Unidad 5: Modelos Lineales Generalizados

Modelamiento estadístico

Modelos Lineales Generalizados (GLM)

Los **Modelos Lineales Generalizados** (GLM, por sus siglas en inglés) son una generalización de los modelos lineales clásicos (como la regresión lineal) que permiten modelar una mayor variedad de *distribuciones de datos* y relaciones *no lineales* entre las variables dependientes e independientes.

Los GLM extienden los modelos lineales tradicionales al permitir que la **variable dependiente siga una distribución de probabilidad diferente a la normal** (por ejemplo, binomial, Poisson, etc.), y que la relación entre la variable dependiente y las variables independientes no sea necesariamente lineal.

La ecuación general de un GLM es la siguiente:

$$E[y_i|x_{ij}] = g(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik})$$

donde $g(.)$ es la función de enlace que **transforma** la media de la variable dependiente en una cantidad lineal.

Utilidad de los GLM

1. Flexibilidad:

Los GLM permiten trabajar con datos que no siguen distribuciones normales. Por ejemplo, cuando la variable dependiente es binaria (como en un modelo de clasificación) o es un conteo (como en el número de eventos ocurridos en un intervalo de tiempo).

2. Modelos adecuados para distintos tipos de datos:

Los GLM pueden modelar datos de distinta naturaleza (binarios, continuos, de conteo) sin tener que hacer suposiciones rígidas sobre la distribución de los datos.

3. Ajuste de la varianza:

Los GLM permiten ajustar la varianza de los errores en función de los predictores, lo que es útil cuando los datos tienen heterocedasticidad (es decir, cuando la varianza de los errores cambia con los valores de las variables independientes).

Componentes de un GLM

Función de enlace (link function)

Esta función define la relación entre la *media* de la variable dependiente y los predictores (variables independientes). Es una **transformación** de la variable dependiente para que la relación con los predictores sea lineal.

En los modelos clásicos de regresión lineal, se asume que la variable dependiente está directamente relacionada con los predictores a través de una combinación lineal. Sin embargo, en los GLM, la variable dependiente puede tener una distribución diferente a la normal y su relación con los predictores no siempre es lineal. La función de enlace transforma la media de la variable dependiente para hacer que la relación sea lineal en términos de los parámetros del modelo.

En términos formales, asegura que el modelo respete las **restricciones de dominio** de la variable dependiente. Por ejemplo, en regresiones logísticas, donde la variable dependiente es binaria, el *"link"* logit previene que las predicciones se salgan de este dominio.

Existen diferentes tipos de funciones de enlace, las más populares:

- Logit
- Probit
- Log
- Inversa
- Identidad (utilizada en MCO)

Distribución de la variable dependiente

A diferencia de la regresión lineal, que asume que la variable dependiente sigue una distribución **normal**, los GLM permiten que la variable dependiente siga distribuciones diferentes, como:

- Distribución binomial (variables binarias)
- Distribución Poisson (conteos)
- Distribución gamma (variables continuas positivas y *sesgadas*)
- Distribución normal

Supuestos

A diferencia de MCO, los modelos lineales generalizados son más flexibles:

1. Se requiere datos iid
2. **No requiere** que los datos sigan una distribución normal
3. **No se asume** linealidad en la relación entre la variable de respuesta y las explicativas
4. **No requiere** homocedasticidad, porque por construcción las varianzas son heterocedásticas.

5. Los errores deben ser **independientes** pero **no normalmente distribuidos**

Tipos de GLM

Regresión lineal simple

La regresión lineal simple es parte de los GLM, su función de enlace es la **identidad**,

$$g(.) = E[y]$$

Esta es la función más simple. La variable dependiente sigue una distribución normal

Regresión Logística Binaria

La regresión logística se utiliza cuando la variable dependiente es binaria. Su principal objetivo es modelar la probabilidad de que un evento ocurra (por ejemplo, que un individuo compre un producto, que un paciente sea diagnosticado con una enfermedad, etc.), en función de las variables explicativas $x_1, .., x_k$.

La función logística o sigmoide, mapea números reales al rango (0,1), y se define como

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Donde z es un número real.

Por tanto, la probabilidad de que $y = 1$ se modela como

$$p(y_i = 1|x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})}}$$

Transformando la combinación lineal de los predictores en un valor entre 0 y 1, que se puede interpretar como una probabilidad.

La función de enlace en la regresión logística es el logit de la probabilidad, que es el logaritmo de la razón de probabilidades (odds). La razón de probabilidades es la relación entre la probabilidad de que el evento ocurra y la probabilidad de que no ocurra:

$$g(.) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki}$$

Donde p es la probabilidad de que $y = 1$ y $1 - P$ es la probabilidad de que $y = 0$.

Esto convierte la probabilid (entre 0 y 1) en un rango de $-\infty$ a $+\infty$, lo que permite ajustar un modelo lineal.

Estimación mediante MLE

La estimación de los coeficientes β se realiza mendiate Máxima verosimilitud (MLE). A diferencia de la regresión lineal, donde los coeficientes se estiman minimizando la suma de los cuadrados de los residuos, en la regresión logística, se busca maximizar la función de verosimilitud.

En otras palabras, el MLE busca el conjunto de parámetros que hace que los datos observados sean más probables bajo el modelo propuesto.

La función de verosimilitud para una muestra con N observaciones es:

$$L(\beta_0, ..., \beta_k) = \prod_{i=1}^{N} [p(y_i|x_i)^{y_i} \cdot (1 - p(y_i|x_i))^{1-y_i}]$$

Es decir, es simplemente el producto de las probabilidades de que cada observación sea 1 o 0.

Log verosimilitud

La log verosimilitud es el logaritmo natural de la función de verosimilitud, ya que al aplicar logaritmos, los productos se convierten en sumas y facilitan los cálculos (propiedades del logaritmo de un producto)

$$l(\beta_0, ..., \beta_k) = \sum_{i=1}^{N} [y_i log(p[y_i|x_i]^{y_i}) + (1 - y_i) log(1 - p(y_i|x_i))^{1-y_i}]$$

El estimador MLE busca el conjunto de parámetros que maximicen la función de log-verosimilitud.

$$\hat{\beta}_{MLE} = argmax_{\beta_0,...,\beta_k} \left(\sum_{i=1}^{N} [y_i log(p[y_i|x_i]^{y_i}) + (1 - y_i) log(1 - p(y_i|x_i))^{1-y_i}] \right)$$

Este proceso generalmente se realiza utilizando algoritmos de optimización numérica, como el algoritmo de **Newton-Raphson** o el algoritmo de **gradiente descendente**.

Los estimadores MLE son consistentes y siguen una distribución asintótica normal y además son asintóticamente eficientes.

Interpretación de los coeficientes

Los coeficientes $\beta_1, ..., \beta_k$ tienen una interpretación más compleja que en MCO. El coeficiente β_j representa el **cambio en la logit de la probabilidad** cuando x_j cambia en una unidad.

Odds Ratio Es la razón entre las probabilidades de que un evento ocurra y no ocurra

$$OR = \frac{p}{1-p}$$

Si el OR de fumar y desarrollar cáncer de pulmón es 2, significa que los fumadores tienen el doble de probabilidades de desarrollar cáncer de pulmón que los no fumadores.

Sin embargo, lo que muestran los coeficientes es el cambio en el logaritmo del OR $log(OR)$, por lo que debemos utilizar la función exponencial con base e para poder interpretarla como OR. Por ejemplo, si $\beta_j = 1.2$, es decir, un incremento en una unidad de x_j aumenta la OR a $e^{1.2} = 3.32$, es decir la probabilidad de que $y = 1$ es 3.2 veces superior a la probabilidad de que $y = 0$.

Aún con la conversión a Odds Ratio la interpretación no permite que sea comparable con el coeficiente estimado por una regresión lineal.

Efecto Marginal Para dar una interpretación en término de probabilidades se calcula el **Efecto Marginal**, que simplemente es la calcular la pendiente de nuestra regresión *logit*.

Dado que no es una linea recta, el valor o efecto marginal, va a depender de en qué parte de la curva estemos.

Por tanto ¿cómo reportamos el efecto marginal?

- Efecto Marginal Promedio (AME): Calcula el efecto marginal de cada observación y toma la media, que es la medida más utilizada.
- Efecto Marginal **en el** Promedio: Calcula la media de todas las variables y luego el efecto marginal para esta *observación hipotética* que tiene valores promedios en todas las variables.

Ejemplo:

Se desea estimar un modelo sobre el efecto de la cercanía de la universidad en la probabilidad de estar o no inscrito.

$enroll = 1$ cuando el individuo está inscrito en la universidad

$$enroll_i = \beta_0 + \beta_1 nearc2 + \beta_2 nearc4 + \beta_3 age + \beta_4 IQ + e_i$$

```r
library(wooldridge)
library(jtools)
data("card")
lpm<- lm(enroll ~ nearc2 + nearc4 + age+ IQ ,
         data = card)

logit<-glm(enroll ~ nearc2 + nearc4 +age + IQ,
         data = card, family = binomial(link = "logit"))
export_summs(lpm,logit)
```

	Model 1	Model 2
(Intercept)	0.21 **	-0.96
	(0.08)	(0.89)
nearc2	0.01	0.13
	(0.01)	(0.15)
nearc4	0.03 *	0.40 *
	(0.02)	(0.18)
age	-0.01 ***	-0.09 ***
	(0.00)	(0.03)
IQ	0.00 *	0.01 *
	(0.00)	(0.00)
N	2061	2061
R2	0.01	
AIC	965.96	1370.40
BIC	999.74	1398.55
Pseudo R2		0.03

*** p < 0.001; ** p < 0.01; * p < 0.05.

El único coeficiente que puede ser interpretado irectamente es el del MCO,

podríamos decir que vivir cerca de una universidad (que ofrece títulos de 4 años), incrementa la probabilidad de estar inscrito en 4%.

La interpretación del modelo **logit** no es directa, sólo podemos decir que vivir cerca de una universidad con carreras de 4 años, incrementa el logaritmo del Odds Ratio $log(p/1-p)$ de estar inscrito en 0.42, o que el Odds Ratio aumenta a 1.52 ($e^{0.42}$).

Ahora calcularemos el AME (Efecto Marginal Promedio) para darle una interpretación.

```
library(tidyverse)
library(broom)
library(margins)
logitmar<-margins(logit) #AME
Mean <- model.frame(logit) %>%
  map_df(mean)
logitmean<-margins(logit, at= Mean)
export_summs(lpm,logitmar, logitmean)
```

"Un estudiante que vive cerca de una universidad con carreras de 4 años, tiene una probabilidad 4% superior de estar inscrito"

Regresión Probit

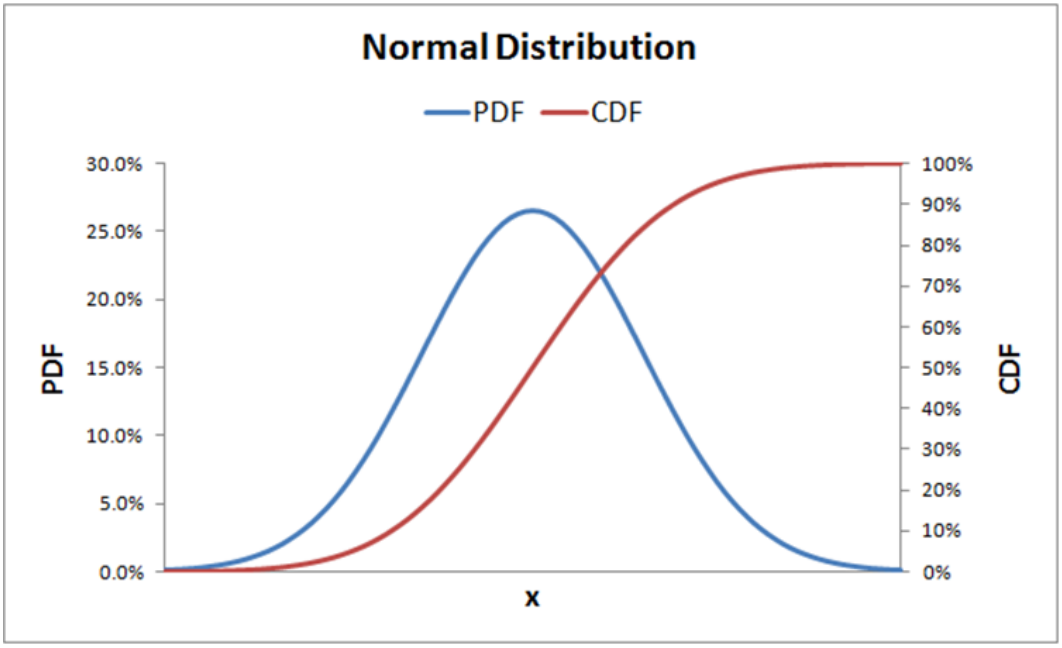
El enlace probit se utiliza cuando la variable dependiente es binaria. La diferencia con el logit es que el probit usa la distribución normal acumulada en lugar de la logaritmo de las probabilidades.

La función de enlace en el modelo probit es la función de distribución acumulada normal (CDF de la normal estándar). En la práctica, esto se usa porque la distribución normal es muy flexible y se ajusta bien a muchos tipos de datos, incluyendo aquellos provenientes de una distribución binomial.

En este contexto, el modelo probit puede ser visto como un modelo de regresión lineal, pero transformando la probabilidad mediante la función de distribución acumulada para garantizar que las predicciones estén entre 0 y 1, lo cual es adecuado para variables binarias.

	Model 1	Model 2	Model 3
(Intercept)	0.21 **		
	(0.08)		
nearc2	0.01	0.01	0.01
	(0.01)	(0.01)	(0.01)
nearc4	0.03 *	0.04 *	0.04 *
	(0.02)	(0.02)	(0.02)
age	-0.01 ***	-0.01 ***	-0.01 ***
	(0.00)	(0.00)	(0.00)
IQ	0.00 *	0.00 *	0.00 *
	(0.00)	(0.00)	(0.00)
N	2061	0	0
R2	0.01		

*** p < 0.001; ** p < 0.01; * p < 0.05.



Si se toma cualquier valor de x, la Función de Distribución Acumulada la convierte en un valor entre 0 y 1.

$$g(.) = \Phi^{-1}(x'\beta)$$

Donde Φ^{-1} es la función **inversa** de la distribución normal acumulada.

```
probit<-glm(enroll ~ nearc2 + nearc4 +age + IQ,
         data = card, family = binomial(link = "probit"))
export_summs(lpm,logit, probit)
```

	Model 1	Model 2	Model 3
(Intercept)	0.21 **	-0.96	-0.62
	(0.08)	(0.89)	(0.46)
nearc2	0.01	0.13	0.07
	(0.01)	(0.15)	(0.08)
nearc4	0.03 *	0.40 *	0.20 *
	(0.02)	(0.18)	(0.09)
age	-0.01 ***	-0.09 ***	-0.05 ***
	(0.00)	(0.03)	(0.01)
IQ	0.00 *	0.01 *	0.01 *
	(0.00)	(0.00)	(0.00)
N	2061	2061	2061
R2	0.01		
AIC	965.96	1370.40	1370.57
BIC	999.74	1398.55	1398.72
Pseudo R2		0.03	0.03

*** p < 0.001; ** p < 0.01; * p < 0.05.

Los resultados difieren entre ambos modelos (probit vs logit).

Ahora tomaremos los efectos marginales promedios y compararemos los tres modelos:

```
logitmar<-margins(logit) #AME
probitmar<-margins(probit)
export_summs(lpm,logitmar, probitmar)
```

```
## Warning in nobs.default(m, use.fallback = TRUE): no 'nobs' method is available
## Warning in nobs.default(m, use.fallback = TRUE): no 'nobs' method is available
```

	Model 1	Model 2	Model 3
(Intercept)	0.21 **		
	(0.08)		
nearc2	0.01	0.01	0.01
	(0.01)	(0.01)	(0.01)
nearc4	0.03 *	0.04 *	0.04 *
	(0.02)	(0.02)	(0.02)
age	-0.01 ***	-0.01 ***	-0.01 ***
	(0.00)	(0.00)	(0.00)
IQ	0.00 *	0.00 *	0.00 *
	(0.00)	(0.00)	(0.00)
N	2061	0	0
R2	0.01		

*** p < 0.001; ** p < 0.01; * p < 0.05.

Medidas de bondad de ajuste

Pseduo R^2

El **Pseduo** R^2 o el R de McFadden, compara la log verosimilitud del modelo que tenemos versus un modelo que solo tiene una constante.

$$R^2_{McF} = 1 - \frac{LL_{ur}}{LL_0}$$

Donde LL_{ur} es la log verosimilitud de nuestro modelo no restringido con todas las variables independientes y LL_0 es la verosimilitud de un modelo con solo una constante.

Si las variables independientes no explican la variable dependiente, entonces la verosimilitud del modelo restringido y no restringido serán similares, por tanto el R^2 será cercano a 0.

Sin embargo, el pseudo R^2 no tiene una interpretación directa como el coeficiente de determinación de MCO. Tan solo se busca que este sea los más alto posible, para comparar modelos.

El pseudo R^2 indica la capacidad del modelo de predecir la variable de resultado, no el ajuste del modelo a los datos.

Matriz de confusión

Muestra el porcentaje de casos correctamente predichos por nuestro modelo.

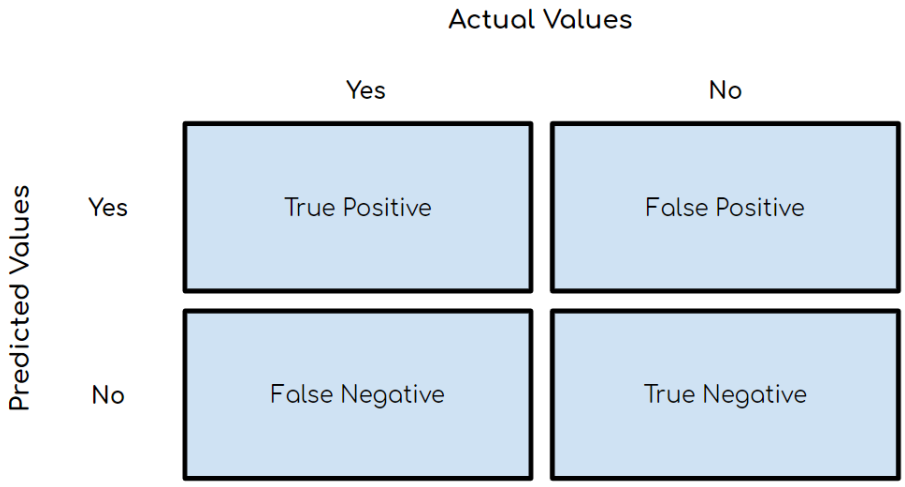


Figure 1: Matriz de Confusión