# T4. Modelos con variables cualitativas

Ana J. López y Rigoberto Pérez

Dpto Economía Aplicada. Universidad de Oviedo

Curso 2010-2011

# Índice


1. Las variables cualitativas en el ámbito económico
2. La "trampa" de las variables ficticias
3. Variables cualitativas dependientes
   - Modelos Logit

# Modelos con variables cualitativas

## Competencias

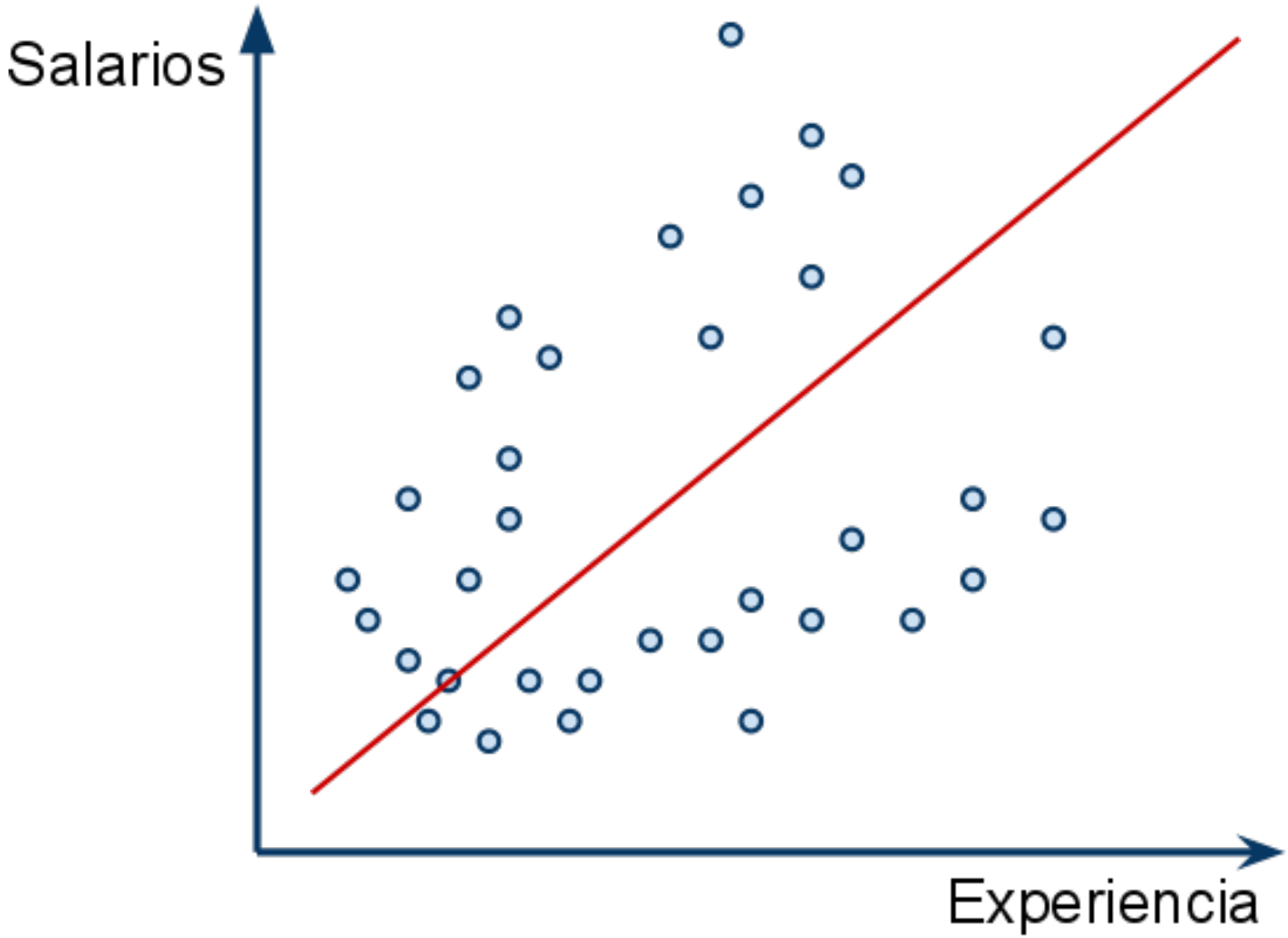
Este tema analiza la posibilidad de incorporar características cualitativas para mejorar la capacidad explicativa de los modelos y presenta a título introductorio los modelos de variable cualitativa dependiente. Se pretende que a su finalización los alumnos hayan adquirido las siguientes competencias:

- Definir e interpretar las variables dummy
- Comprender las razones que impiden plantear modelos de regresión con variables dependientes cualitativas
- Interpretar los coeficientes estimados de un modelo logit

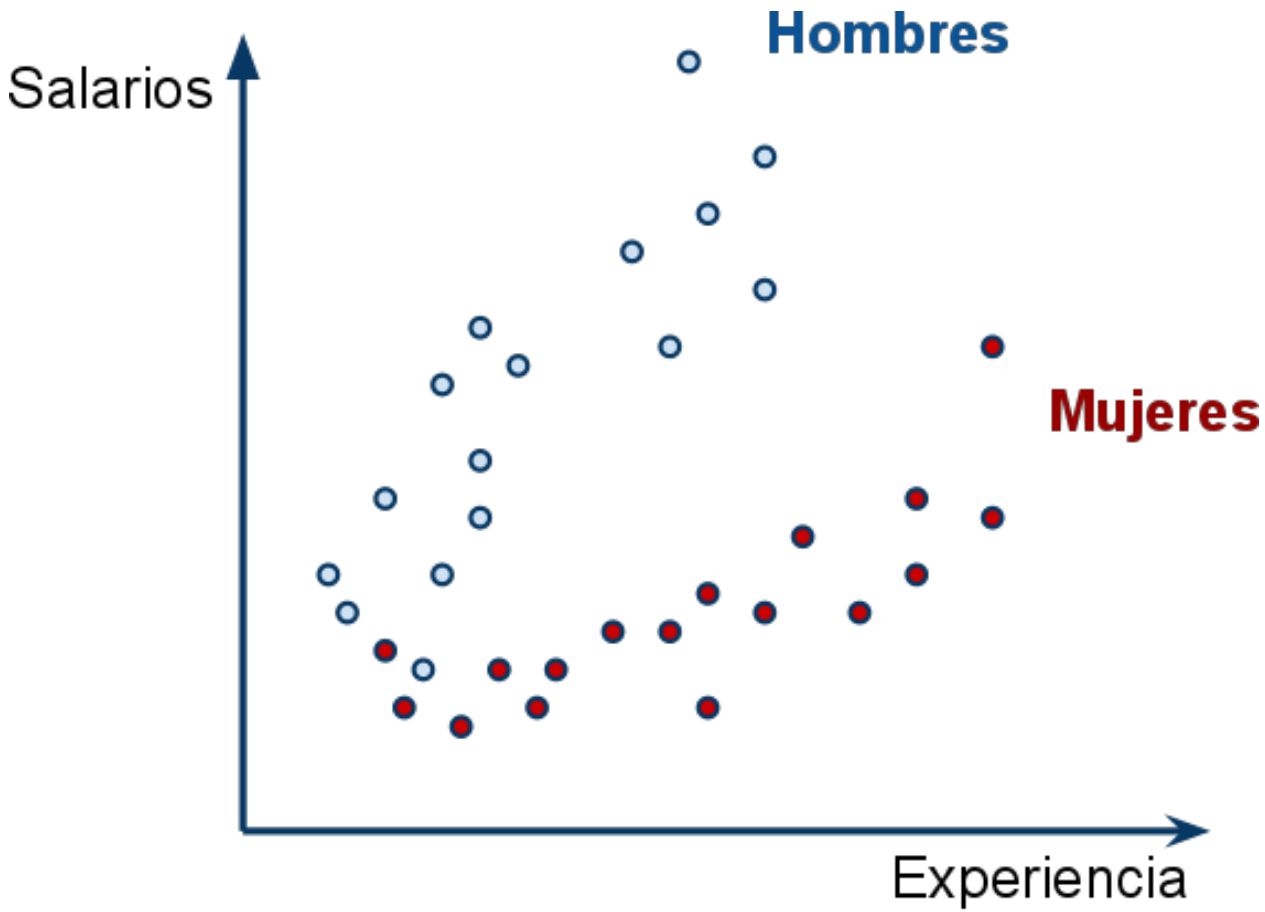
# Las variables cualitativas en el ámbito económico

- Algunas variables económicas pueden depender de características tales como el género, el sector de actividad, el lugar de residencia, la ideología política...
  - Ejemplos: Discriminación salarial por género, impacto sobre el gasto del tipo de gobierno
- En el análisis temporal pueden existir efectos asociados a la estacionalidad, o cambios de tendencia que también serán recogidos mediante variables cualitativas
  - Ejemplos: Estacionalidad en el turismo, impacto de la ampliación de la Unión Europea, ...

# Incorporación de variables cualitativas

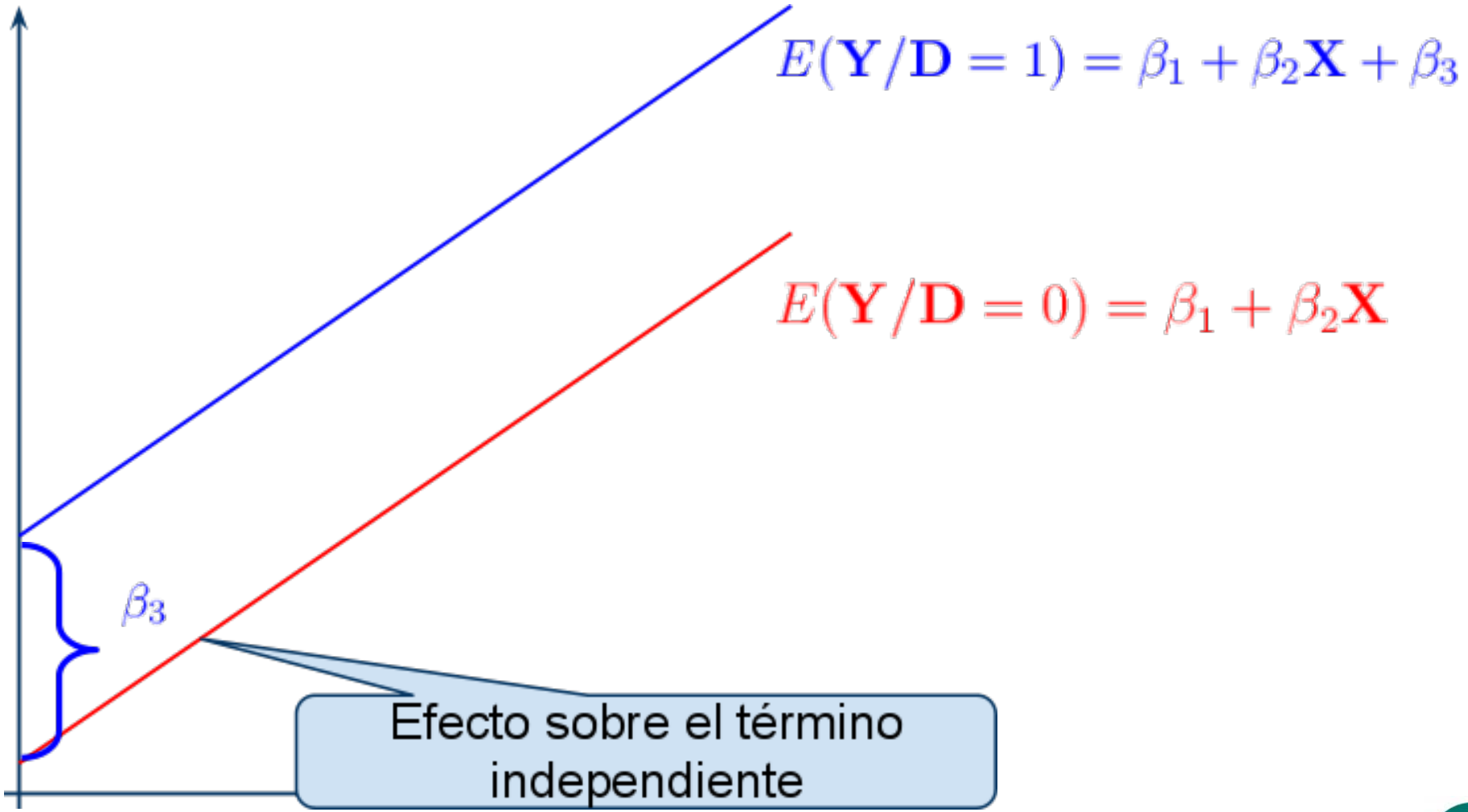


# Incorporación de variables cualitativas



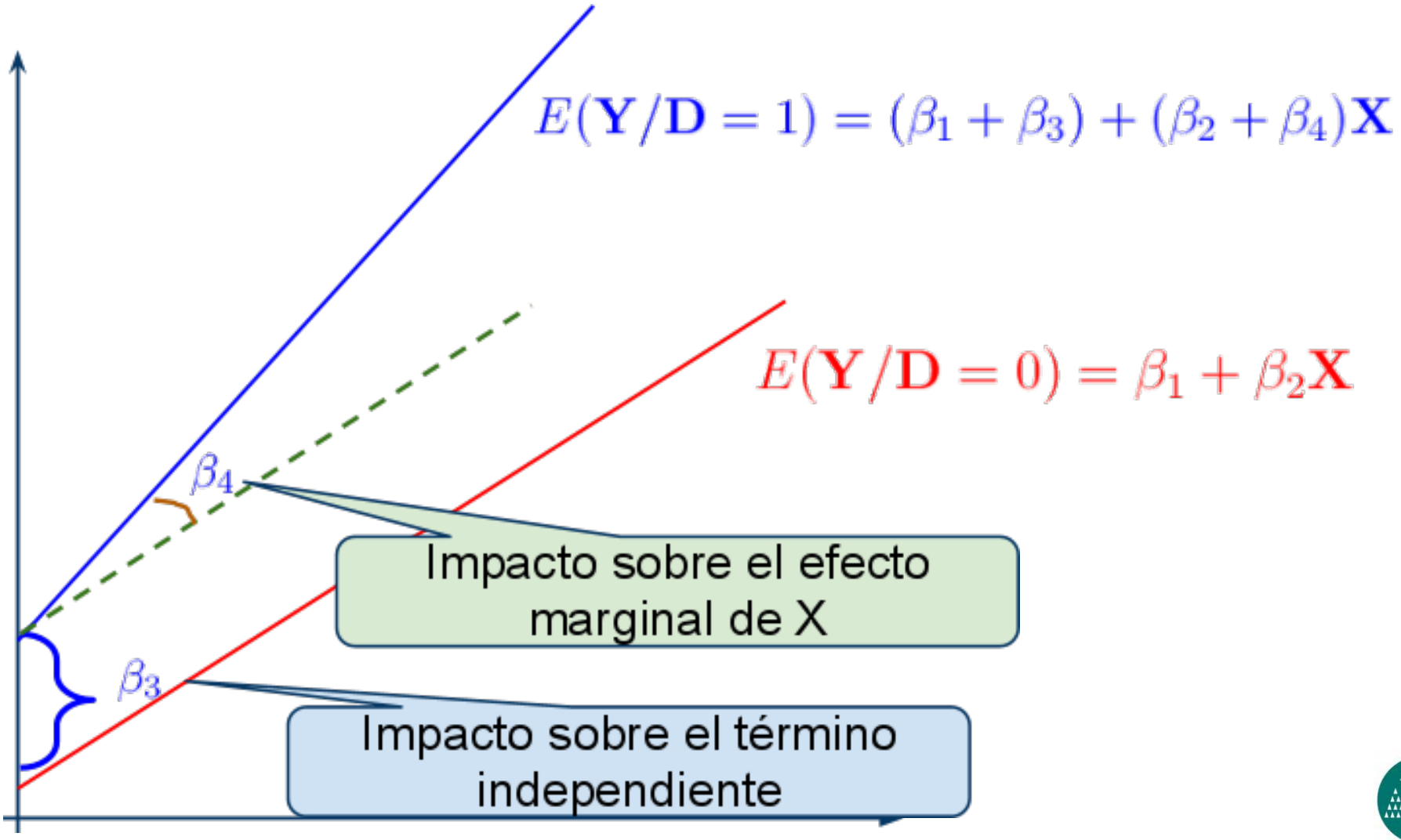

Introducción de variable dummy: $D = \begin{cases} 1 & \text{si el trabajador es hombre} \\ 0 & \text{si el trabajador es mujer} \end{cases}$

# Modelos con variable dummy: $\mathbf{Y} = \beta_1 + \beta_2\mathbf{X} + \beta_3\mathbf{D} + \mathbf{u}$



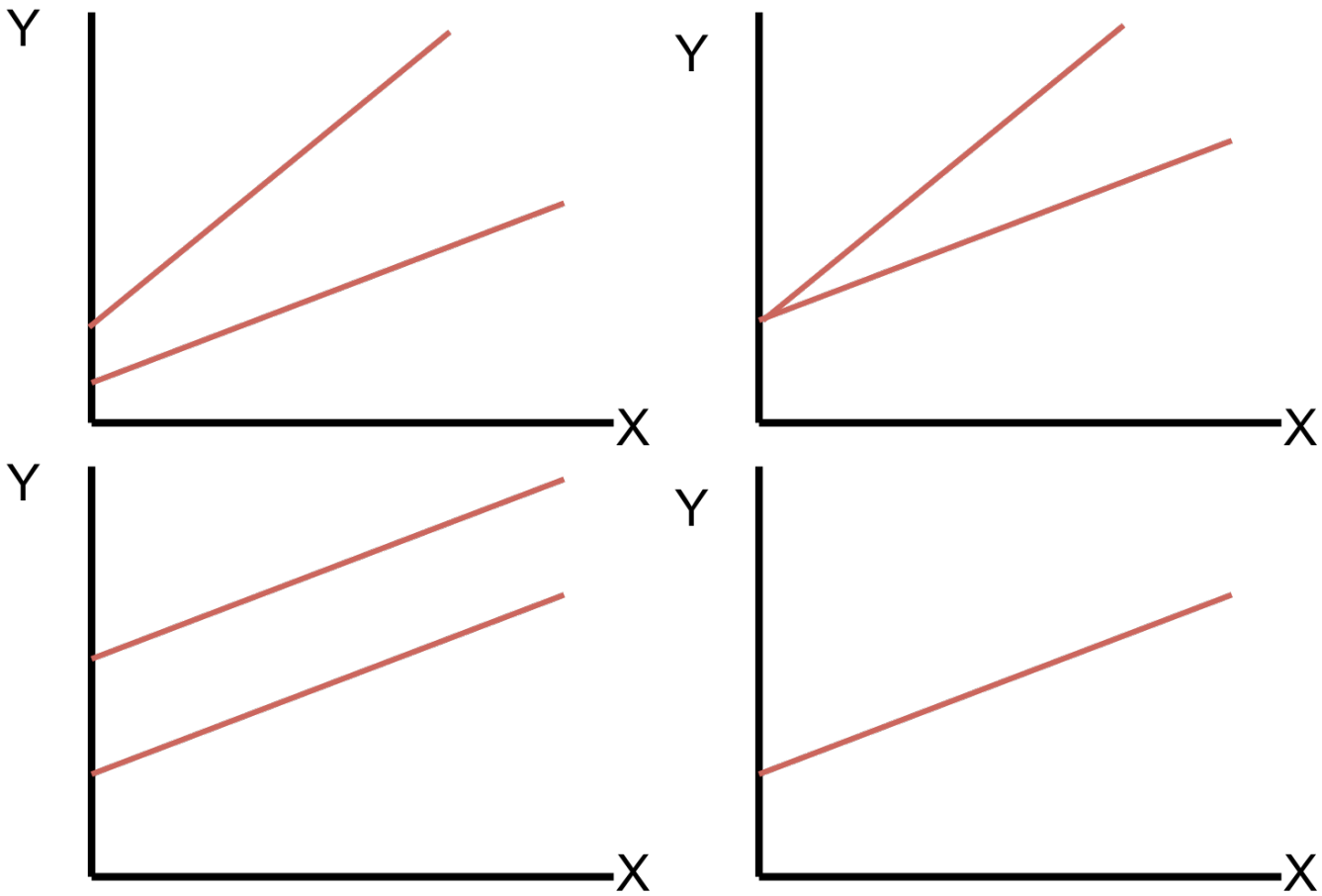
# Modelos con variable dummy:

$$\mathbf{Y} = \beta_1 + \beta_2\mathbf{X} + \beta_3\mathbf{D} + \beta_4\mathbf{DX} + \mathbf{u}$$



# Modelos con variable dummy:

$$\mathbf{Y} = \beta_1 + \beta_2\mathbf{X} + \beta_3\mathbf{D} + \beta_4\mathbf{DX} + \mathbf{u}$$




En estos gráficos, ¿$\beta_3$ y $\beta_4$ son significativos?

# La "trampa"de las variables ficticias

$$D_A = \begin{cases} 1 & \text{si el trabajador pertenece al sector agricultura} \\ 0 & \text{en otro caso} \end{cases}$$

$$D_I = \begin{cases} 1 & \text{si el trabajador pertenece al sector industria} \\ 0 & \text{en otro caso} \end{cases}$$

$$D_C = \begin{cases} 1 & \text{si el trabajador pertenece al sector construcción} \\ 0 & \text{en otro caso} \end{cases}$$

$$D_S = \begin{cases} 1 & \text{si el trabajador pertenece al sector servicios} \\ 0 & \text{en otro caso} \end{cases}$$

$$Y = \beta_1 + \beta_2 X + \beta_3 D_A + \beta_4 D_I + \beta_5 D_C + \beta_6 D_S + u$$

# La "trampa"de las variables ficticias

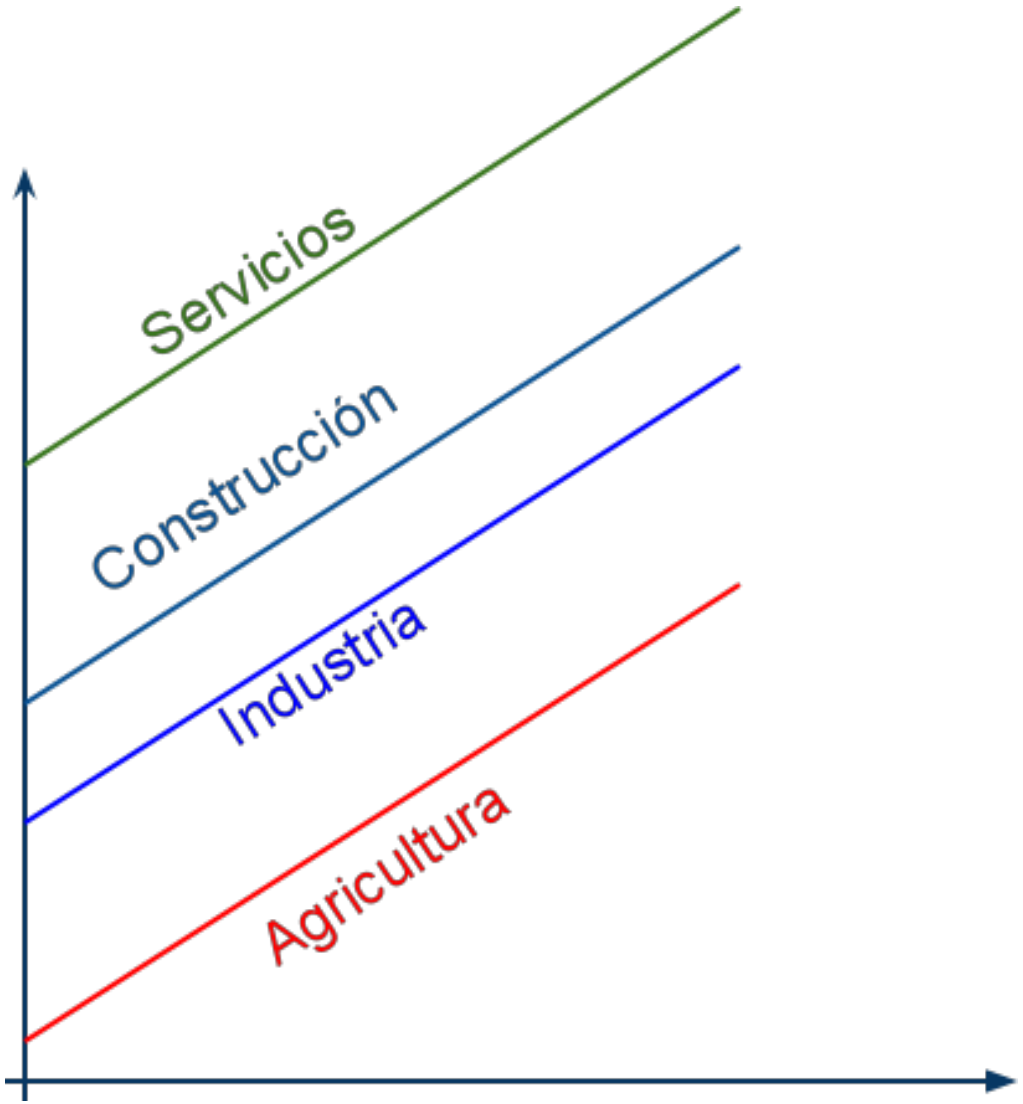
$$Y = \beta_1 + \beta_2 X + \beta_3 D_A + \beta_4 D_I + \beta_5 D_C + \beta_6 D_S + u$$

$$D_{Ai} + D_{Ii} + D_{Ci} + D_{Si} = 1 \ , \ \forall i = 1, \ldots, n$$

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 & D_{A1} & D_{I1} & D_{C1} & D_{S1} \\ 1 & X_2 & D_{A2} & D_{I2} & D_{C2} & D_{S2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_n & D_{An} & D_{In} & D_{Cn} & D_{Sn} \end{pmatrix}$$

- Relación lineal o Multicolinealidad entre las variables explicativas (rango no pleno $\rho(\mathbf{X}) \neq k$; $|\mathbf{X}'\mathbf{X}| = 0 \quad \Rightarrow \quad \mathbf{X}'\mathbf{X}$ no es invertible, EMC no definidos )
- **SOLUCIÓN:** Excluir una de las r categorías consideradas, definiendo r-1 variables dummy (la categoría excluida es la referencia para la interpretación de coeficientes).

# Modelo salarial en función de experiencia y sector económico: $Y = \beta_1 + \beta_2 X + \beta_3 D_I + \beta_4 D_C + \beta_5 D_S + u$




$\beta_1 + \beta_2 X + \beta_5$

$\beta_1 + \beta_2 X + \beta_4$

$\beta_1 + \beta_2 X + \beta_3$

$\beta_1 + \beta_2 X$

# Ilustración: Variable dummy asociada al género

Modelo 1: MCO, usando las observaciones 1--50
Variable dependiente: salario

| | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 831.818 | 310.984 | 2.6748 | 0.0102 |
| experiencia | 36.5540 | 9.19926 | 3.9736 | 0.0002 |

| | | | |
|---|---|---|---|
| Media de la vble. dep. | 2010.320 | D.T. de la vble. dep. | 754.6359 |
| Suma de cuad. residuos | 20997328 | D.T. de la regresión | 661.3958 |
| $R^2$ | **0.247523** | $R^2$ corregido | **0.231847** |
| $F(1, 48)$ | 15.78935 | Valor p (de $F$) | 0.000237 |
| Log-verosimilitud | −394.6440 | Criterio de Akaike | 793.2880 |
| Criterio de Schwarz | 797.1120 | Hannan--Quinn | 794.7442 |

# Ilustración: Variable dummy asociada al género

Modelo 2: MCO, usando las observaciones 1--50
Variable dependiente: salario

| | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 185.252 | 89.3293 | 2.0738 | 0.0436 |
| experiencia | 33.8524 | 2.52513 | 13.4062 | 0.0000 |
| masculino | 1264.94 | 52.0198 | 24.3165 | 0.0000 |

| | | | |
|---|---|---|---|
| Media de la vble. dep. | 2010.320 | D.T. de la vble. dep. | 754.6359 |
| Suma de cuad. residuos | 1546114 | D.T. de la regresión | 181.3726 |
| $R^2$ | **0.944592** | $R^2$ corregido | **0.942234** |
| $F(2,47)$ | 400.6285 | Valor p (de $F$) | 2.98e–30 |
| Log-verosimilitud | −329.4277 | Criterio de Akaike | 664.8555 |
| Criterio de Schwarz | 670.5915 | Hannan--Quinn | 667.0398 |

# Ilustración: Variable dummy asociada al género

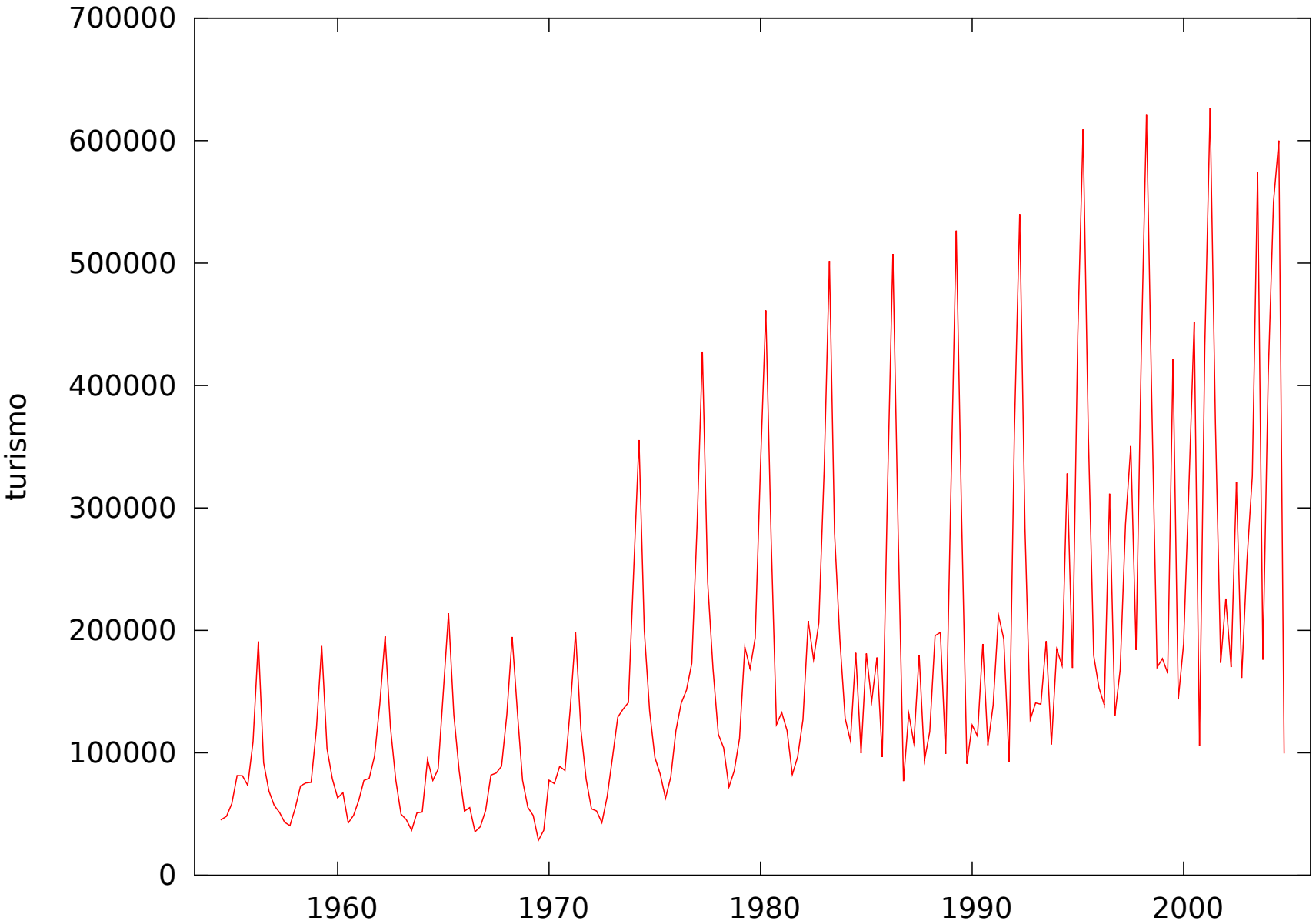
Modelo 3: MCO, usando las observaciones 1--50
Variable dependiente: salario

| | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 857.022 | 64.4528 | 13.2969 | 0.0000 |
| experiencia | 12.6705 | 1.95120 | 6.4937 | 0.0000 |
| masculino | 225.344 | 80.9078 | 2.7852 | 0.0077 |
| exp_masc | 32.4578 | 2.41534 | 13.4382 | 0.0000 |

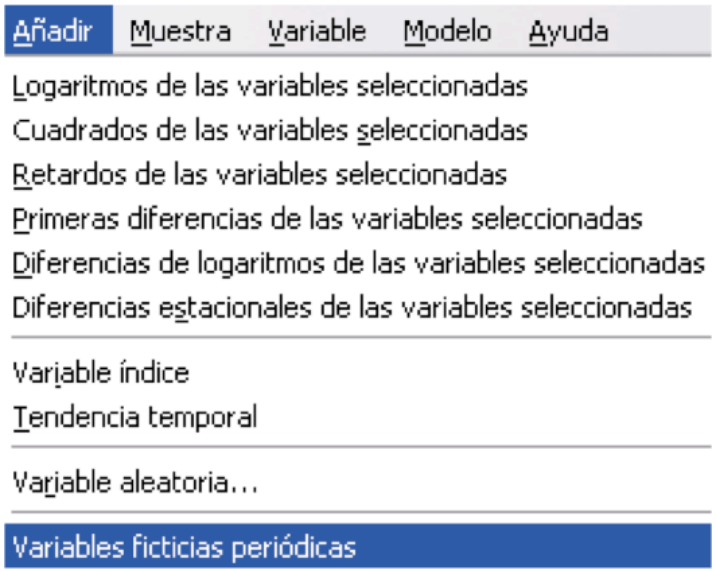
| | | | |
|---|---|---|---|
| Media de la vble. dep. | 2010.320 | D.T. de la vble. dep. | 754.6359 |
| Suma de cuad. residuos | 313882.3 | D.T. de la regresión | 82.60465 |
| $R^2$ | 0.988751 | $R^2$ corregido | **0.988018** |
| $F(3, 46)$ | 1347.808 | Valor p (de $F$) | 8.19e–45 |
| Log-verosimilitud | −289.5657 | Criterio de Akaike | 587.1314 |
| Criterio de Schwarz | 594.7795 | Hannan--Quinn | 590.0438 |

Los trabajadores de género MASCULINO ven aumentado su salario esperado y también el efecto marginal de la experiencia sobre el salario

# Ilustración: Variable dummy asociada a la estacionalidad



# Ilustración: Variable dummy asociada a la estacionalidad




| ID # | Nombre de variable | Etiqueta descriptiva |
|---|---|---|
| 0 | const | constante generada automáticamente |
| 1 | turismo | Entrada de turistas |
| 2 | dq1 | = 1 si el trimestre es = 1, 0 en otro caso |
| 3 | dq2 | = 1 si el trimestre es = 2, 0 en otro caso |
| 4 | dq3 | = 1 si el trimestre es = 3, 0 en otro caso |
| 5 | dq4 | = 1 si el trimestre es = 4, 0 en otro caso |

# Ilustración: Variable dummy asociada a la estacionalidad

Modelo 3: MCO, usando las observaciones 1983:3--2004:4 ($T$ = 86)
Variable dependiente: turismo

| | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 239171. | 28294.6 | 8.4529 | 0.0000 |
| dq2 | 74095.4 | 40014.7 | 1.8517 | 0.0677 |
| dq3 | 75382.9 | 39557.3 | 1.9057 | 0.0602 |
| dq4 | −108457. | 39557.3 | −2.7418 | 0.0075 |

| | | | |
|---|---|---|---|
| Media de la vble. dep. | 248803.2 | D.T. de la vble. dep. | 148313.1 |
| Suma de cuad. residuos | 1.38e+12 | D.T. de la regresión | 129662.3 |
| $R^2$ | 0.262667 | $R^2$ corregido | 0.235691 |
| $F(3, 82)$ | 9.737208 | Valor p (de $F$) | 0.000014 |
| Log-verosimilitud | −1132.432 | Criterio de Akaike | 2272.864 |
| Criterio de Schwarz | 2282.681 | Hannan--Quinn | 2276.815 |
| $\hat{\rho}$ | 0.528413 | Durbin--Watson | 0.942326 |

Respecto al primer trimestre el turismo aumenta sistemáticamente el segundo trimestre y también el tercero. Por el contrario en el cuarto se reduce

# Modelos de variable cualitativa dependiente

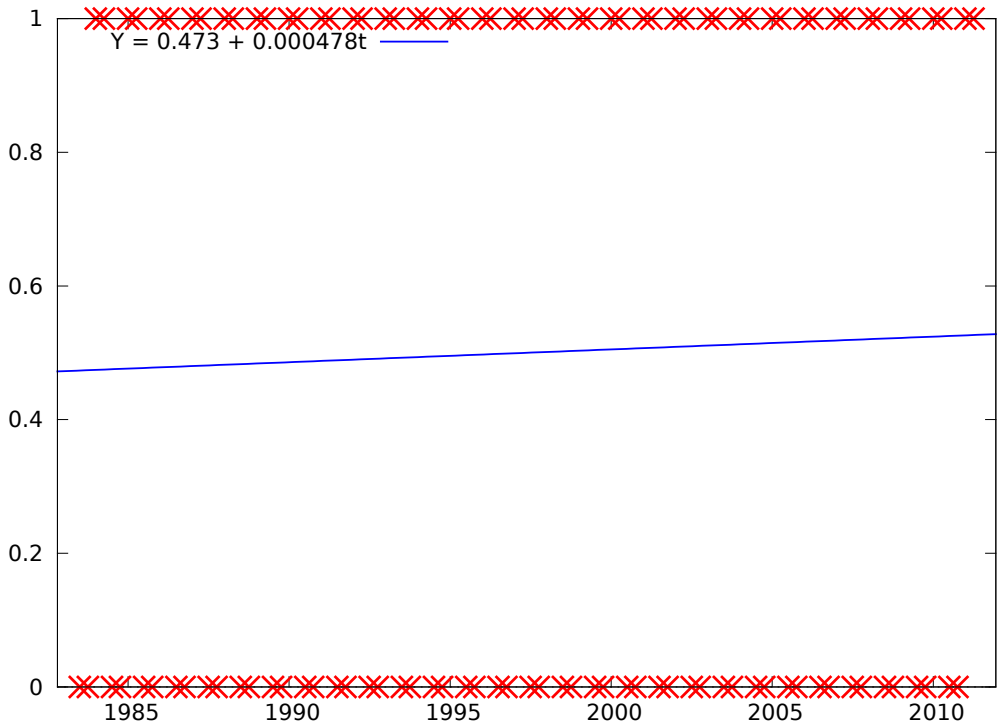
En algunas ocasiones nuestro objetivo es explicar una variable dependiente cualitativa:

- Con dos modalidades: **Modelos binomiales**
- Con más de dos modalidades: **Modelos multinomiales**
- Con varias modalidades que presentan un orden natural: **Modelos ordenados**
- Con modalidades asociadas a una decisión que condiciona las siguientes: **Modelos secuenciales**

# Modelos de variable cualitativa dependiente

El modelo lineal $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ no es aplicable para variables dependientes dicotómicas

- Las perturbaciones $\mathbf{u}$ son dicotómicas y por tanto no normales
- Al ser y dicotómica se cumple $E(\mathbf{y}) = p$
- No está garantizado que $E(\mathbf{y}) = \mathbf{X}\beta$ adopte valores entre 0 y 1



# Modelos de variable cualitativa dependiente

**SOLUCIÓN:**

Introducir una variable auxiliar ("variable índice") Z continua que se interpreta como "propensión" a la categoría investigada (encontrar empleo, afiliarse a un sindicato, realizar una compra, ...)

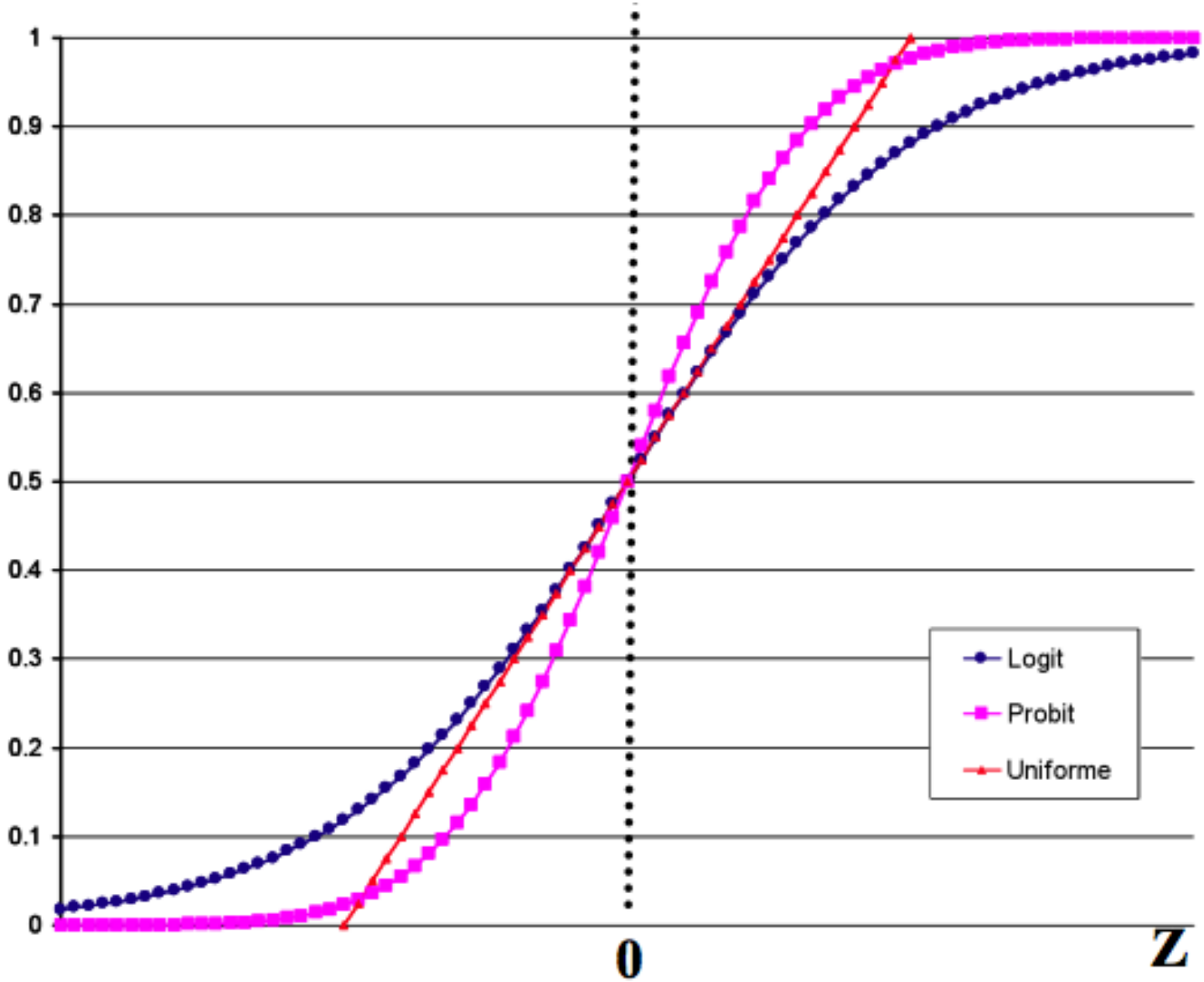
$$Y = \begin{cases} 1, & \text{si } Z > 0 \\ 0, & \text{si } Z \leq 0 \end{cases}$$

$$\begin{aligned} p_i &= P(Y=1) = P(Z>0) = P(\mathbf{x}'\boldsymbol{\beta} + u > 0) = P(u > -\mathbf{x}'\boldsymbol{\beta}) \\ &= 1 - F_u(-\mathbf{x}'\boldsymbol{\beta}) \end{aligned}$$

$$\begin{aligned} 1 - p_i &= P(Y=0) = P(Z \leq 0) = P(\mathbf{x}'\boldsymbol{\beta} + u \leq 0) = P(u \leq -\mathbf{x}'\boldsymbol{\beta}) \\ &= F_u(-\mathbf{x}'\boldsymbol{\beta}) \end{aligned}$$

Asumiendo ciertas distribuciones probabilísticas para u (logística, Normal, uniforme, ... ) es posible conocer la distribución de probabilidad de la variable Y.

# Modelos Logit, Probit y Uniforme



# Modelos Logit

Función logística de distribución de los errores: $F_u(x) = \dfrac{1}{1+e^{-x}}$

$$p_i = P(Y_i = 1) = 1 - F_u(-\mathbf{x}_i'\boldsymbol{\beta}) = 1 - \frac{1}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}$$

$$p_i\left(1+e^{\mathbf{x}_i'\boldsymbol{\beta}}\right) = e^{\mathbf{x}_i'\boldsymbol{\beta}} \Rightarrow e^{\mathbf{x}_i'\boldsymbol{\beta}} = \frac{p_i}{1-p_i}$$

$$\ln\left(e^{\mathbf{x}_i'\boldsymbol{\beta}}\right) = \ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}$$

$$\boxed{\ln\left(\frac{p_i}{1-p_i}\right) = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}}$$

Logit expresados como función lineal de las variables explicativas

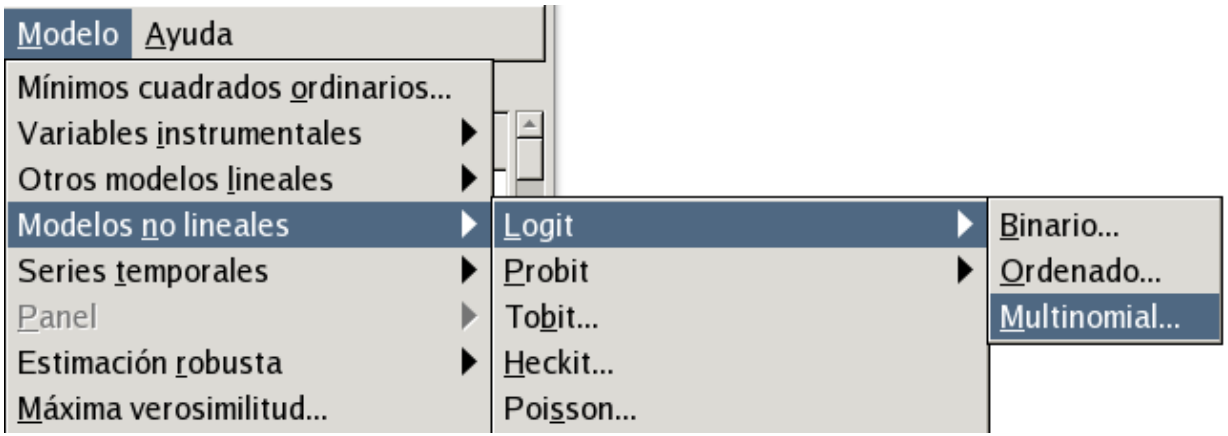
# Ilustración: Modelo logit para el empleo

gretl

Archivo Herramientas Datos Ver Añadir Muestra Variable Modelo Ayuda

Ilustra_logit_empleo.gdt

| ID # | Nombre de variable | Etiqueta descriptiva |
|---|---|---|
| 0 | const | constante generada automáticamente |
| 1 | empleo | =1 si OCUPADO , 0= si NO OCUPADO |
| 2 | estudios | Años de estudios |




Modelo logit para explicar si una persona está ocupada en función de sus estudios

# Ilustración: Modelo logit para el empleo

```
Iteración 0: log-verosimilitud = -504.571221559
Iteración 1: log-verosimilitud = -491.183952849
Iteración 2: log-verosimilitud = -491.172320140
Iteración 3: log-verosimilitud = -491.172320124
Criterio de parada basado en Log-Verosimilitud
```

Modelo 2: Logit, usando las observaciones 1--740
Variable dependiente: empleo

| | Coeficiente | Desv. Típica | *z* | Pendiente* |
|---|---|---|---|---|
| const | −1.74855 | 0.429247 | −4.0735 | . |
| estudios | 0.168757 | 0.0347109 | 4.8618 | 0.0410787 |

| | | | |
|---|---|---|---|
| Media de la vble. dep. | 0.578378 | D.T. de la vble. dep. | 0.243419 |
| $R^2$ de McFadden | 0.025064 | $R^2$ corregido | 0.021094 |
| Log-verosimilitud | −491.1723 | Criterio de Akaike | 986.3446 |
| Criterio de Schwarz | 995.5579 | Hannan--Quinn | 989.8969 |

# Ilustración: Modelo logit para el empleo

| | | | | |
|---|---|---|---|---|
| 402 | 1.000000 | 0.401674 | 0.598326 | Falso negativo |
| 403 | 1.000000 | 0.648864 | 0.351136 | |
| 404 | 1.000000 | 0.568697 | 0.431303 | |
| 405 | 1.000000 | 0.568697 | 0.431303 | |
| | | | | |
| 423 | 1.000000 | 0.754049 | 0.245951 | |
| 424 | 1.000000 | 0.484760 | 0.515240 | Falso negativo |
| | | | | |
| 429 | 0.000000 | 0.568697 | -0.568697 | Falso positivo |
| 430 | 0.000000 | 0.721430 | -0.721430 | Falso positivo |
| 431 | 0.000000 | 0.568697 | -0.568697 | Falso positivo |
| 432 | 0.000000 | 0.568697 | -0.568697 | Falso positivo |
| 433 | 0.000000 | 0.568697 | -0.568697 | Falso positivo |
| 434 | 0.000000 | 0.568697 | -0.568697 | Falso positivo |
| 435 | 0.000000 | 0.609520 | -0.609520 | Falso positivo |
| 436 | 0.000000 | 0.568697 | -0.568697 | Falso positivo |
| 437 | 0.000000 | 0.568697 | -0.568697 | Falso positivo |
| 438 | 0.000000 | 0.484760 | -0.484760 | |

# Bondad de los modelos Logit

| | |
|---|---|
| Medida basada en razón de verosimilitudes | $-2 \ln \frac{L_{NR}}{L_R}$ |
| Medida de Mc Fadden (1974) | $R^2 = 1 - \frac{\ln L_{NR}}{\ln L_R}$ |
| Proporción de aciertos | $\frac{\text{Núm, predicciones correctas}}{\text{Núm. observaciones}}$ |

- LNR: Máx de L respecto a todos los parámetros
- LR: Máximo de L restringido (con $\beta_i = 0, \ \forall i$)
- La razón de verosimilitudes contrasta la nulidad de $\beta$

# Ilustración: Modelo logit para el empleo

| | Coeficiente | Desv. Típica | $z$ | Pendiente* |
|---|---|---|---|---|
| const | −1.74855 | 0.429247 | −4.0735 | . |
| estudios | 0.168757 | 0.0347109 | 4.8618 | 0.0410787 |

| | | | |
|---|---|---|---|
| Media de la vble. dep. | 0.578378 | D.T. de la vble. dep. | 0.243419 |
| $R^2$ de McFadden | 0.025064 | $R^2$ corregido | 0.021094 |
| Log-verosimilitud | −491.1723 | Criterio de Akaike | 986.3446 |
| Criterio de Schwarz | 995.5579 | Hannan--Quinn | 989.8969 |

*Evaluado en la media

Número de casos ´´correctamente predichos´´ = 442 (59.7 percent)
$f(\beta'X)$ en la media de las variables independientes = 0.243 Contraste de razón de verosimilitudes: $\chi^2(1)$ = 25.254 [0.0000]

| | | Predicho | |
|---|---|---|---|
| | | 0 | 1 |
| Observado | 0 | 64 | 248 |
| | 1 | 50 | 378 |

Este modelo logit clasifica correctamente 442 casos (casi el 60 %). Hay 248 falsos positivos (34 %) y 50 falsos negativos (6 %)