

# Clase 8

# Análisis de regresión

# Logística

Análisis Avanzado de Datos

Gabriel Sotomayor



# Modelos de probabilidad lineal



# Regresiones para variables dicotómicas

Las semanas anteriores 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.

Statistical models

|                     | Pobreza según sexo JH |
|---------------------|-----------------------|
| Intercepto          | 0.08***<br>(0.00)     |
| Mujer (ref.hombre)  | 0.04***<br>(0.00)     |
| R <sup>2</sup>      | 0.00                  |
| Adj. R <sup>2</sup> | 0.00                  |
| Num. obs.           | 62911                 |



# 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).

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



# Regresión logística

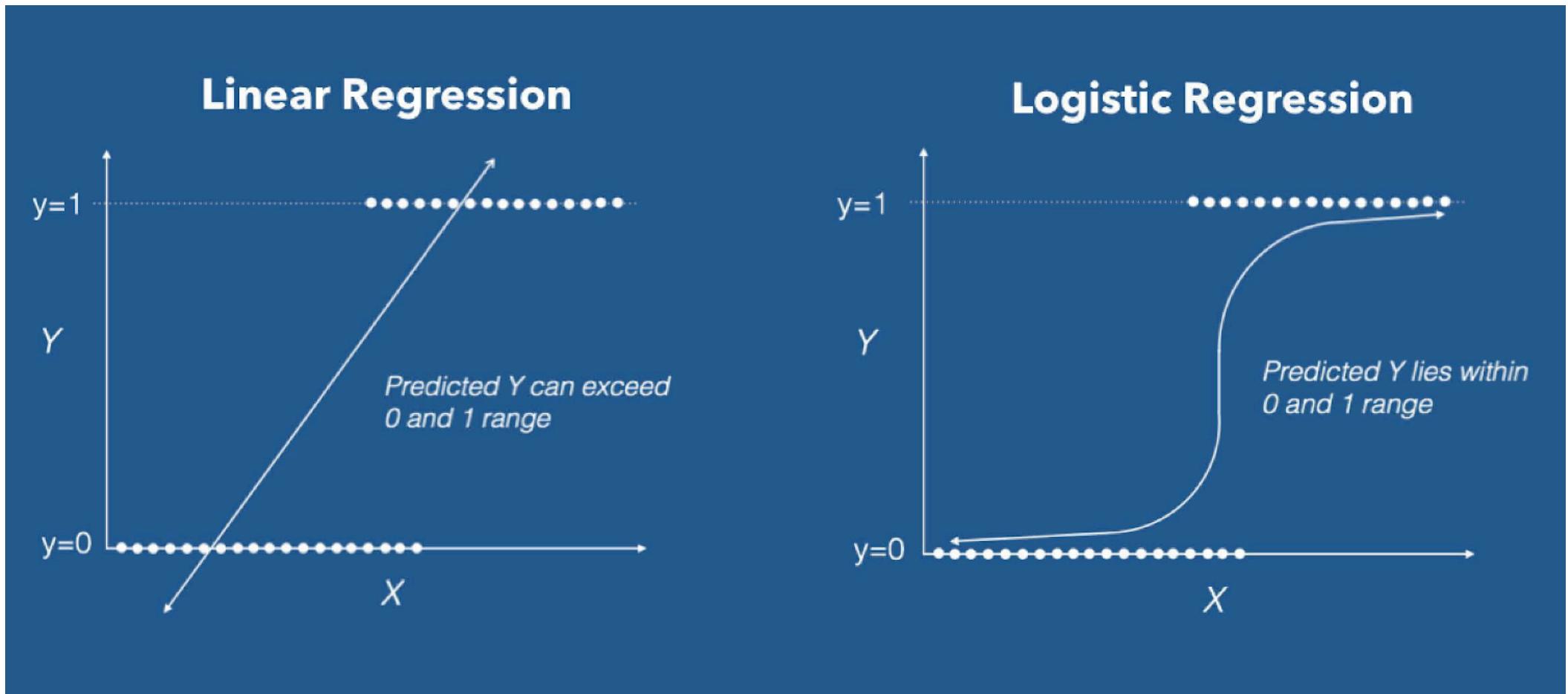
$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X$$

- $\log\left(\frac{P}{1-P}\right)$ : Esto representa el **logit** o **log odds**. El término  $P$  es la probabilidad de que ocurra el evento de interés, y  $1 - P$  es la probabilidad de que no ocurra. Al tomar el logaritmo de las “odds” o probabilidades, transformamos el rango de  $P$  (que va de 0 a 1) a un rango de  $-\infty$  a  $\infty$ .
- $\beta_0$ : Es el **intercepto**. Este valor indica el valor de los log odds cuando la variable independiente  $X$  es igual a 0.
- $\beta_1 X$ : Es el **coeficiente** que acompaña a la variable independiente  $X$ . Representa el cambio en los log odds por cada unidad adicional de  $X$ . Si  $X$  aumenta en una unidad, los log odds se incrementarán o disminuirán dependiendo del valor de  $\beta_1$ .

Este es un modelo de regresión logística, donde la relación entre  $X$  y la probabilidad  $P$  de un evento se modela de manera no lineal, usando la función logit para transformar la probabilidad.



# Regresión logistica



# Probabilidad, odds y odds ratio



# Probabilidad y Odds

- **Probabilidad (p):** Representa la probabilidad de ocurrencia de un evento y toma valores entre 0 y 1.
- **Odds:** Es la razón entre la probabilidad de que ocurra un evento y la probabilidad de que no ocurra, es decir:

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



# Probabilidad y Odds

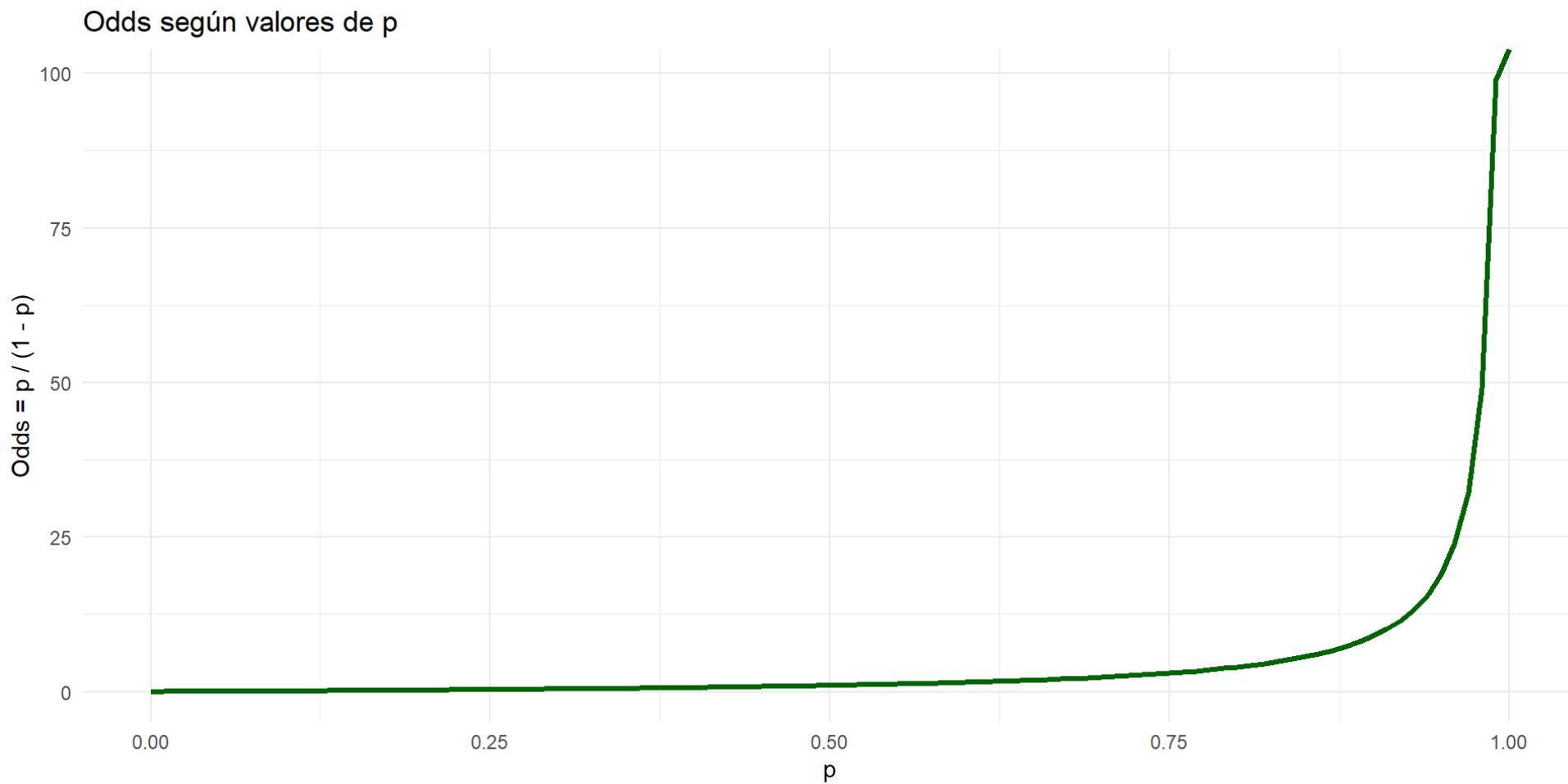
```
1 p <- seq(0, 1, 0.1)
2 odds <- p / (1 - p)
3
4 print(data.frame(p, odds))
```

|    | p   | odds      |
|----|-----|-----------|
| 1  | 0.0 | 0.0000000 |
| 2  | 0.1 | 0.1111111 |
| 3  | 0.2 | 0.2500000 |
| 4  | 0.3 | 0.4285714 |
| 5  | 0.4 | 0.6666667 |
| 6  | 0.5 | 1.0000000 |
| 7  | 0.6 | 1.5000000 |
| 8  | 0.7 | 2.3333333 |
| 9  | 0.8 | 4.0000000 |
| 10 | 0.9 | 9.0000000 |
| 11 | 1.0 | Inf       |



# Gráfico: Odds según valores de p

Este gráfico ilustra cómo varían los odds a medida que cambia la probabilidad ( $p$ ). Podemos observar que, cuando  $p$  se acerca a 1, los odds tienden a crecer exponencialmente.



# Ejemplo 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 CASEN 2020, odds de ser pobre

Odds pobreza = 0.095/0.905 = 0.105

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



# Concepto de Odds

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

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

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



# Logit

- **Logit:** La función logit se define como el logaritmo natural de los odds:

$$\text{Logit} = \ln\left(\frac{p}{1 - p}\right)$$

Es la función de enlace que utilizamos en la regresión logística para transformar la probabilidad en un valor continuo entre  $(-\infty, \infty)$ .



# Probabilidad, Odds y logaritmo de los odds (logit)

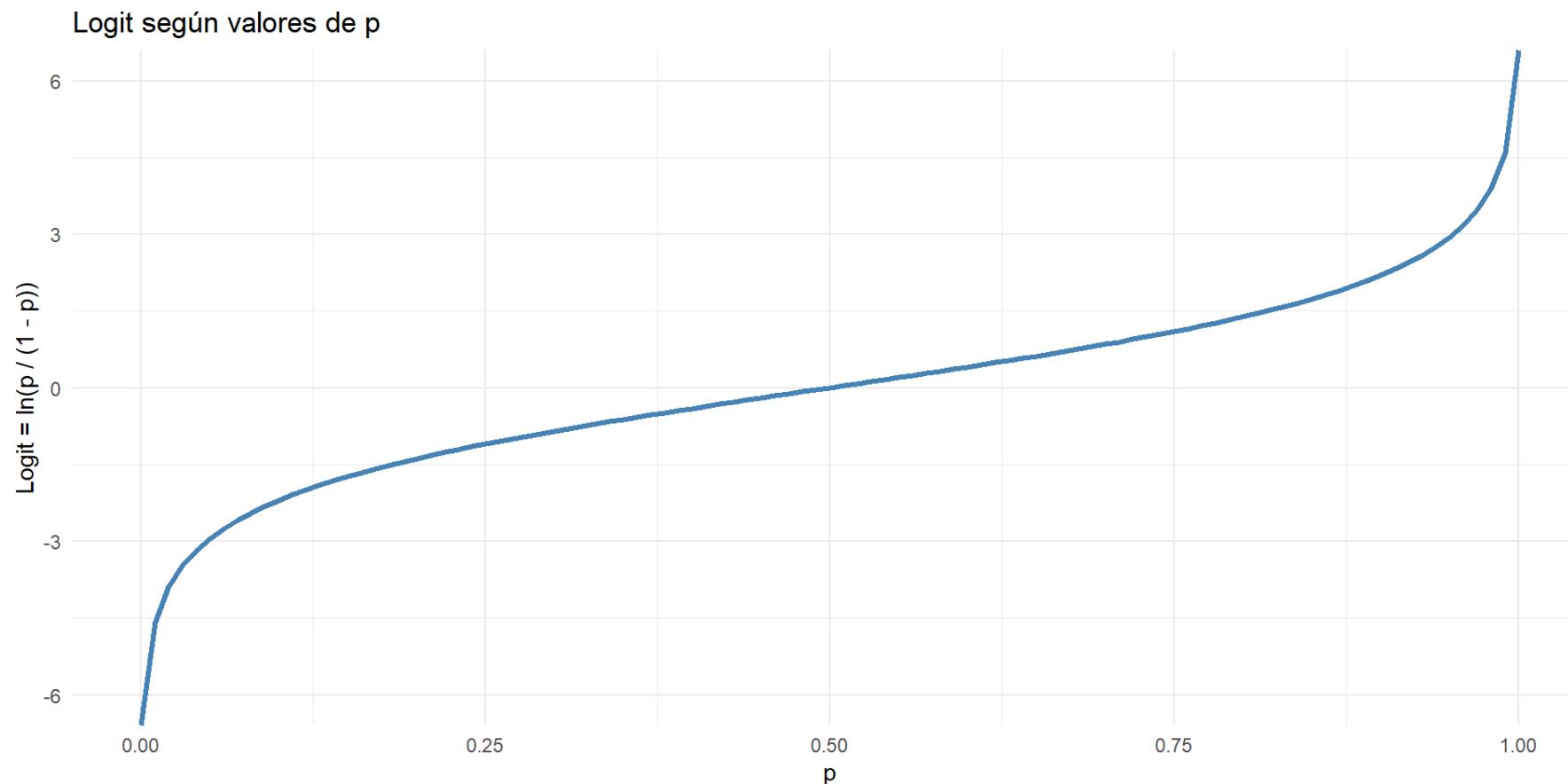
```
1 p <- seq(0, 1, 0.1)
2 odds <- p / (1 - p)
3 logit <- log(p / (1 - p))
4
5 print(data.frame(p, odds, logit))
```

|    | p   | odds      | logit      |
|----|-----|-----------|------------|
| 1  | 0.0 | 0.0000000 | -Inf       |
| 2  | 0.1 | 0.1111111 | -2.1972246 |
| 3  | 0.2 | 0.2500000 | -1.3862944 |
| 4  | 0.3 | 0.4285714 | -0.8472979 |
| 5  | 0.4 | 0.6666667 | -0.4054651 |
| 6  | 0.5 | 1.0000000 | 0.0000000  |
| 7  | 0.6 | 1.5000000 | 0.4054651  |
| 8  | 0.7 | 2.3333333 | 0.8472979  |
| 9  | 0.8 | 4.0000000 | 1.3862944  |
| 10 | 0.9 | 9.0000000 | 2.1972246  |
| 11 | 1.0 | Inf       | Inf        |



# Gráfico: Logit según valores de p

Este gráfico muestra cómo varía el logit con respecto a  $p$ . A diferencia de los odds, el logit transforma la probabilidad en una escala lineal, lo cual es esencial para poder aplicar un modelo lineal en la regresión logística.



# 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

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



# 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`

```
1 glm(pobre~as_factor(sexo),  
2      data=base[base$pco1==1,], # filtro por JH  
3      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).



# Interpretación de coeficiente

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.

| Model 1                                |                     |
|--|---------------------|
| Intercepto                             | -2.40 ***<br>(0.02) |
| Mujer (ref.hombre)                     | 0.41 ***<br>(0.03)  |
| AIC                                    | 41074.58            |
| BIC                                    | 41092.68            |
| Log Likelihood                         | -20535.29           |
| *** p < 0.001; ** p < 0.01; * p < 0.05 |                     |



## Model 1

Deviance 41070.58

Num. obs. 62911

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

Statistical models



# 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.

$$1 \exp(0.41) = 1.506818$$

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.



# Interpretación de coeficientes

| Pobreza según sexo JH |                 |
|-----------------------|-----------------|
| Intercepto            | 0.09 (0.02) *** |
| Mujer (ref.hombre)    | 1.51 (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

Statistical models



# Modelos con múltiples variables independientes

| Pobreza según sexo JH |                  |
|-----------------------|------------------|
| Intercepto            | -1.50 (0.02) *** |
| Mujer (ref.hombre)    | 0.11 (0.01) ***  |
| Edad                  | -0.02 (0.00) *** |
| AIC                   | 129517.95        |
| BIC                   | 129548.34        |
| Log Likelihood        | -64755.98        |
| Deviance              | 129511.95        |
| Num. obs.             | 185339           |

\*\*\* 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**.



# Modelos con múltiples variables independientes

| Pobreza según sexo JH (odd ratios) |                 |
|------------------------------------|-----------------|
| Intercepto                         | 0.22 (0.02) *** |
| Mujer (ref.hombre)                 | 1.12 (0.01) *** |
| Edad                               | 0.98 (0.00) *** |
| AIC                                | 129517.95       |
| BIC                                | 129548.34       |
| Log Likelihood                     | -64755.98       |
| Deviance                           | 129511.95       |
| Num. obs.                          | 185339          |

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

Statistical models



# Interpretación de Resultados de Regresión Logística en Odds Ratios

En el siguiente modelo de regresión logística, se analiza la probabilidad de encontrarse en situación de pobreza según sexo del jefe de hogar (JH) y su edad.

- **Intercepto (0.22):** Los odds de encontrarse en situación de pobreza son 0.22 cuando el jefe de hogar es hombre y tiene una edad de referencia (edad igual a 0).
- **Mujer (ref. hombre, 1.12):** Los odds de estar en situación de pobreza son 1.12 veces mayores en los hogares con jefatura femenina, comparado con los hogares con jefatura masculina, manteniendo constante la edad. Esto indica que las mujeres jefas de hogar tienen una mayor probabilidad de encontrarse en situación de pobreza.
- **Edad (0.98):** Por cada año adicional de edad del jefe de hogar, los odds de encontrarse en situación de pobreza disminuyen un 2%, manteniendo constante el sexo del jefe de hogar.



# Interpretación de los Coeficientes en Términos de Odds Ratios: Comparación de Grupos

Al interpretar un coeficiente de regresión logística en términos de odds ratios para la comparación entre grupos, se deben considerar los siguientes criterios:

- 1. Valor mayor a 1:** Si el odds ratio es mayor a 1, significa que el grupo comparado tiene una mayor probabilidad del evento estudiado en comparación con el grupo de referencia. Por ejemplo, un odds ratio de 1.12 indica que la probabilidad del evento es un 12% mayor en el grupo comparado, manteniendo constantes las demás variables. Esto implica que el grupo comparado tiene un mayor riesgo o probabilidad del evento en relación con el grupo de referencia.
- 2. Valor menor a 1:** Si el odds ratio es menor a 1, indica que el grupo comparado tiene una menor probabilidad del evento estudiado en comparación con el grupo de referencia. Esto implica un efecto protector o una reducción de la probabilidad del evento para el grupo comparado.



# Interpretación de los Coeficientes en Términos de Odds Ratios: Comparación de Grupos

- 3. Valores cercanos a 1:** Valores cercanos a 1, como el 1.01 o 0.99, implican que la variable tiene poco o ningún efecto sobre los odds del evento, es decir, el impacto de dicha variable es mínimo.
- 4. Significancia Estadística:** Para determinar si el coeficiente tiene un efecto significativo sobre la variable dependiente, se consideran los valores p ( $p < 0.001$ ;  $p < 0.01$ ;  $p < 0.05$ ). En este caso, todos los coeficientes son significativos al nivel del 0.001.



# Interpretación de los Coeficientes en Términos de Odds Ratios: Variable Continua

Al interpretar un coeficiente de regresión logística en términos de odds ratios para una variable continua, se deben considerar los siguientes criterios:

- 1. Valor mayor a 1:** Si el odds ratio es mayor a 1, significa que por cada unidad adicional de la variable continua, la probabilidad del evento estudiado aumenta. Esto implica que la variable tiene un efecto positivo en la probabilidad del evento. Un valor mayor que 1 sugiere un incremento proporcional en los odds del evento.
- 2. Valor menor a 1:** Si el odds ratio es menor a 1, indica que por cada unidad adicional de la variable continua, la probabilidad del evento disminuye. Por ejemplo, un odds ratio de 0.98 implica que, por cada año adicional de edad del jefe de hogar, los odds de estar en situación de pobreza disminuyen en un 2%. Un valor menor que 1 indica un efecto protector o una reducción en los odds del evento conforme aumenta la variable.



# Interpretación de los Coeficientes en Términos de Odds Ratios: Variable Continua

- 3. Valores cercanos a 1:** Valores cercanos a 1, como el 1.01 o 0.99, implican que la variable tiene poco o ningún efecto sobre los odds del evento, es decir, el impacto de dicha variable es mínimo.
- 4. Significancia Estadística:** Para determinar si el coeficiente tiene un efecto significativo sobre la variable dependiente, se consideran los valores p ( $p < 0.001$ ;  $p < 0.01$ ;  $p < 0.05$ ). En este caso, todos los coeficientes son significativos al nivel del 0.001.



# Diferencia entre un Coeficiente de Comparación de Grupos y uno de una Variable Continua

En un modelo de regresión logística, los coeficientes pueden representar comparaciones entre grupos o el efecto de variables continuas:

- **Coeficiente de Comparación de Grupos:** Este tipo de coeficiente se utiliza para variables categóricas, donde se compara un grupo específico con un grupo de referencia. Por ejemplo, en el caso de la variable **sexo** del jefe de hogar, el coeficiente de **Mujer (ref. hombre, 1.12)** representa la comparación entre los hogares con jefatura femenina y aquellos con jefatura masculina. En este caso, el odds ratio de 1.12 indica que los hogares con jefatura femenina tienen un 12% más de probabilidad de encontrarse en situación de pobreza comparado con los hogares con jefatura masculina, manteniendo constante la edad. Los coeficientes de comparación de grupos siempre se interpretan en relación con una categoría de referencia.



# Diferencia entre un Coeficiente de Comparación de Grupos y uno de una Variable Continua

- **Coeficiente de una Variable Continua:** Este tipo de coeficiente se utiliza para variables numéricas que pueden tomar un rango de valores, como la **edad** del jefe de hogar. En este caso, el coeficiente de **Edad (0.98)** indica el cambio en los odds de encontrarse en situación de pobreza por cada año adicional de edad del jefe de hogar, manteniendo constante el sexo. Un odds ratio de 0.98 significa que, por cada año adicional de edad, los odds de estar en situación de pobreza disminuyen en un 2%. A diferencia de los coeficientes de comparación de grupos, los coeficientes de variables continuas reflejan un cambio incremental en los odds del evento por cada unidad adicional de la variable independiente.



# 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 R<sup>2</sup>s
- 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: `.red[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 variables 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.



