Modelos de Regresión Aplicados a Datos Reales: Lineal, Logística y Logística Multivariada

Carlos Ernesto

4 de abril de 2025

Índice

1. Introducción

En este documento se exploran distintos modelos de regresión aplicados a datos reales en dos áreas clave: ciencias médicas y ciencias sociales. Se abordan modelos lineales, logísticos y logísticos multivariados utilizando herramientas estadísticas implementadas en el lenguaje de programación R.

2. Marco Teórico

2.1. Regresión Lineal

Explicación de la regresión lineal simple, ecuación del modelo, supuestos y estimación por mínimos cuadrados.

2.2. Regresión Logística

En la regresión logística, las variables predictoras influyen sobre los logaritmos de los odds (probabilidades relativas) del resultado mediante una función de enlace llamada **logit**. La combinación lineal de predictores genera los log-odds, los cuales se transforman nuevamente en probabilidades utilizando la función logística:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$$

Esta relación es no lineal en términos de probabilidad, pero lineal en términos de log-odds. El objetivo del modelo logístico es estimar los parámetros β utilizando el método de máxima verosimilitud (MLE), el cual busca el conjunto de coeficientes que maximiza la probabilidad de observar los datos disponibles.

Los coeficientes de regresión reflejan la fuerza y dirección de la relación entre cada variable independiente y el resultado. Específicamente, indican cuánto cambian los odds del resultado ante un cambio de una unidad en la variable predictora correspondiente.

El modelo predice la categoría del resultado calculando la razón de probabilidades de éxito frente a fracaso, lo cual se expresa comúnmente como una razón de momios u **odds ratio (OR)**. Este valor mide la probabilidad del resultado dado una exposición, comparado con la probabilidad del mismo resultado sin la exposición. Este enfoque es ampliamente utilizado en estudios de casos y controles, así como en estudios transversales y de cohorte.

Estandarización (Z-score)

$$\mu_i = \frac{1}{m}\sum_{j=1}^{m} x_i^{(j)}, \quad \sigma_i = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(x_i^{(j)} - \mu_i\right)^2}$$

$$x_i' = \frac{x_i - \mu_i}{\sigma_i}$$

Normalización

$$x_i' = \frac{x_i - \text{mín}(x_i)}{\text{máx}(x_i) - \text{mín}(x_i)}$$

Evaluación de la adecuación del modelo

La evaluación del ajuste del modelo (goodness-of-fit) permite determinar qué tan bien el modelo refleja los resultados observados. Se pueden utilizar la estadística de Hosmer–Lemeshow, la prueba de Pearson, la desviación y la razón de verosimilitud.

Para clasificación binaria, se utiliza el área bajo la curva ROC (AUC). Este valor mide la tasa de verdaderos positivos frente a falsos positivos para distintos umbrales. Un AUC de 0.5 implica discriminación nula; un AUC de 1.0 indica discriminación perfecta.

La validación del modelo también es crucial, pues establece la credibilidad del modelo ajustado.

3. Metodología

3.1. Fuentes de datos

Descripción de las bases utilizadas: cáncer, enfermedad renal, divorcio y redes sociales.

3.2. Tratamiento y limpieza de datos en R

Explicación del preprocesamiento realizado, valores faltantes, variables transformadas.

3.3. Ajuste de modelos y validación

Técnicas para seleccionar variables, ajuste de modelos y evaluación con métricas.

4. Resultados

4.1. Caso 1: Cáncer de mama (Regresión Logística)

4.2. Caso 2: Enfermedad renal crónica (Logística multivariada)

4.3. Caso 3: Divorcio vs edad del hijo (Regresión logística)

4.4. Caso 4: Redes sociales y estado civil (Modelado exploratorio)

5. Discusión

Comparación entre modelos, interpretabilidad, utilidad práctica en cada área.

6. Conclusiones

Principales hallazgos, limitaciones del estudio y sugerencias para investigaciones futuras.

Referencias