Documento base

Pablo Álvarez Arnedo

2026-02-11

Table of contents

Capítulos

```
1 + 1
```

```
[1] 2
```

1 Introducción

```
1 + 1
```

```
[1] 2
```

2 Capítulo 2: Datos longitudinales

2.1 ¿Qué son los datos longitudinales?

Los **datos longitudinales** son aquellos que obtenemos al realizar distintas medidas a un individuo (individuos, regiones, células, etc.). Dichas medidas se pueden observar repetidamente a lo largo del tiempo (análisis temporal), del espacio (análisis espacial), o a lo largo del espacio y tiempo (análisis espacio-temporal); es por eso que a los datos longitudinales también se les conoce como medidas repetidas. Esta forma de observar las medidas nos permite detectar cambios o tendencias temporales en nuestras variables, lo cual nos puede llevar a observar patrones que nos sería difícil examinar en otro tipo de investigaciones. Este tipo de datos es común en estudios donde se busca evaluar cómo evolucionan ciertas características o mediciones bajo distintas condiciones o tratamientos. En el ámbito biosanitario, los datos longitudinales son fundamentales para investigar la progresión de enfermedades, la efectividad de tratamientos y el impacto de intervenciones médicas.

2.1.1 Características principales

1. **Medidas repetidas**: cada unidad tiene varias observaciones en diferentes momentos temporales.
2. **Estructura jerárquica**: las observaciones están agrupadas por unidades (e.g., pacientes, regiones).
3. **Dependencia entre observaciones**: las mediciones dentro de la misma unidad tienden a estar correlacionadas.
4. **Variables**: como la mayoría de medidas se realizan en distintos del tiempo, diremos que son variables **tiempo-dependientes**; pero también hay que tener en cuenta que hay otras variables que cambian igual en el tiempo para todos los sujetos (como la edad) que **no** consideraremos tiempo-dependientes y otras que directamente consideraremos **constantes** como el sexo.

2.1.2 Componentes de la respuesta de cada individuo

1. **Efecto fijo**: función de las covariables
2. **Efecto aleatorio**: muestra la variación entre individuos
3. **Error**: originado por las mediciones o a variables no registradas

2.1.3 Objetivos

1. Observar la **evolución** de una variable a lo largo del tiempo/espacio
2. Comparar si la **evolución** de una variable a lo largo del tiempo/espacio es **igual** para distintas partes de la población
3. Tratar de observar e identificar **patrones** en el desarrollo de una variable a lo largo del tiempo/espacio

2.1.4 Ejemplos de datos longitudinales

1. **Ámbito biosanitario**: medidas repetidas de presión arterial en un grupo de pacientes durante un tratamiento.
2. **Educación**: evaluación de los puntajes de un estudiante a lo largo de varios exámenes anuales.
3. **Ciencias sociales**: encuestas de opinión realizadas periódicamente a las mismas personas.
4. **Alimentación**: estudio de diferentes dietas a diferentes grupos de la población a lo largo del tiempo a través de medidas tales como actividad física, medidas antropométricas, etc.

2.2 ¿Por qué no se puede usar la estadística clásica?

La **estadística clásica** (e.g., regresión lineal simple) supone que todas las observaciones son independientes entre sí. Sin embargo, en datos longitudinales, esta suposición no se cumple debido a la correlación entre observaciones tomadas de la misma unidad. Pero este no es el único motivo por el cual no podemos usar la estadística clásica únicamente para analizar datos longitudinales.

2.2.1 Problemas al aplicar técnicas clásicas

1. **Dependencia entre observaciones**: como bien habíamos comentado, los datos longitudinales tienen una estructura que lleva a que las observaciones sobre el mismo individuo estén correlacionadas.
2. **Correlación de los errores**: siguiendo el punto anterior, los datos longitudinales contienen una correlación en los errores que no puede ser modelada correctamente a través de modelos de estadística clásica como podría ser un modelo de regresión lineal simple. Esto ocurre porque las medidas repetidas pueden estar influenciadas por factores externos o por variables no registradas en modelos clásicos.

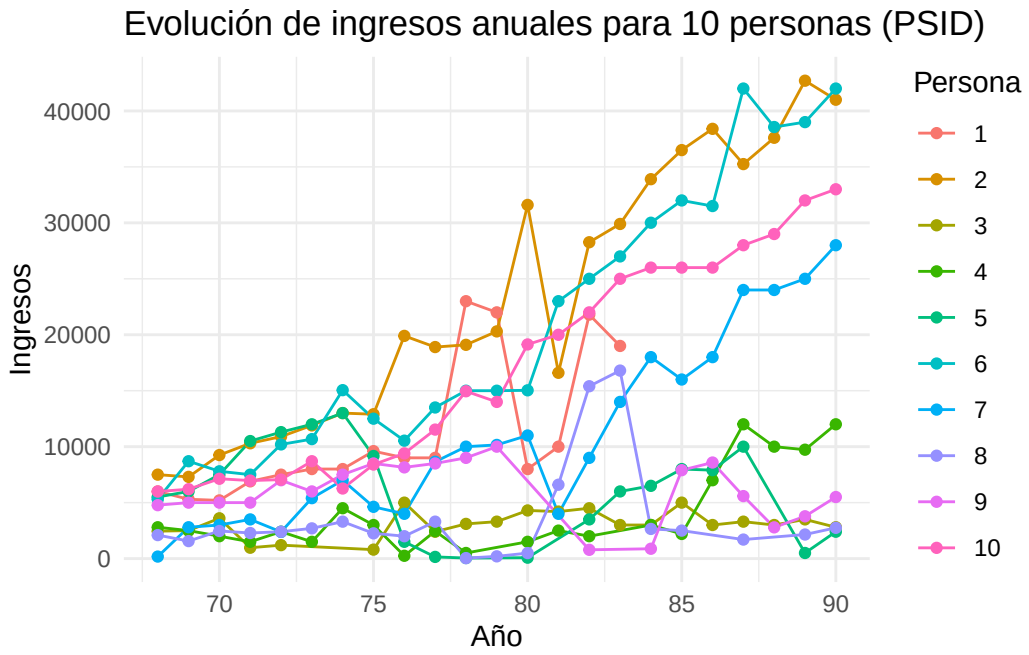
3. **Variabilidad**: otro de los motivos por los que no se pueden usar modelos clásicos para datos longitudinales es que estos modelos no tienen un enfoque apropiado para la variabilidad de los datos, ya que adaptan una estructura homogénea la cual no corresponde con un modelo de datos longitudinales en el cual hay que tener en cuenta las diferencias entre individuos.
4. **Sesgo**: a raíz del punto anterior, surge otro problema que lleva a evitar utilizar estadística clásica para este tipo de datos: los sesgos. Al ignorar dichas diferencias entre individuos y la dependencia entre observaciones, las estimaciones no reflejan correctamente la relación entre variables ya que no cuentan con la existencia de efectos aleatorios, entre otros.

2.2.2 Ejemplo conceptual

Vamos a considerar un conjunto de datos sobre ingresos anuales de personas a lo largo de varios años (psid). Vamos a utilizar un modelo regresión lineal simple para modelar los ingresos en función del tiempo, ignorando la correlación entre mediciones.

```
Warning: package 'faraway' was built under R version 4.4.2
```

```
Warning: package 'ggplot2' was built under R version 4.4.2
```

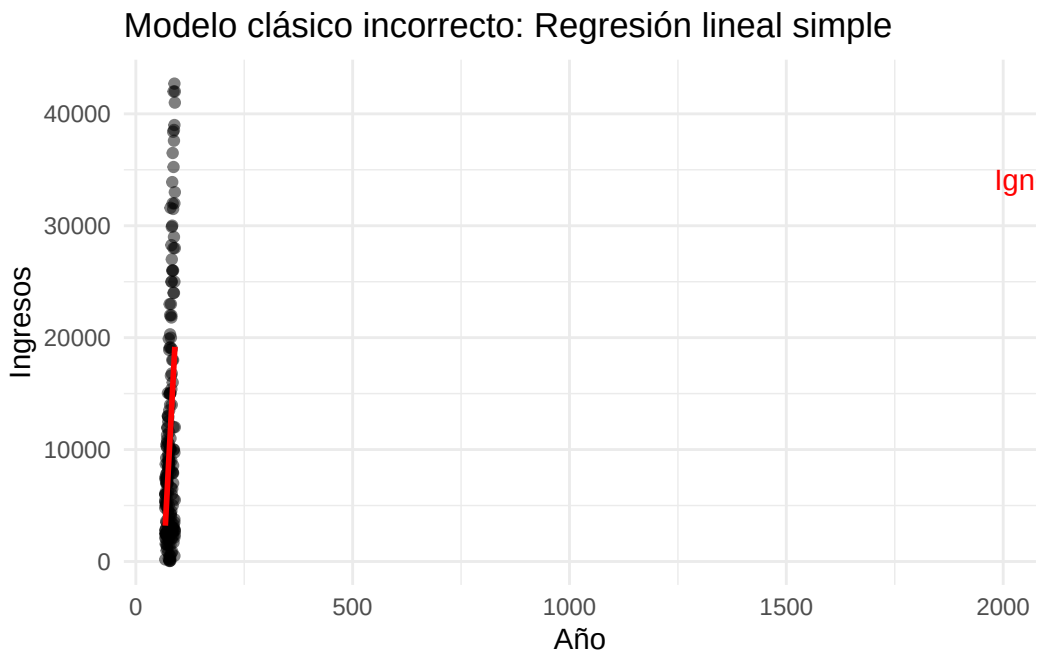


Este gráfico muestra la evolución de los ingresos anuales para diferentes personas a lo largo del tiempo. Se observa que los datos son heterogéneos y varían significativamente entre individuos, lo que muestra la dependencia entre observaciones; algo que viola los supuestos básicos de independencia de las observaciones, fundamentales para modelos clásicos como la regresión lineal simple.

```
Call:
lm(formula = income ~ year, data = psid_subset)

Residuals:
     Min       1Q   Median       3Q      Max
-17956.7  -7314.1   -380.3   4693.2  24996.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -46198.57    7519.91  -6.143 4.02e-09 ***
year           726.46      95.33   7.621 8.65e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9192 on 209 degrees of freedom
Multiple R-squared:  0.2175,    Adjusted R-squared:  0.2137
F-statistic: 58.08 on 1 and 209 DF,  p-value: 8.655e-13
```

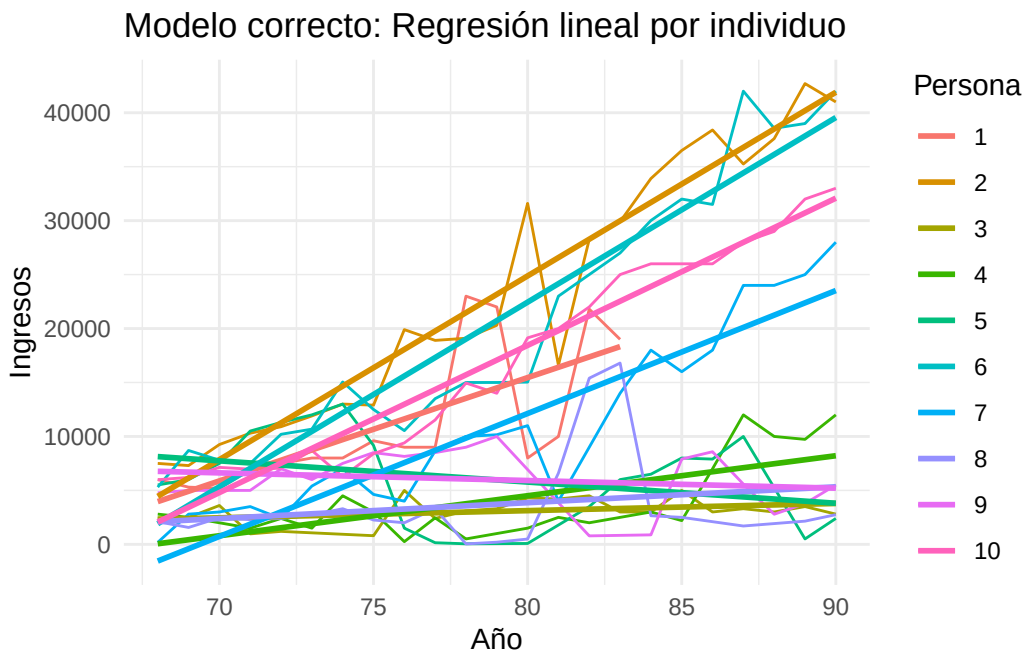
En la salida del modelo, vemos cómo el modelo asume que la variabilidad entre individuos se puede representar con un único coeficiente, ignorando por completo la dependencia entre observaciones. Además, dicho coeficiente tiene un valor muy bajo, mostrando que el modelo explica muy poca variabilidad de los datos y que, por tanto, no nos sirve para analizar datos longitudinales.

```
`geom_smooth()` using formula = 'y ~ x'
```



Este gráfico muestra cómo la regresión lineal simple aplicada a estos datos genera una representación distorsionada, ignorando por completo la correlación de los datos longitudinales; dando lugar a un mal ajuste y a resultados estadísticos inapropiados que demuestran por qué no debemos utilizar estadística clásica para este tipo de datos.

```
`geom_smooth()` using formula = 'y ~ x'
```



En esta gráfica, en la que ajustamos un modelo para cada individuo, mostrando que las pendientes e interceptos varían significativamente, destacando la necesidad de modelos mixtos.

2.3 Modelos mixtos

Para analizar datos longitudinales de manera adecuada, se deben emplear modelos mixtos, que permiten: - Capturar la variabilidad entre individuos mediante efectos aleatorios. - Modelar la correlación entre observaciones dentro de una misma unidad. - Incluir covariables tanto a nivel individual como grupal.

2.3.1 Ventajas de los modelos mixtos

- Flexibilidad para incluir efectos específicos por individuo o grupo.
- Estimación precisa de la incertidumbre, respetando la dependencia entre observaciones.
- Generalización a estructuras de datos complejas.