

Visualización, descriptivas y regresiones

Sesión 3 — Unidades 3 y 4

Prof. Eduard F. Martínez-González

Curso Nivelatorio de Programación en R
CIENFI — Universidad Icesi

Julio de 2026

Tabla de contenidos

- 1 Dónde estamos
- 2 Unidad 3: descriptivas y visualización
- 3 Unidad 4: regresiones básicas
- 4 Cierre y trabajo de la semana

De dónde venimos y qué haremos hoy

- **Sesión 1 (U1):** R y RStudio, objetos, vectores, data frames, proyectos y scripts con estilo.
- **Sesión 2 (U2):** importar, diagnosticar, limpiar, transformar con `dplyr`, agrupar y cruzar bases.
- **Hoy (U3 y U4):** cerramos el flujo de trabajo — **describir, visualizar, estimar y reportar**.

El flujo completo

Importar → Diagnosticar → Limpiar → Transformar → **Describir/Visualizar** → **Estimar** → Comunicar

Hoy también se **lanza el proyecto final**.

Cambiamos de base: de firmas a personas

Hasta ahora trabajamos con `innovacion_empresas.csv` (firmas, base sucia).
Hoy usamos la **otra base del curso**:

`geih_nivelacion.csv` — extracto docente de la GEIH (DANE)

- **21.821 ocupados** (18 a 65 años, ingreso laboral positivo), 24 cabeceras.
- Datos **reales** del DANE, simplificados con fines docentes.
- Viene **limpia**: hoy vamos directo a describir, graficar y estimar.

Variables: `sexo`, `edad`, `anios_educacion`, `nivel_educativo`, `departamento`, `ingreso_laboral` (COP/mes), `horas_semana`.

Punto de partida: importar y diagnosticar

```
library(pacman)
p_load(rio, tidyverse, skimr, fixest, modelsummary)

raw <- "data/raw/geih_nivelacion.csv"
geih <- import(raw)

glimpse(geih)      # tipo de cada variable
skim(geih)         # faltantes, min/max, percentiles
summary(geih$ingreso_laboral)
```

La pregunta que abre la sesión

El ingreso tiene una **cola larguísima** a la derecha: la media queda muy por encima de la mediana. ¿El promedio describe bien a esta población, o la mediana es más honesta?

Descriptivas por grupo: el puente con la Unidad 2

```
geih %>%  
  group_by(sexo) %>%  
  summarise(n = n(),  
            ingreso_mediano = median(ingreso_laboral),  
            educacion_prom = mean(anios_educacion),  
            .groups = "drop")
```

sexo	n	ingreso mediano	educación prom.
hombre	11.786	850.000	10,4
mujer	10.035	781.242	11,3

El hallazgo que guiará la sesión: las mujeres tienen *más* educación promedio y *menos* ingreso mediano (brecha $\approx 8,1\%$).

La gramática de ggplot2: capas que se suman

Un gráfico se construye por capas, y las capas se unen con + (no con %>%).

```
# capa 1: el lienzo (datos + ejes) -- todavía vacío
ggplot(geih, aes(x = años_educación, y = log(ingreso_laboral)))

# capa 2: + la geometría
... + geom_point()

# capa 3: + etiquetas y tema
... + labs(title = "Educación e ingreso laboral",
           caption = "Fuente: GEIH (DANE)") +
  theme_minimal()
```

- `aes()` = **estética**: qué variable va en cada eje, color, relleno.
- Un gráfico sin título ni fuente está incompleto.

La geometría se elige según la pregunta

¿Qué quiero mostrar?	Geometría
La distribución de una variable	<code>geom_histogram()</code>
Comparar la distribución entre grupos	<code>geom_boxplot()</code>
La relación entre dos variables	<code>geom_point()</code> (+ <code>geom_smooth()</code>)
Comparar categorías (valores ya calculados)	<code>geom_col()</code>

Galerías de referencia

<https://www.data-to-viz.com/> · Cheatsheet de ggplot2 (Posit)

Distribución: por qué trabajamos en logaritmos

```
# en NIVELES: asimetria extrema, cola larguísima a la derecha
ggplot(geih, aes(x = ingreso_laboral)) + geom_histogram(bins = 50)

# en LOGS: la distribucion se vuelve legible (casi simetrica)
ggplot(geih, aes(x = log(ingreso_laboral))) + geom_histogram(bins = 50)
```

Por qué importa

No es un truco estético: el log hace legible la distribución **y** permite leer los coeficientes de la regresión como **cambios porcentuales**. Volveremos a esto en la Unidad 4.

Comparar grupos: boxplot

```
ggplot(geih, aes(x = nivel_educativo, y = log(ingreso_laboral))) +  
  geom_boxplot() +  
  labs(x = NULL, y = "Log del ingreso laboral") +  
  theme_minimal()
```

La caja es el 50% central (p25–p75), la línea es la **mediana** y los puntos son valores atípicos.

Ingreso mediano por nivel:

ninguno 400.000 · primaria 600.000 ·
secundaria 720.000 · media 800.000 ·
superior 1.200.000

Ojo: hay que **ordenar** el nivel educativo con `factor(..., levels = ...)`, o R lo ordena alfabéticamente.

Estética: dentro vs. fuera de aes()

```
# FUERA de aes(): un valor fijo para todo el grafico
geom_boxplot(fill = "steelblue")

# DENTRO de aes(): el color MAPEA una variable (y aparece la leyenda)
ggplot(geih, aes(x = nivel_educativo, y = log(ingreso_laboral),
                fill = sexo)) + geom_boxplot()
```

Error clásico

Poner `color = "blue"` *dentro* de `aes()`: R lo interpreta como una **variable** llamada "blue" y termina pintando todo de rojo con una leyenda absurda. Lo provocaremos en clase.

Barras y facetas

```
# barras: primero resumen, LUEGO grafico con geom_col()
ggplot(resumen_dpto, aes(x = reorder(departamento, ingreso_med),
                                   y = ingreso_med)) +
  geom_col(fill = "steelblue") + coord_flip()

# facetas: un panel por grupo, mismo codigo
... + facet_wrap(~ sexo)
```

- `reorder()`: ordena las barras **por su valor**; sin él, R usa orden alfabético.
- `coord_flip()`: barras horizontales para que las etiquetas se lean.
- `facet_wrap()`: varios gráficos comparables entre sí, sin repetir código.

Exportar: la regla del curso

```
ggsave("output/01_graphs/brecha_sexo_educacion.png",  
       plot = fig_brecha, width = 8, height = 5, dpi = 300)  
  
export(tabla_desc, "output/02_tables/descriptivas_sexo_nivel.csv")  
  
datasummary_skim(geih)    # tabla descriptiva en una línea
```

Nada de capturas de pantalla

Toda figura y toda tabla del curso se **exporta desde el código**, con dimensiones y resolución explícitas. Si el resultado no se puede regenerar corriendo el script, no sirve.

De describir a cuantificar

El gráfico de dispersión ya nos mostró que educación e ingreso van juntos. La regresión responde **cuánto**.

```
modelo_1 <- lm(log(ingreso_laboral) ~ anios_educacion, data = geih)
summary(modelo_1)
```

- La fórmula $y \sim x$ se lee: “*log del ingreso explicado por educación*”.
- En la salida: ubique Estimate, Std. Error, $\Pr(>|t|)$ y R-squared.
- El modelo es un **objeto**: `coef()`, `nobs()`, `summary()`.

Resultado con la GEIH

Un año adicional de educación se asocia a un ingreso **9,3 % mayor**, en promedio ($N = 21,821$).

Controles: comparar personas más parecidas

```
modelo_2 <- lm(log(ingreso_laboral) ~ años_educacion + edad, data = geih)
cor(geih$edad, geih$años_educacion) # = -0.237
```

- Al agregar edad, el retorno a la educación compara personas de la **misma edad**.
- ¿El coeficiente subió o bajó? **Pista:** en Colombia las cohortes jóvenes estudian más — la correlación entre edad y educación es **negativa** ($-0,24$).

Esta es la lógica de todo control: mantener fijo algo que, si se mueve, contamina la comparación.

Dummies: variables categóricas en la regresión

```
modelo_3 <- lm(log(ingreso_laboral) ~ anios_educacion + edad + sexo,  
              data = geih)  
  
# para dummies con efectos grandes, el % exacto:  
exp(coef(modelo_3)["sexomujer"]) - 1
```

- sexo es **texto** y R crea la dummy solo: no hay que codificar nada.
- El **nombre** del coeficiente dice quién es la referencia: sexomujer $\Rightarrow$ la referencia es *hombre*.

¿Por qué no basta con lm?

Heterogeneidad no observada entre departamentos

El ingreso mediano va de **700.000** a **1.000.000** según el departamento, y la educación promedio de **9,7** a **12,3** años. Los departamentos difieren en cosas que afectan el ingreso y **no están en la base**; al correlacionarse con la educación, el coeficiente queda **sesgado por variable omitida**.

Solución: efectos fijos de departamento.

Qué hacen los efectos fijos

- Incluir efectos fijos de departamento equivale a meter **una dummy por departamento** (24 en la GEIH).
- Eso **absorbe** todo lo que es constante dentro de cada departamento: costo de vida, estructura productiva, instituciones locales... aunque no lo hayamos medido.
- El coeficiente pasa a comparar personas **dentro del mismo departamento**, no entre departamentos distintos.

Lo que los efectos fijos NO hacen

Solo controlan lo que es **fijo dentro del grupo**. No resuelven la habilidad, la motivación o el contexto familiar, que varían *entre personas del mismo departamento*. Siguen sin darnos causalidad.

fixest: las dos correcciones en una línea

```
modelo_fe <- feols(log(ingreso_laboral) ~ años_educacion + edad + sexo  
                  | departamento,          # efectos fijos tras la barra
```

- Misma fórmula que `lm`, más la barra `|` para los efectos fijos.
- `feols` los **absorbe**: no imprime 23 dummies que nadie quiere leer, sino la línea Fixed-effects: departamento: 24.
- Al comparar con el modelo (3): si el **coeficiente** casi no se mueve, el sesgo por departamento era menor; el **error estándar** sí cambia — esa es la corrección por heterocedasticidad.

La tabla profesional: modelsummary

```
modelos <- list("(1) Simple" = modelo_1, "(2) + Edad" = modelo_2,  
               "(3) + Sexo" = modelo_3, "(4) + EF depto" = modelo_fe)  
  
msummary(modelos,  
          coef_rename = c(anios_educacion = "Años de educacion"),  
          gof_map = c("nobs", "r.squared"), stars = TRUE,  
          output = "output/02_tables/tabla_regresiones.csv")
```

Una sola llamada produce la tabla de todos los modelos, con nombres legibles y lista para el informe. El argumento `output` la escribe en disco (`.csv`, `.docx`, `.tex`...).

Lo que la regresión NO dice

- 1 **“Asociado a”, nunca “causa”.** Las personas con más educación difieren en cosas que no observamos: habilidad, contexto familiar, redes.
- 2 **Significancia no es magnitud.** Con $N = 21,821$ casi todo sale con estrellas; la pregunta relevante es *¿de cuántos pesos estamos hablando?*
- 3 **El R^2 mide ajuste**, no que el modelo esté bien especificado.

Ejercicio de cierre

Escriba en tres líneas la lectura **completa** del coeficiente de educación del modelo con efectos fijos: magnitud, unidades, la condición “controlando por...” y la advertencia causal.

Trabajo asincrónico (27 de julio – 2 de agosto)

- **Unidad 3:** teoría + video → practica_unidad-3.R → **Tarea 3** + cuestionario.
- **Unidad 4:** teoría + video → practica_unidad-4.R → **Tarea 4** + cuestionario.
- **Actividades con IA:** *Mejora esta figura (U3) y Crítica al econometrista artificial (U4).*
- **Proyecto final en parejas:** pipeline + informe + README. Se lanza hoy.

Próxima sesión — Lunes 3 de agosto

Sustentación de los proyectos finales. Se defiende el código **sin IA**: aprobar la sustentación es condición necesaria para aprobar el proyecto.

¿Preguntas?

Material de la sesión: `sesiones/sesion_3/example_vis_reg`
<https://eduard-martinez.github.io/teaching/intro-r/>

`efmartinez@icesi.edu.co`