Laboratorio 8 – Pruebas conjuntas de hipótesis

Table of contents

El propósito de este laboratorio es **practicar la realización de pruebas conjuntas de hipótesis** sobre parámetros de regresión en R. Lo haremos usando pruebas t y pruebas F.

Para empezar

Abre un nuevo script de R y carga los paqeutes

```
# Instalar paquetes solo si faltan
pkgs <- c("tidyverse", "broom", "wooldridge", "car", "magrittr", "kableExtra")
to_install <- pkgs[!pkgs %in% rownames(installed.packages())]
if (length(to_install) > 0) {
  install.packages(to_install, repos = "http://cran.us.r-project.org")
}

library(tidyverse)
library(broom)
library(wooldridge)
library(car)
library(magrittr)
library(kableExtra)
```

Cargar los datos

Usaremos un conjunto de datos sobre ingresos y habilidad, llamado `htv`. El conjunto de datos contiene una muestra de 1,230 trabajadores.

```
df <- as_tibble(htv)
```

Revisa qué contiene el conjunto de datos escribiendo:

```
glimpse(df)
```

```
Rows: 1,230
Columns: 23
$ wage     <dbl> 12.019231, 8.912656, 15.514334, 13.333333, 11.070110, 17.4825~
$ abil     <dbl> 5.0277381, 2.0371704, 2.4758952, 3.6092398, 2.6365459, 3.4743~
$ educ     <int> 15, 13, 15, 15, 13, 18, 13, 12, 13, 12, 12, 12, 17, 13, 17, 1~
$ ne       <int> 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1~
$ nc       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ west     <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0~
$ south    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ exper    <int> 9, 8, 11, 6, 15, 8, 13, 14, 9, 9, 13, 14, 4, 8, 7, 10, 10, 9,~
$ motheduc <int> 12, 12, 12, 12, 12, 12, 13, 12, 10, 14, 9, 12, 17, 16, 16, 5,~
$ fatheduc <int> 12, 10, 16, 12, 15, 12, 12, 12, 12, 12, 10, 16, 16, 16, 18, 4~
$ brkhme14 <int> 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ sibs     <int> 1, 4, 2, 1, 2, 2, 5, 4, 3, 1, 2, 1, 1, 3, 2, 2, 1, 1, 2, 2, 2~
$ urban    <int> 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0~
$ ne18     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1~
$ nc18     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ south18  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ west18   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0~
$ urban18  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ tuit17   <dbl> 7.582914, 8.595144, 7.311346, 9.499537, 7.311346, 7.311346, 7~
$ tuit18   <dbl> 7.260242, 9.499537, 7.311346, 10.162070, 7.311346, 7.311346, ~
$ lwage    <dbl> 2.486508, 2.187472, 2.741764, 2.590267, 2.404249, 2.861201, 3~
$ expersq  <int> 81, 64, 121, 36, 225, 64, 169, 196, 81, 81, 169, 196, 16, 64,~
$ ctuit    <dbl> -0.32267141, 0.90439224, 0.00000000, 0.66253376, 0.00000000, ~
```

Las principales variables que nos interesan son: salarios, educación, habilidad, educación de los padres y región de residencia (`ne`, `nc`, `west` y `south`).

Crear variable factorial de región

Comencemos creando una variable factorial a partir de las cuatro variables dummy regionales. Tomando código del lab 6, tenemos:

```
df %<>% mutate(region = case_when(ne==1 ~ "Northeast",
                                  nc==1 ~ "NorthCentral",
                                  west==1 ~ "West",
                                  south==1 ~ "South")) %>%
        mutate(region = factor(region))

# Verificar que funcionó
table(df$region)
```

```
NorthCentral    Northeast        South         West
         458          260          304          208
```

Regresión y pruebas de hipótesis

Estima el siguiente modelo de regresión:

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + \beta_5 region + u$$

Nota que *abil* está en unidades de desviación estándar. Necesitarás usar una función `mutate()` para crear $abil^2$. Llámala `abilsq`. *region* representa la variable factorial que creaste anteriormente.[1]

```
df %<>% mutate(abilsq = abil^2)
```

```
est <- lm(educ ~ motheduc + fatheduc + abil + abilsq + region, data = df)
```

** Modelo completo: educ ~ motheduc + fatheduc + abil + abil² + region **

Coeficientes estimados:

---

[1] Aquí la notación de $\beta_5 region$ no es del todo correcta. Técnicamente debería escribirse $\beta_5 region.NE + \beta_6 region.S + \beta_7 region.W$, donde cada una de las variables $region.X$ es una dummy. La forma en que está escrita arriba, $\beta_5 region$ implica que β_5 es un vector, no un escalar.

	Estimación	Error estándar	Estadístico t	Valor p
(Intercept)	8.228830	0.292800	28.103924	0.000000
motheduc	0.188975	0.028254	6.688469	0.000000
fatheduc	0.107275	0.019666	5.454933	0.000000
abil	0.400004	0.030384	13.164776	0.000000
abilsq	0.050585	0.008317	6.081899	0.000000
regionNortheast	0.187376	0.136853	1.369171	0.171197
regionSouth	-0.014308	0.130779	-0.109408	0.912897
regionWest	0.075615	0.148390	0.509569	0.610445

Estadísticas del modelo:

Estadístico	Valor
R^2	0.4454
R^2 ajustado	0.4423
Error estándar residual	1.7583
Estadístico F	140.21
N	1230

1. Prueba t: efecto lineal de habilidad

Pregunta: Prueba la hipótesis de que *abil* tiene un efecto lineal sobre *educ*.

Esto significa probar:

$$H_0 : \beta_4 = 0; \quad H_a : \beta_4 \neq 0$$

Si $\beta_4 = 0$, entonces el efecto de habilidad es puramente lineal (sin componente cuadrático).

```
# Extraer información de abilsq
info_abilsq <- tidy(est) %>% filter(term == "abilsq")
t_stat <- info_abilsq$statistic
p_value <- info_abilsq$p.value
beta_abilsq <- info_abilsq$estimate

cat("Prueba de efecto lineal de habilidad:\n")
```

```
Prueba de efecto lineal de habilidad:
```

```r
cat("====================================\n")
```

```
====================================
```

```r
cat("H :   = 0 (efecto puramente lineal)\n")
```

```
H :   = 0 (efecto puramente lineal)
```

```r
cat("H :    0 (efecto cuadrático presente)\n\n")
```

```
H :    0 (efecto cuadrático presente)
```

```r
cat("^  (abilsq) =", sprintf("%.6f", beta_abilsq), "\n")
```

```
^  (abilsq) = 0.050585
```

```r
cat("Error estándar =", sprintf("%.6f", info_abilsq$std.error), "\n")
```

```
Error estándar = 0.008317
```

```r
cat("Estadístico t =", sprintf("%.4f", t_stat), "\n")
```

```
Estadístico t = 6.0819
```

```r
cat("Valor p =", sprintf("%.6f", p_value), "\n\n")
```

```
Valor p = 0.000000
```

```r
niveles <- c(0.01, 0.05, 0.10)
for (nivel in niveles) {
  significativo <- p_value < nivel
  cat("Al nivel   =", nivel, ":",
      ifelse(significativo, "  RECHAZAMOS H ", "  NO RECHAZAMOS H "), "\n")
}
```

```
Al nivel   = 0.01 :   RECHAZAMOS H
Al nivel   = 0.05 :   RECHAZAMOS H
Al nivel   = 0.1 :   RECHAZAMOS H
```

```r
cat("\nConclusión: ")
```

```
Conclusión:
```

```r
if (p_value < 0.05) {
  cat("Hay evidencia significativa de un efecto cuadrático de la habilidad.\n")
  cat("La relación entre habilidad y educación NO es lineal.\n")
} else {
  cat("No hay evidencia suficiente de un efecto cuadrático.\n")
  cat("La relación entre habilidad y educación puede considerarse lineal.\n")
}
```

```
Hay evidencia significativa de un efecto cuadrático de la habilidad.
La relación entre habilidad y educación NO es lineal.
```

2. Prueba F: igualdad de efectos parentales

Pregunta: Ahora prueba que *motheduc* y *fatheduc* tienen efectos iguales sobre *educ*. En otras palabras, prueba:

$$H_0 : \beta_1 = \beta_2; \quad H_a : \beta_1 \neq \beta_2$$

Para esto, necesitarás obtener $se(\beta_1 - \beta_2)$. Afortunadamente, R lo hará por ti con la función `linearHypothesis()` del paquete `car`:

```r
test_padres <- linearHypothesis(est, "motheduc = fatheduc")
```

Prueba F: Igualdad de efectos de educación parental

$H_0 : \beta_1 = \beta_2$ (madre y padre tienen el mismo efecto)

$H_a : \beta_1 \neq \beta_2$ (efectos diferentes)

Table 3: Resultado de la prueba F

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1223	3789.557	NA	NA	NA	NA
1222	3777.902	1	11.65565	3.770137	0.052405

```r
# Extraer información de la prueba
F_stat_padres <- test_padres$F[2]
p_value_padres <- test_padres$`Pr(>F)`[2]

cat("\nResultados de la prueba:\n")
```

```
Resultados de la prueba:
```

```r
cat("========================\n")
```

```
========================
```

```r
cat("Estadístico F =", sprintf("%.4f", F_stat_padres), "\n")
```

```
Estadístico F = 3.7701
```

```r
cat("Valor p =", sprintf("%.6f", p_value_padres), "\n\n")
```

```
Valor p = 0.052405
```

```r
if (p_value_padres < 0.05) {
  cat("Decisión: RECHAZAMOS H  al 5%\n")
  cat("Conclusión: La educación de la madre y del padre tienen\n")
  cat("            efectos DIFERENTES sobre la educación del hijo.\n")
} else {
  cat("Decisión: NO RECHAZAMOS H  al 5%\n")
  cat("Conclusión: No hay evidencia suficiente para decir que los\n")
  cat("            efectos de la educación parental sean diferentes.\n")
}
```

```
Decisión: NO RECHAZAMOS H  al 5%
Conclusión: No hay evidencia suficiente para decir que los
            efectos de la educación parental sean diferentes.
```

```r
# Mostrar los coeficientes individuales para comparación
beta_madre <- coef(est)["motheduc"]
beta_padre <- coef(est)["fatheduc"]
cat("\nCoeficientes individuales:\n")
```

```
Coeficientes individuales:
```

```
cat("^  (motheduc) =", sprintf("%.6f", beta_madre), "\n")
```

```
^  (motheduc) = 0.188975
```

```
cat("^  (fatheduc) =", sprintf("%.6f", beta_padre), "\n")
```

```
^  (fatheduc) = 0.107275
```

```
cat("Diferencia =", sprintf("%.6f", beta_madre - beta_padre), "\n")
```

```
Diferencia = 0.081700
```

Nota: El valor p resultante es de una prueba F, pero se obtendría un resultado idéntico usando una prueba t, ya que esta es una hipótesis simple (ver Wooldridge, pp. 125-126).

3. Prueba F: significancia conjunta de las dummies regionales

Los valores p de la regresión anterior podrían indicar que las tres dummies regionales no contribuyen a la educación.

Pregunta: Prueba la hipótesis de que no contribuyen; es decir, prueba:

$$H_0 : \text{todas las dummies regionales} = 0$$
$$H_a : \text{cualquier dummy regional} \neq 0$$

El código para hacer esto nuevamente proviene de la función `linearHypothesis()`. La sintaxis es encerrar cada hipótesis componente entre comillas y luego rodearlas con `c()`, que es cómo R crea vectores.

```
test_region1 <- linearHypothesis(est, c("regionNortheast=0", "regionSouth=0", "regionWest=0"
```

O, más simplemente:

```
test_region2 <- linearHypothesis(est, matchCoefs(est, "region"))
```

Prueba F: Significancia conjunta de variables regionales

H_0: Todas las dummies regionales = 0 (región no importa)

H_a: Al menos una dummy regional 0 (región sí importa)

Table 4: Resultado de la prueba F conjunta

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1225	3785.243	NA	NA	NA	NA
1222	3777.902	3	7.34082	0.791487	0.498656

```
# Extraer información de la prueba
F_stat_region <- test_region2$F[2]
p_value_region <- test_region2$`Pr(>F)`[2]

cat("\nResultados de la prueba:\n")
```

```
Resultados de la prueba:
```

```
cat("========================\n")
```

```
========================
```

```
cat("Estadístico F =", sprintf("%.4f", F_stat_region), "\n")
```

```
Estadístico F = 0.7915
```

```
cat("Grados de libertad = (3,", test_region2$Df[2], ")\n")
```

```
Grados de libertad = (3, 3 )
```

```
cat("Valor p =", sprintf("%.6f", p_value_region), "\n\n")
```

```
Valor p = 0.498656
```

```r
if (p_value_region < 0.05) {
  cat("Decisión: RECHAZAMOS H  al 5%\n")
  cat("Conclusión: La región de residencia SÍ tiene un efecto\n")
  cat("            significativo sobre la educación.\n")
} else {
  cat("Decisión: NO RECHAZAMOS H  al 5%\n")
  cat("Conclusión: La región de residencia NO tiene un efecto\n")
  cat("            significativo sobre la educación.\n")
  cat("            Podríamos considerar eliminar estas variables del modelo.\n")
}
```

```
Decisión: NO RECHAZAMOS H  al 5%
Conclusión: La región de residencia NO tiene un efecto
            significativo sobre la educación.
            Podríamos considerar eliminar estas variables del modelo.
```

Método alternativo: Prueba F "a mano"

Alternativamente, puedes realizar la prueba F de la siguiente manera (no es necesario poner esto en tu script de R; solo te muestro cómo hacerlo "a mano"):

```r
# Modelo restringido (sin dummies regionales)
est.restrict <- lm(educ ~ motheduc + fatheduc + abil + abilsq, data = df)

# Calcular estadístico F manualmente
Fstat.numerator   <- (deviance(est.restrict) - deviance(est)) / 3
Fstat.denominator <- deviance(est) / (nobs(est) - length(coef(est)))
Fstat <- Fstat.numerator / Fstat.denominator
p.value <- 1 - pf(Fstat, 3, nobs(est) - length(coef(est)))

cat("\nPrueba F calculada manualmente:\n")
```

```
Prueba F calculada manualmente:
```

```r
cat("================================\n")
```

```
================================
```

```r
cat("RSS (restringido) =", sprintf("%.4f", deviance(est.restrict)), "\n")
```

```
RSS (restringido) = 3785.2426
```

```r
cat("RSS (no restringido) =", sprintf("%.4f", deviance(est)), "\n")
```

```
RSS (no restringido) = 3777.9018
```

```r
cat("Estadístico F =", sprintf("%.4f", Fstat), "\n")
```

```
Estadístico F = 0.7915
```

```r
cat("Valor p =", sprintf("%.6f", p.value), "\n")
```

```
Valor p = 0.498656
```

```r
cat("\n¿Coincide con linearHypothesis()? ",
    ifelse(abs(Fstat - F_stat_region) < 0.01, "  SÍ", "  NO"), "\n")
```

```
¿Coincide con linearHypothesis()?   SÍ
```

Esto da exactamente la misma respuesta que el código de `linearHypothesis()`.

Resumen de las pruebas de hipótesis

Tabla resumen de todas las pruebas:

Prueba	Hipótesis nula	Tipo	Estadístico	Valor p	Decisi
1. Efecto lineal de habilidad	H : = 0	Prueba t	t = 6.0819	0.000000	Recha
2. Igualdad efectos parentales	H : =	Prueba F	F = 3.7701	0.052405	No re
3. Significancia conjunta región	H : todas dummies región = 0	Prueba F	F = 0.7915	0.498656	No re