Reg. con Y dicotomica

Unidad 2: Estadística Básica y Aplicada

Nicolás Sidicaro
Octubre 2025

Introducción y Motivación

¿Qué queremos predecir?

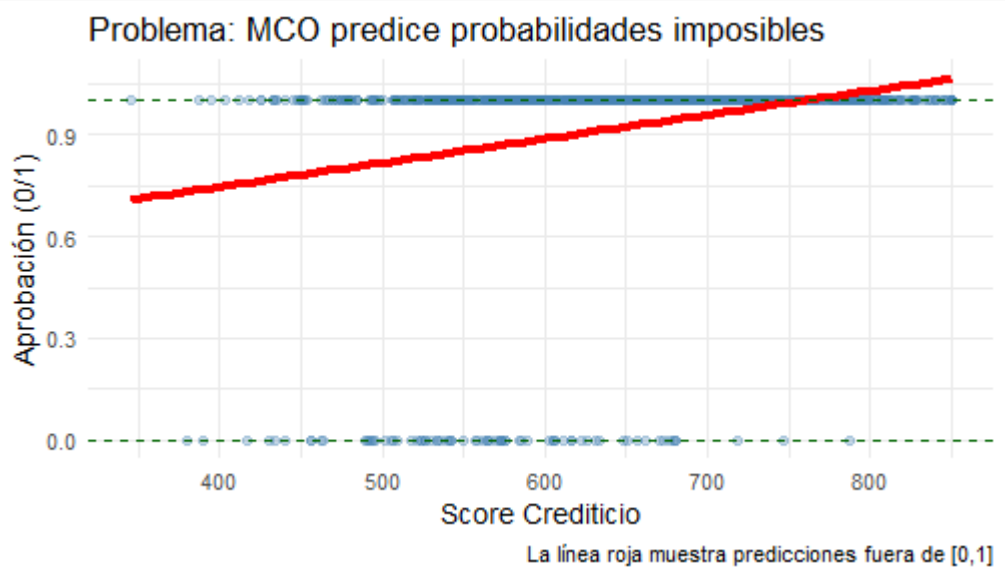
Variable Y que solo toma valores 0 o 1

Ejemplos:

- Default crediticio (sí/no)
- Aprobación de crédito (aprobado/rechazado)
- Compra de producto (compra/no compra)
- Recuperación de paciente (recuperado/no recuperado)
- Participación laboral (trabaja/no trabaja)

Objetivo: Modelar la probabilidad de que Y = 1 dado un conjunto de variables X

El Problema con MCO



No podemos tener 120% de probabilidad de aprobar un crédito

Modelo de Probabilidad Lineal (MPL)

MPL: Idea Básica

Usar MCO con variable dependiente binaria

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

donde $Y_i \in \{0, 1\}$

Interpretación directa y sencilla:

- β_1 = cambio en la probabilidad cuando X_1 aumenta 1 unidad
- $E[Y_i|X_i] = P(Y_i = 1|X_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$

Ventaja principal: Coeficientes = efectos marginales

MPL: Ejemplo Práctico

```
# Estimar MPL
modelo_mpl ← lm(aprobo ~ ingreso + score + edad, data = datos)
summary(modelo_mpl)$coefficients %>%
  round(6) %>%
  kable()
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.233690	0.062265	3.753124	0.000185
ingreso	0.000005	0.000001	9.128993	0.000000
score	0.000644	0.000079	8.158776	0.000000
edad	0.000705	0.000679	1.038645	0.299222

Interpretación: Por cada punto adicional en el score crediticio, la probabilidad de aprobación aumenta 0.8 puntos porcentuales

MPL: Problemas

1. Heteroscedasticidad automática

$$Var(u_i|X_i) = P_i(1 - P_i)$$

- **Consecuencia**: Errores estándar incorrectos → inferencia inválida
- **Solución**: Errores robustos (White/HC)

2. Predicciones imposibles

- Puede predecir $P < 0$ o $P > 1$
- No son probabilidades válidas

```
## Predicciones fuera de rango:

## P < 0: 0 observaciones

## P > 1: 217 observaciones
```

MPL: Solución - Errores Robustos

```r
# MPL con errores robustos
coeftest(modelo_mpl, vcov = vcovHC(modelo_mpl, type = "HC1"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 2.3369e-01 8.5832e-02  2.7227  0.006589 **
## ingreso     4.8248e-06 6.0644e-07  7.9559 4.814e-15 ***
## score       6.4360e-04 9.0122e-05  7.1414 1.780e-12 ***
## edad        7.0476e-04 6.3795e-04  1.1047  0.269547
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nota: Los coeficientes son iguales, pero los errores estándar cambian

MPL: ¿Cuándo Usarlo?

✓ **Aproximación rápida** cuando el tiempo es limitado

✓ **Efectos marginales constantes** son razonables

✓ **Interpretación directa** es prioritaria

✓ **Siempre con errores robustos** para inferencia

✗ Evitar si hay muchas predicciones fuera de [0,1]

✗ No usar si la relación es claramente no lineal

Modelos No Lineales: Logit y Probit

¿Por Qué Modelos No Lineales?

Necesitamos garantizar que $0 \leq P \leq 1$

Solución: Usar una función de distribución que transforme la combinación lineal

$$P(Y_i = 1|X_i) = F(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})$$

donde $F(\cdot)$ es una función de distribución acumulada

Dos opciones principales:

- **Logit**: F es la distribución logística
- **Probit**: F es la distribución normal estándar

Logit: Función Logística

Función de distribución logística:

$$P(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2)}}$$

Características:

- Forma de "S" simétrica
- Siempre entre 0 y 1
- Pendiente máxima en P = 0.5
- Permite interpretar odds-ratios

Probit: Distribución Normal

Función de distribución normal acumulada:

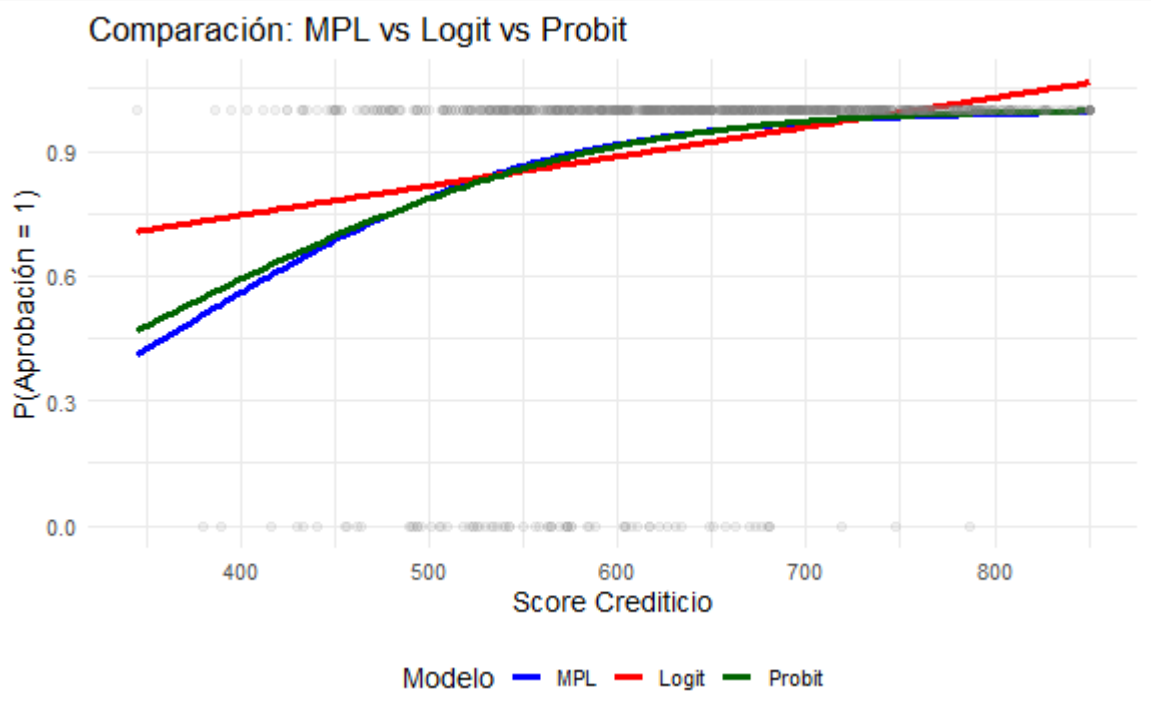
$$P(Y = 1|X) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$$

donde Φ es la función de distribución normal estándar

Características:

- También forma de "S"
- Colas más ligeras que Logit
- Tradición en econometría
- Resultados muy similares a Logit

Comparación Visual



Estimación en R

```
# Logit
modelo_logit ← glm(aprobo ~ ingreso + score + edad,
                   data = datos,
                   family = binomial(link = "logit"))

# Probit
modelo_probit ← glm(aprobo ~ ingreso + score + edad,
                    data = datos,
                    family = binomial(link = "probit"))
```

Variable	Logit	Probit
(Intercept)	-9.7405	-4.8852
ingreso	0.0001	0.0001
score	0.0120	0.0061
edad	0.0196	0.0095

¿Logit o Probit?

Logit:

- ✓ Más común en la práctica
- ✓ Interpretación vía odds-ratios
- ✓ Colas más pesadas (captura eventos extremos)

Probit:

- ✓ Tradición econométrica
- ✓ Útil si asumimos normalidad subyacente
- ✓ Ligeramente más fácil en modelos multivariados

En la práctica: Los resultados son muy similares. La elección depende más de convención del campo que de consideraciones técnicas.

Interpretación de Resultados

⚠ Problema: Los Coeficientes NO son Efectos Marginales

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.74052	1.28486	-7.58101	0.00000
ingreso	0.00010	0.00001	8.21072	0.00000
score	0.01197	0.00162	7.38689	0.00000
edad	0.01962	0.01299	1.50988	0.13107

Solo podemos interpretar:

- ✓ **Signo**: positivo → aumenta P(Y=1); negativo → disminuye P(Y=1)
- ✓ **Significancia**: el coeficiente es estadísticamente distinto de cero

NO interpretar la magnitud directamente

Solución 1: Efectos Marginales (RECOMENDADO)

Efecto marginal: Cambio en la probabilidad cuando X aumenta en 1 unidad

$$\frac{\partial P(Y=1|X)}{\partial X_j} = f(\beta_0 + \beta_1 X_1 + \dots) \cdot \beta_j$$

donde $f(\cdot)$ es la función de densidad (derivada de F)

Dos formas de calcular:

1. Evaluar en la media de X (at means)
2. Calcular para cada observación y promediar (average marginal effects) ← **Mejor**

Efectos Marginales: Ejemplo

```
# Calcular efectos marginales promedio
efectos_logit ← margins(modelo_logit)
summary(efectos_logit) %>%
  select(factor, AME, SE, p) %>%
  kable(digits = 5, col.names = c("Variable", "Efecto Marginal",
                                  "Error Est.", "p-value"))
```

Variable	Efecto Marginal	Error Est.	p-value
edad	0.00103	0.00068	0.13093
ingreso	0.00001	0.00000	0.00000
score	0.00063	0.00008	0.00000

Interpretación práctica:

- Un punto adicional en el score crediticio aumenta la probabilidad de aprobación en **0.8 puntos porcentuales**
- $1,000 adicionales de ingreso aumentan la probabilidad en **0.6 puntos porcentuales**

Solución 2: Odds-Ratios (Solo Logit)

Odds (chances): $\frac{P(Y=1)}{P(Y=0)}$ = "chances de que ocurra vs no ocurra"

Odds-Ratio: $OR = e^{\beta_j}$

```
# Calcular odds-ratios
odds_ratios ← exp(coef(modelo_logit))
tibble(
  Variable = names(odds_ratios),
  `Odds-Ratio` = round(odds_ratios, 4)
) %>% kable()
```

Variable	Odds-Ratio
(Intercept)	0.0001
ingreso	1.0001
score	1.0120
edad	1.0198

Interpretación: Un punto adicional en el score multiplica las chances de aprobación por 1.0083 (aumento del 0.83%)

Variables Dicotómicas

Cuando la variable independiente es dicotómica (ej: sexo, región):

Efecto marginal = Diferencia de probabilidades entre grupos

Variable	Efecto Marginal	Error Est.	p-value
mujer	0.0236	0.0161	0.1415
score	0.0007	0.0001	0.0000

Interpretación: Si el coeficiente de "mujer" fuera -0.15 → "Ser mujer reduce la probabilidad de aprobación en 15 puntos porcentuales"

Código: Efectos Marginales en R

```r
# Instalar paquete si no lo tienes
# install.packages("margins")
library(margins)

# Efectos marginales promedio (AME)
efectos ← margins(modelo_logit)
summary(efectos)

# Efectos marginales evaluados en la media (MEM)
efectos_media ← margins(modelo_logit, at = list(
  ingreso = mean(datos$ingreso),
  score = mean(datos$score),
  edad = mean(datos$edad)
))

# Visualizar efectos marginales
plot(efectos)
```

Evaluación de Modelos

R^2 de McFadden

Análogo al R^2 tradicional para modelos de variable dependiente limitada

$$R^2_{McFadden} = 1 - \frac{\log L(\hat{\beta})}{\log L(\beta_0)}$$

```
# Calcular R² de McFadden
logLik_full ← logLik(modelo_logit)
logLik_null ← logLik(glm(aprobo ~ 1, data = datos,
                         family = binomial(link = "logit")))

r2_mcfadden ← 1 - (as.numeric(logLik_full) / as.numeric(logLik_null))
cat("R² de McFadden:", round(r2_mcfadden, 4))
```

```
## R² de McFadden: 0.3066
```

Interpretación: Valores entre 0.2-0.4 son considerados buenos

Usar R^2 ajustado para comparar modelos con diferente número de variables

Matriz de Confusión

Clasificación: ¿El modelo predice correctamente?

Necesitamos un **punto de corte** (c): si $\hat{P} > c \rightarrow \hat{Y} = 1$

	Predijo Y=0	**Predijo Y=1**
Real Y=0	TN (Verdadero Neg.) ✓	FP (Falso Pos.) ✗
Real Y=1	FN (Falso Neg.) ✗	TP (Verdadero Pos.) ✓

Métricas:

- **Tasa de acierto** (Accuracy) = $\frac{TN+TP}{n}$
- **Sensitividad** (Recall) = $\frac{TP}{TP+FN}$ = % de casos positivos bien clasificados
- **Especificidad** = $\frac{TN}{TN+FP}$ = % de casos negativos bien clasificados

Matriz de Confusión: Ejemplo

```r
# Predicciones
predicciones ← predict(modelo_logit, type = "response")

# Clasificación con punto de corte c = 0.5
clase_pred ← ifelse(predicciones > 0.5, 1, 0)

# Matriz de confusión
conf_matrix ← confusionMatrix(
  factor(clase_pred, levels = c(0,1)),
  factor(datos$aprobo, levels = c(0,1)),
  positive = "1"
)
conf_matrix$table
```

```
##           Reference
## Prediction   0   1
##          0  15  12
##          1  60 913
```

Elección del Punto de Corte

Diferentes puntos de corte según el objetivo:

- **c = 0.5**: Estándar, trata FP y FN como igual de costosos
- **c = proporción de Y=1**: Balancea las clases
- **c personalizado**: Depende del costo relativo de errores

Ejemplo: En detección de fraude, un falso negativo (fraude no detectado) es más costoso que un falso positivo → usar c más bajo (ej: 0.3)

Supuestos, Problemas y Soluciones

Supuestos Clave

1. Independencia de observaciones

- Las observaciones son independientes entre sí
- Violación común: datos en panel, clustering

2. Correcta especificación del modelo

- Forma funcional apropiada
- Variables relevantes incluidas
- No variables irrelevantes que inflen varianza

3. No multicolinealidad perfecta

- Las variables independientes no son combinaciones lineales exactas

Problemas Comunes y Soluciones

Problema	Efecto	Solución
Heteroscedasticidad (MPL)	Errores estándar incorrectos	Errores robustos de White
Predicciones fuera de [0,1] (MPL)	No interpretable como probabilidad	Usar Logit/Probit
Variables omitidas	Sesgo en coeficientes	Agregar variables relevantes
Separación perfecta	Modelo no converge	Penalización (Firth), más datos
Multicolinealidad alta	Coeficientes inestables	Eliminar variables correlacionadas

¡Gracias!

Preguntas