Metodología I

Magíster en Ciencias Sociales

Pablo Pérez Ahumada
Universidad de Chile
Departamento de sociología

Módulo 3
Regresión lineal y logística binaria

REGRESIÓN LOGÍSTICA BINARIA

¿Qué es la regresión logística binaria?

- Tipo de regresión utilizada cuando la **variable dependiente** es **dicotómica** (dummy)
  - ¿Ha participado en una marcha en los últimos 12 meses? 1 = sí, 0 = no
  - ¿Se ha contagiado de COVID alguna vez? 1 = sí, 0 = no
  - ¿Practica deportes al menos una vez por semana? 1 = sí, 0 = no
  - Etc.

- Cuando tenemos este tipo de variables, la estimación estadística es, esencialmente, una estimación de probabilidades (ej., probabilidad de contagiarse de COVID)

- Al igual que en una regresión lineal (OLS), en una regresión logística las variables independientes pueden ser nominales, ordinales, o de intervalo/razón

Ejemplo: Encuesta Mundial de Valores (Chile, 2005 – 2022)

- ¿Existe una relación entre la afiliación a sindicatos y la participación en marchas?

Ejemplo: Encuesta Mundial de Valores (Chile, 2005 – 2022)

- ¿Existe una relación entre la afiliación a sindicatos y la participación en marchas?

Participación en marchas según afiliación sindical

demonstr_dummy	*Unionized*		**Total**
	No	Sí	
No	941 80.6 %	214 74 %	1155 79.3 %
Sí	227 19.4 %	75 26 %	302 20.7 %
Total	1168 100 %	289 100 %	1457 100 %

$\chi^2=5.598 \cdot df=1 \cdot \phi=0.064 \cdot p=0.018$

Modelando una relación con variable dependiente dicotómica

- ¿Es posible llegar a la misma conclusión pero controlando por otros factores relevantes?

- ¿Es posible aprovechar las herramientas de un modelo de regresión convencional (ej., obtener un coeficiente β que muestre el "efecto" de una variable X sobre una variable Y)?

- En principio, eso se *podría* hacer con una regresión lineal (MCO)
  - Esto se conoce como Modelo de Probabilidad Lineal

MODELO DE PROBABILIDAD LINEAL

Modelo de Probabilidad Lineal

- Utilización de una recta de regresión de MCO para estimar la probabilidad de ocurrencia de un fenómeno (en este caso, participar en marchas)

- ¿Qué nos dice el coeficiente β para “Sindicalizado/a”?

	Modelo de Prob Lineal
(Intercept)	0.167 ***
	(0.022)
Unionized	0.072 **
	(0.027)
WaveWave 6	0.079 **
	(0.027)
WaveWave 7	-0.005
	(0.027)
R^2	0.013
Adj. R^2	0.011
Num. obs.	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

Modelo de Probabilidad Lineal: problemas

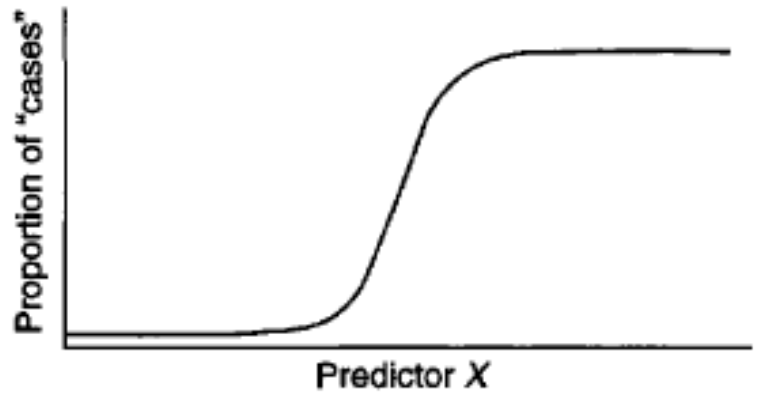
- Se pueden incumplir dos supuestos de la regresión MCO:
  1. Los residuos pueden presentar problemas de heterocedasticidad (su varianza no es constante). Con ello, el coeficiente es insesgado pero su error estándar será incorrecto
  2. Los residuos pueden no estar distribuidos normalmente (lo cual afecta a la prueba T y a la estimación de los intervalos de confianza)

- Además, valores predichos por la recta pueden caer fuera del rango de una probabilidad (ej., pueden ser mayores a 1).
  - Cuando esto pasa, el modelo producido no es un buen estimador de la probabilidad a nivel poblacional
  - ¿Por qué ocurre esto? Porque la distribución de probabilidades *no* es lineal

Diferencia entre una predicción lineal y una de probabilidades

(A) For a continuous outcome variable Y, the numerical value of Y at each value of X.

(B) For a binary outcome variable, the proportion of individuals who are "cases" (exhibit a particular outcome property) at each value of X.

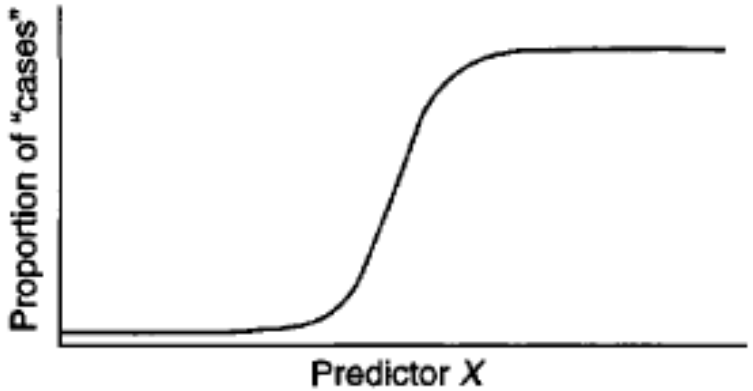


Jacob Cohen, et al. (2014). Applied multiple regression/correlation analysis for the behavioral sciences. Third edition. LEA Publishers, p. 482.

Diferencia entre una predicción lineal y una de probabilidades

(A) For a continuous outcome variable Y, the numerical value of Y at each value of X.

(B) For a binary outcome variable, the proportion of individuals who are "cases" (exhibit a particular outcome property) at each value of X.



En este gráfico, la tasa del cambio de Y respecto a los cambios de X *no* es siempre la misma

Jacob Cohen, et al. (2014). Applied multiple regression/correlation analysis for the behavioral sciences. Third edition. LEA Publishers, p. 482.

MODELO DE REGRESIÓN LOGÍSTICA

Regresión logística

- La regresión logística resuelve estos problemas.

- Para ello, su **punto de partida** es transformación de los coeficientes β en **coeficientes *logit***

Coeficientes *logit y odds* en la regresión logística

- Se conoce como “logit” a la transformación logarítmica de los *odds* (traducidos comúnmente como “chances”)

- ¿Qué son los *odds*? Una razón de probabilidades

Coeficientes *logit y odds* en la regresión logística

- Por lo tanto, para estimar probabilidades a través de una regresión logística hay que seguir estos pasos

1. Estimar los *odds* o razón de probabilidades
2. Estimar *odds ratios* (razones entre odds)
3. Aplicar una transformación logarítmica a esos odds ratios para obtener coeficientes *logit*
4. Calcular las *probabilidades*

1) Odds

- Se define como la probabilidad de que ocurra un evento dividido por la probabilidad de que dicho evento no ocurra

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

Ejemplo participación en marchas

Participación en marchas según afiliación sindical

demonstr_dummy	*Unionized*		***Total***
	No	Sí	
No	941 80.6 %	214 74 %	1155 79.3 %
Sí	227 19.4 %	75 26 %	302 20.7 %
Total	1168 100 %	289 100 %	1457 100 %

$\chi^2=5.598 \cdot df=1 \cdot \phi=0.064 \cdot p=0.018$

Participación en marchas según afiliación sindical

demonstr_dummy	*Unionized*		***Total***
	No	Sí	
No	941 80.6 %	214 74 %	1155 79.3 %
Sí	227 19.4 %	75 26 %	302 20.7 %
Total	1168 100 %	289 100 %	1457 100 %

$\chi^2=5.598 \cdot df=1 \cdot \phi=0.064 \cdot p=0.018$

$$Odds_{participar} = \frac{0{,}207}{0{,}793} = 0{,}26$$

Las chances de participar en una marcha son de 0,26, respecto a las chances de no participar

Participación en marchas según afiliación sindical

demonstr_dummy	*Unionized* No	*Unionized* Sí	***Total***
No	941 80.6 %	214 74 %	1155 79.3 %
Sí	227 19.4 %	75 26 %	302 20.7 %
Total	1168 100 %	289 100 %	1457 100 %

$\chi^2=5.598 \cdot df=1 \cdot \phi=0.064 \cdot p=0.018$

$$Odds_{participar} = \frac{0{,}207}{0{,}793} = 0{,}26$$

Las chances de participar en una marcha son de 0,26, respecto a las chances de no participar

En otras palabras: por cada 1 persona, hay sólo 0,26 personas que participan en marchas.

O más **intuitivamente,** por cada 100 personas, hay sólo 26 personas que participan

Participación en marchas según afiliación sindical

demonstr_dummy	*Unionized*		***Total***
	No	Sí	
No	941 80.6 %	214 74 %	1155 79.3 %
Sí	227 19.4 %	75 26 %	302 20.7 %
Total	1168 100 %	289 100 %	1457 100 %

$\chi^2=5.598 \cdot df=1 \cdot \phi=0.064 \cdot p=0.018$

¿Cambian las chances de participar según se esté afiliado/a a un sindicato?

Participación en marchas según afiliación sindical

demonstr_dummy	*Unionized* No	*Unionized* Sí	***Total***
No	941 80.6 %	214 74 %	1155 79.3 %
Sí	227 19.4 %	75 26 %	302 20.7 %
Total	1168 100 %	289 100 %	1457 100 %

$\chi^2=5.598 \cdot df=1 \cdot \phi=0.064 \cdot p=0.018$

¿Cambian las chances de participar según se esté afiliado/a a un sindicato?

$$Odds_{sindical} = \frac{0{,}26}{0{,}74} = 0{,}35$$

$$Odds_{NOsindical} = \frac{0{,}194}{0{,}806} = 0{,}24$$

Implicancias para la interpretación de los odds

- Valores bajo 1 indican que las chances de que ocurra un evento son negativas
- Valores iguales a 1 indican chances iguales
- Valores sobre 1 indican chances positivas

Implicancias para la interpretación de los odds

- Valores bajo 1 indican que las chances de que ocurra un evento son negativas
- Valores iguales a 1 indican chances iguales
- Valores sobre 1 indican chances positivas
  - Ejemplo, imaginemos que el 65% de los/as afiliados/as a sindicatos ha participado en marchas, y sólo el 35% no lo hecho.

Implicancias para la interpretación de los odds

- Valores bajo 1 indican que las chances de que ocurra un evento son negativas
- Valores iguales a 1 indican chances iguales
- Valores sobre 1 indican chances positivas
  - Ejemplo, imaginemos que el 65% de los/as afiliados/as a sindicatos ha participado en marchas, y sólo el 35% no lo hecho.

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

$$Odds_{sindical} = \frac{0{,}65}{0{,}35} = 1{,}86$$

Implicancias para la interpretación de los odds

- Valores bajo 1 indican que las chances de que ocurra un evento son negativas

- Valores iguales a 1 indican chances iguales

- Valores sobre 1 indican chances positivas
  - Ejemplo, imaginemos que el 65% de los/as afiliados/as a sindicatos ha participado en marchas, y sólo el 35% no lo hecho.

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

$$Odds_{sindical} = \frac{0{,}65}{0{,}35} = 1{,}86$$

Por cada 100 personas sindicalizadas, hay 186 que participan

2) Odds ratios (razones de chances)

- Cálculo que permite reflejar asociación entre dos variables dicotómicas, a partir de una comparación entre chances

- Siguiendo con el ejemplo anterior, ¿tienen los/as sindicalizados más chances de participar en marchas que quienes no están sindicalizados/as?

2) Odds ratios

- Cálculo que permite reflejar asociación entre dos variables dicotómicas, a partir de una comparación entre chances

- Siguiendo con el ejemplo anterior, ¿tienen los/as sindicalizados más chances de participar en marchas que quienes no están sindicalizados/as?

$$OR = \frac{p_{sindical}/(1 - p_{sindical})}{p_{NOsindical}/(1 - p_{NOsindical})}$$

2) Odds ratios

- Cálculo que permite reflejar asociación entre dos variables dicotómicas, a partir de una comparación entre chances

- Siguiendo con el ejemplo anterior, ¿tienen los/as sindicalizados más chances de participar en marchas que quienes no están sindicalizados/as?

$$OR = \frac{p_{sindical}/(1 - p_{sindical})}{p_{NOsindical}/(1 - p_{NOsindical})}$$

$$OR = \frac{0{,}26/0{,}74}{0{,}194/0{,}806} = \frac{0{,}35}{0{,}24} = 1{,}46$$

Las chances de participar en marchas de los/as sindicalizados/as son 1,5 veces más que las de quienes no están sindicalizados/as

2) Odds ratios: implicancias

- El odds ratio o razones de chances es útil porque nos permite expresar en un número la relación entre dos variables categóricas

- En las regresiones logísticas, el odds ratio es la primera manera de aproximarnos a relación entre variables

- Sin embargo, falta un paso más necesario para construir modelos de regresión logística

3) Logit

- Es una unidad de medida de la relación entre dos variables (VD: dicotómica), que en regresión logística se calcula a partir del ***logaritmo natural de los odds***

3) Logit

- Es una unidad de medida de la relación entre dos variables (VD: dicotómica), que en regresión logística se calcula a partir del ***logaritmo natural de los odds***

- Esta transformación logarítmica es la base de la estimación de parámetros en la regresión logística:
  - Con ella se puede modelar la probabilidad de ocurrencia de un fenómeno como la función logística de una *combinación lineal de las variables independientes o predictores*

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n$$

3) Logit

- Es una unidad de medida de la relación entre dos variables (VD: dicotómica), que en regresión logística se calcula a partir del ***logaritmo natural de los odds***

- Esta transformación logarítmica es la base de la estimación de parámetros en la regresión logística:
  - Con ella se puede modelar la probabilidad de ocurrencia de un fenómeno como la función logística de una *combinación lineal de las variables independientes o predictores*

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n$$

La *mejor combinación lineal* de predictores se obtiene no a través de MCO, sino a través del procedimiento de **máxima verosimilitud**

3) Logit

- La estimación de coeficientes *logit* es un gran avance, porque permite obtener coeficientes similares al beta en la regresión MCO

- De hecho, la mayoría de los softwares estadísticos reportan resultados en coeficientes logit (a veces llamados *log odds*)

- A diferencia de los odds ratio, los coeficientes logit tienen valores que van de $-\infty$ a $+\infty$
  - Así, una relación negativa puede ser directamente interpretable (por el signo)

REGRESIÓN LOGÍSTICA: ejemplo

Ejemplo regresión logística (resultados en log odds)

- ¿Existe una relación estadísticamente significativa entre afiliación sindical y participación en marchas?

- Datos: Encuesta Mundial de Valores (Chile 2006, 2012 y 2018)
  - Variable dependiente: ¿Ha participado en marchas en los últimos 12 meses? 1 = sí; 0 = no
  - Variable independiente de interés: ¿Está afiliado/a a un sindicato?
  - Controles demográficos: género (female), edad (X003), educación (3 tramos), sector privado (private_sector) y nivel de politización medido en escala de 0 a 6 (politicization)
  - Otro control: ola de aplicación de la encuesta. Wave: 2006, 2012, 2018

¿Existe una relación estadísticamente significativa entre afiliación sindical y participación en marchas?

	M1 (m prob lineal)	M1 (log odds)	M2 (log odds)	M3 (log odds)
(Intercept)	0.167 ***	-1.601 ***	-1.023 *	-1.379 **
	(0.022)	(0.140)	(0.424)	(0.437)
Unionized	0.072 **	0.418 **	0.390 *	0.312 †
	(0.027)	(0.155)	(0.157)	(0.161)
WaveWave 6	0.079 **	0.470 **	0.498 **	0.496 **
	(0.027)	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.005	-0.031	-0.102	-0.121
	(0.027)	(0.174)	(0.181)	(0.184)
Female			-0.020	0.100
			(0.136)	(0.139)
X003			-0.007	-0.012 *
			(0.006)	(0.006)
Educ2			0.042	-0.053
			(0.268)	(0.272)
Educ3			0.521 †	0.259
			(0.283)	(0.289)
private_sector			-0.516 **	-0.445 *
			(0.175)	(0.180)
politicization				0.282 ***
				(0.042)
R^2	0.013			
Adj. R^2	0.011			
Num. obs.	1457	1457	1457	1457
AIC		1475.887	1460.059	1415.850
BIC		1497.024	1507.616	1468.691
Log Likelihood		-733.944	-721.030	-697.925
Deviance		1467.887	1442.059	1395.850

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

¿Existe una relación estadísticamente significativa entre afiliación sindical y participación en marchas?

Ej. Modelo 2:

	M1 (m prob lineal)	M1 (log odds)	M2 (log odds)	M3 (log odds)
(Intercept)	0.167 ***	-1.601 ***	-1.023 *	-1.379 **
	(0.022)	(0.140)	(0.424)	(0.437)
Unionized	0.072 **	0.418 **	0.390 *	0.312 †
	(0.027)	(0.155)	(0.157)	(0.161)
WaveWave 6	0.079 **	0.470 **	0.498 **	0.496 **
	(0.027)	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.005	-0.031	-0.102	-0.121
	(0.027)	(0.174)	(0.181)	(0.184)
Female			-0.020	0.100
			(0.136)	(0.139)
X003			-0.007	-0.012 *
			(0.006)	(0.006)
Educ2			0.042	-0.053
			(0.268)	(0.272)
Educ3			0.521 †	0.259
			(0.283)	(0.289)
private_sector			-0.516 **	-0.445 *
			(0.175)	(0.180)
politicization				0.282 ***
				(0.042)
R^2	0.013			
Adj. R^2	0.011			
Num. obs.	1457	1457	1457	1457
AIC		1475.887	1460.059	1415.850
BIC		1497.024	1507.616	1468.691
Log Likelihood		-733.944	-721.030	-697.925
Deviance		1467.887	1442.059	1395.850

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

¿Existe una relación estadísticamente significativa entre afiliación sindical y participación en marchas?

Ej. Modelo 2:

- En comparación a los no sindicalizados (categoría de referencia), el log-odds de participación en marchas para afiliados a sindicatos aumenta en 0,39 ($p < 0,05$)

	M1 (m prob lineal)	M1 (log odds)	M2 (log odds)	M3 (log odds)
(Intercept)	0.167 ***	-1.601 ***	-1.023 *	-1.379 **
	(0.022)	(0.140)	(0.424)	(0.437)
Unionized	0.072 **	0.418 **	0.390 *	0.312 †
	(0.027)	(0.155)	(0.157)	(0.161)
WaveWave 6	0.079 **	0.470 **	0.498 **	0.496 **
	(0.027)	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.005	-0.031	-0.102	-0.121
	(0.027)	(0.174)	(0.181)	(0.184)
Female			-0.020	0.100
			(0.136)	(0.139)
X003			-0.007	-0.012 *
			(0.006)	(0.006)
Educ2			0.042	-0.053
			(0.268)	(0.272)
Educ3			0.521 †	0.259
			(0.283)	(0.289)
private_sector			-0.516 **	-0.445 *
			(0.175)	(0.180)
politicization				0.282 ***
				(0.042)
R^2	0.013			
Adj. R^2	0.011			
Num. obs.	1457	1457	1457	1457
AIC		1475.887	1460.059	1415.850
BIC		1497.024	1507.616	1468.691
Log Likelihood		-733.944	-721.030	-697.925
Deviance		1467.887	1442.059	1395.850

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

¿Existe una relación estadísticamente significativa entre afiliación sindical y participación en marchas?

Ej. Modelo 2:

- En comparación a los/as trabajadores/as del sector público (categoría de referencia), el log-odds de participación en marchas para los/as del sector privado *disminuye* en 0,52 (valor – p < 0,01)

	M1 (m prob lineal)	M1 (log odds)	M2 (log odds)	M3 (log odds)
(Intercept)	0.167 ***	-1.601 ***	-1.023 *	-1.379 **
	(0.022)	(0.140)	(0.424)	(0.437)
Unionized	0.072 **	0.418 **	0.390 *	0.312 †
	(0.027)	(0.155)	(0.157)	(0.161)
WaveWave 6	0.079 **	0.470 **	0.498 **	0.496 **
	(0.027)	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.005	-0.031	-0.102	-0.121
	(0.027)	(0.174)	(0.181)	(0.184)
Female			-0.020	0.100
			(0.136)	(0.139)
X003			-0.007	-0.012 *
			(0.006)	(0.006)
Educ2			0.042	-0.053
			(0.268)	(0.272)
Educ3			0.521 †	0.259
			(0.283)	(0.289)
private_sector			-0.516 **	-0.445 *
			(0.175)	(0.180)
politicization				0.282 ***
				(0.042)
R^2	0.013			
Adj. R^2	0.011			
Num. obs.	1457	1457	1457	1457
AIC		1475.887	1460.059	1415.850
BIC		1497.024	1507.616	1468.691
Log Likelihood		-733.944	-721.030	-697.925
Deviance		1467.887	1442.059	1395.850

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

¿Existe una relación estadísticamente significativa entre afiliación sindical y participación en marchas?

Ej. Modelo 3

Politización: variable de intervalo
¿Cómo interpretarla?

	M1 (m prob lineal)	M1 (log odds)	M2 (log odds)	M3 (log odds)
(Intercept)	0.167 ***	-1.601 ***	-1.023 *	-1.379 **
	(0.022)	(0.140)	(0.424)	(0.437)
Unionized	0.072 **	0.418 **	0.390 *	0.312 †
	(0.027)	(0.155)	(0.157)	(0.161)
WaveWave 6	0.079 **	0.470 **	0.498 **	0.496 **
	(0.027)	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.005	-0.031	-0.102	-0.121
	(0.027)	(0.174)	(0.181)	(0.184)
Female			-0.020	0.100
			(0.136)	(0.139)
X003			-0.007	-0.012 *
			(0.006)	(0.006)
Educ2			0.042	-0.053
			(0.268)	(0.272)
Educ3			0.521 †	0.259
			(0.283)	(0.289)
private_sector			-0.516 **	-0.445 *
			(0.175)	(0.180)
politicization				0.282 ***
				(0.042)
R^2	0.013			
Adj. R^2	0.011			
Num. obs.	1457	1457	1457	1457
AIC		1475.887	1460.059	1415.850
BIC		1497.024	1507.616	1468.691
Log Likelihood		-733.944	-721.030	-697.925
Deviance		1467.887	1442.059	1395.850

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

¿Existe una relación estadísticamente significativa entre afiliación sindical y participación en marchas?

	M1 (m prob lineal)	M1 (log odds)	M2 (log odds)	M3 (log odds)
(Intercept)	0.167 ***	-1.601 ***	-1.023 *	-1.379 **
	(0.022)	(0.140)	(0.424)	(0.437)
Unionized	0.072 **	0.418 **	0.390 *	0.312 †
	(0.027)	(0.155)	(0.157)	(0.161)
WaveWave 6	0.079 **	0.470 **	0.498 **	0.496 **
	(0.027)	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.005	-0.031	-0.102	-0.121
	(0.027)	(0.174)	(0.181)	(0.184)
Female			-0.020	0.100
			(0.136)	(0.139)
X003			-0.007	-0.012 *
			(0.006)	(0.006)
Educ2			0.042	-0.053
			(0.268)	(0.272)
Educ3			0.521 †	0.259
			(0.283)	(0.289)
private_sector			-0.516 **	-0.445 *
			(0.175)	(0.180)
politicization				0.282 ***
				(0.042)
R^2	0.013			
Adj. R^2	0.011			
Num. obs.	1457	1457	1457	1457
AIC		1475.887	1460.059	1415.850
BIC		1497.024	1507.616	1468.691
Log Likelihood		-733.944	-721.030	-697.925
Deviance		1467.887	1442.059	1395.850

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

Ej. Modelo 3

Politización: variable de intervalo
¿Cómo interpretarla?

"Por cada unidad de aumento en la escala de politización, el log-odds de participación en marchas aumenta en 0,28 (valor-p < 0,001)".

Medidas de bondad de ajuste de los modelos (hablaremos sobre ellas más adelante)

	M1 (m prob lineal)	M1 (log odds)	M2 (log odds)	M3 (log odds)
(Intercept)	0.167 ***	-1.601 ***	-1.023 *	-1.379 **
	(0.022)	(0.140)	(0.424)	(0.437)
Unionized	0.072 **	0.418 **	0.390 *	0.312 †
	(0.027)	(0.155)	(0.157)	(0.161)
WaveWave 6	0.079 **	0.470 **	0.498 **	0.496 **
	(0.027)	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.005	-0.031	-0.102	-0.121
	(0.027)	(0.174)	(0.181)	(0.184)
Female			-0.020	0.100
			(0.136)	(0.139)
X003			-0.007	-0.012 *
			(0.006)	(0.006)
Educ2			0.042	-0.053
			(0.268)	(0.272)
Educ3			0.521 †	0.259
			(0.283)	(0.289)
private_sector			-0.516 **	-0.445 *
			(0.175)	(0.180)
politicization				0.282 ***
				(0.042)
R^2	0.013			
Adj. R^2	0.011			
Num. obs.	1457	1457	1457	1457
AIC		1475.887	1460.059	1415.850
BIC		1497.024	1507.616	1468.691
Log Likelihood		-733.944	-721.030	-697.925
Deviance		1467.887	1442.059	1395.850

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

Resultado M1 (log odds, comando *summary*)

```
> summary(m1log)

Call:
glm(formula = demonstr_dummy ~ Unionized + Wave, family = binomial(link = "logit"),
    data = WVS_2005_2022_Chl_log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8930  -0.7478  -0.6062  -0.5976   1.9030

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.60103    0.14045 -11.399  < 2e-16 ***
Unionized    0.41774    0.15458   2.702  0.00689 **
WaveWave 6   0.46979    0.16849   2.788  0.00530 **
WaveWave 7  -0.03108    0.17372  -0.179  0.85803
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1487.1  on 1456  degrees of freedom
Residual deviance: 1467.9  on 1453  degrees of freedom
AIC: 1475.9

Number of Fisher Scoring iterations: 4
```

Log-odds – problemas de interpretación

- A pesar de sus ventajas, los coeficientes logit son difíciles de interpretar
  - Los coef. logit son el resultado de una transformación de la escala original
  - Ellos *no* muestran directamente probabilidades (éstas siempre tienen valores entre 0 y 1)

Log-odds – problemas de interpretación

- ¿Qué hacer?
  - Volver a la escala original; ***odds ratios***

- ¿Cómo?
  - Mediante la exponenciación de los coeficientes (la función exponencial es la inversa del logaritmo)

Paso de log-odds a odds

	M1 (m prob lineal)	M1 (log odds)	M2 (log odds)	M3 (log odds)
(Intercept)	0.167 ***	-1.601 ***	-1.023 *	-1.379 **
	(0.022)	(0.140)	(0.424)	(0.437)
Unionized	0.072 **	0.418 **	0.390 *	0.312 †
	(0.027)	(0.155)	(0.157)	(0.161)
WaveWave 6	0.079 **	0.470 **	0.498 **	0.496 **
	(0.027)	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.005	-0.031	-0.102	-0.121
	(0.027)	(0.174)	(0.181)	(0.184)
Female			-0.020	0.100
			(0.136)	(0.139)
X003			-0.007	-0.012 *
			(0.006)	(0.006)
Educ2			0.042	-0.053
			(0.268)	(0.272)
Educ3			0.521 †	0.259
			(0.283)	(0.289)
private_sector			-0.516 **	-0.445 *
			(0.175)	(0.180)
politicization				0.282 ***
				(0.042)
R^2	0.013			
Adj. R^2	0.011			
Num. obs.	1457	1457	1457	1457
AIC		1475.887	1460.059	1415.850
BIC		1497.024	1507.616	1468.691
Log Likelihood		-733.944	-721.030	-697.925
Deviance		1467.887	1442.059	1395.850

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

$$\text{logit}_x = \log(Odds)$$

$$e^{logit} = Odds_x$$

$$e^{0,39} = Odds_x = 1,477$$

Paso de log-odds a odds

	M1 (m prob lineal)	M1 (log odds)	M2 (log odds)	M3 (log odds)
(Intercept)	0.167 ***	-1.601 ***	-1.023 *	-1.379 **
	(0.022)	(0.140)	(0.424)	(0.437)
Unionized	0.072 **	0.418 **	0.390 *	0.312 †
	(0.027)	(0.155)	(0.157)	(0.161)
WaveWave 6	0.079 **	0.470 **	0.498 **	0.496 **
	(0.027)	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.005	-0.031	-0.102	-0.121
	(0.027)	(0.174)	(0.181)	(0.184)
Female			-0.020	0.100
			(0.136)	(0.139)
X003			-0.007	-0.012 *
			(0.006)	(0.006)
Educ2			0.042	-0.053
			(0.268)	(0.272)
Educ3			0.521 †	0.259
			(0.283)	(0.289)
private_sector			-0.516 **	-0.445 *
			(0.175)	(0.180)
politicization				0.282 ***
				(0.042)
R^2	0.013			
Adj. R^2	0.011			
Num. obs.	1457	1457	1457	1457
AIC		1475.887	1460.059	1415.850
BIC		1497.024	1507.616	1468.691
Log Likelihood		-733.944	-721.030	-697.925
Deviance		1467.887	1442.059	1395.850

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

$$\text{logit}_x = \log(Odds)$$

$$e^{logit} = Odds_x$$

$$e^{0,39} = Odds_x = 1,477$$

Las chances (odds) de participar en marchas de los/as sindicalizados/as son 1,5 veces más que las de quienes no están sindicalizados/as, *controlando por las otras variables incluidas en el modelo*

Comparación coeficientes logit v/s odds ratios

Comando básico de R *exp(coef())*

Log odds

	m1 (log odds)	m2 (log odds)	m3 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **
	(0.140)	(0.424)	(0.437)
Unionized	0.418 **	0.390 *	0.312 †
	(0.155)	(0.157)	(0.161)
WaveWave 6	0.470 **	0.498 **	0.496 **
	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121
	(0.174)	(0.181)	(0.184)
Female		-0.020	0.100
		(0.136)	(0.139)
X003		-0.007	-0.012 *
		(0.006)	(0.006)
Educ2		0.042	-0.053
		(0.268)	(0.272)
Educ3		0.521 †	0.259
		(0.283)	(0.289)
private_sector		-0.516 **	-0.445 *
		(0.175)	(0.180)
politicization			0.282 ***
			(0.042)
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

Odds ratio

	m1 (OR)	m2 (OR)	m3 (OR)
(Intercept)	0.202 ***	0.359 *	0.252 **
Unionized	1.519 **	1.477 *	1.366 †
WaveWave 6	1.600 **	1.646 **	1.642 **
WaveWave 7	0.969	0.903	0.886
Female		0.980	1.106
X003		0.993	0.988 *
Educ2		1.042	0.949
Educ3		1.684 †	1.296
private_sector		0.597 **	0.641 *
politicization			1.326 ***
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

¿Qué cambia y no cambia en las tablas?

Comparación coeficientes logit v/s odds ratios

Log odds

	m1 (log odds)	m2 (log odds)	m3 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **
	(0.140)	(0.424)	(0.437)
Unionized	0.418 **	0.390 *	0.312 †
	(0.155)	(0.157)	(0.161)
WaveWave 6	0.470 **	0.498 **	0.496 **
	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121
	(0.174)	(0.181)	(0.184)
Female		-0.020	0.100
		(0.136)	(0.139)
X003		-0.007	-0.012 *
		(0.006)	(0.006)
Educ2		0.042	-0.053
		(0.268)	(0.272)
Educ3		0.521 †	0.259
		(0.283)	(0.289)
private_sector		-0.516 **	-0.445 *
		(0.175)	(0.180)
politicization			0.282 ***
			(0.042)
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

Odds ratio

	m1 (OR)	m2 (OR)	m3 (OR)
(Intercept)	0.202 ***	0.359 *	0.252 **
Unionized	1.519 **	1.477 *	1.366 †
WaveWave 6	1.600 **	1.646 **	1.642 **
WaveWave 7	0.969	0.903	0.886
Female		0.980	1.106
X003		0.993	0.988 *
Educ2		1.042	0.949
Educ3		1.684 †	1.296
private_sector		0.597 **	0.641 *
politicization			1.326 ***
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

¿Qué cambia y no cambia en las tablas?

Comparación coeficientes logit v/s odds ratios

Log odds

	m1 (log odds)	m2 (log odds)	m3 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **
	(0.140)	(0.424)	(0.437)
Unionized	0.418 **	0.390 *	0.312 †
	(0.155)	(0.157)	(0.161)
WaveWave 6	0.470 **	0.498 **	0.496 **
	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121
	(0.174)	(0.181)	(0.184)
Female		-0.020	0.100
		(0.136)	(0.139)
X003		-0.007	-0.012 *
		(0.006)	(0.006)
Educ2		0.042	-0.053
		(0.268)	(0.272)
Educ3		0.521 †	0.259
		(0.283)	(0.289)
private_sector		-0.516 **	-0.445 *
		(0.175)	(0.180)
politicization			0.282 ***
			(0.042)
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

Odds ratio

	m1 (OR)	m2 (OR)	m3 (OR)
(Intercept)	0.202 ***	0.359 *	0.252 **
Unionized	1.519 **	1.477 *	1.366 †
WaveWave 6	1.600 **	1.646 **	1.642 **
WaveWave 7	0.969	0.903	0.886
Female		0.980	1.106
X003		0.993	0.988 *
Educ2		1.042	0.949
Educ3		1.684 †	1.296
private_sector		0.597 **	0.641 *
politicization			1.326 ***
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

¿Qué cambia y no cambia en las tablas?

A pesar de que tenemos los elementos suficientes (coeficientes) para construir tablas de reg. logística, siguen existiendo algunas limitaciones

A pesar de que tenemos los elementos suficientes (coeficientes) para construir tablas de reg. logística, siguen existiendo algunas limitaciones

European Sociological Review VOLUME 26 | NUMBER 1 | 2010 67–82
DOI:10.1093/esr/jcp006, available online at www.esr.oxfordjournals.org
Online publication 9 March 2009

Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It

Carina Mood

Logistic regression estimates do not behave like linear regression estimates in one important respect: They are affected by omitted variables, even when these variables are unrelated to the independent variables in the model. This fact has important implications that have gone largely unnoticed by sociologists. Importantly, we cannot straightforwardly interpret log-odds ratios or odds ratios as effect measures, because they also reflect the degree of unobserved heterogeneity in the model. In addition, we cannot compare log-odds ratios or odds ratios for similar models across groups, samples, or time points, or across models with different independent variables in a sample. This article discusses these problems and possible ways of overcoming them.

European Sociological Review VOLUME 26 | NUMBER 1 | 2010 67–82
DOI:10.1093/esr/jcp006, available online at www.esr.oxfordjournals.org
Online publication 9 March 2009

Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It

Carina Mood

Logistic regression estimates do not behave like linear regression estimates in one important respect: They are affected by omitted variables, even when these variables are unrelated to the independent variables in the model. This fact has important implications that have gone largely unnoticed by sociologists. Importantly, we cannot straightforwardly interpret log-odds ratios or odds ratios as effect measures, because they also reflect the degree of unobserved heterogeneity in the model. In addition, we cannot compare log-odds ratios or odds ratios for similar models across groups, samples, or time points, or across models with different independent variables in a sample. This article discusses these problems and possible ways of overcoming them.

Problema central: Los coeficientes de un modelo de reg. logística (log-odds u odds-ratios) no son comparables con los coeficientes de otro modelo

European Sociological Review VOLUME 26 | NUMBER 1 | 2010 67–82 67
DOI:10.1093/esr/jcp006, available online at www.esr.oxfordjournals.org
Online publication 9 March 2009

Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It

Carina Mood

Logistic regression estimates do not behave like linear regression estimates in one important respect: They are affected by omitted variables, even when these variables are unrelated to the independent variables in the model. This fact has important implications that have gone largely unnoticed by sociologists. Importantly, we cannot straightforwardly interpret log-odds ratios or odds ratios as effect measures, because they also reflect the degree of unobserved heterogeneity in the model. In addition, we cannot compare log-odds ratios or odds ratios for similar models across groups, samples, or time points, or across models with different independent variables in a sample. This article discusses these problems and possible ways of overcoming them.

Problema central: Los coeficientes de un modelo de reg. logística (log-odds u odds-ratios) no son comparables con los coeficientes de otro modelo

Pero hay una solución: calcular las probabilidades predichas (el equivalente a los puntajes predichos en la regresión MCO)

REGRESIÓN LOGÍSTICA: probabilidades predichas

Cálculo de probabilidades predichas

- Ver este modelo simplificado, con una sola variable independiente (página siguiente)

Cálculo de probabilidades predichas

	m0 (log odds)
(Intercept)	-1.422 ***
	(0.074)
Unionized	0.374 *
	(0.153)
AIC	1485.340
BIC	1495.908
Log Likelihood	-740.670
Deviance	1481.340
Num. obs.	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

$$logit(prob\ marcha) = \alpha + \beta_1 X_1$$

Cálculo de probabilidades predichas

A partir de este modelo se pueden predecir log-odds y, más importante aún, probabilidades para personas con distintos atributos controlados en el modelo (ej., sindicalizadas o no)

	m0 (log odds)
(Intercept)	-1.422 ***
	(0.074)
Unionized	0.374 *
	(0.153)
AIC	1485.340
BIC	1495.908
Log Likelihood	-740.670
Deviance	1481.340
Num. obs.	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

$$logit(prob\ marcha) = \alpha + \beta_1 X_1$$

Cálculo de probabilidades predichas

A partir de este modelo se pueden predecir log-odds y, más importante aún, probabilidades para personas con distintos atributos controlados en el modelo (ej., sindicalizadas o no)

	m0 (log odds)
(Intercept)	-1.422 ***
	(0.074)
Unionized	0.374 *
	(0.153)
AIC	1485.340
BIC	1495.908
Log Likelihood	-740.670
Deviance	1481.340
Num. obs.	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

$$logit(prob\ marcha) = \alpha + \beta_1 X_1$$

$$logit(prob\ marcha)_{sindical} = -1{,}422 + (0{,}374 * Unionized = 1) = -1{,}048$$

$$logit(prob\ marcha)_{Nosindical} = -1{,}422 + (0{,}374 * Unionized = 0) = -1{,}422$$

Cálculo de probabilidades predichas

A partir de este modelo se pueden predecir log-odds y, más importante aún, probabilidades para personas con distintos atributos controlados en el modelo (ej., sindicalizadas o no)

	m0 (log odds)
(Intercept)	-1.422 ***
	(0.074)
Unionized	0.374 *
	(0.153)
AIC	1485.340
BIC	1495.908
Log Likelihood	-740.670
Deviance	1481.340
Num. obs.	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

$$logit(prob\ marcha) = \alpha + \beta_1 X_1$$

$$logit(prob\ marcha)_{sindical} = -1{,}422 + (0{,}374 * Unionized = 1) = -1{,}048$$

$$logit(prob\ marcha)_{Nosindical} = -1{,}422 + (0{,}374 * Unionized = 0) = -1{,}422$$

Este “puntaje predicho” (log-odds) no tiene interpretación, por lo que hay que pasarlo a Odds

Cálculo de probabilidades predichas

Transformación de log-odds predichos a odds predichos

$$Odds_x = e^{\alpha+\beta_j X_j}$$

Cálculo de probabilidades predichas

Transformación de log-odds predichos a odds predichos

$$Odds_x = e^{\alpha+\beta_j X_j}$$

$$Odds_{sindicalizados} = e^{-1,048} = 0,35$$

$$Odds_{Nosindicalizados} = e^{-1,422} = 0,24$$

Cálculo de probabilidades predichas

Finalmente, habiendo calculado los odds para cada tipo de persona se pueden calcular sus **probabilidades predichas**

$$p = \frac{e^{\alpha+\beta_j X_j}}{1+e^{\alpha+\beta_j X_j}} = \frac{Odds_{x_j}}{1+Odds_{x_j}}$$

Cálculo de probabilidades predichas

Finalmente, habiendo calculado los odds para cada tipo de persona se pueden calcular sus **probabilidades predichas**

$$p = \frac{e^{\alpha+\beta_j X_j}}{1+e^{\alpha+\beta_j X_j}} = \frac{Odds_{x_j}}{1+Odds_{x_j}}$$

$$p_{sindical} = \frac{0{,}35}{1+0{,}35} = \frac{0{,}35}{1{,}35} = 0{,}26$$

$$p_{nosindical} = \frac{0{,}24}{1+0{,}24} = \frac{0{,}24}{1{,}24} = 0{,}19$$

Cálculo de probabilidades predichas

Finalmente, habiendo calculado los odds para cada tipo de persona se pueden calcular sus **probabilidades predichas**

$$p = \frac{e^{\alpha+\beta_j X_j}}{1+e^{\alpha+\beta_j X_j}} = \frac{Odds_{x_j}}{1+Odds_{x_j}}$$

$$p_{sindical} = \frac{0{,}35}{1+0{,}35} = \frac{0{,}35}{1{,}35} = 0{,}26$$

$$p_{nosindical} = \frac{0{,}24}{1+0{,}24} = \frac{0{,}24}{1{,}24} = 0{,}19$$

La probabilidad de que un/a sindicalizado participe en marchas es del 26%, mientras que la probabilidad de que alguien que no esté sindicalizado/a es del 19%

En resumen…

- La gran dificultad de los modelo de regresión logística está en la transformación que se tiene que hacer de sus coeficientes (de log-odds, a odds y luego a probabilidades)

- Solo luego de realizar esa transformación se pueden estimar probabilidades

- Otra dificultad: la relación entre estas unidades de medida no es intuitiva

Ejemplo ficticio (Cohen et al 2014, p. 489): Predicción de la probabilidad de que un/a académico/a sea ascendido, según el número de publicaciones

Modelo de regresión logística:

$$logit(prob\ ascenso) = \alpha + \beta_1 X_1$$

$$logit(prob\ ascenso) = -6{,}00\ + 0{,}39(num.publicaciones)$$

TABLE 13.2.1
Fictitious Logistic Regression Example Predicting Probability of Promotion to Associate Professor as a Function of Number of Publications

The regression equation is

$$\text{logit(promotion)} = .39\ \text{(publications)} - 6.00.$$

Case	Number of publications	Logit	Odds	Probability	
100	0	−6.00	.00	.00	
101	1	−5.61	.00	.00	
102	2	−5.22	.01	.01	
103	3	−4.83	.01	.01	
104	4	−4.44	.01	.01	
105	5	−4.05	.02	.02	
106	6	−3.66	.03	.03	
107	7	−3.27	.04	.04	
108	8	−2.88	.06	.05	
109	9	−2.49	.08	.08	
110	10	−2.10	.12	.11	
111	11	−1.71	.18	.15	
112	12	−1.32	.27	.21	
113	13	−.93	.39	.28	
114	14	−.54	.58	.37	
115	15	−.15	.86	.46	
	15.38	**.00**	**1.00**	**.50**	hypothetical case with 15.38 publications and exactly .50 probability of promotion.
116	16	.24	1.27	.56	
117	17	.63	1.88	.65	
118	18	1.02	2.77	.73	
119	19	1.41	4.10	.80	
120	20	1.80	6.05	.86	
121	21	2.19	8.94	.90	
122	22	2.58	13.20	.93	
123	23	2.97	19.49	.95	
124	24	3.36	28.79	.97	
125	25	3.75	42.52	.98	
126	26	4.14	62.80	.98	
127	27	4.53	92.76	.99	
128	28	4.92	137.00	.99	
129	29	5.31	202.35	1.00	
130	30	5.70	298.87	1.00	

Ejemplo ficticio (Cohen et al 2014, p. 489): Predicción de la probabilidad de que un/a académico/a sea ascendido, según el número de publicaciones

Modelo de regresión logística:

$$logit(prob\ ascenso) = \alpha + \beta_1 X_1$$

$$logit(prob\ ascenso) = -6{,}00\ + 0{,}39(num.publicaciones)$$

Probabilidades predichas en R (paquete *ggeffects*)

- Paquete *ggeffects* de R: últil para estimar probabilidades predichas a partir de modelos de regresión logísticas

- Combinado con *ggplot2*, se pueden generar gráficos que muestran de modo más intuitivo la relación entre variables

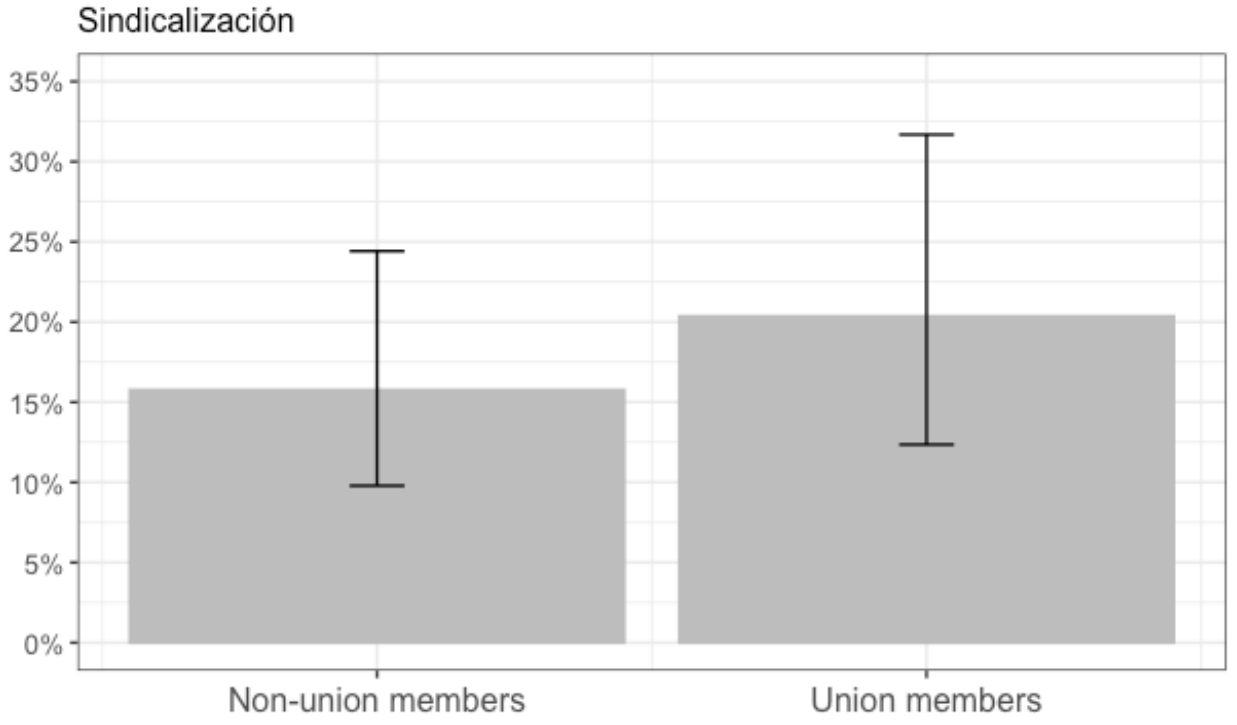
	m1 (log odds)	m2 (log odds)	m3 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **
	(0.140)	(0.424)	(0.437)
Unionized	0.418 **	0.390 *	0.312 †
	(0.155)	(0.157)	(0.161)
WaveWave 6	0.470 **	0.498 **	0.496 **
	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121
	(0.174)	(0.181)	(0.184)
Female		-0.020	0.100
		(0.136)	(0.139)
X003		-0.007	-0.012 *
		(0.006)	(0.006)
Educ2		0.042	-0.053
		(0.268)	(0.272)
Educ3		0.521 †	0.259
		(0.283)	(0.289)
private_sector		-0.516 **	-0.445 *
		(0.175)	(0.180)
politicization			0.282 ***
			(0.042)
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

	m1 (log odds)	m2 (log odds)	m3 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **
	(0.140)	(0.424)	(0.437)
Unionized	0.418 **	0.390 *	0.312 †
	(0.155)	(0.157)	(0.161)
WaveWave 6	0.470 **	0.498 **	0.496 **
	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121
	(0.174)	(0.181)	(0.184)
Female		-0.020	0.100
		(0.136)	(0.139)
X003		-0.007	-0.012 *
		(0.006)	(0.006)
Educ2		0.042	-0.053
		(0.268)	(0.272)
Educ3		0.521 †	0.259
		(0.283)	(0.289)
private_sector		-0.516 **	-0.445 *
		(0.175)	(0.180)
politicization			0.282 ***
			(0.042)
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

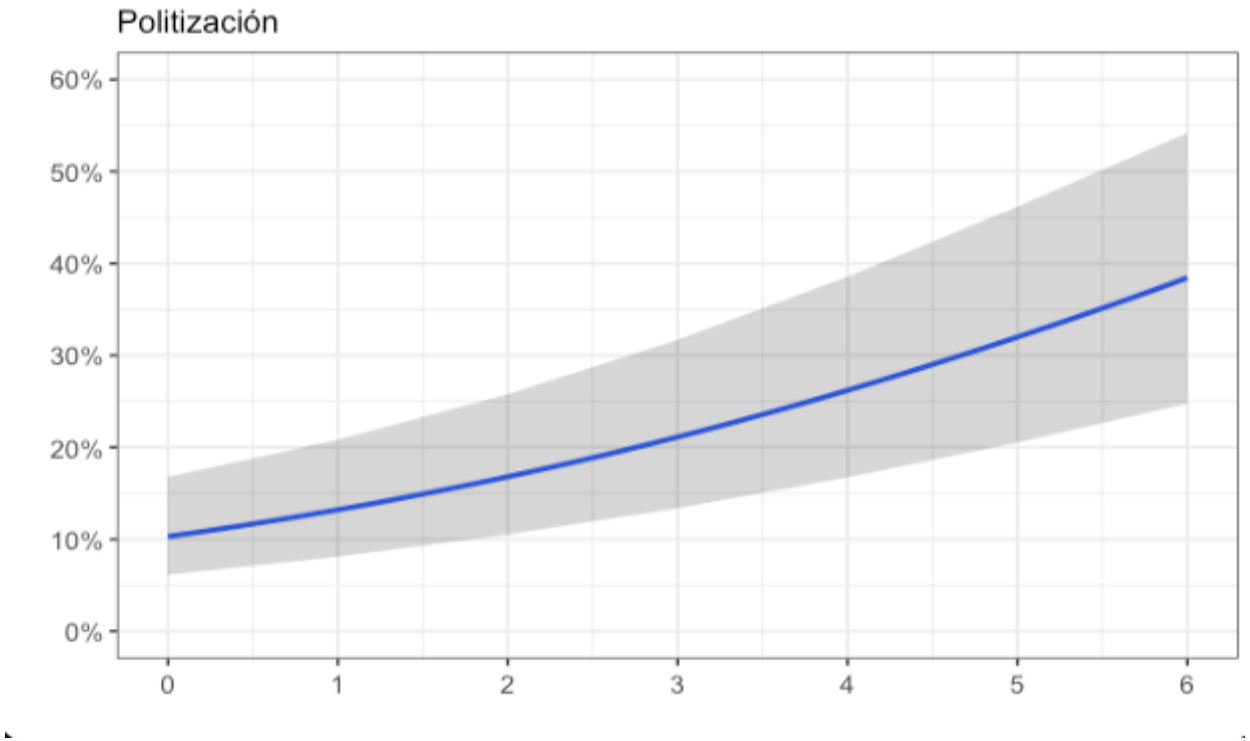
Relación entre sindicalización y prob. participar en marchas (m3)



	m1 (log odds)	m2 (log odds)	m3 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **
	(0.140)	(0.424)	(0.437)
Unionized	0.418 **	0.390 *	0.312 †
	(0.155)	(0.157)	(0.161)
WaveWave 6	0.470 **	0.498 **	0.496 **
	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121
	(0.174)	(0.181)	(0.184)
Female		-0.020	0.100
		(0.136)	(0.139)
X003		-0.007	-0.012 *
		(0.006)	(0.006)
Educ2		0.042	-0.053
		(0.268)	(0.272)
Educ3		0.521 †	0.259
		(0.283)	(0.289)
private_sector		-0.516 **	-0.445 *
		(0.175)	(0.180)
politicization			0.282 ***
			(0.042)
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

Relación entre politización y prob. participar en marchas (m3)

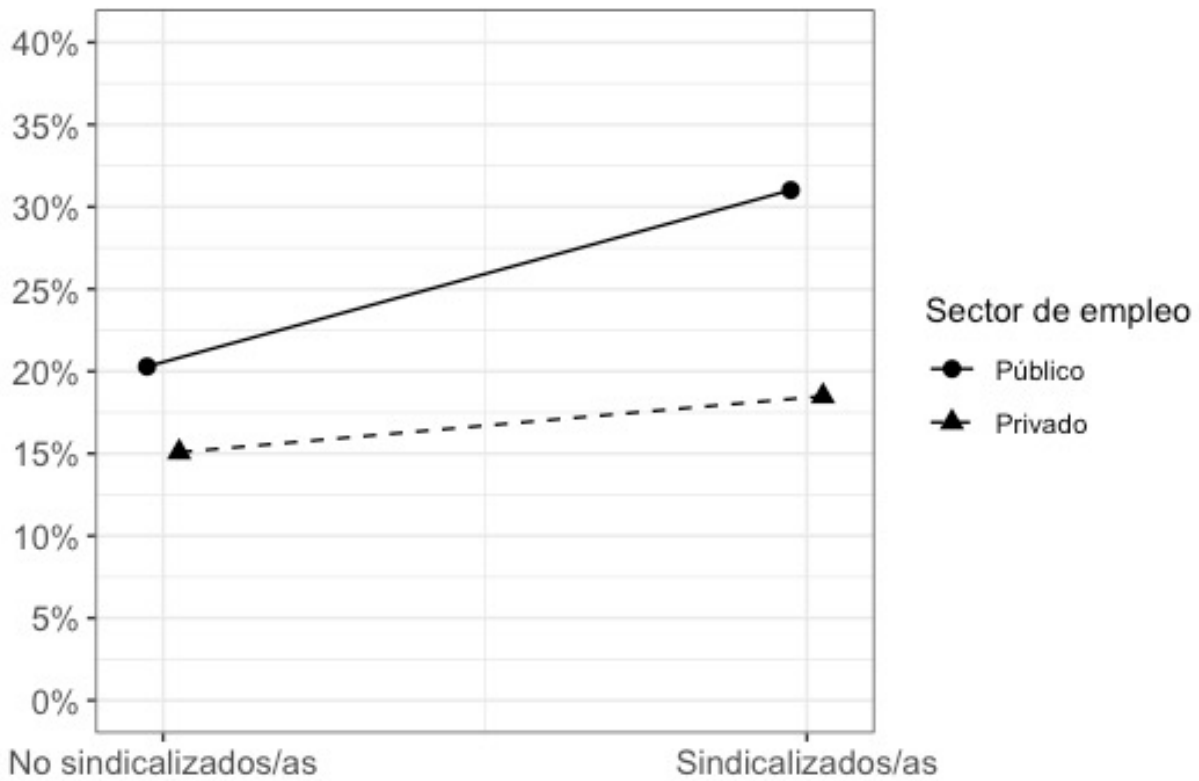


Efecto de interacción sindicalización/sector de empleo (privado)

	M1 (log odds)	M2 (log odds)	M3 (log odds)	M3.1 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **	-1.452 **
	(0.140)	(0.424)	(0.437)	(0.447)
Unionized	0.418 **	0.390 *	0.312 †	0.569
	(0.155)	(0.157)	(0.161)	(0.353)
WaveWave 6	0.470 **	0.498 **	0.496 **	0.496 **
	(0.168)	(0.171)	(0.174)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121	-0.122
	(0.174)	(0.181)	(0.184)	(0.184)
Female		-0.020	0.100	0.094
		(0.136)	(0.139)	(0.140)
X003		-0.007	-0.012 *	-0.012 *
		(0.006)	(0.006)	(0.006)
Educ2		0.042	-0.053	-0.044
		(0.268)	(0.272)	(0.272)
Educ3		0.521 †	0.259	0.259
		(0.283)	(0.289)	(0.289)
private_sector		-0.516 **	-0.445 *	-0.360 †
		(0.175)	(0.180)	(0.210)
politicization			0.282 ***	0.281 ***
			(0.042)	(0.042)
Unionized:private_sector				-0.324
				(0.397)
AIC	1475.887	1460.059	1415.850	1417.185
BIC	1497.024	1507.616	1468.691	1475.310
Log Likelihood	-733.944	-721.030	-697.925	-697.592
Deviance	1467.887	1442.059	1395.850	1395.185
Num. obs.	1457	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

Efecto de interacción sindicalización/sector de empleo (privado)



	M1 (log odds)	M2 (log odds)	M3 (log odds)	M3.1 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **	-1.452 **
	(0.140)	(0.424)	(0.437)	(0.447)
Unionized	0.418 **	0.390 *	0.312 †	0.569
	(0.155)	(0.157)	(0.161)	(0.353)
WaveWave 6	0.470 **	0.498 **	0.496 **	0.496 **
	(0.168)	(0.171)	(0.174)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121	-0.122
	(0.174)	(0.181)	(0.184)	(0.184)
Female		-0.020	0.100	0.094
		(0.136)	(0.139)	(0.140)
X003		-0.007	-0.012 *	-0.012 *
		(0.006)	(0.006)	(0.006)
Educ2		0.042	-0.053	-0.044
		(0.268)	(0.272)	(0.272)
Educ3		0.521 †	0.259	0.259
		(0.283)	(0.289)	(0.289)
private_sector		-0.516 **	-0.445 *	-0.360 †
		(0.175)	(0.180)	(0.210)
politicization			0.282 ***	0.281 ***
			(0.042)	(0.042)
Unionized:private_sector				-0.324
				(0.397)
AIC	1475.887	1460.059	1415.850	1417.185
BIC	1497.024	1507.616	1468.691	1475.310
Log Likelihood	-733.944	-721.030	-697.925	-697.592
Deviance	1467.887	1442.059	1395.850	1395.185
Num. obs.	1457	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

REGRESIÓN LOGÍSTICA: medidas de bondad de ajuste

Medidas de bondad de ajuste de la regresión logística

- A diferencia de la regresión MCO, en la reg. logística no existe un R^2 que muestre la cantidad de varianza explicada por el modelo
  - Por ello, en la reg. logística se usan otras medidas para evaluar en la calidad del modelo

- Estas medidas se calculan a partir del proceso a través del cual se estiman los coeficientes de la regresión logística: *máxima verosimilitud*

- En otras palabras, estas medidas de ajuste se basan en el concepto de **log-verosimilitud (LL o *log-likelihood*)**, que evalúa un modelo en función de sus **residuos** (lo no explicado por el modelo)
  - Hay *varias* medidas complementarias que se analizan *comparativamente*

	m1 (log odds)	m2 (log odds)	m3 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **
	(0.140)	(0.424)	(0.437)
Unionized	0.418 **	0.390 *	0.312 †
	(0.155)	(0.157)	(0.161)
WaveWave 6	0.470 **	0.498 **	0.496 **
	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121
	(0.174)	(0.181)	(0.184)
Female		-0.020	0.100
		(0.136)	(0.139)
X003		-0.007	-0.012 *
		(0.006)	(0.006)
Educ2		0.042	-0.053
		(0.268)	(0.272)
Educ3		0.521 †	0.259
		(0.283)	(0.289)
private_sector		-0.516 **	-0.445 *
		(0.175)	(0.180)
politicization			0.282 ***
			(0.042)
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

LL y Devianza

Log-likelihood (razón de verosimilitud)

Medida que indica el grado de ajuste de cada modelo.

Tiene valores $-\infty$ a $+\infty$. Mayor puntaje indica mejor ajuste

	m1 (log odds)	m2 (log odds)	m3 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **
	(0.140)	(0.424)	(0.437)
Unionized	0.418 **	0.390 *	0.312 †
	(0.155)	(0.157)	(0.161)
WaveWave 6	0.470 **	0.498 **	0.496 **
	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121
	(0.174)	(0.181)	(0.184)
Female		-0.020	0.100
		(0.136)	(0.139)
X003		-0.007	-0.012 *
		(0.006)	(0.006)
Educ2		0.042	-0.053
		(0.268)	(0.272)
Educ3		0.521 †	0.259
		(0.283)	(0.289)
private_sector		-0.516 **	-0.445 *
		(0.175)	(0.180)
politicization			0.282 ***
			(0.042)
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

LL y Devianza

Log-likelihood

Medida que indica el grado de ajuste de cada modelo.

Tiene valores $-\infty$ a $+\infty$. Mayor puntaje indica mejor ajuste

Devianza

Medida de la distancia entre el ajuste del modelo y una situación ideal de ajuste perfecto

Devianza = -2*log-likelihood

Tiene valores de 0 a $+\infty$. *Menor* puntaje indica mejor ajuste

LL y Devianza

	m1 (log odds)	m2 (log odds)	m3 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **
	(0.140)	(0.424)	(0.437)
Unionized	0.418 **	0.390 *	0.312 †
	(0.155)	(0.157)	(0.161)
WaveWave 6	0.470 **	0.498 **	0.496 **
	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121
	(0.174)	(0.181)	(0.184)
Female		-0.020	0.100
		(0.136)	(0.139)
X003		-0.007	-0.012 *
		(0.006)	(0.006)
Educ2		0.042	-0.053
		(0.268)	(0.272)
Educ3		0.521 †	0.259
		(0.283)	(0.289)
private_sector		-0.516 **	-0.445 *
		(0.175)	(0.180)
politicization			0.282 ***
			(0.042)
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

Log-likelihood

Medida que indica el grado de ajuste de cada modelo.

Tiene valores $-\infty$ a $+\infty$. Mayor puntaje indica mejor ajuste

Devianza

Medida de la distancia entre el ajuste del modelo y una situación ideal de ajuste perfecto

Devianza = -2*log-likelihood

Tiene valores de 0 a $+\infty$. *Menor* puntaje indica mejor ajuste

Problema: más variables en un modelo van a aumentar siempre el ajuste del modelo

LL y Devianza

	m1 (log odds)	m2 (log odds)	m3 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **
	(0.140)	(0.424)	(0.437)
Unionized	0.418 **	0.390 *	0.312 †
	(0.155)	(0.157)	(0.161)
WaveWave 6	0.470 **	0.498 **	0.496 **
	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121
	(0.174)	(0.181)	(0.184)
Female		-0.020	0.100
		(0.136)	(0.139)
X003		-0.007	-0.012 *
		(0.006)	(0.006)
Educ2		0.042	-0.053
		(0.268)	(0.272)
Educ3		0.521 †	0.259
		(0.283)	(0.289)
private_sector		-0.516 **	-0.445 *
		(0.175)	(0.180)
politicization			0.282 ***
			(0.042)
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

Log-likelihood

Medida que indica el grado de ajuste de cada modelo.

Tiene valores $-\infty$ a $+\infty$. Mayor puntaje indica mejor ajuste

Devianza

Medida de la distancia entre el ajuste del modelo y una situación ideal de ajuste perfecto

Devianza = -2*log-likelihood

Tiene valores de 0 a $+\infty$. *Menor* puntaje indica mejor ajuste

Problema: más variables en un modelo van a aumentar siempre el ajuste del modelo

Solución: la Prueba de la Razón de Verosimilitud – Likelihood Ratio Test (comando *anova* en R; también se puede con paquete *lmtest*)

Likelihood Ratio Test

Comparación m1 – m2

```
> anova(m1log, m2log, test = "Chisq")
Analysis of Deviance Table

Model 1: demonstr_dummy ~ Unionized + Wave
Model 2: demonstr_dummy ~ Unionized + Female + X003 + Educ + private_sector +
    Wave
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1453     1467.9
2      1448     1442.1  5   25.828 9.635e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ho = no existen diferencias significativas entre los modelos

Ha = el modelo 2 se ajusta sign. mejor a los datos que el modelo 1

Likelihood Ratio Test

Comparación m2 – m3

```
> anova(m2log, m3log, test = "Chisq")
Analysis of Deviance Table

Model 1: demonstr_dummy ~ Unionized + Female + X003 + Educ + private_sector +
    Wave
Model 2: demonstr_dummy ~ Unionized + Female + X003 + Educ + private_sector +
    politicization + Wave
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1448     1442.1
2      1447     1395.8  1    46.21 1.063e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ho = no existen diferencias significativas entre los modelos

Ha = el modelo 3 se ajusta sign. mejor a los datos que el modelo 2

	m1 (log odds)	m2 (log odds)	m3 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **
	(0.140)	(0.424)	(0.437)
Unionized	0.418 **	0.390 *	0.312 †
	(0.155)	(0.157)	(0.161)
WaveWave 6	0.470 **	0.498 **	0.496 **
	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121
	(0.174)	(0.181)	(0.184)
Female		-0.020	0.100
		(0.136)	(0.139)
X003		-0.007	-0.012 *
		(0.006)	(0.006)
Educ2		0.042	-0.053
		(0.268)	(0.272)
Educ3		0.521 †	0.259
		(0.283)	(0.289)
private_sector		-0.516 **	-0.445 *
		(0.175)	(0.180)
politicization			0.282 ***
			(0.042)
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

Akaike information criterion (AIC)

AIC

Compara calidad de ajuste de modelos, pero corrige por la inclusión de variables

$$\textbf{AIC} = 2K - 2*(\log\text{-likelihood})$$

donde

K = cantidad de parámetros del modelo (variables + intercepto)

Menor AIC indica mejor ajuste

Bayesian information criterion (BIC)

	m1 (log odds)	m2 (log odds)	m3 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **
	(0.140)	(0.424)	(0.437)
Unionized	0.418 **	0.390 *	0.312 †
	(0.155)	(0.157)	(0.161)
WaveWave 6	0.470 **	0.498 **	0.496 **
	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121
	(0.174)	(0.181)	(0.184)
Female		-0.020	0.100
		(0.136)	(0.139)
X003		-0.007	-0.012 *
		(0.006)	(0.006)
Educ2		0.042	-0.053
		(0.268)	(0.272)
Educ3		0.521 †	0.259
		(0.283)	(0.289)
private_sector		-0.516 **	-0.445 *
		(0.175)	(0.180)
politicization			0.282 ***
			(0.042)
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

BIC

Similar al AIC, compara calidad de ajuste de modelos corrigiendo por la cantidad de variables incluidas y tamaño de la muestra

$$\textbf{BIC} = K*\log(N) - 2*(\log\text{-likelihood})$$

donde

K = cantidad de parámetros del modelo (variables + intercepto)

N= tamaño muestra

*Menor B*IC indica mejor ajuste. En BIC la penalización por cantidad de parámetros es más alta, por lo que el ajuste reportado es *menor*

Otra medida de ajuste: Pseudo R^2 (McFadden)

- El Pseudo R^2 de McFadden es *una* de las formas de calcular R^2 en regresión logística (hay más, como el R^2 de Cox & Snell y el R^2 de Nagelkerke)

$$Pseudo\ R^2 = 1 - \frac{\ln(L_M)}{\ln(L_0)}$$

Donde:

ln(LM) = log-likelihood del modelo que está siendo evaluado

ln(L0) = log-likelihood del modelo nulo (sin predictores)

- En R este Pseudo R^2 (y los otros) se pueden obtener con el paquete *DescTool*

Tabla de regresión con Pseudo R^2 (McFadden)

	m1 (log odds)	m2 (log odds)	m3 (log odds)
(Intercept)	-1.601 ***	-1.023 *	-1.379 **
	(0.140)	(0.424)	(0.437)
Unionized	0.418 **	0.390 *	0.312 †
	(0.155)	(0.157)	(0.161)
WaveWave 6	0.470 **	0.498 **	0.496 **
	(0.168)	(0.171)	(0.174)
WaveWave 7	-0.031	-0.102	-0.121
	(0.174)	(0.181)	(0.184)
Female		-0.020	0.100
		(0.136)	(0.139)
X003		-0.007	-0.012 *
		(0.006)	(0.006)
Educ2		0.042	-0.053
		(0.268)	(0.272)
Educ3		0.521 †	0.259
		(0.283)	(0.289)
private_sector		-0.516 **	-0.445 *
		(0.175)	(0.180)
politicization			0.282 ***
			(0.042)
Pseudo R2	0.013	0.030	0.061
AIC	1475.887	1460.059	1415.850
BIC	1497.024	1507.616	1468.691
Log Likelihood	-733.944	-721.030	-697.925
Deviance	1467.887	1442.059	1395.850
Num. obs.	1457	1457	1457

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.1

En resumen

- Existen varias formas para chequear la calidad de los modelos

- Todas ellas funcionan comparativamente

- Regla general (según mi experiencia): reportar
  - LL
  - Devianza
  - AIC/BIC
  - Cuando sea necesario, Incluir alguna discusión sobre el *likelihood ratio test*
  - Pseudo R^2, solo como complemento a las otras medidas (nunca por sí solo)