

# Talleres de Análisis Político I

Sesión 4

11/12/2023

Pau Vall-Prat

[pau.vall@uc3m.es](mailto:pau.vall@uc3m.es)

# Ejercicios

- Combinad la base de datos de resultados electorales con información sociodemográfica a nivel municipal del INE
  - Probad con distintos tipos de join()
  - ¿Por qué los resultados son distintos?
- Combinad datos de censo y votantes *a nivel provincial* con información a nivel provincial sobre votos por correo

# Modificar la estructura de una base de datos

El paquete tidyr

# Datos tidy

“**TIDY DATA** is a standard way of mapping the meaning of a dataset to its structure.”  
—HADLEY WICKHAM

## In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

each column a variable

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

each row an observation

- En la filosofía *tidy* cada fila debe representar una observación.
  - Muchas veces la estructura de los datos no es esta
  - En una encuesta panel
    - Fila: individuo y columnas variables para cada ola
    - Fila: individuo y ola: columnas variables
- Idealmente, *con filosofía tidy*
- Fila: individuo, ola y pregunta: una única columna con datos

# Reestructurar los datos

- A veces nos puede interesar tener los datos en formato
- Ancho (wide)
  - Una fila, una unidad/observación
- Largo (long)
  - Una fila, un dato

id	r1	r2	r3
A1	1	3	5
A2	2	4	6



id	r	valor
A1	r1	1
A1	r2	3
A1	r3	5
A2	r1	2
A2	r2	4
A2	r3	6

- Se usa el paquete tidyr
- Especialmente las funciones
  - `pivot_longer()`: de wide a long
  - `pivot_wider()`: de long a wide

# pivot\_longer

- Principales argumentos
  - data: indicar nombre del objeto con los datos
  - cols: indicar los nombres de las variables/columnas que queremos que pasen a ser filas
    - En el ejemplo anterior: r1 y r2
  - names\_to: especificar el nombre nuevo de la variable que identificará valores de columnas
    - En el ejemplo anterior: "r"
  - values\_to: especificar el nombre de la variable con los valores
    - En el ejemplo anterior: "valor"

```
pivot_longer(  
  data,  
  cols,  
  names_to = "name",  
  names_prefix = NULL,  
  names_sep = NULL,  
  names_pattern = NULL,  
  names_ptypes = list(),  
  names_transform = list(),  
  names_repair = "check_unique",  
  values_to = "value",  
  values_drop_na = FALSE,  
  values_ptypes = list(),  
  values_transform = list(),  
  ...  
)
```

id	r1	r2
a1	1	3
a2	2	4

id	name	value
a1	r1	1
a1	r2	2
a2	r1	3
a2	r2	4

# Cómo seleccionar columnas

- Según rango de la r1 a la r3 con: `r1:r3`
- Según característica de la columna: `starts_with("...")`
- Según las posiciones: `2:4`
- Según un vector: `c(r1, r2, r3)`
- Según la clase de la columna: `where(is.numeric)`
- Combinando criterios con los operadores habituales (`!` `&` `|`)



# pivot\_wider

- Principales argumentos
  - data: indicar nombre del objeto con los datos
  - names\_from: especificar el nombre la variable cuyos distintos valores identificarán nuevos nombres de columnas
    - En el ejemplo anterior: "r"
  - values\_from: especificar el nombre de la variable con los valores que deben rellenar las celdas de las columnas
    - En el ejemplo anterior: "valor"

```
pivot_wider(  
  data,
```

```
  id_cols = NULL,
```

```
  names_from = name,
```

```
  names_prefix = "",  
  names_sep = "_",
```

```
  names_glue = NULL,  
  names_sort = FALSE,
```

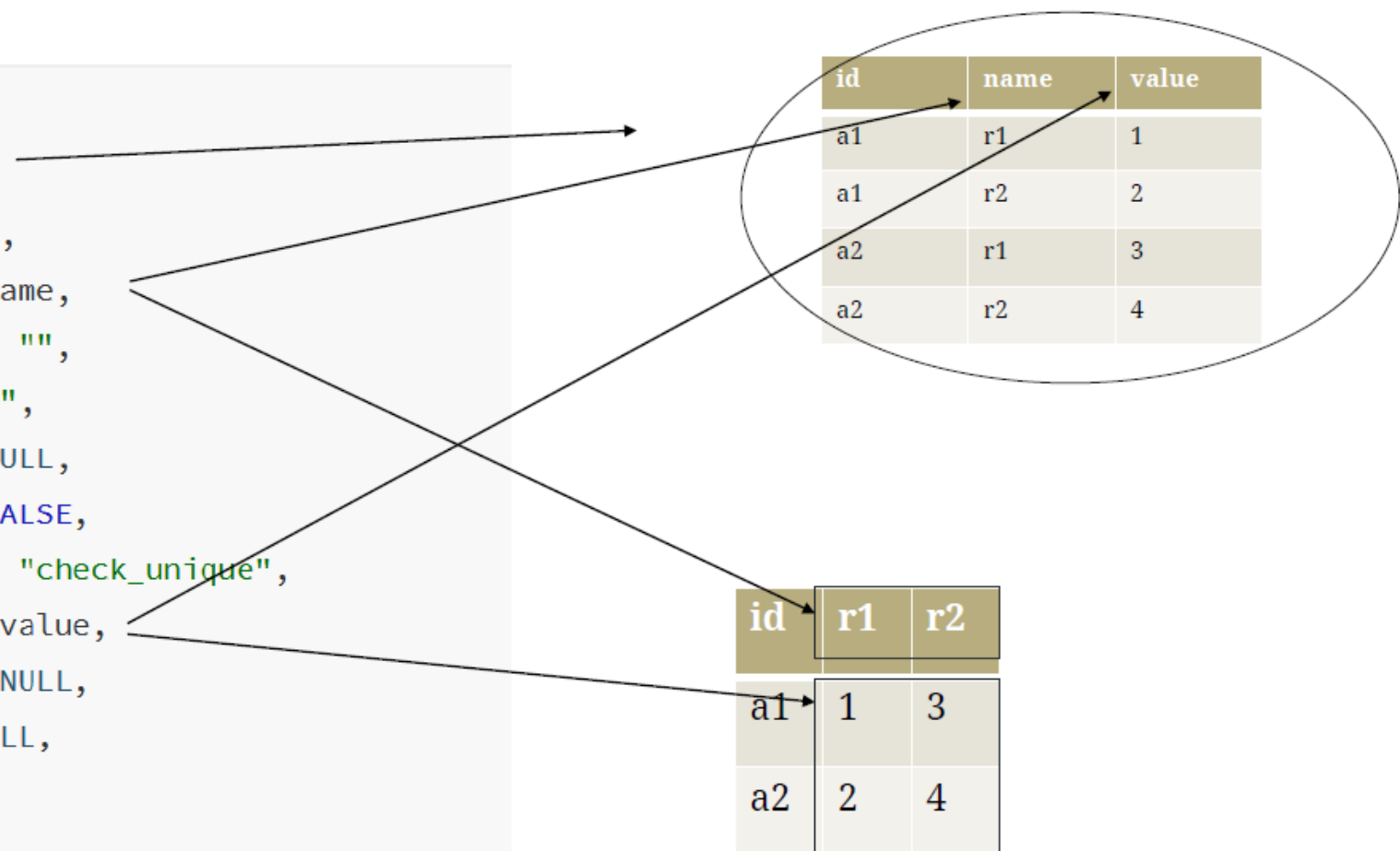
```
  names_repair = "check_unique",
```

```
  values_from = value,
```

```
  values_fill = NULL,  
  values_fn = NULL,
```

```
  ...
```

```
)
```



id	name	value
a1	r1	1
a1	r2	2
a2	r1	3
a2	r2	4

id	r1	r2
a1	1	3
a2	2	4

## *Ejercicio (15-20')*

- Queremos conocer la posición del PP en el ranquin de partido más votado para cada municipio. Queremos una variable que lo indique.

Recomendación: seleccionad las variables imprescindibles

Pistas:

- Reshape wide → long
- Group\_by, arrange, mutate [Pista: row\_number()]
- Reshape long → wide
- Join!

# Regresiones

# Definición

- Una regresión consiste en buscar la línea que pase lo más cerca posible de los puntos en un diagrama de dispersión
- Muestra el vínculo en términos matemáticos entre una variable dependiente y una variable independiente
- Este vínculo es lineal
- Permite estimar un valor promedio condicional en la variable dependiente a partir de ciertos valores de las variables independientes.

# Regresión: representación matemática

Idealmente

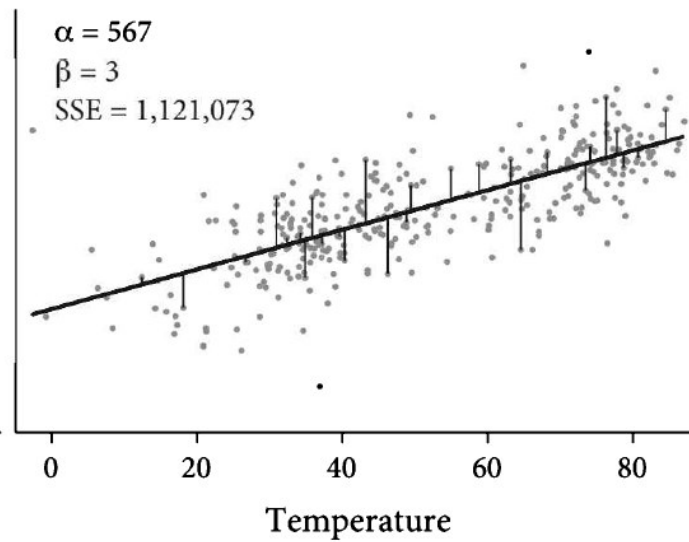
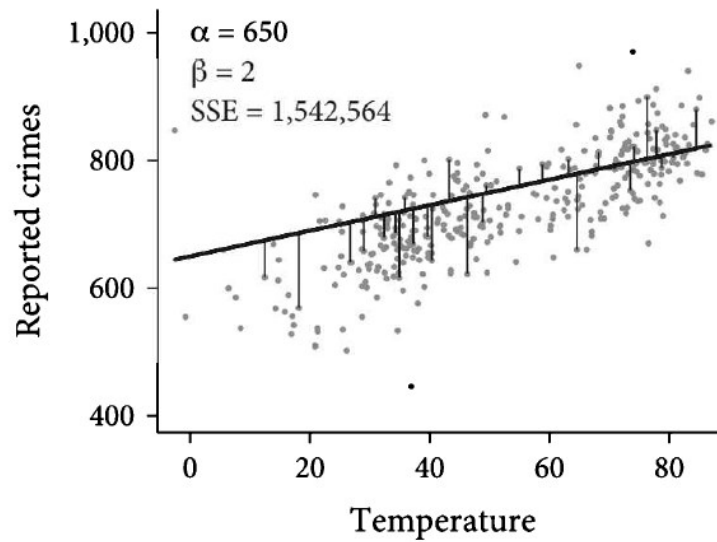
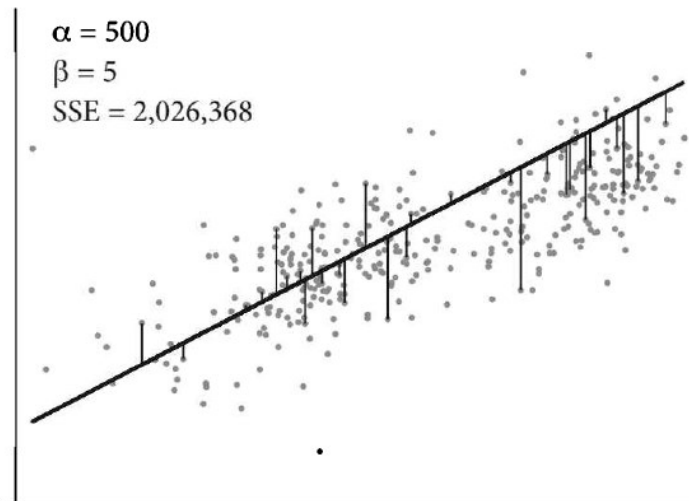
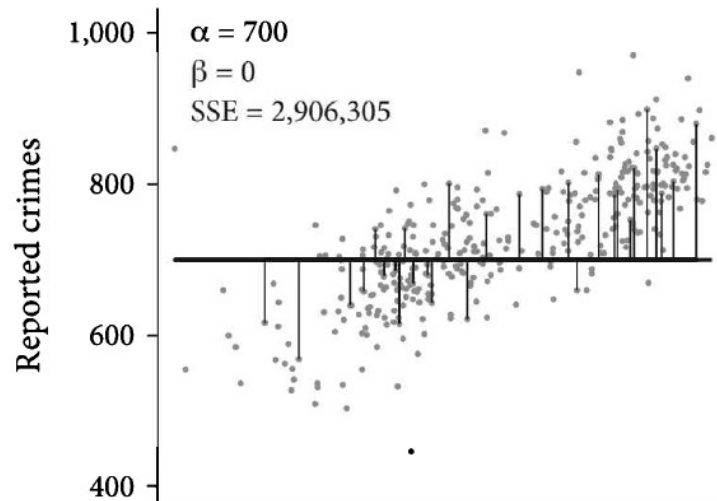
$$Y = \alpha + \beta_1 \cdot x_1 + \epsilon$$

*Parámetros de regresión*

En la práctica

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 \cdot x_1$$

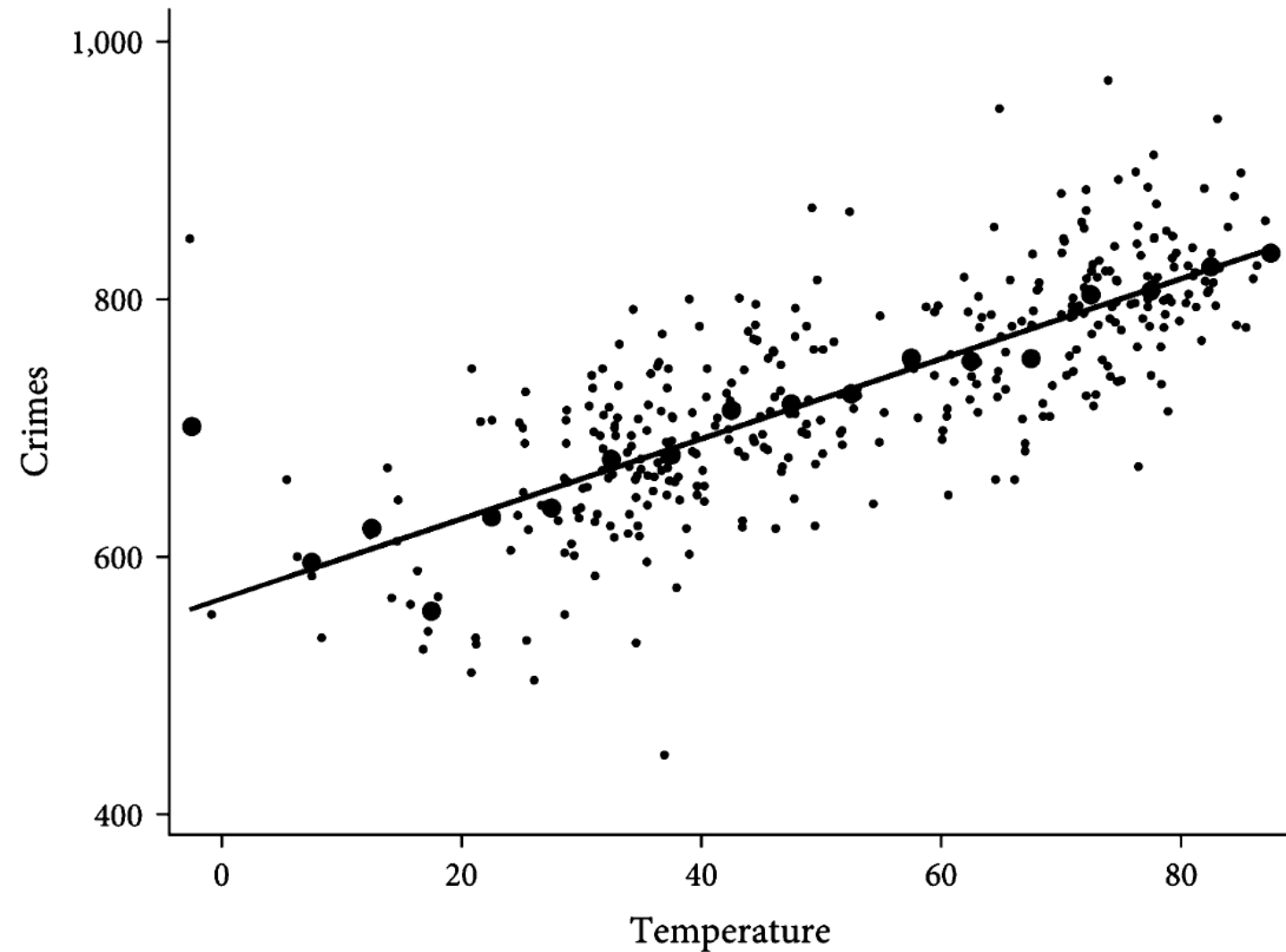
Todos estos elementos van a aparecer en los resultados de una tabla de regresión



# Ventajas de la regresión OLS

- Informa de muchos aspectos de la relación entre variables
  - Dirección
  - Fuerza
  - Magnitud
  - Significación estadística
- Permite estimar rápidamente valores para cualquier rango
- Supuestos básicos fáciles de entender y comunicar
- Es parsimoniosa





Una tabla de medias condicionales también sería informativa, pero menos parsimoniosa

# Regresiones con VI dicotómica

```
colony      peace
<fct>      <dbl>
Never colony 1.85
Colony      2.20
```

peace: indicador de “pacificidad” (presencia de conflicto en un país, en escala 1-5)

```
Call:
lm(formula = gpi_gpi ~ colony, data = qog)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.8920	-0.3134	-0.1074	0.2016	1.7446

colony: indica si el país fue o es una colonia (variable dicotómica: 0 no, 1 sí)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.84740	0.05793	31.892	< 2e-16 ***
colonyColony	0.34859	0.07502	4.647	7.03e-06 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.467 on 159 degrees of freedom  
(33 observations deleted due to missingness)

Multiple R-squared: 0.1196, Adjusted R-squared: 0.114

F-statistic: 21.59 on 1 and 159 DF, p-value: 7.031e-06

# Interpretar una tabla de regresión

	(1) Left-Right	(2) Left-Right
Woman	-0.274*** (0.0282)	
Age		0.000211 (0.000780)
Constant	5.372*** (0.0205)	5.217*** (0.0422)
N	28445	28254
R-Square	0.00332	0.00000259
Adj. R-Square	0.00329	-0.0000328

Standard errors in parentheses

\* p<0.05, \*\* p<0.01, \*\*\* p<0.001

Datos en bruto de la ESS10

# Regresión multivariante

- Los modelos de regresión multivariante modelizan una relación lineal entre
  - Una variable dependiente
  - Dos o más variables independientes

$$Y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \epsilon$$

- Recordad, en la práctica

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2$$

# Interpretación

- Al añadir más elementos en los modelos de regresión la interpretación de los coeficientes cambia ligeramente
  - Constante: Valor de  $y$  cuando  $x_1$  y  $x_2$  son iguales a 0
  - $\beta_1$ : efecto de  $x_1$  sobre  $y$  siempre que  $x_2$  se mantenga constante
  - $\beta_2$ : efecto de  $x_2$  sobre  $y$  siempre que  $x_1$  se mantenga constante
- \* se mantenga constante = cláusula *ceteris paribus*
- Hay que cuidar el lenguaje y distinguir la interpretación de coeficientes en función de regresiones bivariadas o multivariantes

# Ejemplo

Queremos entender por qué hay  
variación en el número de  
consejerías de las CCAA

**H1:** A mayor número de partidos,  
más consejerías  
*Para acomodar cargos para todos  
los partidos*

**H2:** Los gobiernos de izquierdas  
tendrán más consejerías  
*Tienden a gastar más*

```
Call:
lm(formula = cabinet_size ~ num_parties + left_cabinet, data = rcs)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.6269	-1.6269	-0.2862	1.1358	5.3731

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	10.2050	0.3061	33.343	<2e-16 ***
num_parties	0.4220	0.1872	2.254	0.0249 *
left_cabinet	0.2373	0.2504	0.948	0.3441

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.08 on 278 degrees of freedom

Multiple R-squared: 0.02088, Adjusted R-squared: 0.01384

F-statistic: 2.964 on 2 and 278 DF, p-value: 0.05324

Fuente de los datos: Vall-Prat & Rodon (2017)

# Ejercicios

- Haced regresiones con la base de datos de resultados electorales de 2019 combinada con datos sociodemográficos
  1. Una regresión simple
  2. Una regresión múltiple

# En la próxima sesión

Interpretar interacciones...