Tema 6. Formas funcionales del modelo de RLM y variables binarias o *dummies*

Gustavo A. García

ggarci24@eafit.edu.co

Econometría para la Toma de Decisiones

Maestría en Economía Aplicada

Escuela de Finanzas, Economía y Gobierno

Universidad EAFIT

Link slides en formato html

Link slides en formato PDF

En este tema

- Formas funcionales del modelo de RLM
- Variables binarias o *dummies*
- Modelación de factores y categorías
- Ejercicio aplicado en R

Lecturas

- Wooldridge, Jeffrey (2013). *Introducción a la econometría*. 5a edición, Cengage Learning. Cap. 2.4, 6, 7
- Gujarati, D. y Porter, D. (2010). *Econometría*. 5a edición, Mc Graw Hill. Cap. 6 y 9

Formas funcionales del modelo de RLM

Variable dependiente Y_i	Regresor X_{ij}	β_j	Interpretación
Niveles	Niveles	Efecto mg en Y ante un cambio unitario en X_j	Y aumenta o disminuye β_j veces cuando aumenta X_j
Logaritmo	Logaritmo	Elasticidad X_j de Y	Un incremento en 1 % en X_j genera un incremento o disminución de β_j % en Y
Logaritmo	Niveles	Tasas de crecimiento o retorno	Un incremento en una unidad de X_j genera un incremento o disminución en $\beta_j * 100\,\%$ en Y
Niveles	Logaritmo	Respuesta de Y ante una variación de X_j	Un incremento en 1 % en X_j genera un incremento o disminución en $\beta_j/100$ en Y

Variables binarias o *dummies*: conceptualización general

- La inclusión de variables binarias (también llamadas *dummy* o falsas) en los modelos de regresión, obedece a la necesidad de incorporar factores de naturaleza cualitativa que se traducen en cambios paramétricos. Uno de estos cambios puede ser:
    - La ecuación de Mincer o de ingresos laborales puede ser para hombres y mujeres (diferencias en el salario de reserva por discriminación) y el log del ingreso mínimo (o intercepto) puede ser diferente para cada género
    - La demanda por carne puede variar según los grupos religiosos, las elasticidades precio e ingreso de cada grupo pueden ser diferentes
    - Un cambio estructural en el tiempo puede ser el resultado de un factor cualitativo que induce el cambio paramétrico
- Si se piensa en la función de consumo para Colombia de 1950 a 2000, es intuitivo afirmar que debido a migración campo-ciudad, transición demográfica o modernización del aparato financiero, la función de consumo de 1950 a 1970 no debe ser la misma que la correspondiente de 1971 a 2000
- El consumo autónomo (intercepto) y la propensión marginal a consumir (la pendiente) de los dos períodos puede haber cambiado. Igual sucedería con los parámetros de la función de importaciones antes y después de la apertura económica en 1990

Variables binarias o *dummies*: conceptualización general

- La forma de incluir estos factores cualitativos es usando una variables que sólo tomen el valor 0 y 1, y se denominan falsas, dicótomas binarias o *dummies* $\Longrightarrow$ variables indiciadores

- La escogencia de 0 y 1 no es arbitraria, proviene de la esencia del conteo. Cuando se esta contando algo, se suma 1 si ese algo esta y se suma 0 si ese algo no esta

$$\text{se puede asociar} = \begin{cases} 0 & \text{Ausencia} \\ 1 & \text{Presencia} \end{cases}$$

- Otro par de números (3 y 7 por ejemplo) no servirían para lo mismo, lo que puede ser arbitrario es la asignación del 0 y el 1

- Cuando se usan variables binarias en los modelos se producen cambios en
  - el intercepto
  - la pendiente
  - intercepto y pendiente

Modelación de factores y categorías

i. Un factor dos categorías

Supóngase que se quiere incorporar al modelo de Mincer (ecuación salarial) el factor cualitativo género. Existen tres posibilidades según el efecto que se quiere modelar

- cambio en el intercepto (en el log del salario mínimo)
- cambio en la pendiente (en la tasa de retorno de la educación)
- cambio de ambos, intercepto y pendiente

Lo que se intenta incorporar es una hipótesis de diferenciación por género en la ecuación de ingresos. Se define una variable binaria de la forma

$$bsexo_i = \begin{cases} 0 & \text{hombre} \\ 1 & \text{mujer} \end{cases}$$

Modelación de factores y categorías

i. Un factor dos categorías

1. Cambio en el intercepto
Sea $lwage_i =$ log de los salarios y $Educ_{i2} =$ Años de educación aprobados

En el modelo $lwage_i = \beta_1 + \beta_2 Educ_i + u_i$

β_1 : log tasa de salario mínima
β_2 : tasa de retorno de la educación
u_i : perturbación aleatoria con supuestos estándar

Al incorporar la variable binaria de género se tendría

$$lwage_i = \beta_1 + \beta_2 Educ_i + \beta_3 bsexo_i + u_i$$

Es como si el modelo se convirtiese en dos submodelos

Mujeres $(bsexo_i = 1) \Longrightarrow lwage_i = (\beta_1 + \beta_3) + \beta_2 Educ_i + u_i$
Hombres $(bsexo_i = 0 \Longrightarrow lwage_i = \beta_1 + \beta_2 Educ_i + u_i$

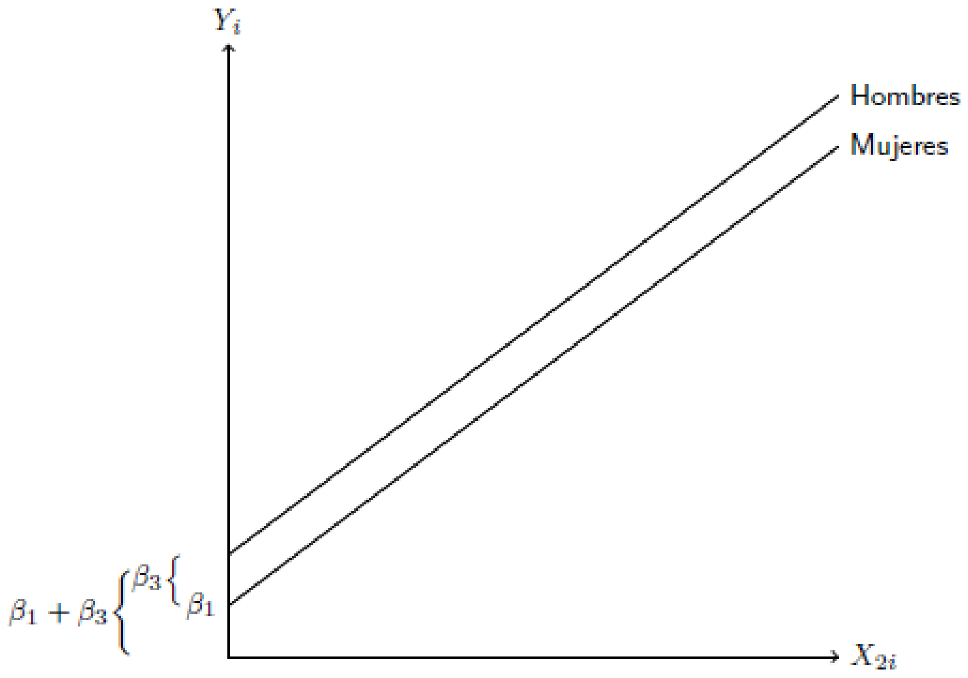
En esta situación
β_1 : log de la tasa salaria mínima de los hombres
β_3 : cambio en log de la tasa salarial mínima de las mujeres respecto a los hombres
$\beta_1 + \beta_3$: log de la tasa salarial mínima de las mujeres

Modelación de factores y categorías

i. Un factor dos categorías

1. Cambio en el intercepto

Gráficamente tenemos



Lo que se esta modelando es un cambio en el intercepto manteniendo constante la pendiente

Lo que se hizo fue conservar el intercepto (β_1) y agregar una variable falsa $(bsexo_i)$

Modelación de factores y categorías

i. Un factor dos categorías

1. Cambio en el intercepto

Alternativamente se puede eliminar el intercepto e incluir dos variables binarias

$$bmujer_i = \begin{cases} 0 & \text{Hombre} \\ 1 & \text{Mujer} \end{cases} \qquad bhombre_i = \begin{cases} 0 & \text{Mujer} \\ 1 & \text{Hombre} \end{cases}$$

Observe que $bhombre_i + bmujer_i = 1$

El modelo queda de la forma

$$lwage_i = \gamma_2 Educ_{i2} + \gamma_3 bhombre_i + \gamma_4 bmujer_i + u_i$$

Nuevamente se tienen dos modelos
Mujeres $(bhombre_i = 0, bmujer_i = 1) \Longrightarrow lwage_i = \gamma_4 + \gamma_2 Educ_{i2} + u_i$
Hombres $(bhombre_i = 1, bmujer_i = 0) \Longrightarrow lwage_i = \gamma_3 + \gamma_2 Educ_{i2} + u_i$

En esta situación\ γ_2 : tasa de retorno de la educación, se supone igual para hombres y mujeres
γ_3 : log de la tasa salarial mínima para hombres
γ_4 : log de la tasa salarial mínima para mujeres
$\gamma_4 - \gamma_3$: diferencial del log de la tasa mínima de salario de mujeres frente a hombres

Modelación de factores y categorías

i. Un factor dos categorías

1. Cambio en el intercepto
Qué sucede si se utilizan las dos opciones anteriores al mismo tiempo: conservar el intercepto e incluir las dos variables binarias

$$lwage_i = \gamma_1 + \gamma_2 Educ_i + \gamma_3 bhombre_i + \gamma_4 bmujer_i + u_i$$

La matriz $\mathbf{X}$ del modelo tendría la siguiente estructura (suponemos primero mujeres (M) y después hombres $(N-M)$)

$$\mathbf{X}_{N\text{x}4} = \begin{bmatrix} 1 & Educ_1 & 0 & 1 \\ 1 & \vdots & 0 & 1 \\ 1 & Educ_M & 0 & 1 \\ \dots & \dots & \dots & \dots \\ 1 & Educ_{M+1} & 1 & 0 \\ 1 & \vdots & 1 & 0 \\ 1 & Educ_N & 1 & 0 \end{bmatrix}$$

Se observa que Col(1)=Col(3)+Col(4), lo cual implica que rango de la matriz $\mathbf{X}$ no es de 4 sino de 3, así que hay un problema de multicolinealidad perfecta:

$$(X'X)_{4\text{x}4} \text{ es singular}$$

$$(X'X)^{-1}_{4\text{x}4} \text{ no existe}$$

Este caso se conoce como **la trampa de las variables *dummies***

Modelación de factores y categorías

i. Un factor dos categorías

2. Cambio en pendiente
La manera de incorporar cambios en la pendiente es agregar el producto de la variable binaria por la correspondiente variable explicativa. Por ejemplo, modelando diferentes tasas de retornos a la educación por género, el modelo queda de la forma:

$$lwage_i = \beta_1 + \beta_2 Educ_i + \beta_3 Educ_i bsexo_i + u_i$$

Nuevamente es un modelo que contiene dos
Hombres $(bsexo_i = 0) \implies lwage_i = \beta_1 + \beta_2 Educ_i + u_i$
Mujeres $(bsexo_i = 1) \implies lwage_i = \beta_1 + (\beta_2 + \beta_3) Educ_i + u_i$

En esta situación
β_1 : log de la tasa salarial mínima, se supone igual para hombres y mujeres
β_2 : tasa de retorno de la educación de las mujeres
β_3 : cambio en la tasa de retorno de la educación de hombres respecto a mujeres
$\beta_2 + \beta_3$: tasa de retorno de la educación de los hombres

Modelación de factores y categorías

i. Un factor dos categorías

3. Cambio en el intercepto y la pendiente

La intuición indica que se debe reunir los dos casos anteriores: agregar una variable binaria (o eliminar el intercepto y agregar dos binarias) y la binaria multiplicada por la variable independiente. El modelo queda de la forma:

$$lwage_i = \beta_1 + \beta_2 Educ_i + \beta_3 bsexo_i + \beta_4 Educ_i bsexo_i + u_i$$

Hombres $(bsexo_i = 0) \implies Y_i = \beta_1 + \beta_2 X_{i2} + u_i$
Mujeres $(bsexo_i = 1) \implies Y_i = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) X_{i2} + u_i$

En esta situación
β_1 : log de la tasa salarial mínima de las mujeres
β_2 : tasa de retorno de la educación de las mujeres
β_3 : cambio en log de la tasa salarial mínima de hombres respecto a mujeres
β_4 : cambio en la tasa de retorno de la educación de hombres respecto a mujeres
$\beta_1 + \beta_3$: log de la tasa salarial mínima de los hombres
$\beta_2 + \beta_4$: Tasa de retorno de la educación de los hombres

El modelo conjunto es equivalente a estimar dos regresiones por separado

Modelación de factores y categorías

ii. Un factor varias categorías
Se tienen tres niveles educatiovs, así que se definen las siguientes variables binarias

$$bpri_i = \begin{cases} 1 & \text{primaria} \\ 0 & \text{otro caso} \end{cases} \qquad bsec_i = \begin{cases} 1 & \text{secundaria} \\ 0 & \text{otro caso} \end{cases} \qquad bsup_i = \begin{cases} 1 & \text{superior} \\ 0 & \text{otro caso} \end{cases}$$

En el modelo de RLM se conserva el intercepto y al haber 3 categorías se incluyen 2 variables binarias. La categoría a la cual no se le incluye la variable binaria se vuelve el patrón de referencia del modelo. El modelo queda de la forma:

$$lwage_i = \beta_1 + \beta_2 Exper_i + \beta_3 bsec_i + \beta_4 bsup_i + u_i$$

El modelo incluye 3 submodelos:
Secundaria $(bsec_i = 1, bsup_i = 0) \Longrightarrow lwage_i = (\beta_1 + \beta_3) + \beta_2 Exper_i + u_i$
Superior $(bsec_i = 0, bsup_i = 1) \Longrightarrow lwage_i = (\beta_1 + \beta_4) + \beta_2 Exper_{i2} + u_i$
Primaria $(bsec_i = 0, bsup_i = 0) \Longrightarrow lwage_i = \beta_1 + \beta_2 Exper_{i2} + u_i$

En esta situación
β_1 : log de la tasa salarial mínima de los individuos con primaria
β_2 : tasa de retorno de la experiencia, asumida igual independiente del nivel educativo
β_3 : diferencia en log de la tasa salarial mínima de individuos con secundaria respecto a los de primaria
β_4 : diferencia en log de la tasa salarial mínima de individuos con superior respecto a los de primaria
$\beta_1 + \beta_3$: log de la tasa salarial mínima de los individuos con secundaria
$\beta_1 + \beta_4$: log de la tasa salarial mínima de los individuos con superior

Ejemplo aplicado en R

Se tiene una base de datos de corte transversal de 526 trabajadores correspondientes a 1976 para los Estados unidos. $wage$ son los salarios en dólares por hora y $educ$ los años de educación

i. Un factor dos categorías

1. Cambio en el intercepto (intercepto + una binaria)

$$lwage = \beta_1 + \beta_2 educ + \beta_3 female + u$$

$$female_i = \begin{cases} 1 & \text{mujer} \\ 0 & \text{hombre} \end{cases}$$

```r
library(haven); library(tidyverse); library(summarytools)
data <- read_stata("http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.dta")

modelo1 <- lm(lwage ~ educ + female, data=data)
summary(modelo1)
```

```
Call:
lm(formula = lwage ~ educ + female, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.02672 -0.27470 -0.03731  0.26219  1.34738 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.826269   0.094054   8.785   <2e-16 ***
educ         0.077203   0.007047  10.955   <2e-16 ***
female      -0.360865   0.039024  -9.247   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4455 on 523 degrees of freedom
```

Ejemplo aplicado en R

i. Un factor dos categorías

1. Cambio en el intercepto (no intercepto y dos binarias)

$$female_i = \begin{cases} 1 & \text{mujer} \\ 0 & \text{hombre} \end{cases} \qquad male_i = \begin{cases} 1 & \text{hombre} \\ 0 & \text{mujer} \end{cases}$$

```r
data <- read_stata("http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.dta") |>
  mutate(male = case_when(female==1~0,
                          female==0~1))

modelo2 <- lm(lwage ~ 0 + educ + female + male , data=data)
summary(modelo2)
```

```

Call:
lm(formula = lwage ~ 0 + educ + female + male, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.02672 -0.27470 -0.03731  0.26219  1.34738 

Coefficients:
       Estimate Std. Error t value Pr(>|t|)    
educ   0.077203   0.007047  10.955  < 2e-16 ***
female 0.465404   0.091227   5.102 4.72e-07 ***
male   0.826269   0.094054   8.785  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4455 on 523 degrees of freedom
Multiple R-squared:  0.9323,	Adjusted R-squared:  0.932 
F-statistic:  2403 on 3 and 523 DF,  p-value: < 2.2e-16
```

Ejemplo aplicado en R

i. Un factor dos categorías

2. Cambio en pendiente

$$lwage = \beta_1 + \beta_2 educ + \beta_3 educ * female + u$$

```
modelo3 <- lm(lwage ~ educ + educ:female, data=data)
summary(modelo3)
```

```
Call:
lm(formula = lwage ~ educ + educ:female, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.04030 -0.28526 -0.03285  0.27044  1.36353 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.680021   0.091277   7.450 3.88e-13 ***
educ         0.088045   0.007071  12.451  < 2e-16 ***
educ:female -0.027595   0.003063  -9.008  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4471 on 523 degrees of freedom
Multiple R-squared:  0.2952,	Adjusted R-squared:  0.2925 
F-statistic: 109.5 on 2 and 523 DF,  p-value: < 2.2e-16
```

Ejemplo aplicado en R

i. Un factor dos categorías

2. Cambio en intercepto y pendiente

$$lwage = \beta_1 + \beta_2 educ + \beta_3 female + \beta_4 educ * female + u$$

```
modelo4 <- lm(lwage ~ educ + educ*female, data=data)
summary(modelo4)
```

```
Call:
lm(formula = lwage ~ educ + educ * female, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.02673 -0.27468 -0.03721  0.26221  1.34740 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.260e-01  1.181e-01   6.997 8.08e-12 ***
educ         7.723e-02  8.988e-03   8.593  < 2e-16 ***
female      -3.601e-01  1.854e-01  -1.942   0.0527 .  
educ:female -6.408e-05  1.450e-02  -0.004   0.9965    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4459 on 522 degrees of freedom
Multiple R-squared:  0.3002,	Adjusted R-squared:  0.2962 
F-statistic: 74.65 on 3 and 522 DF,  p-value: < 2.2e-16
```

Ejemplo aplicado en R

i. Un factor dos categorías

2. Cambio en intercepto y pendiente

Lo anterior es equivalente a estimar dos regresiones por separado, una cuando $female = 1$ y otra cuando $female = 0$

```r
modelo5 <- lm(lwage ~ educ, data=subset(data,female==1))
summary(modelo5)
```

```
Call:
lm(formula = lwage ~ educ, data = subset(data, female == 1))

Residuals:
     Min       1Q   Median       3Q      Max 
-2.02673 -0.24397 -0.06163  0.21415  1.21924 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.46589    0.12890   3.614 0.000364 ***
educ         0.07716    0.01026   7.520 9.82e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.402 on 250 degrees of freedom
Multiple R-squared:  0.1845,	Adjusted R-squared:  0.1812 
F-statistic: 56.55 on 1 and 250 DF,  p-value: 9.824e-13
```

```r
modelo6 <- lm(lwage ~ educ, data=subset(data,female==0))
summary(modelo6)
```

```
Call:
lm(formula = lwage ~ educ, data = subset(data, female == 0))

Residuals:
     Min       1Q   Median       3Q      Max 
-1.11585 -0.34240 -0.01708  0.32659  1.34740 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.825955   0.127813   6.462 4.75e-10 ***
educ        0.077228   0.009731   7.936 5.47e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4828 on 272 degrees of freedom
Multiple R-squared:  0.188,	Adjusted R-squared:  0.185 
F-statistic: 62.99 on 1 and 272 DF,  p-value: 5.471e-14
```

Ejemplo aplicado en R

ii. Un factor varias categorías

$$lwage = \beta_1 + \beta_2 exper + \beta_3 bsec + \beta_4 bsup + u$$

$$bpri_i = \begin{cases} 1 & \text{primaria} \\ 0 & \text{otro caso} \end{cases} \qquad bsec_i = \begin{cases} 1 & \text{secundaria} \\ 0 & \text{otro caso} \end{cases} \qquad bsup_i = \begin{cases} 1 & \text{superior} \\ 0 & \text{otro caso} \end{cases}$$

```
freq(data$educ, headings=F)
```

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
0	2	0.38	0.38	0.38	0.38
2	1	0.19	0.57	0.19	0.57
3	1	0.19	0.76	0.19	0.76
4	3	0.57	1.33	0.57	1.33
5	1	0.19	1.52	0.19	1.52
6	6	1.14	2.66	1.14	2.66
7	4	0.76	3.42	0.76	3.42
8	22	4.18	7.60	4.18	7.60
9	17	3.23	10.84	3.23	10.84
10	30	5.70	16.54	5.70	16.54
11	29	5.51	22.05	5.51	22.05
12	198	37.64	59.70	37.64	59.70
13	39	7.41	67.11	7.41	67.11
14	53	10.08	77.19	10.08	77.19
15	21	3.99	81.18	3.99	81.18
16	68	12.93	94.11	12.93	94.11
17	12	2.28	96.39	2.28	96.39
18	19	3.61	100.00	3.61	100.00
<NA>	0			0.00	100.00
Total	526	100.00	100.00	100.00	100.00

```
data <- data |>
  mutate(educ_n = case_when(educ>=0 & educ<=5 ~ 1,
                            educ>=6 & educ<=13 ~ 2,
                            educ>=14 & educ<=18 ~ 3))
freq(data$educ_n, headings=F)
```

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
1	8	1.52	1.52	1.52	1.52
2	345	65.59	67.11	65.59	67.11
3	173	32.89	100.00	32.89	100.00
<NA>	0			0.00	100.00
Total	526	100.00	100.00	100.00	100.00

Ejemplo aplicado en R

ii. Un factor varias categorías

```r
modelo7 <- lm(lwage ~ exper + factor(educ_n), data=data)
summary(modelo7)
```

```

Call:
lm(formula = lwage ~ exper + factor(educ_n), data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.99901 -0.31438 -0.07762  0.32933  1.55045 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     0.860162   0.180326   4.770 2.39e-06 ***
exper           0.008099   0.001594   5.081 5.23e-07 ***
factor(educ_n)2 0.479675   0.174428   2.750  0.00617 ** 
factor(educ_n)3 0.944581   0.177853   5.311 1.62e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.479 on 522 degrees of freedom
Multiple R-squared:  0.1926,	Adjusted R-squared:  0.188 
F-statistic: 41.51 on 3 and 522 DF,  p-value: < 2.2e-16
```

```r
data <- data |>
  mutate(bpri = case_when(educ>=0 & educ<=5 ~ 1,
                          TRUE ~ 0),
         bsec = case_when(educ>=6 & educ<=13 ~ 1,
                          TRUE ~ 0),
         bsup = case_when(educ>=14 & educ<=18 ~ 1,
                          TRUE ~ 0))

modelo8 <- lm(lwage ~ exper + bsec + bsup, data=data)
summary(modelo8)
```

```

Call:
lm(formula = lwage ~ exper + bsec + bsup, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.99901 -0.31438 -0.07762  0.32933  1.55045 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.860162   0.180326   4.770 2.39e-06 ***
exper       0.008099   0.001594   5.081 5.23e-07 ***
bsec        0.479675   0.174428   2.750  0.00617 ** 
bsup        0.944581   0.177853   5.311 1.62e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.479 on 522 degrees of freedom
Multiple R-squared:  0.1926,	Adjusted R-squared:  0.188 
F-statistic: 41.51 on 3 and 522 DF,  p-value: < 2.2e-16
```