Regresión lineal aplicada

Unidad 2: Estadística Básica y Aplicada

Nicolás Sidicaro
Octubre 2025

Bloque 1

Fundamentos de Regresión Lineal

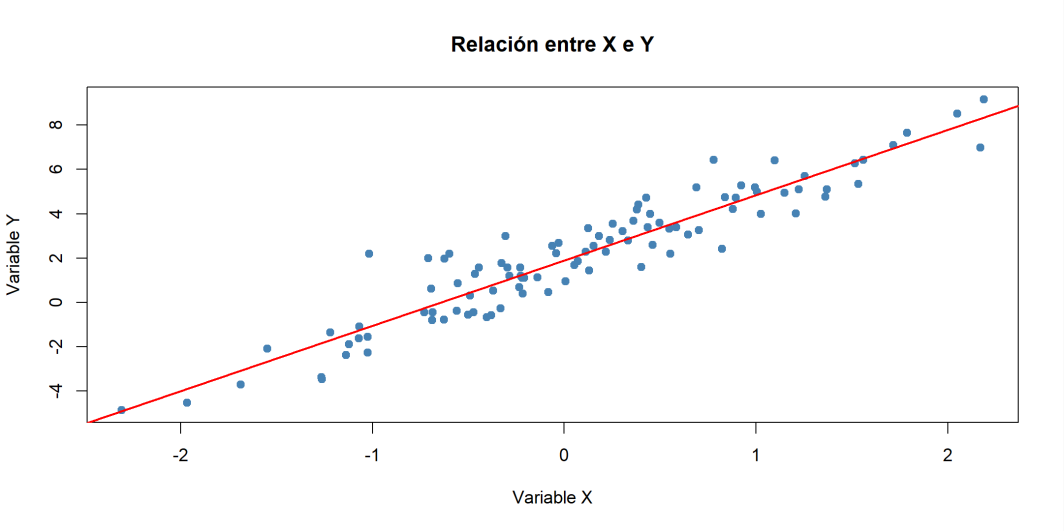
¿Qué estamos haciendo con regresión?

Objetivo principal:

- Modelar relaciones entre variables
- Describir asociaciones
- Hacer predicciones

No necesariamente:

- Establecer causalidad
- Demostrar que X "causa" Y



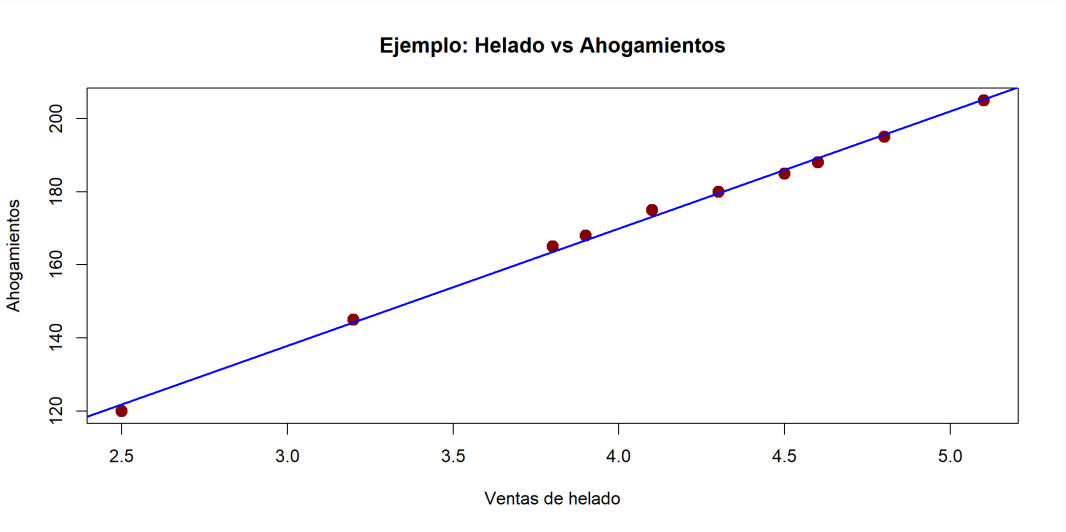
⚠️ Correlación ≠ Causalidad

Correlaciones espurias:

Ejemplos clásicos:

- Consumo de helado y ahogamientos
- Número de películas de Nicolas Cage y ahogamientos en piscinas
- Divorcio en Maine y consumo de margarina

Mensaje clave: Una relación estadística fuerte NO implica que una variable cause la otra



El Modelo de Regresión Lineal Simple

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Componentes:

- Y_i: Variable dependiente (lo que queremos explicar)
- X_i: Variable independiente (explicativa)
- β_0: Intercepto (valor de Y cuando X=0)
- β_1: Pendiente (efecto marginal de X sobre Y)
- u_i: Error o residuo (lo que no podemos explicar)

MCO (Mínimos Cuadrados Ordinarios):

- Minimiza la suma de errores al cuadrado: $\min \sum u_i^2$
- Encuentra la "mejor" línea que pasa por los datos

Primer Ejemplo en R: Salarios

```r
# Cargar datos de salarios
data(wage1)
head(wage1[, c("wage", "educ", "exper", "tenure")], 8)
```

```
##    wage educ exper tenure
## 1  3.10   11     2      0
## 2  3.24   12    22      2
## 3  3.00   11     2      0
## 4  6.00    8    44     28
## 5  5.30   12     7      2
## 6  8.75   16     9      8
## 7 11.25   18    15      7
## 8  5.00   12     5      3
```

Modelo Simple: Salario ~ Educación

```r
# Estimar modelo
modelo1 ← lm(wage ~ educ, data = wage1)
summary(modelo1)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = wage1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3396 -2.1501 -0.9674  1.1921 16.6085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.90485    0.68497  -1.321    0.187
## educ         0.54136    0.05325  10.167   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.378 on 524 degrees of freedom
## Multiple R-squared:  0.1648,    Adjusted R-squared:  0.1632
## F-statistic: 103.4 on 1 and 524 DF,  p-value: < 2.2e-16
```

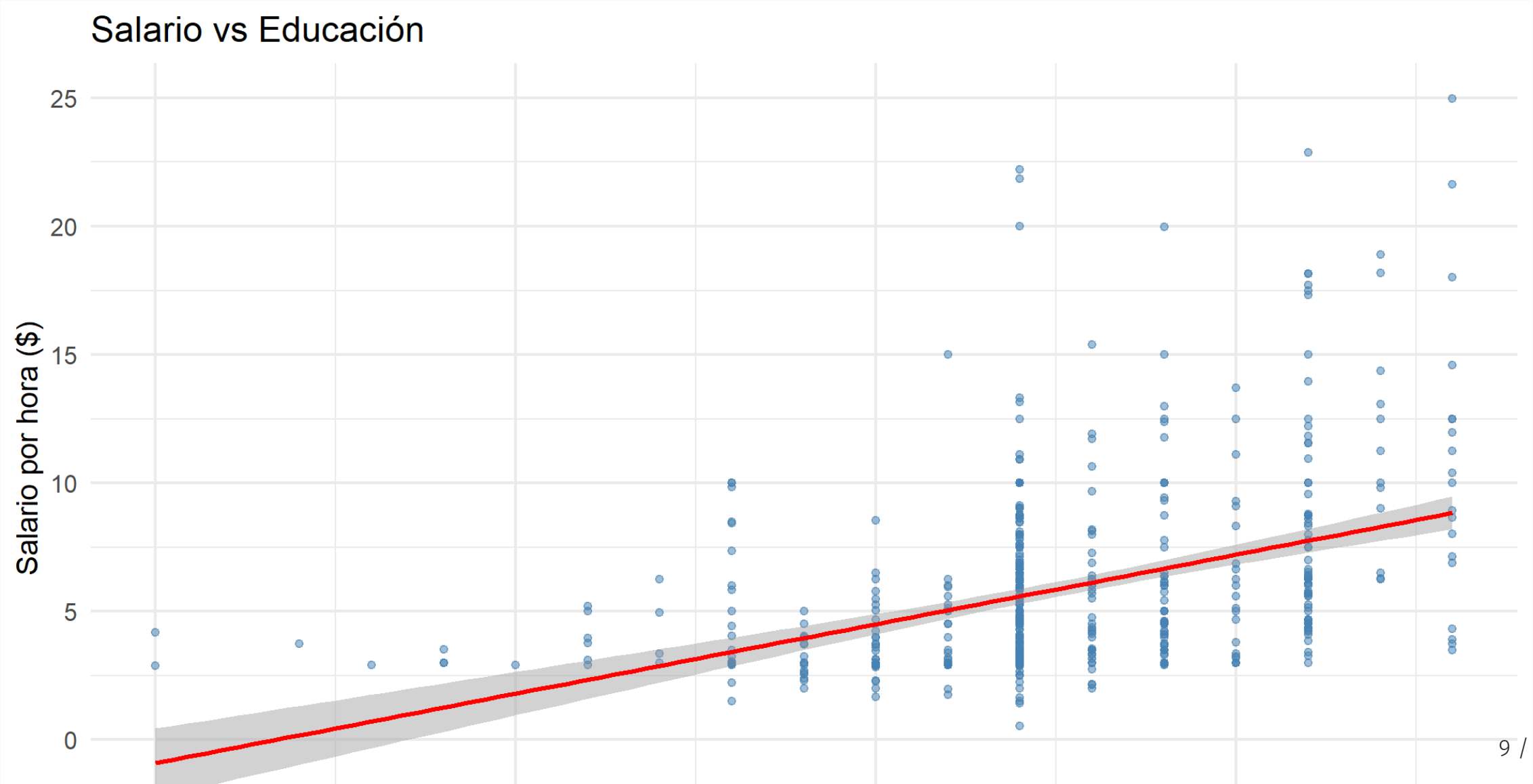
Interpretación del Output

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -0.9048516 0.68496782 -1.321013 1.870735e-01
## educ         0.5413593 0.05324804 10.166746 2.782599e-22
```

Interpretación:

- **Intercepto (β_0):** -0.90 → Una persona con 0 años de educación ganaría -$0.90/hora (no tiene sentido real, es extrapolación)
- **Educación (β_1):** 0.54 → Por cada año adicional de educación, el salario aumenta en $0.54/hora **en promedio**
- **P-valores < 0.001:** Ambos coeficientes son altamente significativos (*)
- **R^2 = 0.165:** La educación explica solo el 16.5% de la variación en salarios

Visualización del Modelo



Modelos Logarítmicos: ¿Por qué?

Problema: Las relaciones económicas rara vez son lineales

Soluciones: Transformaciones logarítmicas

Modelo	**Ecuación**	**Interpretación de β_1**
Lin-Lin	$Y = \beta_0 + \beta_1 X$	Cambio en unidades
Log-Lin	$\log(Y) = \beta_0 + \beta_1 X$	$\beta_1 \times 100\%$ cambio en Y por unidad de X
Lin-Log	$Y = \beta_0 + \beta_1 \log(X)$	$\beta_1/100$ unidades de Y por 1% cambio en X
Log-Log	$\log(Y) = \beta_0 + \beta_1 \log(X)$	Elasticidad: β_1% cambio en Y por 1% cambio en X

Ejemplo: Modelo Log-Lin

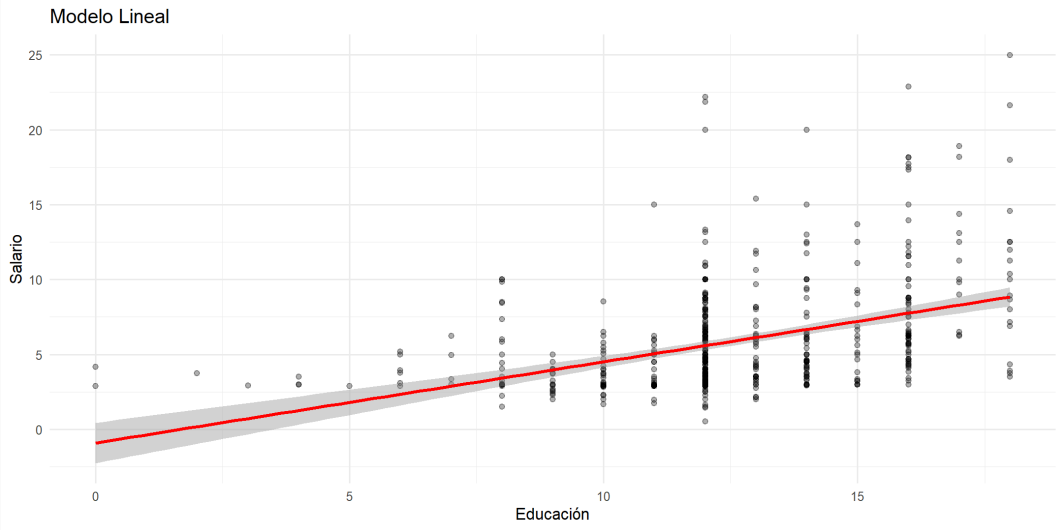
```
# Modelo con logaritmo del salario
modelo_log ← lm(log(wage) ~ educ, data = wage1)
summary(modelo_log)$coefficients
```

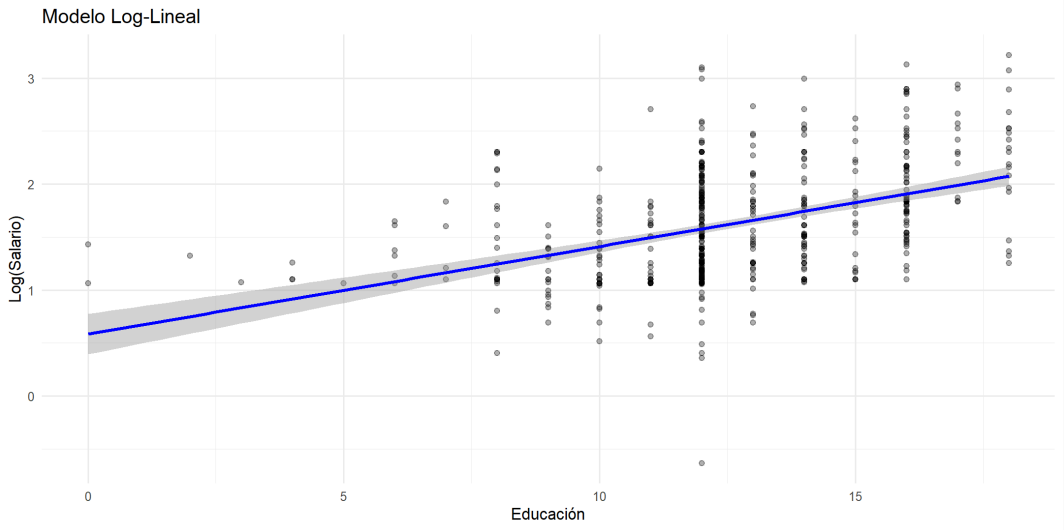
```
##                Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept) 0.58377267 0.097335834  5.99751 3.736702e-09
## educ        0.08274437 0.007566694 10.93534 3.270645e-25
```

Interpretación:

- $\beta_1 = 0.083 \rightarrow$ Por cada año adicional de educación, el salario aumenta aproximadamente **8.3%**
- Cálculo: $e^{0.083} - 1 = 0.0865 \approx 8.65\%$
- Este modelo suele ajustar mejor para salarios ($R^2 = 0.186$ vs 0.165)

Comparación Visual: Lineal vs Log





Modelo Log-Log: Elasticidades

```r
# Precios de casas y tamaño
data(hprice1)

# Modelo log-log
modelo_elasticidad ← lm(log(price) ~ log(sqrft), data = hprice1)
summary(modelo_elasticidad)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -0.9751302 0.64104852 -1.521149 1.318909e-01
## log(sqrft)   0.8726596 0.08460479 10.314541 1.051418e-16
```

Interpretación:

- $\beta_1 = 0.70 \rightarrow$ Si el tamaño de la casa aumenta 1%, el precio aumenta 0.70% (elasticidad)
- Útil para comparar variables en diferentes escalas

Bloque 2

Regresión Múltiple y Variables Categóricas

Regresión Múltiple

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$

Ventajas:

- Controlar por múltiples factores simultáneamente
- Reducir el sesgo por variables omitidas
- Interpretación **ceteris paribus** (manteniendo todo lo demás constante)

Interpretación de β_j:

- El cambio en Y cuando X_j aumenta en una unidad, **manteniendo las demás variables constantes**

Ejemplo: Salario con Múltiples Variables

```r
# Modelo múltiple
modelo_multiple ← lm(wage ~ educ + exper + tenure, data = wage1)
summary(modelo_multiple)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + tenure, data = wage1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6068 -1.7747 -0.6279  1.1969 14.6536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.87273    0.72896  -3.941 9.22e-05 ***
## educ         0.59897    0.05128  11.679  < 2e-16 ***
## exper        0.02234    0.01206   1.853   0.0645 .
## tenure       0.16927    0.02164   7.820 2.93e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.084 on 522 degrees of freedom
## Multiple R-squared:  0.3064,    Adjusted R-squared:  0.3024
## F-statistic: 76.87 on 3 and 522 DF,  p-value: < 2.2e-16
```

Interpretación del Modelo Múltiple

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -2.87273482 0.72896429 -3.940844 9.224742e-05
## educ         0.59896507 0.05128355 11.679478 3.681353e-28
## exper        0.02233952 0.01205685  1.852849 6.446818e-02
## tenure       0.16926865 0.02164461  7.820361 2.934527e-14
```

Interpretaciones (ceteris paribus):

- **Educación:** Por cada año adicional de educación, el salario aumenta $0.60/hora, manteniendo experiencia y antigüedad constantes
- **Experiencia:** Por cada año de experiencia, el salario aumenta $0.014/hora
- **Antigüedad (tenure):** Por cada año en el trabajo actual, el salario aumenta $0.17/hora

Nota: Los coeficientes cambiaron respecto al modelo simple (educación era 0.54, ahora es 0.60)

Bondad de Ajuste: R^2 y R^2 Ajustado

```
## R² = 0.3064
```

```
## R² ajustado = 0.3024
```

R^2 (coeficiente de determinación):

- Proporción de la varianza de Y explicada por el modelo
- Rango: 0 a 1
- **Problema:** Siempre aumenta al agregar variables (aunque no sean relevantes)

R^2 ajustado:

- Penaliza por agregar variables
- Puede disminuir si agregamos variables irrelevantes
- **Mejor para comparar modelos con distinto número de variables**

Tests de Significancia

Test t individual: ¿Es β_j significativamente distinto de cero?

- H_0: β_j = 0 (la variable no tiene efecto)
- H_i: β_j != 0 (la variable sí tiene efecto)
- Miramos el p-valor:
  - p < 0.01 $\rightarrow$ $***$ (muy significativo)
  - p < 0.05 $\rightarrow$ $**$ (significativo)
  - p < 0.10 $\rightarrow$ $*$ (marginalmente significativo)

Test F global: ¿El modelo completo es significativo?

- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ (ninguna variable importa)
- H_1 : Al menos un β_k != 0

```
## F-statistic: 76.87 con p-valor < 0.001
```

Variables Dicotómicas (Dummies)

¿Qué son?

- Variables que toman valores 0 o 1
- Representan categorías: género, región, tratamiento, etc.

Codificación:

```
wage1 ← wage1 %>%
  mutate(mujer = ifelse(female == 1, 1, 0))

# Ver distribución
table(wage1$mujer)
```

```
##
##   0   1
## 274 252
```

Ejemplo: Brecha de Género Salarial

```r
# Modelo con dummy de género
modelo_genero ← lm(wage ~ educ + exper + mujer, data = wage1)
summary(modelo_genero)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -1.73448100 0.75362027 -2.301532 2.175453e-02
## educ         0.60258016 0.05111738 11.788166 1.332335e-28
## exper        0.06424172 0.01040033  6.176894 1.316295e-09
## mujer       -2.15551716 0.27030549 -7.974374 9.735838e-15
```

Interpretación:

- **mujer = -1.81** → Las mujeres ganan $1.81/hora **menos** que los hombres, manteniendo educación y experiencia constantes
- Altamente significativo ($p < 0.001$)

Brecha de Género en Modelo Log-Lineal

```
# Modelo log para interpretar en porcentajes
modelo_genero_log ← lm(log(wage) ~ educ + exper + mujer, data = wage1)
summary(modelo_genero_log)$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)  0.480835706 0.105016267  4.578678 5.856166e-06
## educ         0.091289739 0.007123158 12.815907 6.939276e-33
## exper        0.009413852 0.001449276  6.495556 1.931108e-10
## mujer       -0.343596717 0.037666812 -9.122002 1.596051e-18
```

Interpretación:

- **mujer = -0.297** → Las mujeres ganan aproximadamente **29.7% menos**, manteniendo educación y experiencia constantes
- Más fácil de comunicar que diferencias absolutas

Variables Categóricas con Múltiples Categorías

Regla m-1: Si una variable tiene m categorías, incluimos m-1 dummies

Ejemplo: Ocupación (8 categorías en wage1)

```
# Ver ocupaciones
table(wage1$occupation)
```

```
## < table of extent 0 >
```

¿Por qué m-1?

- Evitar multicolinealidad perfecta
- La categoría omitida es la **categoría base** (referencia)
- Las dummies se interpretan respecto a la base

Ejemplo: Ocupación y Salarios

```
wage1 ← wage1 %>%
  mutate(ocupacion = case_when(construc == 1 ~ 'Construccion',
                               ndurman == 1 ~ 'Industria',
                               trcommpu == 1 ~ 'Transporte y comunicaciones',
                               trade == 1 ~ 'Comercio',
                               services == 1 ~ 'Servicios',
                               TRUE ~ 'Otros servicios'
                               ),
         ocupacion_f = factor (ocupacion))
modelo_ocupacion ← lm(log(wage) ~ educ + exper + ocupacion_f, data = wage1)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper + ocupacion_f, data = wage1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81330 -0.32272 -0.03493  0.29406  1.58048
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        0.165269   0.107252   1.541 0.123942
## educ                               0.092331   0.007455  12.386  < 2e-16 ***
## exper                              0.009623   0.001519   6.333 5.22e-10 ***
```

Bloque 3

Supuestos y Diagnósticos

Supuestos de Gauss-Markov

Para que MCO sea el **Mejor Estimador Lineal Insesgado (MELI/BLUE):**

1. **Linealidad en parámetros:** El modelo es lineal en los β
2. **Exogeneidad:** $E[u|X] = 0$ (el error no está correlacionado con X)
3. **No multicolinealidad perfecta:** Las X no están perfectamente correlacionadas
4. **Homocedasticidad:** $Var(u|X) = \sigma^2$ (varianza constante del error)
5. **No autocorrelación:** $Cov(u_i, u_j) = 0$ para i ≠ j

¿Qué pasa si se violan?

- Estimadores dejan de ser óptimos
- Pueden ser sesgados o ineficientes
- Necesitamos diagnósticos y correcciones

PROBLEMA 1: Multicolinealidad

¿Qué es?

- Alta correlación entre variables independientes
- Las X están "diciendo lo mismo"

Síntomas:

- R^2 alto pero pocos coeficientes significativos
- Coeficientes con signos inesperados
- Errores estándar muy grandes
- Coeficientes muy sensibles a pequeños cambios en datos

Consecuencia:

- Estimadores siguen siendo insesgados
- Pero tienen varianzas infladas → dificulta detectar efectos reales

Detección: Matriz de Correlación

```r
# Seleccionar variables numéricas
vars_numericas ← wage1 %>% select(wage, educ, exper, tenure)

# Matriz de correlación
cor(vars_numericas) %>% round(3)
```

```
##         wage   educ  exper tenure
## wage   1.000  0.406  0.113  0.347
## educ   0.406  1.000 -0.300 -0.056
## exper  0.113 -0.300  1.000  0.499
## tenure 0.347 -0.056  0.499  1.000
```

Interpretación:

- Experiencia y antigüedad tienen correlación moderada (0.459)
- Educación tiene baja correlación con las demás (bueno)

Detección: Factor de Inflación de Varianza (VIF)

```
# Calcular VIF
vif(modelo_multiple)
```

```
##     educ    exper   tenure
## 1.112771 1.477618 1.349296
```

Regla general:

- VIF = 1: No correlación
- VIF entre 1-5: Correlación moderada (aceptable)
- VIF entre 5-10: Correlación alta (preocupante)
- VIF > 10: Multicolinealidad severa

En nuestro caso: No hay problema serio de multicolinealidad

Ejemplo con Multicolinealidad Alta

```r
# Crear variable altamente correlacionada
wage1$exper_mas_tenure ← wage1$exper + wage1$tenure + rnorm(nrow(wage1), 0, 0.5)

modelo_multicol ← lm(wage ~ educ + exper + tenure + exper_mas_tenure, data = wage1)
```

```
##                      Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)      -2.871557638  0.7310832 -3.92781233 9.726054e-05
## educ              0.598877459  0.0514445 11.64123477 5.323768e-28
## exper             0.015224322  0.2755578  0.05524911 9.559612e-01
## tenure            0.162209615  0.2739784  0.59205267 5.540721e-01
## exper_mas_tenure  0.007099474  0.2746850  0.02584588 9.793902e-01
```

```r
vif(modelo_multicol)
```

```
##             educ            exper           tenure exper_mas_tenure
##         1.117623       770.349147       215.778390      1388.498321
```

¡VIF altísimos! exper_mas_tenure está causando multicolinealidad

Soluciones a Multicolinealidad

1. Eliminar una de las variables correlacionadas

```
# Quitar la variable problemática
modelo_sin_multicol ← lm(wage ~ educ + exper + tenure, data = wage1)
```

2. Transformar variables

- Crear ratios (ej: PBI per cápita en vez de PBI y Población)
- Usar diferencias

3. Aceptarlo (si el objetivo es predicción)

- Si no nos interesa interpretar coeficientes individuales
- Si queremos solo predecir Y

4. Obtener más datos

- A veces el problema es muestra pequeña

PROBLEMA 2: Heteroscedasticidad

¿Qué es?

- La varianza del error NO es constante
- $Var(u|X) = \sigma_i^2$ (varía con i)

¿Por qué importa?

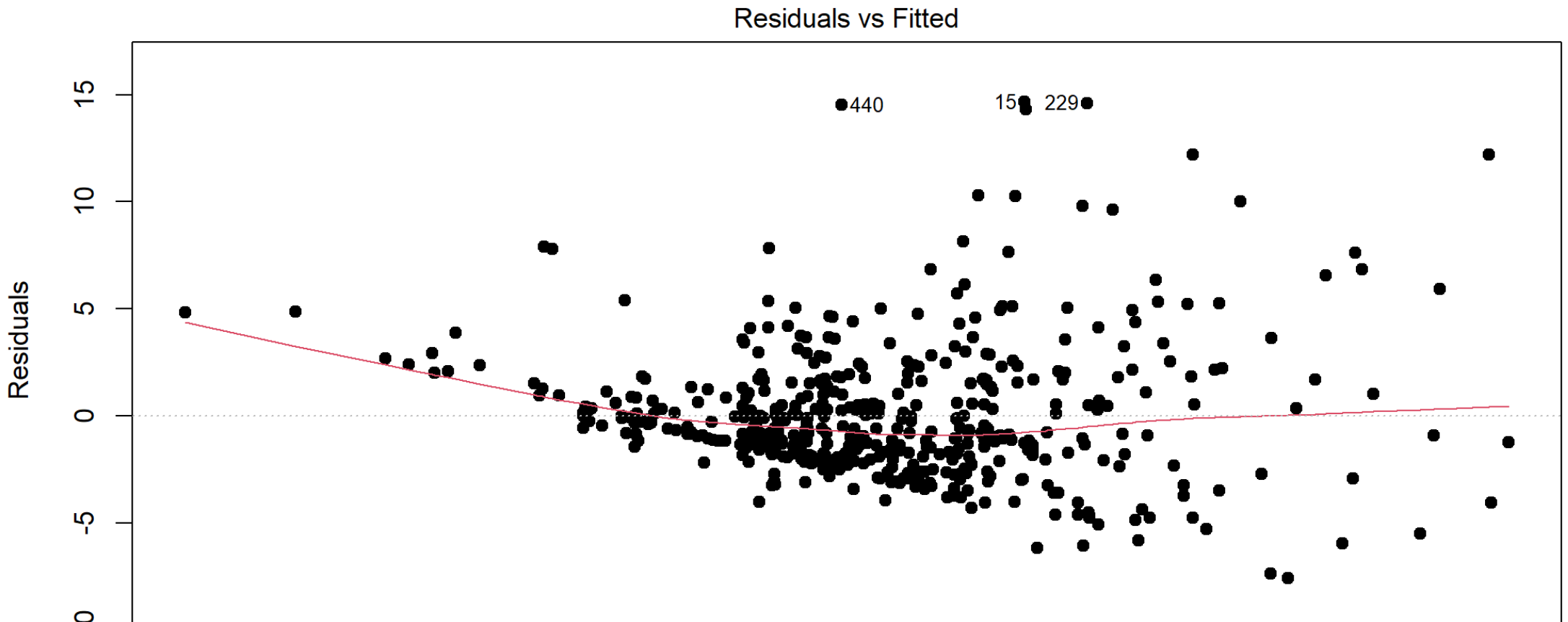
- Estimadores de MCO siguen siendo **insesgados**
- Pero los **errores estándar están mal calculados**
- Consecuencia: inferencia incorrecta (p-valores erróneos)

Causas comunes:

- Variables con escalas muy diferentes
- Relaciones que se amplifican (ej: ingreso alto → mayor variación en consumo)
- Variables omitidas

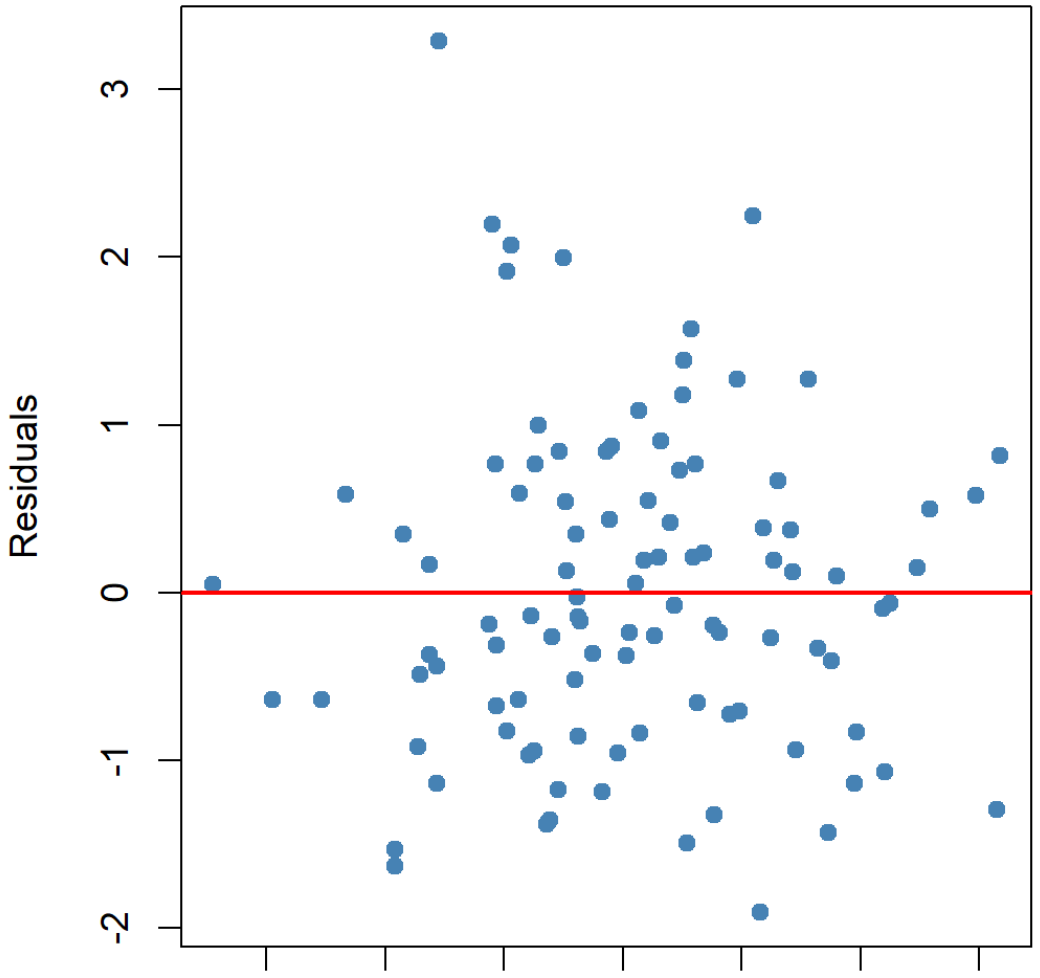
Detección Visual: Gráfico de Residuos

```r
# Residuos vs valores ajustados
plot(modelo_multiple, which = 1, pch = 19)
```

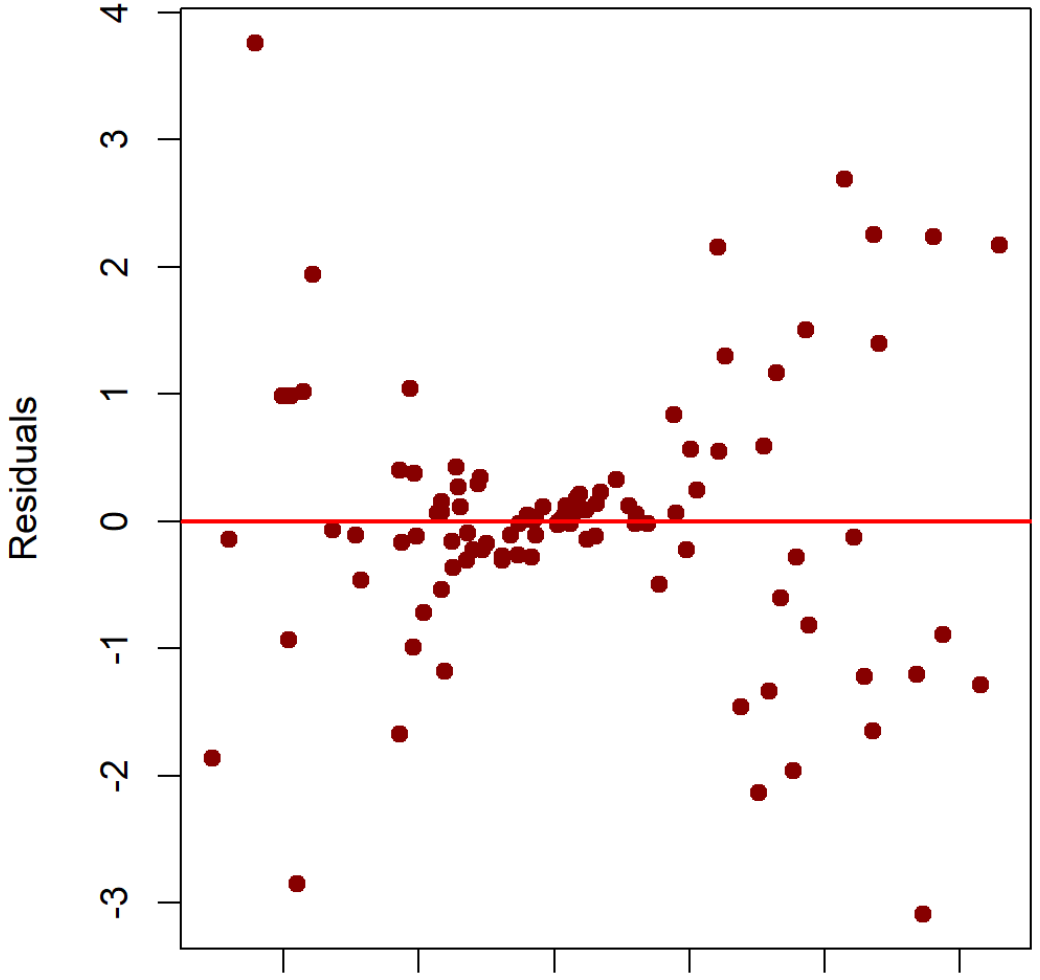


Interpretación de Gráficos de Residuos

✓ **Homocedasticidad**



✗ **Heteroscedasticidad**



Tests Formales de Heteroscedasticidad

Test de Breusch-Pagan:

```
# Test de Breusch-Pagan
bptest(modelo_multiple)
```

```
##
##     studentized Breusch-Pagan test
##
## data:  modelo_multiple
## BP = 43.096, df = 3, p-value = 2.349e-09
```

Interpretación:

- H_0: Homocedasticidad (varianza constante)
- H_1: Heteroscedasticidad
- p-valor = 0.014 < 0.05 $\rightarrow$ **Rechazamos H_0**
- **Hay evidencia de heteroscedasticidad**

Test de White (más general)

```
# Test de White (incluye términos cuadráticos)
bptest(modelo_multiple, ~ educ + exper + tenure +
         I(educ^2) + I(exper^2) + I(tenure^2), data = wage1)
```

```
##
##     studentized Breusch-Pagan test
##
## data:  modelo_multiple
## BP = 57.052, df = 6, p-value = 1.783e-10
```

Interpretación:

- Test más general que Breusch-Pagan
- También detecta formas no lineales de heteroscedasticidad
- p-valor < 0.05 → Confirmamos heteroscedasticidad

SOLUCIÓN: Errores Robustos (Errores de White)

Idea: Corregir los errores estándar sin cambiar los coeficientes

```
# Modelo original (errores normales)
summary(modelo_multiple)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -2.87273482 0.72896429 -3.940844 9.224742e-05
## educ         0.59896507 0.05128355 11.679478 3.681353e-28
## exper        0.02233952 0.01205685  1.852849 6.446818e-02
## tenure       0.16926865 0.02164461  7.820361 2.934527e-14
```

Comparación: Errores Normales vs Robustos

```r
# Errores robustos usando sandwich
coeftest(modelo_multiple, vcov = vcovHC(modelo_multiple, type = "HC1"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -2.872735   0.807415 -3.5579 0.0004078 ***
## educ         0.598965   0.061014  9.8169 < 2.2e-16 ***
## exper        0.022340   0.010555  2.1165 0.0347731 *
## tenure       0.169269   0.029278  5.7814 1.277e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

¿Qué cambió?

- Los coeficientes son idénticos
- Los errores estándar son diferentes
- Los p-valores cambian ligeramente
- La significancia puede cambiar en casos borderline

Visualización de la Diferencia

Comparación de Errores Estándar

	Variable	Coeficiente	SE_Normal	SE_Robusto	Diferencia
(Intercept)	(Intercept)	-2.873	0.729	0.807	10.762
educ	educ	0.599	0.051	0.061	18.974
exper	exper	0.022	0.012	0.011	-12.458
tenure	tenure	0.169	0.022	0.029	35.269

Los errores robustos son generalmente mayores (más conservadores)

¿Cuándo Usar Errores Robustos?

Recomendación práctica:

✅ SIEMPRE úsalos en datos de corte transversal

- Es la práctica estándar en economía
- No se pierde nada si hay homocedasticidad
- Proteges tu inferencia si hay heteroscedasticidad

```
# Tu workflow estándar debería ser:
modelo ← lm(Y ~ X1 + X2, data = datos)
coeftest(modelo, vcov = vcovHC(modelo, type = "HC1"))
```

⚠️ No son necesarios para:

- Series de tiempo (hay otras correcciones)
- Datos de panel (efectos fijos/aleatorios)

PROBLEMA 3: Endogeneidad (Mención Breve)

¿Qué es?

- Violación del supuesto de exogeneidad: $E[u|X] \neq 0$
- Las X están correlacionadas con el error

Causas principales:

1. **Variables omitidas:** Olvidamos incluir Z relevante que se correlaciona con X
2. **Error de medición:** X está medido con error
3. **Simultaneidad:** Y también afecta a X (causalidad reversa)

Consecuencia: Estimadores son **SESGADOS** (no solo ineficientes)

En esta clase: Solo lo reconocemos, no lo resolvemos

- Solución requiere Variables Instrumentales (otra clase)
- Nuestro β describe asociación, no causalidad

Ejemplo Conceptual: Variable Omitida

Modelo verdadero:

$$Salario = \beta_0 + \beta_1 Educacion + \beta_2 Habilidad + u$$

Modelo estimado (habilidad no observable):

$$Salario = \alpha_0 + \alpha_1 Educacion + e$$

Problema:

- Habilidad está en el error (e)
- Habilidad se correlaciona con Educación (personas más hábiles estudian más)
- $Cov(Educacion, e) \neq 0 \rightarrow$ ¡Endogeneidad!
- α_1 está **sesgado** (probablemente sobreestima el efecto de educación)

Mensaje clave: Nuestras regresiones describen asociaciones. La causalidad requiere más supuestos o diseño experimental.

Bloque 4

Workflow y Buenas Prácticas

Flujo de Trabajo Recomendado

1. EXPLORACIÓN

```
# Estadísticas descriptivas
summary(datos)
# Correlaciones
cor(datos)
# Gráficos exploratorios
plot(datos)
```

2. ESPECIFICACIÓN

- Elegir variables relevantes (teoría económica)
- Decidir forma funcional (lineal, log, cuadrática)

3. ESTIMACIÓN

```
modelo ← lm(Y ~ X1 + X2, data = datos)
```

Flujo de Trabajo (continuación)

4. DIAGNÓSTICO

```
# Gráficos de residuos
plot(modelo)
# Multicolinealidad
vif(modelo)
# Heteroscedasticidad
bptest(modelo)
```

5. CORRECCIÓN (si es necesario)

```
# Errores robustos
coeftest(modelo, vcov = vcovHC(modelo, type = "HC1"))
```

6. INTERPRETACIÓN Y REPORTE

- Presentar resultados de forma clara
- Ser honesto sobre limitaciones

Selección Entre Modelos

```r
# Tres modelos candidatos
m1 ← lm(log(wage) ~ educ, data = wage1)
m2 ← lm(log(wage) ~ educ + exper, data = wage1)
m3 ← lm(log(wage) ~ educ + exper + tenure, data = wage1)

# Comparar con criterios de información
data.frame(
  Modelo = c("M1: educ", "M2: educ + exper", "M3: educ + exper + tenure"),
  R2_adj = c(summary(m1)$adj.r.squared,
             summary(m2)$adj.r.squared,
             summary(m3)$adj.r.squared),
  AIC = c(AIC(m1), AIC(m2), AIC(m3)),
  BIC = c(BIC(m1), BIC(m2), BIC(m3))
) %>%
  kable(digits = 3) %>%
  kable_styling(font_size = 14)
```

Modelo	R2_adj	AIC	BIC
M1: educ	0.184	724.756	737.552
M2: educ + exper	0.246	684.019	701.080
M3: educ + exper + tenure	0.312	637.096	658.422

Test de Ramsey (RESET)

¿Está bien especificado el modelo?

```
# Test RESET
resettest(modelo_multiple, power = 2:3)
```

```
##
##     RESET test
##
## data:  modelo_multiple
## RESET = 11.566, df1 = 2, df2 = 520, p-value = 1.217e-05
```

Interpretación:

- H_0: El modelo está correctamente especificado
- H_1: Hay problemas de especificación (forma funcional incorrecta, **variables omitidas -depende el manual-**)
- p-valor < 0.05 → Posible problema de especificación
- **Acción:** Considerar transformaciones o variables adicionales

Reporte Profesional: stargazer

```
stargazer(m1, m2, m3,
          type = "text",
          title = "Determinantes del Salario",
          dep.var.labels = "Log(Salario)",
          covariate.labels = c("Educación", "Experiencia", "Antigüedad"),
          se = list(
            sqrt(diag(vcovHC(m1, type="HC1"))),
            sqrt(diag(vcovHC(m2, type="HC1"))),
            sqrt(diag(vcovHC(m3, type="HC1")))
          ),
          notes = "Errores estándar robustos entre paréntesis",
          notes.append = FALSE)
```

Ventajas:

- Formato profesional
- Múltiples modelos lado a lado
- Fácil de exportar a LaTeX, HTML, texto

Tabla de Resultados

```
library(modelsummary)
modelos ← list(
  "Modelo 1" = m1,
  "Modelo 2" = m2,
  "Modelo 3" = m3
)

modelsummary(modelos,output = '',
             vcov = "HC1",
             stars = TRUE,
             gof_map = c("nobs", "r.squared", "adj.r.squared"),
             coef_rename = c("educ" = "Educación",
                             "exper" = "Experiencia",
                             "tenure" = "Antigüedad"),
             title = "Determinantes del Log(Salario)")
```

Checklist de Buenas Prácticas

✅ **SIEMPRE hacer:**

1. Explorar datos antes de modelar
2. Justificar elección de variables (teoría)
3. Reportar R^2 ajustado (no solo R^2)
4. Verificar multicolinealidad (VIF)
5. Hacer gráficos de residuos
6. Testear heteroscedasticidad
7. **Usar errores robustos en corte transversal**
8. Interpretar magnitud Y significancia
9. Ser honesto sobre limitaciones
10. Documentar todo tu código

Checklist de Buenas Prácticas (cont.)

❌ **NUNCA hacer:**

1. Agregar variables solo para mejorar R^2
2. Eliminar observaciones sin justificación
3. Omitir variables relevantes conocidas
4. Ignorar diagnósticos
5. Interpretar correlación como causalidad sin justificación
6. Reportar solo resultados significativos (sesgo de publicación)
7. Olvidar la interpretación económica

Ejemplo Integrador Completo

```r
# 1. DATOS
data(wage1)

# 2. EXPLORACIÓN
summary(wage1[, c("wage", "educ", "exper", "female")])
```

```
##       wage             educ           exper           female      
##  Min.   : 0.530   Min.   : 0.00   Min.   : 1.00   Min.   :0.0000  
##  1st Qu.: 3.330   1st Qu.:12.00   1st Qu.: 5.00   1st Qu.:0.0000  
##  Median : 4.650   Median :12.00   Median :13.50   Median :0.0000  
##  Mean   : 5.896   Mean   :12.56   Mean   :17.02   Mean   :0.4791  
##  3rd Qu.: 6.880   3rd Qu.:14.00   3rd Qu.:26.00   3rd Qu.:1.0000  
##  Max.   :24.980   Max.   :18.00   Max.   :51.00   Max.   :1.0000
```

Ejemplo Integrador (cont.)

```r
# 3. ESPECIFICACIÓN Y ESTIMACIÓN
modelo_final ← lm(log(wage) ~ educ + exper + I(exper^2) + female,
                  data = wage1)

# 4. DIAGNÓSTICO
vif(modelo_final)
```

```
##       educ      exper I(exper^2)     female
##   1.139708  13.162876  13.439344   1.013099
```

Ejemplo Integrador (cont.)

```r
# Test de heteroscedasticidad
bptest(modelo_final)
```

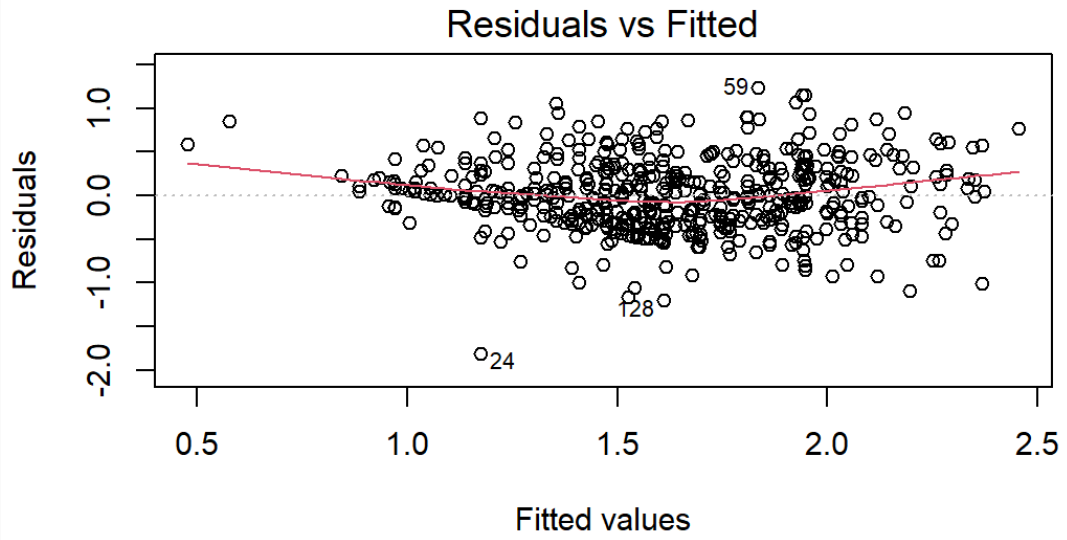
```
##
##     studentized Breusch-Pagan test
##
## data:  modelo_final
## BP = 9.9839, df = 4, p-value = 0.0407
```

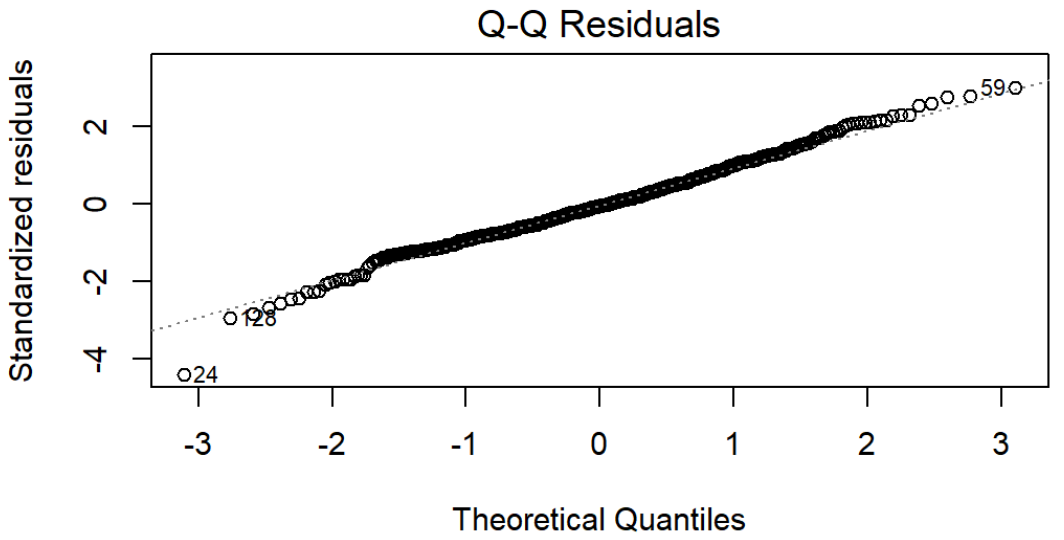
```r
# 5. RESULTADOS CON ERRORES ROBUSTOS
coeftest(modelo_final, vcov = vcovHC(modelo_final, type = "HC1"))
```

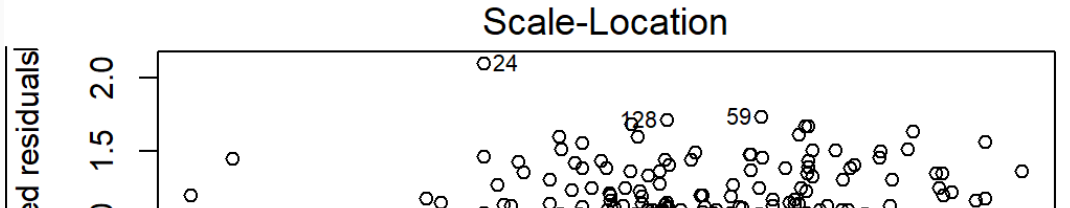
```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  0.39048305  0.10859848  3.5957 0.0003544 ***
## educ         0.08413608  0.00768995 10.9410 < 2.2e-16 ***
## exper        0.03890997  0.00467524  8.3226 7.603e-16 ***
## I(exper^2)  -0.00068602  0.00010046 -6.8288 2.393e-11 ***
## female      -0.33718676  0.03618383 -9.3187 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

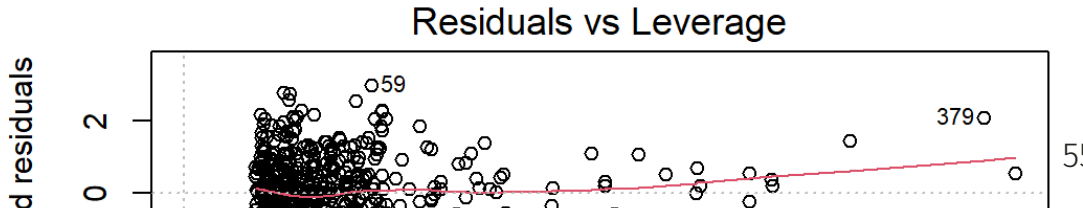
Ejemplo Integrador: Gráficos de Diagnóstico

```
par(mfrow = c(2, 2))
plot(modelo_final)
```









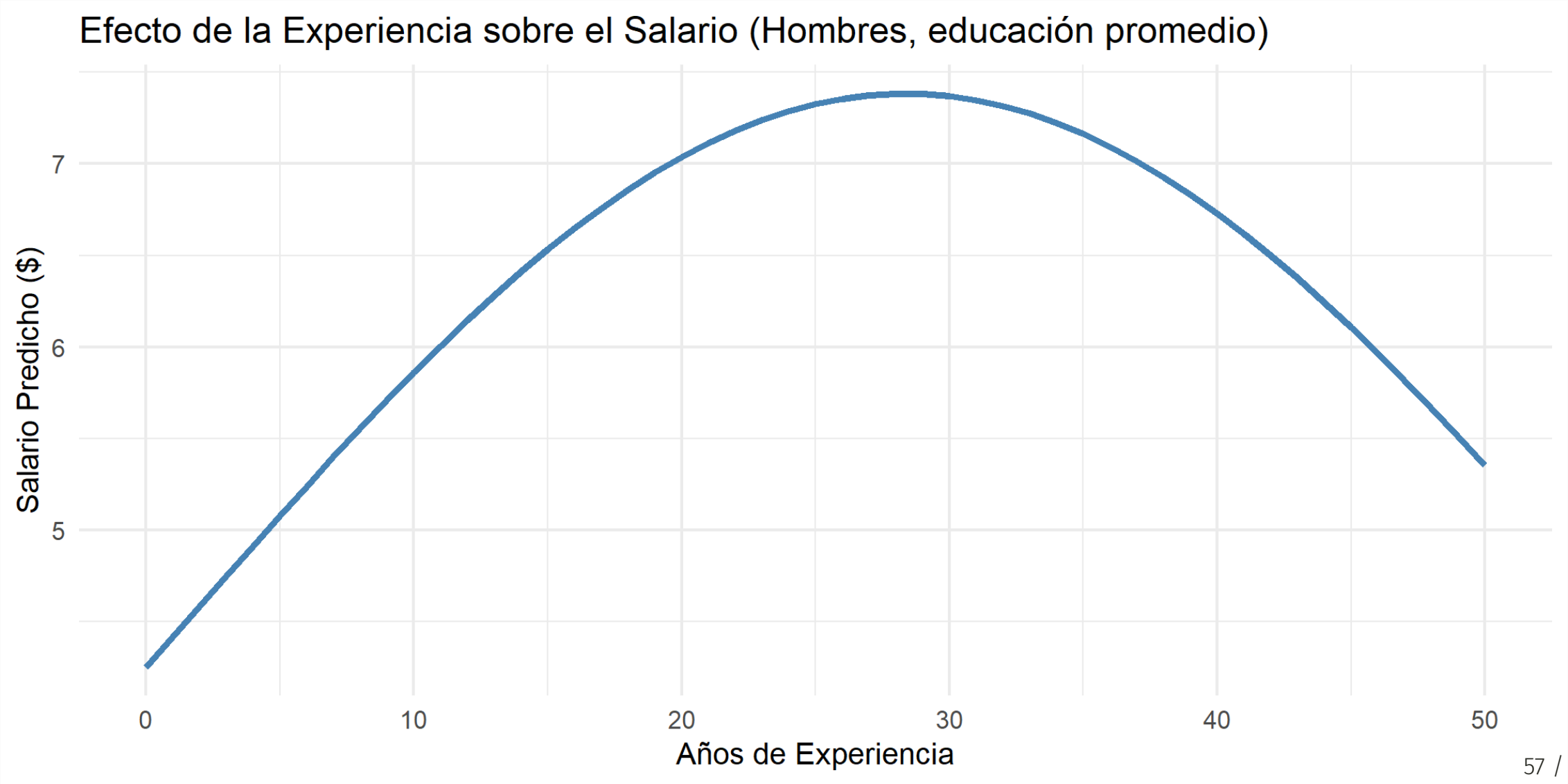
Interpretación Final del Modelo

```
##                 Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)  0.3904830518 0.1022096385  3.820413 1.492951e-04
## educ         0.0841360752 0.0069568041 12.094070 7.520205e-30
## exper        0.0389099671 0.0048235402  8.066682 5.003819e-15
## I(exper^2)  -0.0006860225 0.0001073782 -6.388842 3.709946e-10
## female      -0.3371867567 0.0363213775 -9.283424 4.405993e-19
```

Interpretaciones (en términos porcentuales):

- **Educación:** +1 año → salario aumenta ~9.2%
- **Experiencia:** Efecto no lineal (cuadrático)
  - Inicialmente positivo, luego se aplana
  - Máximo en: $-\beta_{exper}/(2 \times \beta_{exper^2}) = -0.041/(2 \times -0.0007) \approx 29$ años
- **Género:** Mujeres ganan ~30% menos (manteniendo educación y experiencia constantes)

Visualización del Efecto de Experiencia



Recursos Adicionales

Paquetes útiles:

- `car`: Diagnósticos (VIF, etc.)
- `lmtest`: Tests de especificación
- `sandwich`: Errores robustos
- `stargazer` / `modelsummary`: Tablas bonitas
- `ggplot2`: Visualizaciones
- `wooldridge`: Datasets de práctica

Libros recomendados:

- Wooldridge: "Introducción a la Econometría"
- Gujarati: "Econometría"
- Stock & Watson: "Introduction to Econometrics"

Online:

- R for Data Science: https://r4ds.had.co.nz/
- Cross Validated (Stack Exchange para estadística)

Resumen Final

Puntos Clave para Recordar

1. Interpretación:

- Los coeficientes describen **asociaciones**, no necesariamente causalidad
- Diferentes formas funcionales tienen diferentes interpretaciones
- Mantener todo lo demás constante (ceteris paribus)

2. Supuestos:

- MCO requiere varios supuestos para ser óptimo
- Verificar siempre con diagnósticos
- Corregir cuando sea necesario (especialmente heteroscedasticidad)

3. Práctica:

- Explorar → Especificar → Estimar → Diagnosticar → Corregir → Interpretar
- Usar errores robustos por defecto en corte transversal
- Reportar de forma honesta y completa

¿Preguntas?

Contacto: nsidicaro.fce@gmail.com