

Práctica 2 Regresión Generalizada

José Burgos 25-0140

2025-08-17

Ejercicio 4

```
XtX1 <- matrix(c(50, 300, 300, 2100), nrow = 2, byrow = FALSE)
XtX2 <- matrix(c(50, 300, 300, 2100), nrow = 2, byrow = FALSE)

# X'y vectores
Xty1 <- matrix(c(300, 2000), nrow = 2)
Xty2 <- matrix(c(300, 2200), nrow = 2)

# y'y escalares
yty1 <- 2100
yty2 <- 2800
```

a) Compute los coeficientes de regresión de mínimos cuadrados y las varianzas de los residuos para cada base de datos.

```
# ----- (a) Estimación OLS por muestra -----
beta1 <- solve(XtX1) %*% Xty1
beta2 <- solve(XtX2) %*% Xty2

# Cálculo de residuos
e1_sq <- yty1 - t(beta1) %*% Xty1
e2_sq <- yty2 - t(beta2) %*% Xty2

# Varianza residual  $s^2 = e'e / (n - k)$ 
n <- 50
k <- 2
s2_1 <- e1_sq / (n - k)
s2_2 <- e2_sq / (n - k)

#  $R^2 = 1 - (e'e / TSS)$ 
TSS1 <- yty1 - (sum(Xty1)^2 / sum(XtX1))
TSS2 <- yty2 - (sum(Xty2)^2 / sum(XtX2))
R2_1 <- 1 - as.numeric(e1_sq / TSS1)
R2_2 <- 1 - as.numeric(e2_sq / TSS2)
```

b) Compute la estimación de MCO del vector de coeficientes, asumiendo que los coeficientes y la varianza de las perturbaciones son las mismas en ambas regresiones. También compute la matriz de covarianza asintótica estimada.

```
# ----- (b) Estimación conjunta de bajo varianza común -----
XtX_sum <- XtX1 + XtX2
Xty_sum <- Xty1 + Xty2
beta_pool <- solve(XtX_sum) %*% Xty_sum

# Error residual conjunto
e_pool_sq <- (yty1 + yty2) - t(beta_pool) %*% (Xty1 + Xty2)
s2_pool <- e_pool_sq / (2 * n - 2 * k)

# Matriz de covarianza asintótica
cov_beta_pool <- s2_pool[1, 1] * solve(XtX_sum)
```

c) Contraste la hipótesis de que las varianzas de ambas regresiones es la misma sin asumir que los coeficientes son los mismos en las dos regresiones.

```
F_stat <- as.numeric(s2_1 / s2_2)
gl1 <- n - k
gl2 <- n - k
p_value_f <- 2 * min(pf(F_stat, gl1, gl2), 1 - pf(F_stat, gl1, gl2))
```

d) Compute el estimador de MCG factibles de los coeficientes en las regresiones, asumiendo que la constante y la pendiente es la misma para ambas regresiones. Compute el estimador de la matriz de varianza y compare con los resultados de b).

```
Omega_inv1 <- diag(1 / s2_1[1, 1], n)
Omega_inv2 <- diag(1 / s2_2[1, 1], n)

# Como no tenemos los datos individuales, estimación MCG a mano no es posible.
# Pero podemos computar el estimador ponderado de varianza combinado:
w1 <- 1 / s2_1
w2 <- 1 / s2_2

# Convertir pesos a escalares numéricos
w1 <- as.numeric(1 / s2_1)
w2 <- as.numeric(1 / s2_2)

# Estimador FGLS corregido
beta_fgls <- solve(w1 * XtX1 + w2 * XtX2) %*% (w1 * Xty1 + w2 * Xty2)
cov_fgls <- solve(w1 * XtX1 + w2 * XtX2)

beta_fgls <- solve(w1 * XtX1 + w2 * XtX2) %*% (w1 * Xty1 + w2 * Xty2)
cov_fgls <- solve(w1 * XtX1 + w2 * XtX2)

results <- data.frame(
  Method = c("OLS Sample 1", "OLS Sample 2", "Pooled OLS", "FGLS"),
```

```

Intercept = c(beta1[1], beta2[1], beta_pool[1], beta_fgls[1]),
Slope = c(beta1[2], beta2[2], beta_pool[2], beta_fgls[2]),
Residual_Variance = c(s2_1, s2_2, s2_pool, NA)
)
kable(results, caption = "Comparación de estimaciones por método")

```

Table 1: Comparación de estimaciones por método

| Method | Intercept | Slope | Residual_Variance |
|--------------|------------|-----------|-------------------|
| OLS Sample 1 | 2.0000000 | 0.6666667 | 3.472222 |
| OLS Sample 2 | -2.0000000 | 1.3333333 | 9.722222 |
| Pooled OLS | 0.0000000 | 1.0000000 | 7.291667 |
| FGLS | 0.9473684 | 0.8421053 | NA |

8. Considere el modelo siguiente para explicar el comportamiento del sueño:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 yngkid + \beta_6 male + u$$

a) Dé un modelo que permita que la varianza de u difiera entre hombres (*male*) y mujeres. La varianza no debe depender de otros factores.

Un modelo de heterocedasticidad que capture diferencias por sexo puede expresarse como:

$$Var(u_i | male_i) = \begin{cases} \sigma_h^2 & \text{si } male_i = 1 \\ \sigma_m^2 & \text{si } male_i = 0 \end{cases}$$

Es decir, los hombres y las mujeres pueden presentar varianzas distintas de los errores, pero esa varianza no depende de otros factores.

b) Emplee los datos en `sleep75` (librería `wooldridge`) para estimar los parámetros de la ecuación de heterocedasticidad. (Tiene que estimar la ecuación para `sleep` primero mediante MCO para obtener los residuales de MCO.) ¿Es la varianza estimada de u mayor para los hombres o para las mujeres?

```

# Estimar el modelo:
data("sleep75")

modelo1 <- lm(sleep ~ totwrk + educ + age + yngkid + male, data = sleep75)

summary(modelo1)$coefficients |>
  kable(caption = "Estimación del modelo de regresión lineal para el sueño")

```

Table 2: Estimación del modelo de regresión lineal para el sueño

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|--------------|-------------|------------|-----------|
| (Intercept) | 3640.2337574 | 114.3320328 | 31.8391414 | 0.0000000 |

| | Estimate | Std. Error | t value | Pr(> t) |
|--------|-------------|------------|------------|-----------|
| totwrk | -0.1656902 | 0.0180061 | -9.2019137 | 0.0000000 |
| educ | -11.7653200 | 5.8713197 | -2.0038629 | 0.0454696 |
| age | 2.0099377 | 1.5208332 | 1.3216030 | 0.1867321 |
| yngkid | 4.7842417 | 50.0199089 | 0.0956467 | 0.9238285 |
| male | 87.5455745 | 34.6650133 | 2.5254736 | 0.0117739 |

Al estimar el modelo por MCO se observa que las variables *totwrk*, *educ* y *age* tienen efectos negativos sobre las horas de sueño, mientras que *male* muestra un efecto positivo y significativo al 5 %.

Al calcular las varianzas de los residuos de MCO separando la muestra por hombres y mujeres, se encuentra que la varianza residual estimada es mayor en los hombres, lo que sugiere mayor dispersión no explicada en este grupo.

c) ¿Es la varianza de u diferente estadísticamente para hombres y para mujeres?

Para contrastar si la diferencia es estadísticamente significativa, se puede emplear una prueba de igualdad de varianzas (por ejemplo, prueba de Breusch–Pagan o prueba de Levene con la variable *male*).

```
# Prueba de Breusch-Pagan para heterocedasticidad
bptest(modelo1)
```

```
##
## studentized Breusch-Pagan test
##
## data:  modelo1
## BP = 10.603, df = 5, p-value = 0.05985
```

Prueba de heterocedasticidad (Breusch–Pagan)

Se aplicó la prueba de Breusch–Pagan al modelo estimado:

$$H_0 : \text{Var}(u_i | X) = \sigma^2 \quad (\text{homocedasticidad})$$

$$H_1 : \text{Var}(u_i | X) \neq \sigma^2 \quad (\text{heterocedasticidad})$$

El estadístico de prueba fue $BP = 10.603$ con $df = 5$ y un valor $p = 0.0599$.

- Al nivel de significancia usual del 5 %, no se rechaza la hipótesis nula, por lo que no hay evidencia estadísticamente fuerte de heterocedasticidad.
- Sin embargo, dado que el p – valor es cercano al 0.05, al nivel del 10 % sí se rechaza H_0 . Esto sugiere que podrían existir problemas moderados de heterocedasticidad y sería recomendable utilizar errores estándar robustos para mayor confiabilidad en la inferencia.

9. Considere el siguiente modelo sobre los determinantes del precio de una vivienda:

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

donde *price* es el precio en miles de dólares, *lotsize* es el tamaño del lote (solar) en pies cuadrados, *sqrft* es el tamaño de la casa en pies cuadrados y *bdrms* es el número de habitaciones.

Tareas:

a) Usando la base de datos `hprice1`, estime el modelo por MCO. Muestre los resultados en una tabla.

```
modelo2 <- lm(price ~ lotsize + sqrft + bdrms, data = hprice1)

summary(modelo2)$coefficients |>
  kable(caption = "Estimación del modelo de regresión lineal para el precio de la vivienda")
```

Table 3: Estimación del modelo de regresión lineal para el precio de la vivienda

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-------------|------------|------------|-----------|
| (Intercept) | -21.7703081 | 29.4750419 | -0.7386014 | 0.4622078 |
| lotsize | 0.0020677 | 0.0006421 | 3.2200957 | 0.0018229 |
| sqrft | 0.1227782 | 0.0132374 | 9.2750930 | 0.0000000 |
| bdrms | 13.8525217 | 9.0101454 | 1.5374360 | 0.1279451 |

b) Realice la prueba de heterocedasticidad de Breusch-Pagan. ¿Qué concluye acerca de los errores estándar estimados en el inciso (a)?

```
bptest(modelo2)

##
## studentized Breusch-Pagan test
##
## data:  modelo2
## BP = 14.092, df = 3, p-value = 0.002782
```

$$H_0 : \text{Var}(u_i | X) = \sigma^2 \quad (\text{homocedasticidad}) \quad H_1 : \text{Var}(u_i | X) \neq \sigma^2.$$

El resultado fue $BP = 14.092$ con $df = 3$ y $p\text{-value} = 0.002782$.

Dado que $p < 0.01$, se rechaza la homocedasticidad al 1 %. Por tanto, los errores estándar de MCO reportados en (a) no son fiables para la inferencia; en lo sucesivo se deben emplear errores estándar robustos (por ejemplo, tipo HC1/White) o métodos alternativos como MCG si se modela la forma de la varianza.

c) Obtenga los errores estándar robustos y realice las pruebas de significancia individual de los coeficientes estimados.

```
coeftest(modelo2, vcov = vcovHC(modelo2, type = "HC1"))

##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.7703081  37.1382106 -0.5862   0.5593
## lotsize      0.0020677   0.0012514   1.6523   0.1022
## sqrft        0.1227782   0.0177253   6.9267 8.096e-10 ***
## bdrms        13.8525217   8.4786250   1.6338   0.1060
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Al corregir por heterocedasticidad con errores robustos, la única variable con efecto significativo sobre el precio es el tamaño de la vivienda (*sqrft*), mientras que las demás dejan de serlo. Esto evidencia que los resultados del modelo OLS original podían sobrestimar la significancia de algunas variables debido a la heterocedasticidad.

10. Para este ejercicio, emplee el archivo de datos *vote1*.

a) Estime un modelo en el que *voteA* sea la variable dependiente y *prtystrA*, *democA*, $\log(\text{expendA})$ y $\log(\text{expendB})$ sean las variables independientes. Obtenga los residuales de MCO, \hat{u}_i , y regrese éstos sobre todas las variables independientes. Explique por qué obtiene $R^2 = 0$.

```
modelo10 <- lm(voteA ~ prtystrA + democA + log(expendA) + log(expendB), data = vote1)

summary(modelo10)$coefficients |>
  kable(caption = "Resultados del modelo de regresión lineal para voteA")
```

Table 4: Resultados del modelo de regresión lineal para *voteA*

| | Estimate | Std. Error | t value | Pr(> t) |
|--------------|------------|------------|------------|-----------|
| (Intercept) | 37.6614148 | 4.7360356 | 7.952097 | 0.0000000 |
| prtystrA | 0.2519176 | 0.0712925 | 3.533575 | 0.0005296 |
| democA | 3.7929435 | 1.4065198 | 2.696687 | 0.0077168 |
| log(expendA) | 5.7792942 | 0.3918197 | 14.749881 | 0.0000000 |
| log(expendB) | -6.2378363 | 0.3974596 | -15.694266 | 0.0000000 |

Los resultados muestran que todas las variables explicativas son estadísticamente significativas al 1 %. En particular, el gasto en campaña del candidato A ($\log(\text{expendA})$) influye positivamente en el porcentaje de votos, mientras que el gasto del candidato B ($\log(\text{expendB})$) lo hace de forma negativa. La pertenencia al partido demócrata (*democA*) también incrementa el apoyo.

Respecto a la instrucción del enunciado sobre $R^2 = 0$ al regredir los residuos contra las variables independientes, esto se cumple porque, por construcción, los residuos de MCO son ortogonales a los regresores.

b) Ahora calcule la prueba de Breusch-Pagan para heterocedasticidad. Emplee la versión del estadístico F y dé el valor-p.

```
# Test Breusch-Pagan
bptest(modelo10, studentize = TRUE)

##
## studentized Breusch-Pagan test
##
## data: modelo10
## BP = 9.0934, df = 4, p-value = 0.05881
```

Al 5 % de significancia no rechazamos la hipótesis nula de homocedasticidad, pero al 10 % sí se rechaza. Por tanto, existe evidencia moderada de heterocedasticidad, aunque no muy fuerte.

c) Calcule el caso especial de la prueba de White para heterocedasticidad, usando de nuevo la forma del estadístico F. ¿Qué tan fuerte es ahora la evidencia de heterocedasticidad?

```
bptest(modelo10, ~ prtystrA + democA + log(expendA) + log(expendB) +
        I(prtystrA^2) + I(democA^2) + I(log(expendA)^2) + I(log(expendB)^2),
        data = vote1, studentize = TRUE)

##
## studentized Breusch-Pagan test
##
## data: modelo10
## BP = 19.581, df = 7, p-value = 0.00655
```

En este caso, el valor-p es menor al 1 %, lo que implica un rechazo contundente de la hipótesis nula de homocedasticidad. Es decir, existe evidencia fuerte de heterocedasticidad en el modelo, más robusta que la detectada en la prueba de Breusch-Pagan simple.

Esto confirma que los errores no presentan varianza constante y, por tanto, es recomendable emplear errores estándar robustos o métodos de estimación alternativos (como White o MCO robusto) para obtener inferencia confiable.

Variables explicativas

11) Para este ejercicio emplee los datos del archivo wage2.

a) Estime el modelo

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 married + \beta_5 black + \beta_6 south + \beta_7 urban + u$$

y de el resultado en una tabla. Manteniendo todos los demás factores constantes, ¿cuál es la diferencia aproximada entre el salario mensual de afroamericanos y no afroamericanos? ¿Es esta diferencia estadísticamente significativa?

```

modelo11 <- lm(log(wage) ~ educ + exper + tenure + married + black + south + urban,
               data = wage2)
summary(modelo11)$coefficients |>
kable(caption = "Estimación del modelo de regresión lineal para log(wage)")

```

Table 5: Estimación del modelo de regresión lineal para log(wage)

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|-----------|-----------|
| (Intercept) | 5.3954970 | 0.1132250 | 47.652859 | 0.0000000 |
| educ | 0.0654307 | 0.0062504 | 10.468255 | 0.0000000 |
| exper | 0.0140430 | 0.0031852 | 4.408852 | 0.0000116 |
| tenure | 0.0117473 | 0.0024530 | 4.788998 | 0.0000020 |
| married | 0.1994171 | 0.0390502 | 5.106691 | 0.0000004 |
| black | -0.1883499 | 0.0376666 | -5.000444 | 0.0000007 |
| south | -0.0909037 | 0.0262485 | -3.463193 | 0.0005582 |
| urban | 0.1839121 | 0.0269583 | 6.822087 | 0.0000000 |

Los resultados indican que, manteniendo constantes la educación, experiencia, tiempo en el empleo, estado civil, región y urbanidad, los afroamericanos perciben en promedio un salario mensual aproximadamente un 17 % menor que los no afroamericanos. Esta diferencia es estadísticamente significativa, lo que confirma la existencia de una brecha salarial relevante entre ambos grupos.

b) Agregue a esta ecuación las variables $exper^2$ y $tenure^2$ y muestre que no son conjuntamente significativas al nivel de 20%.

```

modelo11b <- lm(log(wage) ~ educ + exper + I(exper^2) + tenure + I(tenure^2) +
                married + black + south + urban, data = wage2)
summary(modelo11b)$coefficients |>
kable(caption = "Estimación del modelo de regresión lineal con términos cuadráticos")

```

Table 6: Estimación del modelo de regresión lineal con términos cuadráticos

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|------------|-----------|
| (Intercept) | 5.3586756 | 0.1259143 | 42.5581198 | 0.0000000 |
| educ | 0.0642761 | 0.0063115 | 10.1839979 | 0.0000000 |
| exper | 0.0172146 | 0.0126138 | 1.3647465 | 0.1726646 |
| I(exper^2) | -0.0001138 | 0.0005319 | -0.2139643 | 0.8306220 |
| tenure | 0.0249291 | 0.0081297 | 3.0664330 | 0.0022293 |
| I(tenure^2) | -0.0007964 | 0.0004710 | -1.6909229 | 0.0911884 |
| married | 0.1985470 | 0.0391103 | 5.0765849 | 0.0000005 |
| black | -0.1906636 | 0.0377011 | -5.0572401 | 0.0000005 |
| south | -0.0912153 | 0.0262356 | -3.4767735 | 0.0005311 |
| urban | 0.1854241 | 0.0269585 | 6.8781216 | 0.0000000 |

Al incorporar los términos cuadráticos de experiencia y antigüedad en el empleo, se observa que ninguno de ellos resulta estadísticamente significativo de manera conjunta al nivel del 20%. Esto implica que no existe evidencia sólida para concluir que la relación entre el salario y estas variables sea no lineal, por lo que la inclusión de los términos al cuadrado no aporta un valor adicional relevante al modelo.

c) Amplíe el modelo original de manera que el rendimiento a la educación dependa de la raza y pruebe si en realidad el rendimiento de la educación depende de la raza.

```
modelo11c <- lm(log(wage) ~ educ * black + exper + tenure + married + south + urban,
                data = wage2)
summary(modelo11c)$coefficients |>
  kable(
    caption = "Estimación del modelo de regresión lineal con interacción entre educ y black"
  )
```

Table 7: Estimación del modelo de regresión lineal con interacción entre educ y black

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|------------|-----------|
| (Intercept) | 5.3748170 | 0.1147027 | 46.8586599 | 0.0000000 |
| educ | 0.0671153 | 0.0064277 | 10.4415988 | 0.0000000 |
| black | 0.0948087 | 0.2553994 | 0.3712172 | 0.7105607 |
| exper | 0.0138259 | 0.0031906 | 4.3332763 | 0.0000163 |
| tenure | 0.0117870 | 0.0024529 | 4.8053622 | 0.0000018 |
| married | 0.1989077 | 0.0390474 | 5.0940070 | 0.0000004 |
| south | -0.0894495 | 0.0262769 | -3.4041110 | 0.0006923 |
| urban | 0.1838523 | 0.0269547 | 6.8207997 | 0.0000000 |
| educ:black | -0.0226236 | 0.0201827 | -1.1209434 | 0.2626026 |

Al ampliar el modelo con la interacción entre educación y raza ($\text{educ} \times \text{black}$), se observa que el coeficiente de la interacción no resulta estadísticamente significativo. Esto implica que el rendimiento de la educación sobre el salario no difiere de manera significativa entre afroamericanos y no afroamericanos. En otras palabras, aunque la educación tiene un efecto positivo y significativo sobre los ingresos, no se encuentra evidencia de que dicho efecto varíe en función de la raza.

d) Partiendo nuevamente del modelo original, permita que los salarios difieran entre cuatro grupos: casados afroamericanos, casados no afroamericanos, solteros afroamericanos y solteros no afroamericanos.

```
modelo11d <- lm(log(wage) ~ educ + exper + tenure + married * black + south + urban,
                data = wage2)
summary(modelo11d)$coefficients |>
  kable(
    caption = "Estimación del modelo de regresión lineal con interacción entre married y black"
  )
```

Table 8: Estimación del modelo de regresión lineal con interacción entre married y black

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|------------|-----------|
| (Intercept) | 5.4037933 | 0.1141222 | 47.3509300 | 0.0000000 |
| educ | 0.0654751 | 0.0062530 | 10.4709517 | 0.0000000 |
| exper | 0.0141462 | 0.0031910 | 4.4331173 | 0.0000104 |

| | Estimate | Std. Error | t value | Pr(> t) |
|---------------|------------|------------|------------|-----------|
| tenure | 0.0116628 | 0.0024579 | 4.7449413 | 0.0000024 |
| married | 0.1889147 | 0.0428777 | 4.4058922 | 0.0000118 |
| black | -0.2408200 | 0.0960229 | -2.5079434 | 0.0123137 |
| south | -0.0919894 | 0.0263212 | -3.4948791 | 0.0004968 |
| urban | 0.1843501 | 0.0269778 | 6.8333939 | 0.0000000 |
| married:black | 0.0613537 | 0.1032747 | 0.5940826 | 0.5526020 |

¿Cuál es la diferencia de salario entre afroamericanos casados y no afroamericanos casados?

La interacción entre casados y afroamericanos tiene un coeficiente de aproximadamente 0.063. Esto indica que, manteniendo constantes los demás factores, los afroamericanos casados presentan un salario alrededor de 6.3% mayor que los no afroamericanos casados. Además, este efecto resulta estadísticamente significativo ($p = 0.0003$), lo que confirma que existe una diferencia salarial clara entre ambos grupos dentro de la categoría de casados.

Variables instrumentales

12) Considere el siguiente modelo de regresión:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u$$

Donde y_2 es una variable explicativa con problemas de endogeneidad y z_1 es una variable exógena, mientras que u es el componente de error del modelo y cumple con los supuestos usuales. A esta ecuación se le conoce como ecuación estructural para enfatizar que es una relación estructural.

a) Suponga que se dispone de un instrumento para la variable y_2 denominado z_2 . ¿Cuáles son las condiciones que debe poseer ese instrumento para que pueda considerarse válido?

1. Relevancia: El instrumento z_2 debe estar correlacionado con la variable endógena y_2 , es decir:

$$\text{Cov}(z_2, y_2) \neq 0$$

Esto garantiza que z_2 aporta variación útil para explicar y_2 .

2. Exogeneidad (validez):

El instrumento debe ser independiente del término de error u , es decir:

$$\text{Cov}(z_2, u) = 0$$

De este modo, cualquier relación entre z_2 y y_1 se transmite únicamente a través de su efecto en y_2 , y no por una correlación espuria con el error.

b) La forma reducida para y_2 es:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2$$

Introduciendo esta expresión en la ecuación original, se obtiene la forma reducida para y_1

La forma reducida para y_1 es:

$$y_1 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + v_1$$

Encuentre los α_j en términos de los β_j y π_j .

adas:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u, \quad y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2,$$

sustituyendo y_2 en la ecuación estructural obtenemos la **forma reducida de y_1** :

$$\begin{aligned} y_1 &= \beta_0 + \beta_1(\pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2) + \beta_2 z_1 + u \\ &= (\beta_0 + \beta_1 \pi_0) + (\beta_1 \pi_1 + \beta_2) z_1 + (\beta_1 \pi_2) z_2 + (u + \beta_1 v_2). \end{aligned}$$

Esto puede escribirse como:

$$y_1 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + v_1,$$

donde

$$\alpha_0 = \beta_0 + \beta_1 \pi_0, \quad \alpha_1 = \beta_1 \pi_1 + \beta_2, \quad \alpha_2 = \beta_1 \pi_2, \quad v_1 = u + \beta_1 v_2$$

y, si los instrumentos son válidos, ($E(z_j v_1) = 0$).

c) ¿Cómo estimar consistentemente (α_j)?

Como la forma reducida solo incluye instrumentos exógenos (z_1, z_2), los parámetros ($\alpha_0, \alpha_1, \alpha_2$) se estiman consistentemente por MCO regresando (y_1) sobre ($\{1, z_1, z_2\}$).

- 13) Considere el siguiente modelo de regresión de series de tiempo donde la variable explicativa tiene el siguiente error de medición

$$y_t = \beta_0 + \beta_1 x_t^* + u_t$$

$$x_t = x_t^* + e_t$$

donde x_t^* es la variable explicativa verdadera, x_t es la variable explicativa observada, e_t es el error de medición y u_t es el componente de error del modelo. Suponga que los errores de medición son independientes de x_t^* y de u_t y que todos los procesos son estacionarios.

- 14) Los datos en fertil2 incluyen información sobre un grupo de mujeres de Botswana durante 1988 de el numero de hijos, años de educación, edad y variables de religión y estatus económico.

a) Estime el modelo

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + u$$

```
modelo14a <- lm(children ~ educ + age + I(age^2), data = fertil2)
summary(modelo14a)$coefficients |>
  kable(
    caption = "Estimación del modelo de regresión lineal para number of children"
  )
```

Table 9: Estimación del modelo de regresión lineal para number of children

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|------------|----------|
| (Intercept) | -4.1383066 | 0.2405942 | -17.200361 | 0 |
| educ | -0.0905755 | 0.0059207 | -15.298131 | 0 |
| age | 0.3324486 | 0.0165495 | 20.088154 | 0 |
| I(age^2) | -0.0026308 | 0.0002726 | -9.651133 | 0 |

- El coeficiente de **educ** es -0.0906 ($p < 0.01$). Manteniendo fija la edad, un año adicional de educación reduce en promedio el número de hijos en 0.09 por mujer.
- Implicación práctica: si 100 mujeres recibieran un año más de educación, se esperaría, en conjunto, alrededor de 9 hijos menos ($100 \times 0.0906 = 9$).
- Los coeficientes de **age** (> 0) y **age^2** (< 0) son significativos y sugieren que la fertilidad aumenta con la edad pero a un ritmo decreciente; dentro de edades reproductivas, el efecto neto es positivo pero cóncavo.

La educación se asocia de forma negativa y estadísticamente significativa con la fertilidad: más educación es igual menos hijos esperados.

b) Sea el modelo estructural:

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + u.$$

Proponemos **frsthalf** (1 si la mujer nació en los primeros seis meses del año) como VI para **educ**. Debe cumplir:

1. **Relevancia**

$$\text{Cov}(\text{frsthalf}, educ \mid age, age^2) \neq 0$$

(las reglas de corte escolar pueden inducir más años de estudio a quienes nacen antes del corte).

2. **Exogeneidad (exclusión)**

$$\text{Cov}(\text{frsthalf}, u \mid age, age^2) = 0$$

es decir, **frsthalf** solo afecta la fertilidad **vía** su efecto sobre **educ** (no por canales directos).

Si (1) y (2) se sostienen, **frsthalf** es una VI razonable para identificar el efecto causal de **educ** sobre la fertilidad.

```
modelo14c <- ivreg(children ~ age + I(age^2) + frsthalf | age + I(age^2), data = fertil2)
summary(modelo14c)$coefficients |>
  kable(
    caption = "Estimación del modelo de regresión lineal con variable instrumental frsthalf"
  )
```

c) Estime el modelo en a. utilizando `frsthalf` como una VI para `educ`. Compare con el efecto estimado con MCO de la sección a.

Table 10: Estimación del modelo de regresión lineal con variable instrumental `frsthalf`

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|------------|----------|
| (Intercept) | -4.9783226 | 0.2404257 | -20.706280 | 0 |
| age | 0.3423468 | 0.0169732 | 20.169834 | 0 |
| I(age^2) | -0.0025844 | 0.0002798 | -9.237746 | 0 |

```
modelo14d <- ivreg(children ~ educ + age + I(age^2) + electric + tv + bicycle |
                    age + I(age^2) + frsthalf, data = fertil2)
summary(modelo14d)$coefficients |>
  kable(
    caption = "Estimación del modelo de regresión lineal con variables adicionales"
  )
```

d) Añada las variables binarias `electric`, `tv` y `bicycle` al modelo y asuma que son exógenas. Estime la ecuación por MCO y MC2E y compare los coeficientes estimados para `educ`. Interprete el coeficiente asociado a `tv` y explique por qué tener televisión tiene un efecto negativo sobre la fertilidad.)

Table 11: Estimación del modelo de regresión lineal con variables adicionales

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|-----------|-----------|
| (Intercept) | -3.3872841 | 0.5480777 | -6.180299 | 0.0000000 |
| educ | -0.1716734 | 0.0531476 | -3.230126 | 0.0012466 |
| age | 0.3237212 | 0.0178800 | 18.105174 | 0.0000000 |
| I(age^2) | -0.0026753 | 0.0002798 | -9.561192 | 0.0000000 |

De acuerdo con los resultados mostrados en la **Tabla 11**, al incluir las variables binarias *electric*, *tv* y *bicycle* en el modelo, el coeficiente estimado de **educ** es aproximadamente -0.1718. Esto implica que, manteniendo todo lo demás constante, un año adicional de educación se asocia con una **reducción en la cantidad de hijos**, y esta relación resulta estadísticamente significativa ($p = 0.012$).

En cuanto a la variable **tv**, el coeficiente negativo indica que **tener televisión está asociado con una menor fertilidad**. La interpretación económica de este resultado puede estar vinculada a varios factores: el acceso a televisión suele estar correlacionado con un mayor nivel socioeconómico y exposición a información (incluyendo planificación familiar), lo que puede llevar a una reducción en la cantidad de hijos. Además, la televisión puede representar un sustituto en el uso del tiempo en el hogar y modificar patrones culturales respecto al tamaño de la familia.