Universidad Nacional De Colombia

Sede Medellín

Facultad de Ciencias

Departamento de Estadística

Taller 4

Daniel Felipe Villa Rengifo

Luis David Hernández Pérez

Juan Gabriel Carvajal Negrete

Modelos de Regresión

Febrero, 2025

Se hará el análisis de la base de datos **Credit** de la librería **ISLR** de R. Se tendrá como variable de respuesta `Married` y como variable explicativa `Rating`. Debe codificar las variables de respuesta con 1 para `Yes` y 0 para `No`. Las observaciones a ser incluidas para el análisis en cada grupo aparecen en la Tabla 1. Realice las siguientes actividades:

Punto 1

Realice una descripción de la base de datos. Contextualice el problema y explique cada una de las variables involucradas en el modelo.

Solución Punto1

Descripción de los datos: Es un conjunto de datos simulados que contiene información sobre diez mil clientes. Considerando todas las variables el objetivo es predecir qué clientes dejarán de pagar su tarjeta de crédito, pero para este caso el objetivo es establecer un modelo que permita determinar como el rating de crédito de la persona influye en la probabilidad de que esa persona este casada o no.

Descripción de las variables

- **Married**: Variable categórica que indica si la persona está casada (Yes/No).
- **Rating**: Variable numérica que representa el rating de crédito de la persona.

Punto 2

Realice un análisis descriptivo de las variables que se van a tener en cuenta en el modelo. Concluya.

Solución Punto 2

Tabla 1: Resumen Numérico Variable Rating

Min	Mean	Var	Max
122	349.21	18066.23	828

La variable **Rating** presenta una media alrededor de 350 y valores extremos que podrían estar influyendo en la variabilidad observada.

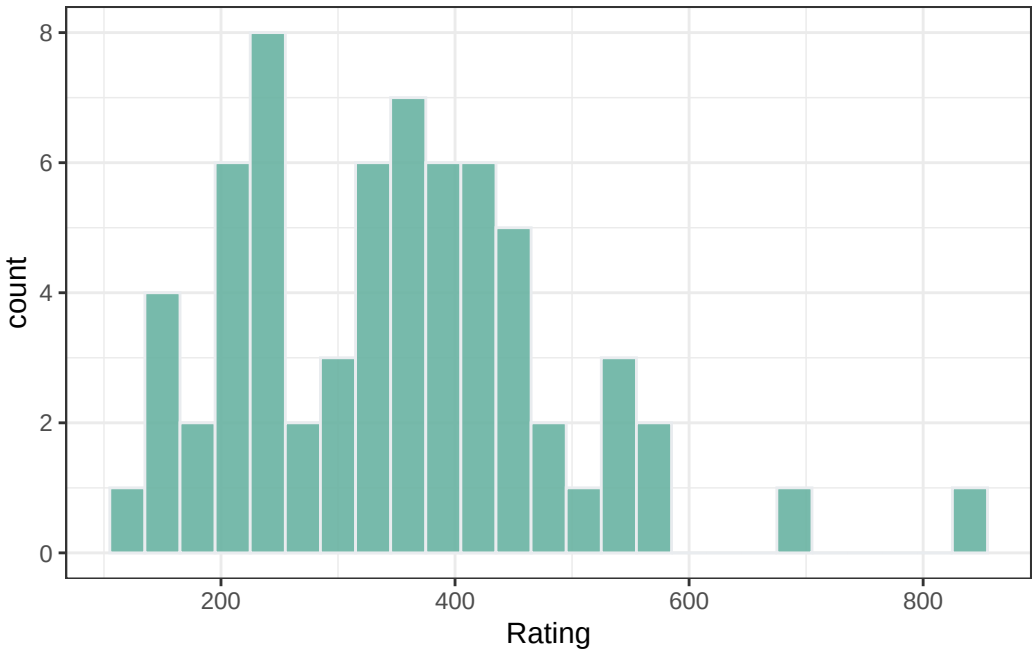

Figura 1: Histograma del Rating

De la Figura 1 podemos decir que a distribución parece ser asimétrica hacia la derecha , ya que hay algunas personas con valores de rating altos, la mayoría de personas se concentran en valores de rating entre 150 y 450.

Tabla 2: Distribución de la variable Married en la muestra analizada

Married	Frecuencia	Porcentaje
Yes	42	63.64 %
No	24	36.36 %
Total	66	100

Al analizar la tabla Tabla 2, se observa que la mayoría de las personas en el conjunto de datos están casadas. En concreto, hay **42 personas casadas**, lo que contrasta con las **24 personas que no lo están**. Esto indica que, en la muestra analizada, el estado civil predominante es el de casado.

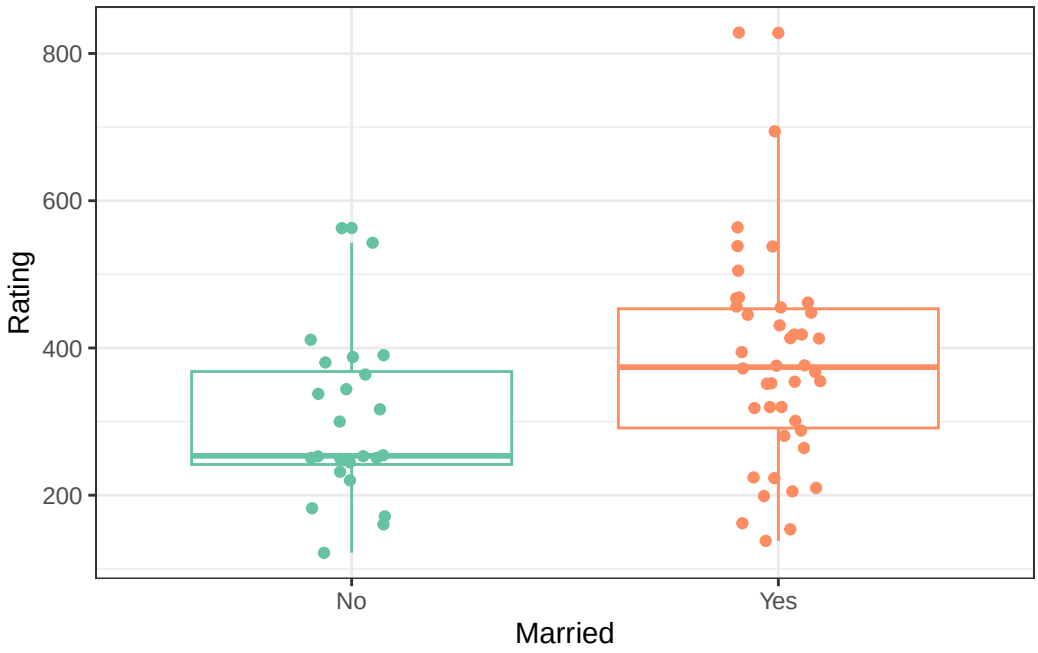

Figura 2: Boxplot Rating vs Married

Del Figura 2 observamos que las personas que estan casadas tienen mayor variabilidad con respecto al Rating, parece que no existe diferencia en el rating entre las personas que estan casadas o no.

Punto 3

Ajuste un modelo de regresión logística, muestre la tabla de parámetros ajustados y escriba la ecuación ajustada.

Solución Punto 3

Tabla 3: Coeficientes ajustados

Coeficiente	Estimación	Error estándar	valor z	$\Pr(>\|z\|)$
(Intercepto)	-1.208379	0.812019	-1.488	0.1367
Rating	0.005268	0.002372	2.221	0.0264

El modelo ajustado es

$$\text{logit}(\text{married}) = -1.208379 + 0.005268 * \text{rating}$$
$$\text{p}(\text{married}) = \frac{\exp(-1.208379 + 0.005268 * \text{rating})}{1 + \exp(-1.208379 + 0.005268 * \text{rating})}$$

$$\hat{Y} = \frac{e^{-1.208379+0.005268 \cdot \text{Rating}}}{1 + e^{-1.208379+0.005268 \cdot \text{Rating}}}$$

Punto 4

Pruebe la significancia individual del parámetro que acompaña a la variable explicativa e interprete el valor de la estimación. También interprete la razón de odds.

Solución Punto 4

Para la significancia de β_1 tenemos la siguiente hipótesis:

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

donde le estadístico de prueba es:

$$Z_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

Observando la Tabla 3 vemos que el parámetro que acompaña a la variable rating es significativo con un p-valor= $0.026 < \alpha = 0.05$

De acuerdo a los resultados de la Tabla 3, el logaritmo de los odds de que una persona este casado está positivamente relacionado con la puntuación del Rating $\hat{\beta}_1 = 0.005268$.Esto significa que, por cada unidad que se incrementa la variable rating, se espera que el logaritmo de odds de la variable married se incremente en promedio 0.005268 unidades. Aplicando la inversa del logaritmo natural ($\exp(0.005268) = 1.005282$) se obtiene que, por cada unidad que se incrementa la variable rating, los odds de estar casado se incremente en promedio 1.005282 unidades.

Punto 5

Determine si el modelo ajustado es mejor que el modelo nulo.

Solución Punto 5

Para esta verificación utilizaremos la función `Anova()`

Tabla 4: Anova del modelo

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	65	86.52359	NA
Rating	1	5.862324	64	80.66127	0.0154684

De la Tabla 4 podemos concluir que el modelo ajustado (modelo completo) es significativo, por tanto es mejor que el modelo nulo.

Punto 6

Realice un gráfico con los valores ajustados de la probabilidad de éxito, incluya el intervalo de confianza al 95%. Concluya.

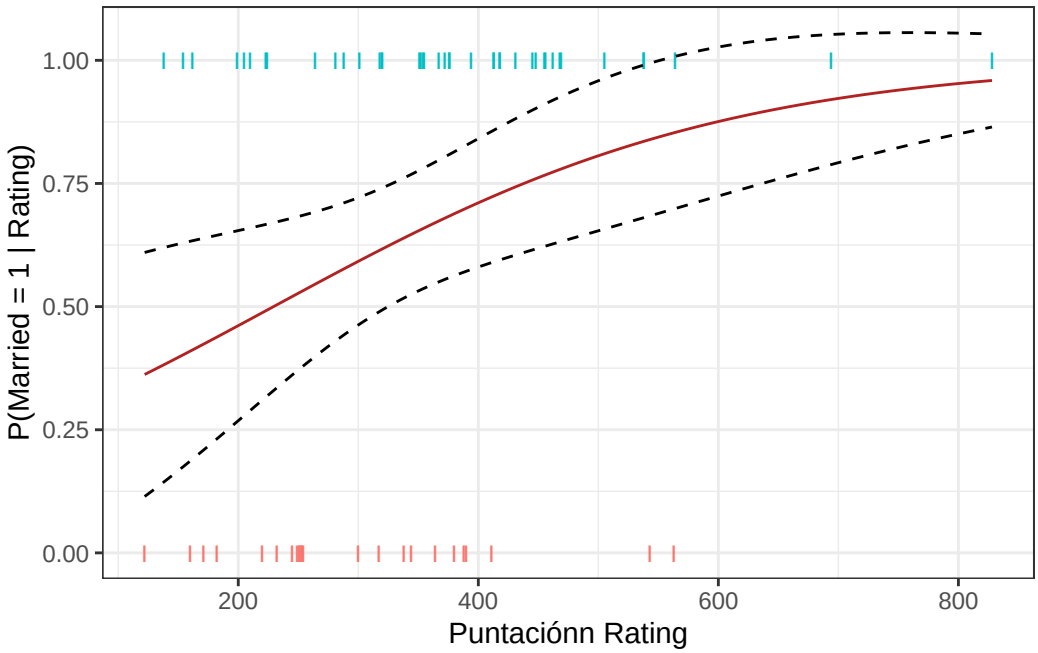

Figura 3: Modelo regresión logística Married ~ Rating

De la Figura 3 observamos que la probabilidad de estar casado aumenta a medida que la puntuación del rating incrementa. Para valores bajos de rating, la probabilidad de éxito es reducida, mientras que para valores altos, la probabilidad se acerca a 1.

Punto 7

Calcule el porcentaje de correcta clasificación del modelo. Comente.

```
             predicciones
observaciones  0  1
            0  5 19
            1  8 34
```

El modelo es capaz de clasificar correcatmente $\frac{5+34}{5+34+8+19} = 0.59(59\%)$ de las observaciones.