

Análisis de regresión logística

Sesión N° 5
Análisis Avanzado de Datos II

Profesor Gabriel Sotomayor López

Contenidos Sesión N°5

Modelos de probabilidad lineal

Modelo de regresión logística binaria

Odds y odds ratio

Contenidos Sesión N°5

Cálculo de modelos e interpretación de coeficientes

Estadísticos de ajuste y selección de modelos

Modelos de probabilidad lineal

Regresiones para variables dicotómicas

La semana pasada revisamos los modelos de regresión lineal múltiple, que nos permiten analizar la relación de una variable dependiente continua y variables independientes de cualquier nivel de medida. Ahora cabe la pregunta:

¿cómo podemos modelar variables dicotómicas?

Modelos de probabilidad lineal

Una opción son los modelos de probabilidad lineal, los cuales consisten en usar una regresión estimada mediante mínimos cuadrados ordinarios para una variable dicotómica (valores 0 y 1). En estos los valores beta pueden interpretarse como cambios promedio en la probabilidad.

Pobreza según sexo JH	
Intercepto	0.08 ^{***}
	(0.00)
Mujer (ref.hombre)	0.04 ^{***}
	(0.00)
R ²	0.00
Adj. R ²	0.00
Num. obs.	62911

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

Modelos de probabilidad lineal

A pesar de la simpleza de su interpretación, los modelos de probabilidad lineal cuentan con dos problemas:

- 1) Pueden entregar variables predichos más allá del rango 0-1 lo cual no tiene sentido en el caso de una probabilidad**
- 2) No entregan un buen ajuste en términos de cumplimiento de los supuestos del modelo y ajuste a los datos**

Modelo de regresión logística binaria

Regresión logística

Una solución a los problemas anteriormente revisado es utilizar otro tipo de modelo: un modelo de regresión logística binaria. En lugar de modelar la probabilidad directamente, hacemos una transformación la variable dependiente: modelamos el logaritmo de los odds (chances).

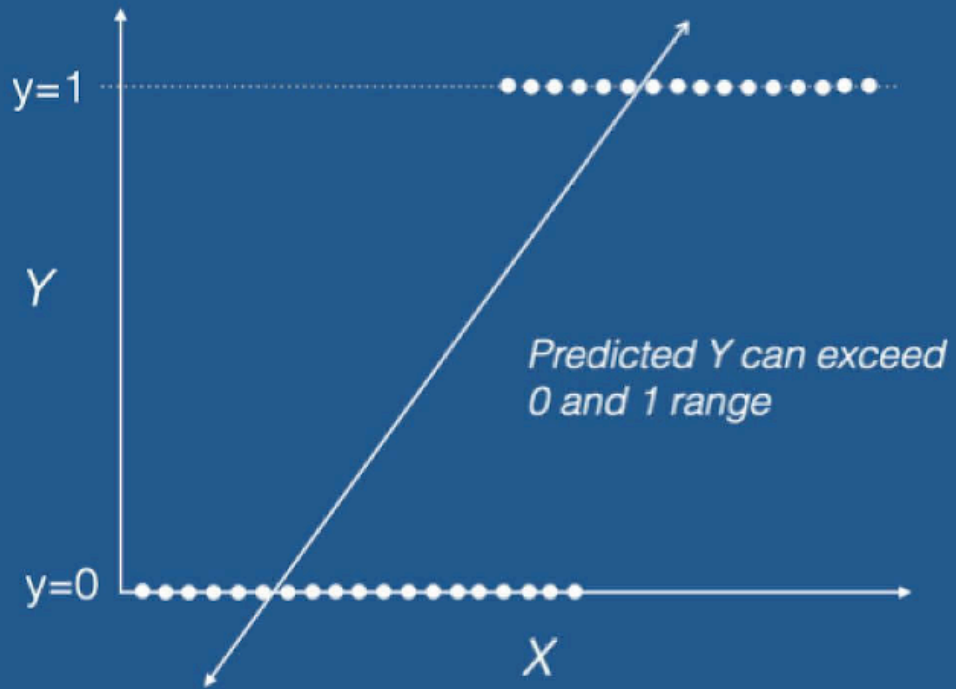
a.k.a. **Log Odds**
or **Logit**

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X$$

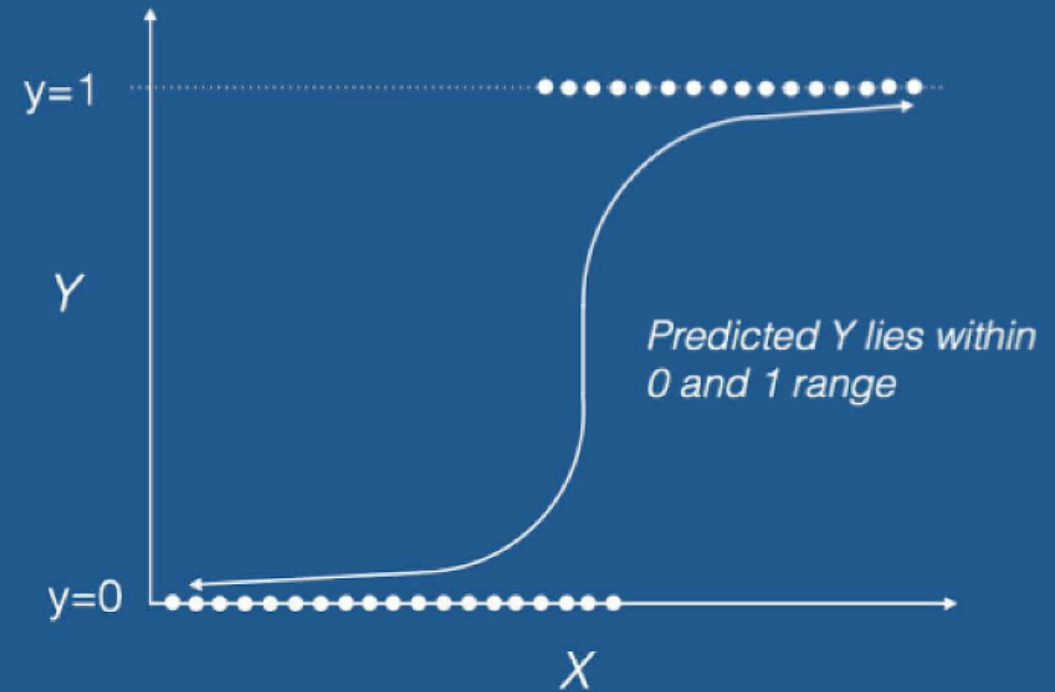
Intercept

Regresión logística

Linear Regression



Logistic Regression



Odds y odds ratio

Concepto de Odds

Los odds (chances) corresponden a la razón entre la probabilidad de que algo ocurra dividido por la probabilidad de que algo no ocurra.

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

Ejemplo en CASON 2020, odds de ser pobre

Odds pobreza = $0.095/0.905 = 0.105$

Es decir, de acuerdo a la CASON 2020, las chances de ser pobre son de 0.105

Concepto de Odds

Odds de **1** significa que existen chances iguales de la correnencia o no ocurrencia de cierto hecho.

Odds **menores de 1** dan cuenta de chances negativas (es más probable que ocurra a que no ocurra)

Odds **mayores a 1** dan cuenta de chances positivas (es más probable que ocurra a que no ocurra)

Odds de dos grupos

En nuestro ejemplo original queremos ver como cambia la probabilidad de que un hogar este en la pobreza según sexo del jefe de hogar.

$$Odds_{JH-Hombre} = 0.076/0.924 = 0.08225$$

$$Odds_{JH-Mujer} = 0.114/0.886 = 0.1286$$

Es decir, existen 8,22 hogares con jefatura masculina en situación de pobreza por cada 100 que no lo están, mientras que 12,86 hogares con jefatura femenina en situación de pobreza por cada 100 que no lo están.

Concepto de Odds ratio

Los odds ratio resultan útiles para comparar la asociación entre las chances de dos variables dicotómicas.

OR de pobreza de un hogar con jefatura masculina / OR de pobreza de un hogar con jefatura masculina

$$\begin{aligned} Odds\ ratio &= \frac{p_m(1 - p_m)}{p_h(1 - p_h)} \\ &= \frac{0.114/0.886}{0.076/0.924} = \frac{0.1286}{0.08225} = 1.564 \end{aligned}$$

Concepto de Odds ratio

Las chances de un hogar con jefatura femenina de encontrarse en situación de pobreza son 1,564 veces mayores a las chances de un hogar con jefatura masculina.

Los odds ratio nos permiten resumir en un número la relación entre dos variables categóricas

Ahora con estos conceptos en mente pasemos a ver como se ajusta un modelo de regresión logística.

Cálculo de modelos e interpretación de coeficientes

Cálculo de modelo en R: función GLM

Para estimar un modelo de regresión logística binaria en R debemos usar la función `glm`, incluida en `r base`

```
glm(pobre~as_factor(sexo),  
    data=base[base$pcol==1,], # filtro por JH  
    family="binomial")
```

Especificamos la formula igual que en una regresión lineal (la variable dependiente debe estar en formato 0-1). Debemos especificar la familia de modelos (ya que la función `glm` sirve para calcular distintos tipos de modelos lineales generalizados).

Los beta de un modelo de regresión logística están puestos en términos del logaritmo de los odds. Es decir, el beta de mujer nos indica que los log-odds de encontrarse en situación de pobreza aumentan en 0.41 en las mujeres en relación a los hombres.

	Pobreza según sexo JH
Intercepto	-2.40*** (0.02)
Mujer (ref.hombre)	0.41*** (0.03)
AIC	41074.58
BIC	41092.68
Log Likelihood	-20535.29
Deviance	41070.58
Num. obs.	62911

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

Interpretación de coeficientes

Para poder realizar una interpretación con sentido de los coeficientes del modelo debemos realizar una transformación, de forma que el beta quede expresado como odds. Para esto debemos hacer una **exponenciación** de los coeficientes.

$$\exp(0.41) = 1.506818$$

Es decir, los odds (chances) de ser pobre para un hogar con jefatura femenina son 1,506818 veces más que las de uno con jefatura masculina.

Modelos con múltiples variables independientes

Pobreza según sexo JH	
Intercepto	-1.07 (0.05) ^{***}
Mujer (ref.hombre)	0.39 (0.03) ^{***}
Edad	-0.03 (0.00) ^{***}
AIC	40182.30
BIC	40209.45
Log Likelihood	-20088.15
Deviance	40176.30
Num. obs.	62911
*** p < 0.001; ** p < 0.01; * p < 0.05	

Statistical models

Modelos con múltiples variables independientes

En el caso de un modelo de regresión logística con múltiples predictores o variables independientes la interpretación es similar, pero integrando el concepto de control estadístico.

En este caso podemos decir que los log-odds predichos de encontrarse en situación de pobreza aumentan en 0,39 en los hogares con jefatura femenina respecto aquellos con jefatura masculina **controlando por edad**

En el mismo sentido, los log-odds predichos de ser encontrarse en situación de pobreza disminuyen en 0.03 por cada año más de edad del jefe de hogar, **controlando por sexo.**

Estadísticos de ajuste y selección de modelos

Estadísticos de ajuste

En el caso de los modelos de regresión logística no contamos con una sola medida de ajuste de los modelos y esta suele interpretarse principalmente en términos comparativos.

4 principales aproximaciones

- Devianza
- Test de Razón de Verosimilitud
- Pseudo R²s
- Criterios de Información (AIC y BIC)

Devianza

- La devianza es una medida de ajuste que se utiliza para evaluar la calidad de un modelo de regresión logística binaria. Es la diferencia entre la log-verosimilitud del modelo ajustado y la log-verosimilitud del modelo nulo, multiplicada por -2. Por eso también se conoce como devianza residual.
- Formula: **Devianza = -2*log likelihood**

Test de razón de verosimilitud

Test de razón de verosimilitud: es un test estadístico utilizado para comparar dos modelos, uno más simple (modelo nulo) y otro más complejo (modelo ajustado). El test de LR se basa en la comparación de la devianza de ambos modelos, y se utiliza para determinar si el modelo ajustado mejora significativamente la predicción en comparación con el modelo nulo.

Criterios de información (AIC y BIC)

Criterios de información: son medidas utilizadas para comparar diferentes modelos y seleccionar el modelo más adecuado. Los dos criterios de información más comunes son el Akaike Information Criterion (AIC) y el **Bayesian Information Criterion (BIC)**. Estos criterios toman en cuenta tanto la bondad de ajuste del modelo como la complejidad del modelo. Un modelo con un valor de AIC o BIC más bajo se considera mejor ajustado que un modelo con un valor más alto.

Pseudo R2s

Son medidas de la variabilidad explicada por el modelo. No pueden interpretarse como proporción de la varianza explicada, como en el caso de los modelos con variable dependientes continuas.

Existen varios tipos de PseudoR2, cada uno con una interpretación diferente. Uno de los PseudoR2 más comunes es el McFadden's PseudoR2. Se define como: $1 - [LL(LM)/LL(L0)]$, donde

- LL es el log likelihood del modelo
- LM es el modelo posterior (con más predictores)
- L0 es el modelo nulo

Un valor cercano a 1 indica que el modelo explica una gran cantidad de la variabilidad en los datos, mientras que un valor cercano a 0 indica que el modelo no explica mucha variabilidad.

Análisis de regresión logística

Sesión N° 5
Análisis Avanzado de Datos II

Profesor Gabriel Sotomayor López