Clase 6
Regresión Lineal Múltiple II

Análisis Avanzado de Datos

Gabriel Sotomayor

Evaluaciones

Tarea 2: 9 de octubre

- Regresión lineal múltiple

Informe 1: 30 de Octubre

- Regresión lineal múltiple o regresión logística

Recordatorio clase anterior

El problema del control estadístico.

El control estadístico consiste en ajustar los análisis para “controlar” el efecto de otras variables (covariadas) que podrían estar influyendo en la relación entre las variables de interés.

Ejemplo: En un estudio sobre diferencias salariales entre hombres y mujeres, las covariadas pueden incluir años de empleo o nivel educativo.

Por otro lado podemos controlar estadísticamente: Control Ajuste matemático que no requiere manipulación directa de datos o exclusión de casos. Es lo que comunmente tendremos que hacer en el contexto de estudios observacionales.

Introducción al Modelo de Regresión Múltiple

Un modelo de regresión múltiple examina la relación entre una **variable dependiente** y **varias variables independientes o predictores**.
La regresión múltiple permite **controlar otras variables** mientras se evalúa el efecto de una variable predictora específica.
Ejemplo: Si estudiamos la relación entre el ejercicio y la pérdida de peso, también podemos controlar la cantidad de alimentos consumidos para aislar su efecto.

Ecuación básica del modelo:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + \epsilon$$

Donde Y es la variable dependiente, b_0 es la constante o intercepto, $X_1, X_2, \ldots X_k$ son las variables independientes, $b_1, b_2, \ldots b_k$ son los coeficientes de regresión, y ϵ es el error.

Asociación Parcial

La asociación parcial mide la relación entre dos variables manteniendo constantes otras variables.

Ejemplo: En un estudio sobre pérdida de peso, podemos medir la relación entre la ingesta de alimentos y la pérdida de peso, controlando la cantidad de ejercicio realizado.

Objetivo de la sesión

Profundizar en el concepto de control estadístico y el uso de regresión lineal múltiple con distintos tipos de variables independientes.

Ejemplo con CASEN

Trabajaremos con CASEN utilizando como variable dependiente los ingresos del trabajo, y como variable predictora la edad y los años de escolaridad. Más adelante se utilizará el sexo y el nivel educacional por tramos para ejemplificar la introudcción de variables categóricas en modelos de regresión lineal múltiple.

```
# A tibble: 6 × 5
  ytrabajocor   esc  edad sexo           educ_simple_factor
        <dbl> <dbl> <dbl> <dbl+lbl>     <dbl+lbl>
1      411242    15    40 2 [2. Mujer]  3 [Superior Técnica]
2      590000     5    64 1 [1. Hombre] 1 [Menos que Media]
3      520000    12    34 1 [1. Hombre] 2 [Media Completa]
4      450000    12    30 2 [2. Mujer]  2 [Media Completa]
5      160000    10    68 2 [2. Mujer]  1 [Menos que Media]
6      580000     8    56 2 [2. Mujer]  1 [Menos que Media]
```

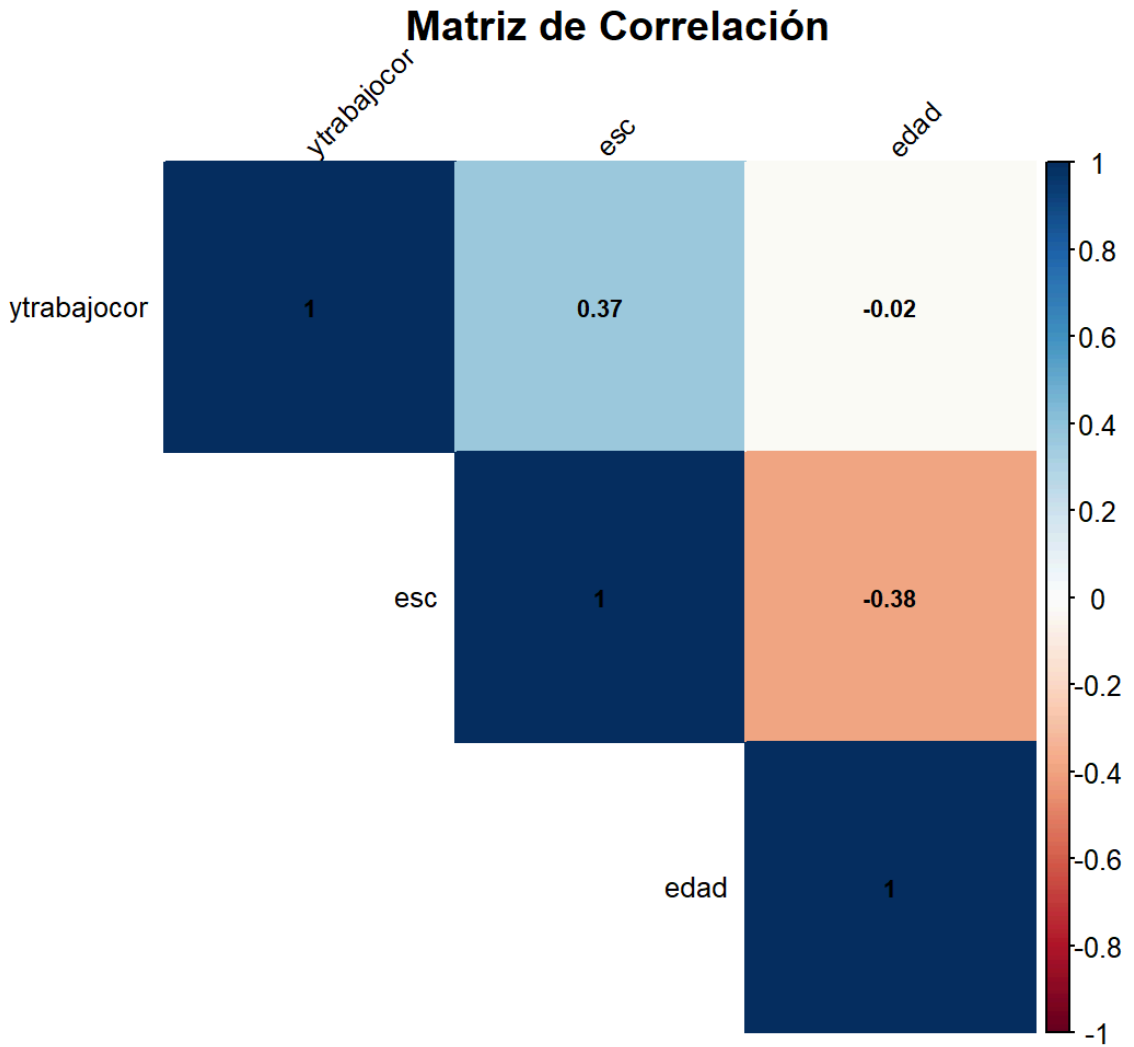
Ejemplo con CASEN

	Model 1	Model 2	Model 3
(Intercept)	-272321.52***	729366.06***	-742740.49***
	(8405.83)	(8757.98)	(14303.97)
esc	77114.51***		88022.20***
	(656.62)		(704.29)
edad		-1403.43***	7681.28***
		(189.32)	(189.88)
R^2	0.13	0.00	0.15
Adj. R^2	0.13	0.00	0.15
Num. obs.	88391	88976	88391

***p < 0.001; **p < 0.01; *p < 0.05

Statistical models

Ejemplo con CASEN



Comparación entre regresiones simples y múltiples

- **Modelo 1 (Años de Escolaridad)**:
  - El coeficiente beta para **Años de Escolaridad** en el modelo simple es **77,114**. Indica que por cada año adicional de escolaridad, los **Ingresos del Trabajo** esperados aumentan en 77,114 unidades.
- **Modelo 2 (Edad)**:
  - El coeficiente beta para **Edad** en el modelo simple es **-1,403**. Indica que por cada año adicional de edad, los **Ingresos del Trabajo** esperados disminuyen en 1,403 unidades.
- **Modelo 3 (Años de Escolaridad + Edad)**:
  - Al incluir ambas variables, el coeficiente para **Años de Escolaridad** aumenta a **88,022**, mientras que el coeficiente para **Edad** cambia drásticamente a **7,681**.

Efecto de la correlación entre Años de Escolaridad y Edad

- La correlación entre **Años de Escolaridad** y **Edad** es **-0.38**, lo que indica que a mayor edad, en promedio, los años de escolaridad son menores.
- En el **Modelo 1**, el coeficiente de **Años de Escolaridad** es **77,114**, pero en el **Modelo 3** aumenta a **88,022** al incluir **Edad**, debido al ajuste por la correlación entre ambas variables.
- En el **Modelo 2**, **Edad** tiene un coeficiente negativo (**-1,403**), pero en el **Modelo 3**, tras ajustar por los **Años de Escolaridad**, el coeficiente de **Edad** se vuelve positivo (**7,681**), lo que muestra que parte del efecto negativo inicial de la edad era por su relación con la escolaridad.

Control estadístico y parcialización

- **Control estadístico**:
  - El proceso de **parcialización** consiste en aislar el efecto de una variable independiente sobre los **Ingresos del Trabajo**, controlando por las otras variables incluidas en el modelo.
  - En el modelo múltiple, el coeficiente de cada variable refleja su **efecto neto** después de ajustar por las demás variables en el modelo.
- **Comparación de los modelos**:
  - Los modelos simples no muestran apropiadamente el efecto de las variables al no tener en cuenta las relaciones entre ellas.
  - El modelo múltiple, al incluir ambas variables, da una imagen más completa y precisa del impacto de cada predictor, lo que se refleja en los coeficientes beta y el incremento en el valor de R^2 ajustado.

Predictores categóricos en RLM

- **Predictores dicotómicos**:
  - Las variables **dicotómicas** se integran directamente en el modelo. Deben ser recodificadas a valores 0 y 1.
  - Cada valor representa una categoría, por ejemplo, 0 para “No” y 1 para “Sí”.
  - **Interpretación**: El coeficiente beta asociado a esta variable representa el **cambio promedio** en la variable dependiente al cambiar de la categoría 0 a la categoría 1, **controlando por las demás variables** del modelo.
- **Transformación de predictores politómicos**:
  - Las variables **politómicas** (más de dos categorías) deben transformarse en un conjunto de **variables dicotómicas** o **dummies**.
  - Se crea una variable dummy para cada categoría, excepto una que se toma como **categoría de referencia**.

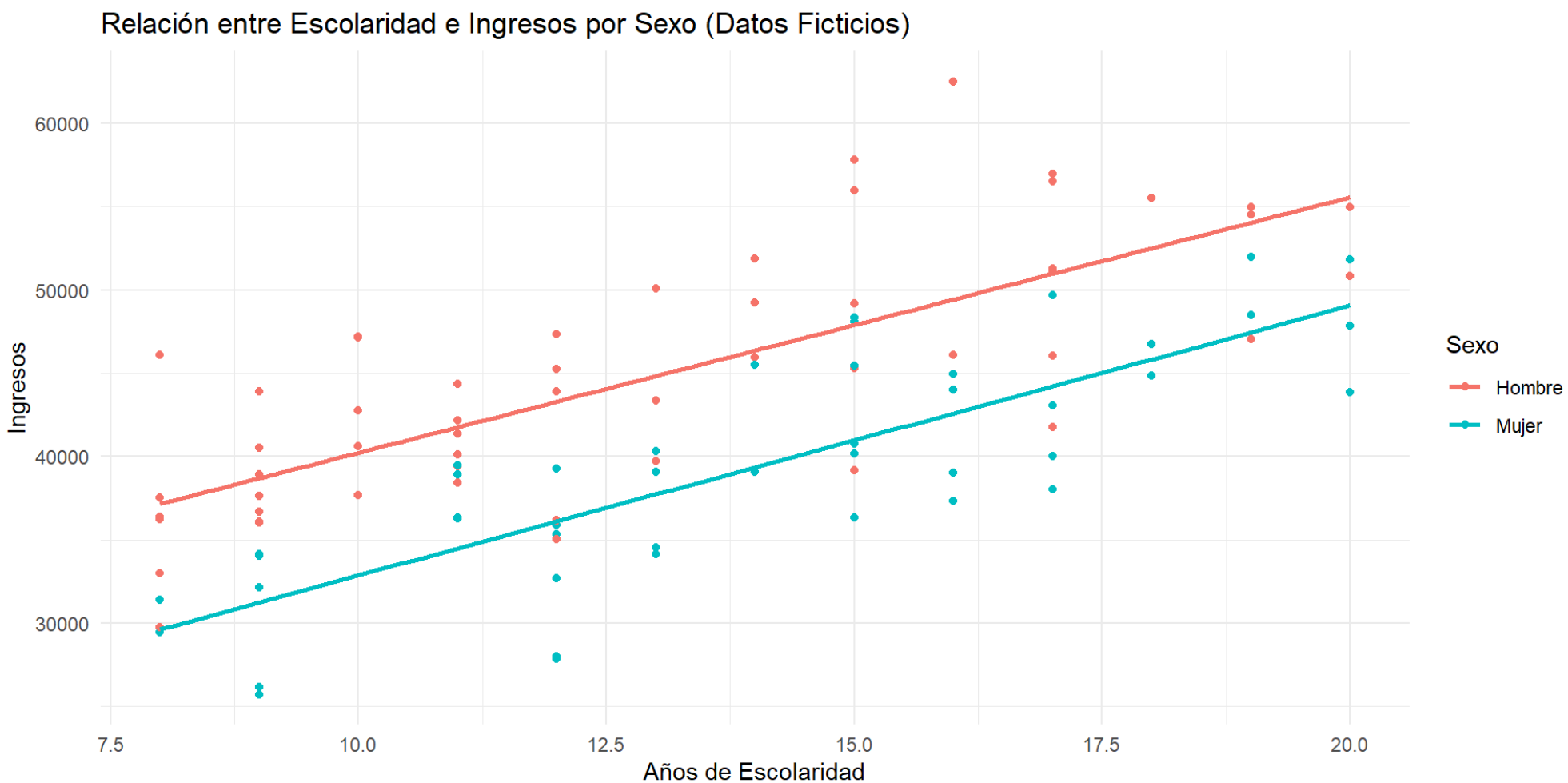
Ejemplo de sexo e ingresos del trabajo

	Model 1	Model 2	Model 3
(Intercept)	-272321.52***	741995.12***	-207194.28***
	(8405.83)	(3648.82)	(8426.08)
esc	77114.51***		80356.85***
	(656.62)		(652.68)
factor(sexo)2		-167442.22***	-235612.34***
		(5477.45)	(5113.14)
R^2	0.13	0.01	0.16
Adj. R^2	0.13	0.01	0.16
Num. obs.	88391	88976	88391

*** p < 0.001; ** p < 0.01; * p < 0.05

Statistical models

Ejemplo de sexo e ingresos del trabajo (datos simulados)



Lógica y interpretación de predictores categóricos

Ejemplo de niveles de escolaridad

educ_simple_factor	dummy_media_completa	dummy_superior_tecnica	dummy_sup
Menos que Media	0	0	0
Media Completa	1	0	0
Superior Técnica	0	1	0
Superior Profesional	0	0	1

Ejemplo de niveles de escolaridad

	Model 1	Model 2	Model 3
Intercepto	729366.06 (8757.98)***	396916.32 (4767.94)***	123364.01 (11171.52)***
Edad	-1403.43 (189.32)***		5178.45 (191.43)***
Media completa (ref. menos que media)		128948.37 (6428.87)***	185782.88 (6738.36)***
Superior Técnica		271608.14 (8629.17)***	344560.44 (9006.92)***
Superior Profesional		797327.92 (7042.71)***	868457.15 (7490.44)***
R^2	0.00	0.14	0.15
Adj. R^2	0.00	0.14	0.15

***p < 0.001; **p < 0.01; *p < 0.05

	Model 1	Model 2	Model 3
Num. obs.	88976	88406	88406

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Statistical models

Interpretación de los Coeficientes Beta en Regresión Lineal Múltiple

- **Tamaño y Dirección del Efecto**:
  - Cada coeficiente beta (β) indica el **cambio en la variable dependiente** por **cada unidad de cambio en la variable independiente** correspondiente.
  - **Dirección**:
    - $\beta > 0$: Indica un **efecto positivo** (la variable dependiente aumenta).
    - $\beta < 0$: Indica un **efecto negativo** (la variable dependiente disminuye).
- **Controlando por las demás variables del modelo**:
  - El valor de β refleja el **efecto neto** de la variable independiente, es decir, **ajustado por todas las otras variables** incluidas en el modelo.
  - Permite evaluar el efecto **aislado** de cada variable independiente mientras se **controlan** los posibles efectos de las demás.

Interpretación de Coeficientes Beta para Variables Categóricas

- **Diferencia con la Categoría de Referencia**:
  - El coeficiente beta para una variable categórica representa la **diferencia promedio** en la variable dependiente entre el grupo correspondiente y la **categoría de referencia**.
  - Si β es positivo, indica que el grupo en cuestión tiene una **mayor** media en comparación con la categoría de referencia. Si β es negativo, la media es **menor**.
- **Controlando por las demás variables del modelo**:
  - Al igual que en las variables continuas, los efectos están **ajustados** por las demás variables independientes, lo que permite interpretar el efecto de la categoría como si las otras variables permanecieran **constantes**.

R^2 en Regresión Lineal Múltiple

- **Definición de R^2**:
  - El R^2 mide la proporción de la **variabilidad explicada** por el modelo en relación a la variabilidad total.
  - Se interpreta como el porcentaje de la variación en la variable dependiente que es explicado por las variables independientes.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

- Donde: - $\hat{y}_i$ son los valores predichos. - y_i son los valores observados. - $\bar{y}$ es el promedio de la variable dependiente.

- **Limitaciones**:
  - El R^2 puede aumentar al agregar más predictores, incluso si no aportan significativamente al modelo.

R^2 Ajustado y su Utilidad

- El R^2 ajustado corrige la sobreestimación del R^2 al penalizar por el número de predictores en el modelo.
- Tiene en cuenta tanto el **número de predictores** como el **tamaño de la muestra**.
- **Cálculo**:
  - $$R^2_{ajustado} = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1} \right)$$
  - Donde:
    - n es el número de observaciones.
    - k es el número de predictores en el modelo.
  - A diferencia del R^2, el R^2 ajustado **disminuye** si se agregan predictores que no mejoran el modelo, ayudando a evitar el sobreajuste.
  - Es más útil cuando se compara la calidad de diferentes modelos con un número distinto de predictores.

Evaluación docente intermedia

Se está realizando el proceso de evaluación docente intermedia, por lo que les solicitamos que puedan contestar el siguiente formulario:

https://forms.gle/xDGewTRR5yHZkZZt9

Profundización sobre el proceso de parcialización {.smaller background-color="white"}

En un **modelo de regresión múltiple**, queremos entender cómo varios predictores afectan una variable dependiente, **manteniendo constante** el efecto de los demás.

¿Qué es la Parcialización?

Cuando dos variables predictoras están relacionadas entre sí (por ejemplo, **escolaridad** y **edad**), sus efectos pueden estar mezclados.

- **Parcialización** es el proceso de “separar” el efecto de una variable del efecto de las otras.
- El resultado es un **residuo parcial**, que refleja solo la parte de la variable que no está explicada por los otros predictores.

Ejemplo con CASEN

```r
casen2p<-casen2 %>%
  select(ytrabajocor, edad, esc) %>%
  filter(!is.na(ytrabajocor) & !is.na(esc) & !is.na(edad))

mod_esc <- lm(esc ~ edad, data = casen2p)
mod_edad <- lm(edad ~ esc, data = casen2p)

casen2pp<-data.frame(casen2p,  mod_esc$residuals, mod_edad$residuals)
print(casen2pp)
```

Ejemplo con CASEN

	ytrabajocor	edad	esc	mod_esc.residuals	mod_edad.residuals
1	411242	40	15	2.39949762	0.058270024
2	590000	64	5	-5.12323797	9.857931758
3	520000	34	12	-1.21981848	-10.201831457
4	450000	30	12	-1.63269588	-14.201831457
5	160000	68	10	0.28963943	20.958100890
6	580000	56	8	-2.94899277	6.118033237
7	19167	68	3	-6.71036057	11.017864105
8	300000	73	3	-6.19426382	16.017864105
9	400000	54	8	-3.15543147	4.118033237
10	400000	35	8	-5.11659913	-14.881966763
11	500000	62	7	-3.32967667	10.697999411
12	300000	63	10	-0.22645732	15.958100890
13	300000	29	12	-1.73591523	-15.201831457
14	300000	66	8	-1.91679927	16.118033237
15	41667	47	12	0.12203307	2.798168543
16	300000	53	8	-3.25865082	3.118033237
17	405000	47	3	-8.87796693	-9.982135895
18	400000	49	8	-3.67152822	-0.881966763
19	500000	28	12	-1.83913458	-16.201831457
20	600000	27	17	3.05764607	-10.101662325
21	1500000	56	12	1.05100723	11.798168543

Ejemplo con CASEN

```
htmlreg(list(lm(ytrabajocor ~ mod_esc.residuals  , data = casen2pp),
             lm(ytrabajocor ~ mod_edad.residuals  , data = casen2pp),
             lm(ytrabajocor ~  esc + edad , data = casen2p)))
```

Ejemplo con CASEN

	Model 1	Model 2	Model 3
(Intercept)	668123.79***	668123.79***	-742740.49***
	(2533.54)	(2726.46)	(14303.97)
mod_esc.residuals	88022.20***		
	(704.54)		
mod_edad.residuals		7681.28***	
		(204.41)	
esc			88022.20***
			(704.29)
edad			7681.28***
			(189.88)
R^2	0.15	0.02	0.15

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

	Model 1	Model 2	Model 3
Adj. R^2	0.15	0.02	0.15
Num. obs.	88391	88391	88391

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Statistical models

Ejemplo con CASEN

Con el fin de comprender la lógica del proceso de parcialziación, hacemos una regresión simple de cada predictor (e.g., escolaridad) sobre los otros predictores (e.g., edad), obteniendo los residuos de estas regresiones.

Los **residuos** obtenidos son la parte de la variable que no está correlacionada con los otros predictores. Estos se utilizan en lugar de las variables originales en la regresión múltiple.

Al realizar una regresión de la variable dependiente (e.g., ingresos) sobre los residuos de los predictores ajustados, obtenemos los mismos coeficientes y resultados que en una regresión múltiple tradicional.