: Este coeficiente indica el efecto del precio del billete en las probabilidades de supervivencia. Un coeficiente positivo sugiere que pagar más por el billete se asocia con mayores probabilidades de supervivencia.

Estadísticas de Ajuste del Modelo

El resumen del modelo (*summary(model)*) incluye varias estadísticas importantes:

- **Estadísticos z y p-valores**: Estas estadísticas indican la significancia de cada coeficiente. Un p-valor bajo (generalmente ¡0.05) sugiere que la variable es un predictor significativo de la variable dependiente.
- **Desviación Residual**: La desviación residual mide la calidad del ajuste del modelo. Valores más bajos indican un mejor ajuste.
- **AIC (Akaike Information Criterion)**: El AIC es una medida de la calidad del modelo que toma en cuenta tanto la bondad del ajuste como la complejidad del modelo. Modelos con AIC más bajo son preferidos.

Precisión del Modelo

La precisión del modelo en el conjunto de prueba es una métrica importante para evaluar el rendimiento del modelo. La precisión se calcula como el número de predicciones correctas dividido por el número total de predicciones.

Ejemplo de Resultados

Supongamos que la precisión del modelo es 0.78 (78 %). Esto significa que el modelo correctamente predijo el estado de supervivencia del 78 % de los pasajeros en el conjunto de prueba.

Matriz de Confusión y Otras Métricas

Además de la precisión, otras métricas como la matriz de confusión, la sensibilidad, la especificidad, y el área bajo la curva ROC (AUC-ROC) también pueden proporcionar una visión más completa del rendimiento del modelo.

Matriz de Confusión

- **Verdaderos Positivos (TP)**: Número de pasajeros que sobrevivieron y fueron predichos como sobrevivientes.
- **Verdaderos Negativos (TN)**: Número de pasajeros que no sobrevivieron y fueron predichos como no sobrevivientes.
- **Falsos Positivos (FP)**: Número de pasajeros que no sobrevivieron pero fueron predichos como sobrevivientes.
- **Falsos Negativos (FN)**: Número de pasajeros que sobrevivieron pero fueron predichos como no sobrevivientes.

Ejemplo de Cálculo de Métricas

```
# Calcular la matriz de confusión
table(test_set$Survived, test_set$Survived_pred)

# Calcular sensibilidad y especificidad
sensitivity <- sum(test_set$Survived == 1 & test_set$Survived_pred == 1) / sum(test_set$Survived == 
specificity <- sum(test_set$Survived == 0 & test_set$Survived_pred == 0) / sum(test_set$Survived == 

# Calcular AUC-ROC
library(pROC)
roc_curve <- roc(test_set$Survived, test_set$prob_pred)
auc(roc_curve)
```

Visualización de Resultados

Graficar los coeficientes del modelo, la curva ROC y otras visualizaciones ayudan a entender mejor el rendimiento y la importancia de cada variable en el modelo.

```
# Graficar la curva ROC
plot(roc_curve, main = "Curva ROC para el Modelo de Regresión Logística")
```

Resumen Final

El modelo de regresión logística aplicado al conjunto de datos del Titanic proporciona una forma de entender cómo diferentes características de los pasajeros influyen en sus probabilidades de supervivencia. La interpretación de los coeficientes del modelo, las estadísticas de ajuste, y la precisión del modelo en el conjunto de prueba son fundamentales para evaluar el rendimiento y la utilidad del modelo en hacer predicciones sobre la supervivencia de los pasajeros del Titanic.

21.1.3. Simulación de Datos de Cáncer - Parte II

Aquí se presenta un ejemplo de cómo realizar una regresión logística utilizando datos simulados de pacientes con cáncer.

```
#---- Archivo: cancerLogRegSimulado.R ----

# Instalación del paquete necesario
if (!requireNamespace("dplyr", quietly = TRUE)) {
  install.packages("dplyr")
}

# Carga del paquete
library(dplyr)
```

```
# Fijar la semilla para reproducibilidad
set.seed(123)

# Número de observaciones
n <- 150

# Generar las variables independientes X1, X2, ..., X15
# Creamos una matriz de tamaño n x 15 con valores generados aleatoriamente de una
distribución normal
X <- as.data.frame(matrix(rnorm(n * 15), nrow = n, ncol = 15))
colnames(X) <- paste0("X", 1:15)  # Nombramos las columnas como X1, X2, ..., X15

# Coeficientes verdaderos para las variables independientes
# Generamos un vector de 16 coeficientes (incluyendo el intercepto) aleatorios entre -1 y 1
beta <- runif(16, -1, 1)  # 15 coeficientes más el intercepto

# Generar el término lineal
# Calculamos el término lineal utilizando los coeficientes y las variables independientes
linear_term <- beta[1] + as.matrix(X) %*% beta[-1]

# Generar la probabilidad utilizando la función logística
# Calculamos las probabilidades utilizando la función logística
p <- 1 / (1 + exp(-linear_term))

# Generar la variable dependiente binaria Y
# Generamos valores binarios (0 o 1) utilizando las probabilidades calculadas
Y <- rbinom(n, 1, p)

# Combinar las variables independientes y la variable dependiente en un data frame
data <- cbind(Y, X)

# Dividir el conjunto de datos en entrenamiento y prueba
set.seed(123)  # Fijar la semilla para reproducibilidad
train_indices <- sample(1:n, size = 0.7 * n)  # 70% de los datos para entrenamiento
train_set <- data[train_indices, ]  # Conjunto de entrenamiento
test_set <- data[-train_indices, ]  # Conjunto de prueba

# Ajuste del modelo de regresión logística en el conjunto de entrenamiento
# Ajustamos un modelo de regresión logística utilizando las variables independientes
para predecir Y
model <- glm(Y ~ ., data = train_set, family = binomial)

# Resumen del modelo
# Mostramos un resumen del modelo ajustado
summary(model)

# Guardar el modelo y los resultados en un archivo
# Guardamos el modelo ajustado en un archivo .RData
save(model, file = "regresion_logistica_cancer_modelo_simulado.RData")

# Guardar los datos simulados en archivos CSV
# Guardamos los conjuntos de datos de entrenamiento y prueba en archivos CSV
write.csv(train_set, "datos_entrenamiento_cancer_simulado.csv", row.names = FALSE)
write.csv(test_set, "datos_prueba_cancer_simulado.csv", row.names = FALSE)

# Hacer predicciones en el conjunto de prueba
```

```
# Utilizamos el modelo ajustado para hacer predicciones en el conjunto de prueba
test_set$prob_pred <- predict(model, newdata = test_set, type = "response")
test_set$Y_pred <- ifelse(test_set$prob_pred > 0.5, 1, 0)
# Convertimos probabilidades a clases binarias

# Calcular la precisión de las predicciones
# Calculamos la precisión de las predicciones comparando con los valores reales de Y
accuracy <- mean(test_set$Y_pred == test_set$Y)
cat("La precisión del modelo en el conjunto de prueba es:", accuracy, "\n")

# Guardar las predicciones en un archivo CSV
# Guardamos las predicciones y las probabilidades predichas en un archivo CSV
write.csv(test_set, "predicciones_cancer_simulado.csv", row.names = FALSE)

# Graficar los coeficientes estimados
# Graficamos los coeficientes estimados del modelo ajustado
plot(coef(model), main = "Coeficientes Estimados del Modelo de Regresión Logística",
     xlab = "Variables", ylab = "Coeficientes", type = "h", col = "blue")
abline(h = 0, col = "red", lwd = 2)

# Mostrar un mensaje indicando que el proceso ha finalizado
cat("El modelo de regresión logística se ha ajustado, se han hecho predicciones
y los resultados se han guardado en 'regresion_logistica_cancer_modelo_simulado.RData'.\n")
```

21.1.4. Simulación de Datos de Cáncer - Parte III

En un estudio sobre cáncer, especialmente en el contexto del cáncer de mama, las principales mediciones suelen incluir una variedad de características clínicas y patológicas. Aquí hay algunas de las principales mediciones que se tienen en cuenta:

- **Tamaño del Tumor**: Medición del diámetro del tumor.
- **Estado de los Ganglios Linfáticos**: Número de ganglios linfáticos afectados.
- **Grado del Tumor**: Clasificación del tumor basada en la apariencia de las células cancerosas.
- **Receptores Hormonales**: Estado de los receptores de estrógeno y progesterona.
- **Estado HER2**: Expresión del receptor 2 del factor de crecimiento epidérmico humano.
- **Ki-67**: Índice de proliferación celular.
- **Edad del Paciente**: Edad en el momento del diagnóstico.
- **Histopatología**: Tipo y subtipo histológico del cáncer.
- **Márgenes Quirúrgicos**: Estado de los márgenes después de la cirugía (si están libres de cáncer o no).
- **Invasión Linfovascular**: Presencia de células cancerosas en los vasos linfáticos o sanguíneos.
- **Tratamientos Previos**: Tipos de tratamientos recibidos antes del diagnóstico (quimioterapia, radioterapia, etc.).
- **Tipo de Cirugía**: Tipo de procedimiento quirúrgico realizado (mastectomía, lumpectomía, etc.).
- **Metástasis**: Presencia de metástasis y ubicación de las mismas.
- **Índice de Masa Corporal (IMC)**: Relación entre el peso y la altura del paciente.
- **Marcadores Genéticos**: Presencia de mutaciones genéticas específicas (BRCA1, BRCA2, etc.).

Estas mediciones proporcionan una visión integral del estado del cáncer y se utilizan para planificar el tratamiento y predecir el pronóstico.

A continuación, se muestra un ejemplo de cómo ajustar un modelo de regresión logística en R utilizando un conjunto de datos simulado con estas mediciones.

```
# Archivo: simulcorrectedCancer.R

# Instalación del paquete necesario
if (!requireNamespace("dplyr", quietly = TRUE)) {
  install.packages("dplyr")
}

# Carga del paquete
library(dplyr)

# Fijar la semilla para reproducibilidad
set.seed(123)

# Número de observaciones
n <- 1500

# Simulación de las variables independientes
# Tamaño del Tumor (en cm)
Tumor_Size <- rnorm(n, mean = 3, sd = 1.5)

# Estado de los Ganglios Linfáticos (número de ganglios afectados)
Lymph_Nodes <- rpois(n, lambda = 3)

# Grado del Tumor (1 a 3)
Tumor_Grade <- sample(1:3, n, replace = TRUE)

# Receptores Hormonales (0: negativo, 1: positivo)
Estrogen_Receptor <- rbinom(n, 1, 0.7)
Progesterone_Receptor <- rbinom(n, 1, 0.7)

# Estado HER2 (0: negativo, 1: positivo)
HER2_Status <- rbinom(n, 1, 0.3)

# Ki-67 (% de células proliferativas)
Ki_67 <- rnorm(n, mean = 20, sd = 10)

# Edad del Paciente (años)
Age <- rnorm(n, mean = 50, sd = 10)

# Histopatología (1: ductal, 2: lobular, 3: otros)
Histopathology <- sample(1:3, n, replace = TRUE)

# Márgenes Quirúrgicos (0: positivo, 1: negativo)
Surgical_Margins <- rbinom(n, 1, 0.8)

# Invasión Linfovascular (0: no, 1: sí)
Lymphovascular_Invasion <- rbinom(n, 1, 0.4)

# Tratamientos Previos (0: no, 1: sí)
Prior_Treatments <- rbinom(n, 1, 0.5)
```

```
# Tipo de Cirugía (0: mastectomía, 1: lumpectomía)
Surgery_Type <- rbinom(n, 1, 0.5)

# Metástasis (0: no, 1: sí)
Metastasis <- rbinom(n, 1, 0.2)

# Índice de Masa Corporal (IMC)
BMI <- rnorm(n, mean = 25, sd = 5)

# Marcadores Genéticos (0: negativo, 1: positivo)
Genetic_Markers <- rbinom(n, 1, 0.1)

# Generar la variable dependiente binaria Y (sobrevivencia 0: no, 1: sí)
# Utilizaremos una combinación arbitraria de las variables para generar Y
linear_term <- -1 + 0.5 * Tumor_Size - 0.3 * Lymph_Nodes + 0.2 * Tumor_Grade +
  0.4 * Estrogen_Receptor + 0.3 * Progesterone_Receptor - 0.2 * HER2_Status +
  0.1 * Ki_67 - 0.05 * Age + 0.3 * Surgical_Margins - 0.4 * Lymphovascular_Invasion +
  0.2 * Prior_Treatments + 0.1 * Surgery_Type - 0.5 * Metastasis + 0.01 * BMI +
  0.2 * Genetic_Markers
p <- 1 / (1 + exp(-linear_term))
Y <- rbinom(n, 1, p)

# Combinar las variables independientes y la variable dependiente en un data frame
data <- data.frame(Y, Tumor_Size, Lymph_Nodes, Tumor_Grade, Estrogen_Receptor,
                   Progesterone_Receptor, HER2_Status, Ki_67, Age, Histopathology,
                   Surgical_Margins, Lymphovascular_Invasion, Prior_Treatments,
                   Surgery_Type, Metastasis, BMI, Genetic_Markers)

# Dividir el conjunto de datos en entrenamiento y prueba
set.seed(123)  # Fijar la semilla para reproducibilidad
train_indices <- sample(1:n, size = 0.7 * n)  # 70% de los datos para entrenamiento
train_set <- data[train_indices, ]  # Conjunto de entrenamiento
test_set <- data[-train_indices, ]  # Conjunto de prueba

# Ajuste del modelo de regresión logística en el conjunto de entrenamiento
# Ajustamos un modelo de regresión logística utilizando las variables independientes para
 predecir Y
model <- glm(Y ~ ., data = train_set, family = binomial)

# Resumen del modelo
# Mostramos un resumen del modelo ajustado
summary(model)

# Guardar el modelo y los resultados en un archivo
# Guardamos el modelo ajustado en un archivo .RData
save(model, file = "regresion_logistica_cancer_modelo_simulado.RData")

# Guardar los datos simulados en archivos CSV
# Guardamos los conjuntos de datos de entrenamiento y prueba en archivos CSV
write.csv(train_set, "datos_entrenamiento_cancer_simulado.csv", row.names = FALSE)
write.csv(test_set, "datos_prueba_cancer_simulado.csv", row.names = FALSE)

# Hacer predicciones en el conjunto de prueba
# Utilizamos el modelo ajustado para hacer predicciones en el conjunto de prueba
test_set$prob_pred <- predict(model, newdata = test_set, type = "response")
test_set$Y_pred <- ifelse(test_set$prob_pred > 0.5, 1, 0)
```

```
# Convertimos probabilidades a clases binarias

# Calcular la precisión de las predicciones
# Calculamos la precisión de las predicciones comparando con los valores reales de Y
accuracy <- mean(test_set$Y_pred == test_set$Y)
cat("La precisión del modelo en el conjunto de prueba es:", accuracy, "\n")

# Guardar las predicciones en un archivo CSV
# Guardamos las predicciones y las probabilidades predichas en un archivo CSV
write.csv(test_set, "predicciones_cancer_simulado.csv", row.names = FALSE)

# Graficar los coeficientes estimados
# Graficamos los coeficientes estimados del modelo ajustado
plot(coef(model), main = "Coeficientes Estimados del Modelo de Regresión Logística",
     xlab = "Variables", ylab = "Coeficientes", type = "h", col = "blue")
abline(h = 0, col = "red", lwd = 2)

# Mostrar un mensaje indicando que el proceso ha finalizado
cat("El modelo de regresión logística se ha ajustado, se han hecho predicciones
y los resultados se han guardado en 'regresion_logistica_cancer_modelo_simulado.RData'.\n")
```

CAPÍTULO 22

Bibliografía

Bibliografía

[1] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

[2] Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

[3] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[4] Harrell, F. E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer.

[5] R Documentation and Tutorials: `https://cran.r-project.org/manuals.html`

[6] Tutorials on R-bloggers: `https://www.r-bloggers.com/`

[7] Coursera: *Machine Learning* by Andrew Ng.

[8] edX: *Data Science and Machine Learning Essentials* by Microsoft.

[9] Ross, S. M. (2014). *Introduction to Probability and Statistics for Engineers and Scientists*. Academic Press.

[10] DeGroot, M. H., and Schervish, M. J. (2012). *Probability and Statistics* (4th ed.). Pearson.

[11] Hogg, R. V., McKean, J., and Craig, A. T. (2019). *Introduction to Mathematical Statistics* (8th ed.). Pearson.

[12] Kleinbaum, D. G., and Klein, M. (2010). *Logistic Regression: A Self-Learning Text* (3rd ed.). Springer.

[13] Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer.

[14] Probability and Statistics Tutorials on Khan Academy: `https://www.khanacademy.org/math/statistics-probability`

[15] Online Statistics Education: `http://onlinestatbook.com/`

[16] Peng, C. Y. J., Lee, K. L., and Ingersoll, G. M. (2002). *An Introduction to Logistic Regression Analysis and Reporting*. The Journal of Educational Research.

[17] Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (2nd ed.). Wiley.

[18] Han, J., Pei, J., and Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

[19] Data Cleaning and Preprocessing on Towards Data Science: `https://towardsdatascience.com/data-cleaning-and-preprocessing`

[20] Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005). *Prediction error estimation: a comparison of resampling methods*. Bioinformatics.

[21] Evaluating Machine Learning Models on Towards Data Science: `https://towardsdatascience.com/evaluating-machine-learning-models`

[22] Practical Guide to Logistic Regression in R on Towards Data Science: `https://towardsdatascience.com/practical-guide-to-logistic-regression-in-r`

[23] Coursera: *Statistics with R* by Duke University.

[24] edX: *Data Science: Probability* by Harvard University.

[25] Coursera: *Logistic Regression* by Stanford University.

[26] edX: *Data Science: Inference and Modeling* by Harvard University.

[27] Coursera: *Data Science: Wrangling and Cleaning* by Johns Hopkins University.

[28] edX: *Data Science: R Basics* by Harvard University.

[29] Coursera: *Regression Models* by Johns Hopkins University.

[30] edX: *Data Science: Statistical Inference* by Harvard University.

[31] An Introduction to Survival Analysis on Towards Data Science: `https://towardsdatascience.com/an-introduction-to-survival-analysis`

[32] Multinomial Logistic Regression on DataCamp: `https://www.datacamp.com/community/tutorials/multinomial-logistic-regression-R`

[33] Coursera: *Survival Analysis* by Johns Hopkins University.

[34] edX: *Data Science: Statistical Inference and Modeling for High-throughput Experiments* by Harvard University.