

Regresión Múltiple

```
b=read.table("datoEmpleado.DAT",header=T,sep="\t")
```

Ejemplo: (archivo datoEmpleado.DAT)

id	sexo	fechnac	educ	catlab	salario	salini	tiempemp
1	h	2/3/1952	15	3	57000	27000	98
2	h	5/23/1958	16	1	40200	18750	98
3	m	7/26/1929	12	1	21450	12000	98
4	m	4/15/1947	8	1	21900	13200	98
5	h	2/9/1955	15	1	45000	21000	98
6	h	8/22/1958	15	1	32100	13500	98
7	h	4/26/1956	15	1	36000	18750	98
8	m	5/6/1966	12	1	21900	9750	98
9	m	1/23/1946	15	1	27900	12750	98
10	m	2/13/1946	12	1	24000	13500	98
11	m	2/7/1950	16	1	30300	16500	98
12	h	1/11/1966	8	1	28350	12000	98
13	h	7/17/1960	15	1	27750	14250	98
14	m	2/26/1949	15	1	35100	16800	98
15	h	8/29/1962	12	1	27300	13500	97
16	h	11/17/1964	12	1	40800	15000	97
17	h	7/18/1962	15	1	46000	14250	97
18	h	3/20/1956	16	3	103750	27510	97
19	h	8/19/1962	12	1	42300	14250	97
20	m	1/23/1940	12	1	26250	11550	97
21	m	2/19/1963	16	1	38850	15000	97
22	h	9/24/1940	12	1	21750	12750	97
23	m	3/15/1965	15	1	24000	11100	97
24	m	3/27/1933	12	1	16950	9000	97

variables en el archivo

names(b)

```
[1] "id"      "sexo"    "fechnac" "educ"    "catlab"  "salario"
[7] "salini"  "tiempemp" "expprev" "minoría"
```

`b=b[, -c(1, 3)]` #Se elimina la columna identificación del trabajador y la fecha

`names(b)` #variables que intervendrán en el análisis

```
"sexo"      "educ"    "catlab"  "salario" "salini"  "tiempemp" "expprev"
"minoría"
```

Regresión Múltiple

Especificación del modelo:

Variable dependiente: salario (salario actual)

Variables independientes:

Cualitativas: sexo
 catlab (Categoría laboral)
 minoría (Pertenencia a grupo de minoría étnica)

Cuantitativas: salini (salario inicial)
 expprev (Experiencia previa)
 educ (Nivel educativo en años de estudio)
 tiempemp

Declaración de variables cualitativas como factores.

```
b$sexo=factor(b$sexo)
b$minoría=factor(b$minoría, labels=c("NO", "SI"))
b$catlab=factor(b$catlab, labels=c("Administrativo", "Seguro", "Directivo"))
```

Variables ficticias derivadas de las cualitativas

CatlabSeguro
catlabDirectivo (Referencia =Administrativo)

Sexom (Referencia=Hombre)

MinoríaSI (Referencia =No es minoría étnica)

Regresión Múltiple

Especificación del modelo:

$$\text{Salario} = \beta_0 + \beta_1 \text{sexom} + \beta_2 \text{educ} + \beta_3 \text{catlabSegur} + \beta_4 \text{catlabDirec} + \beta_5 \text{salini} + \beta_6 \text{tiempemp} + \beta_7 \text{expprev} + \beta_8 \text{minoría} + \varepsilon$$

Estimación de parámetros del modelo o ajuste del modelo:

Procedimiento: Mínimos cuadrados

Resultados:

Modelo con todas las variables pero sin interacción

```
lm(salario ~ ., data=b)
```

Call:

```
lm(formula = salario ~ ., data = b)
```

Coefficients:

(Intercept)	sexom	educ	catlabSegur	catlabDirec	salini
-6365.122	-1988.078	502.104	6853.342	11107.683	1.334
tiempemp	expprev	minoríaSI			
150.157	-21.907	-1033.569			

Regresión Múltiple

Resultados:

Residual standard error: 6808 on 465 degrees of freedom
Multiple R-Squared: 0.8437, Adjusted R-squared: 0.841
F-statistic: 313.8 on 8 and 465 DF, p-value: $< 2.2e-16$

Proporción de variación explicada por el modelo

Estimación de la desviación típica del término de error del modelo

Resultados:

Regresión Múltiple

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-6.365e+03	3.399e+03	-1.873	0.06175	.
sexom	-1.988e+03	7.652e+02	-2.598	0.00967	**
educ	5.021e+02	1.599e+02	3.141	0.00179	**
catlabSegur	6.853e+03	1.631e+03	4.203	3.16e-05	***
catlabDirec	1.111e+04	1.368e+03	8.121	4.19e-15	***
salini	1.334e+00	7.305e-02	18.267	< 2e-16	***
tiempemp	1.502e+02	3.134e+01	4.791	2.24e-06	***
expprev	-2.191e+01	3.575e+00	-6.127	1.91e-09	***
minoríaSI	-1.034e+03	7.872e+02	-1.313	0.18982	

Coefficientes
estimados

La variable minoría no tiene una pendiente significativa al nivel alfa del 5%

De las variables implicadas en el modelo la menos importante (menos significativa) es la variable minoría

P-valor o sig.=0,1898 > alfa=0,05 Por tanto se acepta $H_0 : \beta_8=0$
Se concluye que la variable minoría no sirve para explicar el Salario

Ajuste por pasos:

Regresión Múltiple

step()

```
> regT=(lm(salario~., data=b))
> step(regT)
Start: AIC=8375.86
salario ~ sexo + educ + catlab + salini + tiempemp + expprev +
mi noría
```

	Df	Sum of Sq	RSS	AIC
- mi noría	1	7.9914e+07	2.1634e+10	8.3760e+03
<none>			2.1554e+10	8.3760e+03
- sexo	1	3.1290e+08	2.1867e+10	8.3810e+03
- educ	1	4.5718e+08	2.2011e+10	8.3840e+03
- tiempemp	1	1.0638e+09	2.2618e+10	8.3970e+03
- expprev	1	1.7402e+09	2.3294e+10	8.4110e+03
- catlab	2	4.0028e+09	2.5557e+10	8.4530e+03
- salini	1	1.5467e+10	3.7021e+10	8.6300e+03

```
Step: AIC=8375.62
salario ~ sexo + educ + catlab + salini + tiempemp + expprev
```

	Df	Sum of Sq	RSS	AIC
<none>			2.1634e+10	8.3760e+03
- sexo	1	2.8073e+08	2.1915e+10	8.3800e+03
- educ	1	4.5218e+08	2.2086e+10	8.3830e+03
- tiempemp	1	1.0435e+09	2.2677e+10	8.3960e+03
- expprev	1	1.8121e+09	2.3446e+10	8.4120e+03
- catlab	2	4.0662e+09	2.5700e+10	8.4530e+03
- salini	1	1.5713e+10	3.7347e+10	8.6320e+03

```
Call:
lm(formula = salario ~ sexo + educ + catlab + salini + tiempemp +
data = b)
```

Coefficients:

(Intercept)	sexom	educ	catlabSegur	catlabDi rec	salini
-6585.523	-1870.063	499.308	6736.872	11255.523	1.341
tiempemp	expprev				
148.615	-22.283				

Automatiza el proceso de selección del modelo más adecuado por pasos. El argumento dirección permite indicar si se desea que el proceso sólo añada términos, sólo quite términos, o haga ambas operaciones.

El criterio para añadir o quitar términos es el valor del descenso sufrido por el estadístico AIC.

A menor valor de AIC mejor ajuste.

Tras una primera aproximación por pasos, podría someterse a un procedimiento anova el modelo, con aquellos términos menos significativos. O bien ajustar de nuevo el modelo con los términos seleccionados en step y contrastar la significatividad de los términos mediante test t.

Regresión Múltiple

Se elimina la variable minoría del ajuste y se ejecuta de nuevo el análisis de regresión sin esta variable

Ecuación del modelo a ajustar:

$$\text{Salario} = \beta_0 + \beta_1 \text{sexom} + \beta_2 \text{educ} + \beta_3 \text{catlabSegur} + \beta_4 \text{catlabDirec} + \beta_5 \text{salini} + \beta_6 \text{tiempemp} + \beta_7 \text{expprev} + \varepsilon$$

```
> lm(formula = salario ~ sexo + educ + catlab + salini + tiempemp + expprev, data = b)
```

Call:

```
lm(formula = salario ~ sexo + educ + catlab + salini + tiempemp + expprev, data = b)
```

Coefficients:

(Intercept)	sexom	educ	catlabSegur	catlabDirec	salini
-6585.523	-1870.063	499.308	6736.872	11255.523	1.341
tiempemp	expprev				
148.615	-22.283				

Comprobaremos si las variables incluidas en el modelo son significativas al nivel $\alpha=0,05$. Si la respuesta es afirmativa, concluiremos que todas sirven para explicar el Salario

Regresión Múltiple

Inferencia:

Contraste de hipótesis de nulidad de cada una de las pendientes del modelo nuevo propuesto (sin variable minoría)

$$\text{Salario} = \beta_0 + \beta_1 \text{sexom} + \beta_2 \text{educ} + \beta_3 \text{catlabSegur} + \beta_4 \text{catlabDirec} + \beta_5 \text{salini} + \beta_6 \text{tiempemp} + \beta_7 \text{expprev} + \varepsilon$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-6.586e+03	3.398e+03	-1.938	0.05319	.
sexom	-1.870e+03	7.605e+02	-2.459	0.01429	*
educ	4.993e+02	1.600e+02	3.121	0.00191	**
catlabSegur	6.737e+03	1.630e+03	4.134	4.22e-05	***
catlabDirec	1.126e+04	1.364e+03	8.251	1.63e-15	***
salini	1.341e+00	7.291e-02	18.397	< 2e-16	***
tiempemp	1.486e+02	3.135e+01	4.741	2.83e-06	***
expprev	-2.228e+01	3.567e+00	-6.248	9.41e-10	***

Todos los pvalores son inferiores a 0,05

P-valor o sig < alfa = 0,05

Variable sexo

$H_0: \beta_1 = 0$ (se asume que la variable sexom no es importante)

$H_1: \beta_1 \neq 0$ (la variable es importante para explicar el salario)



P-valor o sig. = 0,0143 < alfa = 0,05

Por tanto se rechaza H_0 y se acepta H_1

Se concluye que la variable Sexo sirve para explicar el Salario

Variable educ

$H_0: \beta_2 = 0$ (se asume que la variable educ no es importante)

$H_1: \beta_2 \neq 0$ (la variable es importante para explicar el salario)



Variable educ

P-valor o sig. = 0,00191 < alfa = 0,05

Por tanto se rechaza H_0 y se acepta H_1

Se concluye que la variable educ sirve para explicar el Salario

Regresión Múltiple

Todas las variables incluidas en el modelo son significativas al nivel $\alpha=0,05$. Todas sirven para explicar el Salario

Ecuación del modelo propuesto:

$$\text{Salario} = \beta_0 + \beta_1 \text{sexom} + \beta_2 \text{educ} + \beta_3 \text{catlabSegur} + \beta_4 \text{catlabDirec} + \beta_5 \text{salini} + \beta_6 \text{tiempemp} + \beta_7 \text{expprev} + \varepsilon$$

Ecuación del modelo ajustado:

$$\text{Salario} = -6586 - 1870 \text{sexom} + 499,3 \text{educ} + 6737 \text{CatlabSegur} + 11260 \text{catlabDirec} + 1,341 \text{salini} + 148,6 \text{tiempemp} - 22,28 \text{expprev}$$

Regresión Múltiple

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-6.586e+03	3.398e+03	-1.938	0.05319	.
sexom	-1.870e+03	7.605e+02	-2.459	0.01429	*
educ	4.993e+02	1.600e+02	3.121	0.00191	**
catlabSegur	6.737e+03	1.630e+03	4.134	4.22e-05	***
catlabDirec	1.126e+04	1.364e+03	8.251	1.63e-15	***
salini	1.341e+00	7.291e-02	18.397	< 2e-16	***
tiempemp	1.486e+02	3.135e+01	4.741	2.83e-06	***
expprev	-2.228e+01	3.567e+00	-6.248	9.41e-10	***

Coefficientes
estimados

Interpretación de coeficientes de variables continuas:

Por cada unidad que se incrementa el nivel educativo (años estudio) se espera un aumento del salario de 499,3 unidades en el salario, supuestos estables los valores en las otras variables

Interpretación de los coeficientes de variables ficticias:

El aumento esperado en el salario al pasar de un trabajador con categoría laboral administrativo (referencia) a otro de seguridad es de 6737, supuestos estables los valores en las otras variables.

El descenso esperado en el salario al pasar de un hombre (referencia) a una mujer es de 1870 unidades, supuestos estables los valores en las otras variables