

Sesión 25 (A) — Laboratorio integrador I: logaritmos, indicadoras, interacciones y el contraste de Chow con el Boston original

Marcos Bujosa

Descripción de la práctica

Esta práctica y la de la sesión 26 cierran el curso con un solo conjunto de datos reales y, de una vez, las herramientas de las lecciones 15 y 16: un modelo con el regresando en logaritmos, una variable indicadora, una interacción y el contraste de Chow. Los datos son el Boston original, el fichero `boston.csv` que se importó en la [actividad 4 \(html\)](#) de la práctica B de la sesión 4: 506 barrios del área de Boston con el precio mediano de la vivienda (`medv`, en miles de dólares), el número medio de habitaciones por vivienda (`rm`, la `rooms` de `hprice2`), la distancia ponderada a cinco centros de empleo (`dis`, la `dist` de `hprice2`), el porcentaje de población de estatus bajo del barrio (`lstat`) y la indicadora `chas`, que vale 1 en los 35 barrios que limitan con el río Charles. La lección 16 usó `chas` en niveles y con `rm` como único regresor; aquí entra en un modelo log-nivel con tres regresores, que es el modelo hilo de las dos prácticas.

La segunda mitad de la práctica lleva a Gretl el contraste de Chow de la lección 16: primero con `wage1`, para reproducir las cifras de la lección con la orden `chow`, y después con el Boston original por `chas`, donde el resultado es distinto: con `wage1` el contraste sugiere pendientes distintas y con Boston solo un intercepto distinto.

Objetivos

1. Estimar el modelo log-nivel del precio mediano sobre `rm`, `lstat` y `dis` y leer sus coeficientes como en la lección 15.
2. Añadir la indicadora `chas` y leer su coeficiente como diferencia porcentual exacta entre barrios con las mismas características, con la categoría base y la cautela de la lección 16; y comparar el ajuste de los modelos en logaritmos y en niveles sobre las previsiones del precio en dólares, como en la lección 15.
3. Añadir la interacción `chas*rm` y decidir, con su t , si los datos sugieren una pendiente distinta para los barrios junto al río.
4. Hacer en Gretl el contraste de Chow de la lección 16 con `wage1`, con la orden `chow` y con `restrict`, y obtener las mismas cifras que la lección.
5. Repetir el contraste con Boston por `chas` y distinguir lo que dice el F conjunto de lo que dicen las preguntas una a una.

Comandos nuevos de esta práctica. Uno solo:

`chow` tras estimar un modelo con `ols`, contrasta si la ecuación es la misma en dos grupos definidos por una indicadora (`chow female --dummy`) o antes y después de una observación (`chow 300`). Imprime la regresión aumentada con la indicadora y sus interacciones, y el F de la lección 13.

Los demás ya se han usado: `open` y `labels` (sesión 4), `summary --simple` (sesión 7), `ols` y sus accesorios, `$yhat` y `$rsq` incluidos (sesiones 11 a 18), `series` (sesión 7), la función `corr` (la correlación entre dos series, como el comando `corr` de la sesión 7), `restrict` y `$test` ([actividad 2 \(html\)](#)) de la práctica A de la sesión 20), `smpl --restrict` y `smpl full` (sesión 14) y `printf` (sesión 7).

Actividad 1 - Los datos y el modelo log-nivel

Importe `boston.csv` (Archivo ->Abrir datos ->Usuario, o desde línea de comandos con la ruta relativa al fichero). Cree el logaritmo del precio mediano y estime el modelo con `rm`, `lstat` y `dis` (Modelo ->Mínimos cuadrados ordinarios).

`open` carga el fichero y `summary --simple` da los estadísticos principales de las cinco variables que se usan. `series lmedv = log(medv)` crea el logaritmo del precio, como `lprice` en la lección 15. El modelo se estima con `ols`, y el `printf` final imprime la lectura exacta del coeficiente de `rm` con $e^{\beta} - 1$.

o bien teclee en línea de comandos:

```
open ../../data/boston.csv --quiet
```

```
summary medv rm lstat dis chas --simple
series lmedv = log(medv)
```

	Media	Mediana	D. T.	Mín	Máx
medv	22,53	21,20	9,197	5,000	50,00
rm	6,285	6,208	0,7026	3,561	8,780
lstat	12,65	11,36	7,141	1,730	37,97
dis	3,795	3,207	2,106	1,130	12,13
chas	0,06917	0,000	0,2540	0,000	1,000

```
ols lmedv const rm lstat dis
printf "Una habitacion mas por vivienda: %.1f%% aproximado, %.1f%% exacto\n", 100*$coeff(rm), 100*(exp($coeff(rm))-1)
```

Modelo 1: MCO, usando las observaciones 1-506
Variable dependiente: `lmedv`

	coeficiente	Desv. típica	Estadístico t	valor p
const	2,79416	0,142415	19,62	4,86e-64 ***
rm	0,124254	0,0187983	6,610	9,85e-11 ***
lstat	-0,0399364	0,00208609	-19,14	9,62e-62 ***
dis	-0,00928057	0,00570650	-1,626	0,1045

Media de la vble. dep.	3,034513	D.T. de la vble. dep.	0,408757
Suma de cuad. residuos	26,97770	D.T. de la regresión	0,231820
R-cuadrado	0,680270	R-cuadrado corregido	0,678359
F(3, 502)	356,0248	Valor p (de F)	7,4e-124
Log-verosimilitud	23,69317	Criterio de Akaike	-39,38633
Criterio de Schwarz	-22,48018	Crit. de Hannan-Quinn	-32,75577

Log-verosimilitud para `medv` = -1511,77

Sin considerar la constante, el valor p más alto fue el de la variable 8 (`dis`)

Una habitacion mas por vivienda: 12,4% aproximado, 13,2% exacto

Lo que debe observar

El modelo de `lmedv` sobre `rm`, `lstat` y `dis` es log-nivel y se lee como en la lección 15, en tanto por uno. El coeficiente de `rm` es 0,124: a una habitación más de media por vivienda le acompaña un precio mediano mayor en la fracción 0,124, un 12,4% aproximado y un 13,2% exacto, con `lstat` y `dis` fijos. El de `lstat`, -0,040: un punto más de porcentaje de población de estatus bajo se acompaña de un precio un 4% menor. El de `dis` es -0,009 con $t = -1,6$: datos como estos no serían raros si el coeficiente de `dis` valiese cero, es

decir, si con `rm` y `lstat` fijos la distancia a los centros de empleo no se acompañase de ningún cambio del precio. El R^2 es 0,680 y la suma de cuadrados de los residuos, 26,98.

La constante, 2,79, es el valor ajustado del logaritmo del precio en el punto en que `rm`, `lstat` y `dis` valen cero a la vez, y no se interpreta: `rm` va de 3,56 a 8,78 y ningún barrio está cerca de ese punto.

Actividad 2 - La indicadora del río Charles

Añada `chas` al modelo. Su coeficiente es la diferencia, en logaritmos, entre un barrio junto al río y otro sin río con las mismas `rm`, `lstat` y `dis`; la lectura exacta es $e^\beta - 1$. Estime también el mismo modelo con `medv` en niveles para comparar.

Los dos `printf` calculan la diferencia porcentual exacta en los dos sentidos, respecto de los barrios sin río (la categoría base, los que tienen `chas` igual a 0) y respecto de los barrios junto al río, como en la transparencia “En log-nivel: una diferencia porcentual” de la lección 16. El tercer `ols` estima el modelo en niveles.

```
ols lmedv const rm lstat dis chas
printf "Junto al rio respecto de sin rio: %.1f%% exacto (aproximado %.1f%%)\n", 100*(exp($coeff(chas))-1), 100*$coeff(chas)
printf "Sin rio respecto de junto al rio: %.1f%% exacto\n", 100*(exp(-$coeff(chas))-1)
ols medv const rm lstat dis chas --quiet
printf "En niveles: chas = %.2f mil dolares (ee %.2f, t = %.2f)\n", $coeff(chas), $stderr(chas), $coeff(chas)/$stderr(chas)
```

Modelo 2: MCO, usando las observaciones 1-506

Variable dependiente: `lmedv`

	coeficiente	Desv. típica	Estadístico t	valor p
const	2,78966	0,140410	19,87	3,29e-65 ***
rm	0,120332	0,0185598	6,483	2,15e-10 ***
lstat	-0,0394202	0,00206083	-19,13	1,22e-61 ***
dis	-0,00623192	0,00567910	-1,097	0,2730
chas	0,159666	0,0405890	3,934	9,54e-05 ***

Media de la vble. dep.	3,034513	D.T. de la vble. dep.	0,408757
Suma de cuad. residuos	26,16942	D.T. de la regresión	0,228548
R-cuadrado	0,689849	R-cuadrado corregido	0,687373
F(4, 501)	278,5861	Valor p (de F)	7,6e-126
Log-verosimilitud	31,38925	Criterio de Akaike	-52,77850
Criterio de Schwarz	-31,64582	Crit. de Hannan-Quinn	-44,49030

Log-verosimilitud para `medv` = -1504,07

Sin considerar la constante, el valor p más alto fue el de la variable 8 (`dis`)

Junto al rio respecto de sin rio: 17,3% exacto (aproximado 16,0%)

Sin rio respecto de junto al rio: -14,8% exacto

En niveles: `chas` = 3,74 mil dolares (ee 0,96, t = 3,90)

La lección 15 (“Previsión con y sin logaritmos”) dejó dicho cómo comparar el ajuste del modelo en logaritmos con el del modelo en niveles: no con sus R^2 , que miden ajustes a regresandos distintos, sino pasando a dólares las previsiones del modelo en logaritmos y midiendo cuánto se parecen al precio observado. `$yhat` es la serie de valores ajustados del último modelo; `exp($yhat)` deshace el logaritmo y predice la mediana del precio, y multiplicar por $e^{s^2/2}$, con `$sigma` la desviación típica de la regresión, `s`, da la previsión de la esperanza bajo normalidad. La función `corr` calcula la correlación entre dos series; su cuadrado, guardado en el escalar `r2log`, es la medida de ajuste sobre la misma variable, que para el modelo en niveles coincide con su R^2 (`$rsq`).

```
ols lmedv const rm lstat dis chas --quiet
series prevLog = exp($yhat) * exp($sigma^2/2)
printf "Factor de correccion exp(s^2/2) = %.4f\n", exp($sigma^2/2)
```

```
printf "Precio medio observado %.2f ; media de exp(yhat) %.2f ; media de la prevision corregida %.2f\n", mean(medv),
↪ mean(exp($yhat)), mean(prevLog)
scalar r2log = corr(medv, prevLog)^2
ols medv const rm lstat dis chas --quiet
printf "Correlacion al cuadrado con medv: modelo en logaritmos %.3f ; modelo en niveles %.3f (su R2: %.3f)\n", r2log,
↪ corr(medv, $yhat)^2, $rsq
```

Factor de correccion $\exp(s^2/2) = 1,0265$

Precio medio observado 22,53 ; media de $\exp(\hat{y})$ 21,94 ; media de la prevision corregida 22,52

Correlacion al cuadrado con medv: modelo en logaritmos 0,728 ; modelo en niveles 0,657 (su R2: 0,657)

Lo que debe observar

El coeficiente de **chas** es 0,160 con error estándar 0,041 y $t = 3,93$: datos como estos serían muy raros si el coeficiente de **chas** valiese cero, es decir, si con las mismas **rm**, **lstat** y **dis** el precio mediano fuese el mismo junto al río y lejos de él. La diferencia exacta es $e^{0,160} - 1 = 0,173$: a un barrio junto al río le corresponde, en promedio, un precio mediano un 17% mayor que a otro sin río con el mismo número medio de habitaciones, el mismo **lstat** y la misma distancia. En el otro sentido, a un barrio sin río le corresponde un precio un 15% menor que a uno junto al río: los dos porcentajes se calculan sobre precios distintos, y por eso no coinciden.

Tres cautelas de la lección 16. La comparación es entre barrios, no entre viviendas: el modelo no dice cuánto más vale una vivienda por estar junto al río. “Con lo demás fijo” cubre tres variables; la lección 16 obtuvo 6,35 mil dólares como diferencia bruta y 4,08 con **rm** fijo, y aquí, en niveles y con tres regresores fijos, la diferencia es de 3,74 mil dólares, con $t = 3,90$. Y los coeficientes de **rm** y **lstat** apenas cambian al añadir **chas**, y el de **dis** pasa de $-0,009$ a $-0,006$: la indicadora está poco correlada con los tres regresores.

Las previsiones en dólares, como en la lección 15. El factor de corrección es $e^{s^2/2} = 1,027$: la media de $e^{\hat{y}}$ es 21,94 mil dólares, un 2,6% por debajo del precio medio observado, 22,53, y la media de las previsiones corregidas, 22,52, lo reproduce. Sobre la misma variable, el modelo en logaritmos ajusta mejor que el modelo en niveles: la correlación al cuadrado entre **medv** y su previsión es 0,728 frente a 0,657, que es el R^2 del modelo en niveles. Los R^2 de los dos modelos, 0,690 y 0,657, no eran comparables, porque miden el ajuste a regresandos distintos.

Actividad 3 - Una interacción: ¿la misma pendiente de **rm** junto al río?

Construya la interacción **chas*rm** y añádala al modelo. Su coeficiente es la diferencia entre la pendiente de **rm** en los barrios junto al río y en los demás, como en la transparencia “Interacciones: la pendiente cambia de grupo” de la lección 16.

series chasrm = chas*rm crea la columna que vale **rm** en los barrios junto al río y 0 en los demás. El **printf** imprime las dos pendientes.

```
series chasrm = chas*rm
ols lmedv const rm lstat dis chas chasrm
printf "Pendiente de rm sin rio: %.4f ; junto al rio: %.4f ; t de la interaccion: %.2f\n", $coeff(rm),
↪ $coeff(rm)+$coeff(chasrm), $coeff(chasrm)/$stderr(chasrm)
```

Modelo 3: MCO, usando las observaciones 1-506

Variable dependiente: **lmedv**

	coeficiente	Desv. típica	Estadístico t	valor p
const	2,74854	0,143754	19,12	1,43e-61 ***
rm	0,127060	0,0192398	6,604	1,02e-10 ***

lstat	-0,0394219	0,00205934	-19,14	1,11e-61 ***
dis	-0,00649824	0,00567860	-1,144	0,2530
chas	0,563014	0,309549	1,819	0,0695 *
chasrm	-0,0621615	0,0472945	-1,314	0,1893

Media de la vble. dep.	3,034513	D.T. de la vble. dep.	0,408757
Suma de cuad. residuos	26,07931	D.T. de la regresión	0,228383
R-cuadrado	0,690917	R-cuadrado corregido	0,687827
F(5, 500)	223,5380	Valor p (de F)	5,7e-125
Log-verosimilitud	32,26187	Criterio de Akaike	-52,52373
Criterio de Schwarz	-27,16451	Crit. de Hannan-Quinn	-42,57789

Log-verosimilitud para medv = -1503,2

Sin considerar la constante, el valor p más alto fue el de la variable 8 (dis)

Pendiente de rm sin rio: 0,1271 ; junto al rio: 0,0649 ; t de la interaccion: -1,31

Lo que debe observar

El coeficiente de la interacción es $-0,062$ con error estándar $0,047$ y $t = -1,31$: datos como estos no serían raros si la pendiente de **rm** fuese la misma en los dos grupos, y el contraste al 5% no rechaza que el coeficiente de la interacción valga cero. Con este contraste no hay motivo para preferir el modelo con la interacción al de la Actividad 2, que es más sencillo.

Fíjese en lo que ocurre con el coeficiente de **chas**: pasa de $0,160$ a $0,563$ y su error estándar de $0,041$ a $0,310$. Con la interacción en el modelo, ese coeficiente es la diferencia entre grupos en el punto **rm** igual a cero, donde no hay barrios, y por eso se estima con tan poca precisión. Es la misma situación que la de **female*educ** en los apuntes de “Interacciones: la pendiente cambia de grupo” de la lección 16, y el remedio es el mismo que allí: centrar **rm** en su media antes de construir la interacción.

Actividad 4 - El contraste de Chow en Gretl: wage1, como en la lección 16

Abra **wage1** (Archivo → Abrir datos → Archivo de muestra, pestaña Wooldridge) y reproduzca el contraste de la transparencia “El contraste de Chow: ¿la misma ecuación?” de la lección 16: el modelo común es **lwage** sobre **educ**, **exper** y **tenure**, y la pregunta es si hombres y mujeres comparten la ecuación.

La orden **chow female --dummy** se ejecuta después de estimar el modelo común (en la ventana del modelo, **Contrastes** → **Contraste de Chow**, eligiendo la variable ficticia **female**). Gretl estima la regresión aumentada con **female** y sus tres interacciones, la imprime, y calcula el F de la lección 13 con el modelo común como restringido: cuatro restricciones. La orden exige que la indicadora no esté ya en el modelo común: si se estima **lwage** sobre **educ**, **exper**, **tenure** y **female** y después se pide **chow female --dummy**, Gretl responde «El modelo ya contiene female». Para preguntar solo por las pendientes, dejando que cada sexo tenga su intercepto, se estima el modelo con las ocho columnas y se imponen con **restrict** las tres restricciones de las interacciones, como en la [actividad 2 \(html\)](#) de la práctica A de la sesión 20. El último bloque comprueba el caso $q = 1$: si solo puede diferir el intercepto, el F con una restricción es el cuadrado del t de **female**.

o bien teclee en línea de comandos:

```
open wage1 --quiet
```

```
ols lwage const educ exper tenure --quiet
chow female --dummy
```

Regresión aumentada para el contraste de Chow
MCO, usando las observaciones 1-526
Variable dependiente: lwage

	coeficiente	Desv. típica	Estadístico t	valor p	
const	0,321899	0,134627	2,391	0,0172	**
educ	0,0962640	0,00902753	10,66	3,90e-24	***
exper	0,00812704	0,00240969	3,373	0,0008	***
tenure	0,0182129	0,00363843	5,006	7,64e-07	***
female	0,0342173	0,199358	0,1716	0,8638	
fe_educ	-0,0162279	0,0141570	-1,146	0,2522	
fe_exper	-0,00585954	0,00325865	-1,798	0,0727	*
fe_tenure	-0,00793844	0,00654819	-1,212	0,2259	
Media de la vble. dep.	1,623268	D.T. de la vble. dep.	0,531538		
Suma de cuad. residuos	88,58250	D.T. de la regresión	0,413532		
R-cuadrado	0,402800	R-cuadrado corregido	0,394730		
F(7, 518)	49,91164	Valor p (de F)	3,57e-54		
Log-verosimilitud	-277,8622	Criterio de Akaike	571,7243		
Criterio de Schwarz	605,8467	Crit. de Hannan-Quinn	585,0848		

Contraste de Chow de diferencia estructural con respecto a female
F(4, 518) = 18,8193 con valor p 0,0000

series construye las tres interacciones; **restrict** impone que sus coeficientes valgan cero y **\$test** y **\$pvalue** recogen el F y su valor p .

```
series femeduc = female*educ
series femexper = female*exper
series femtenure = female*tenure
ols lwage const educ exper tenure female femeduc femexper femtenure --quiet
restrict --quiet
    b[femeduc] = 0
    b[femexper] = 0
    b[femtenure] = 0
end restrict
printf "Solo las pendientes (q = 3): F(3,%d) = %.2f   valor p = %.3f\n", $df, $test, $pvalue
ols lwage const educ exper tenure female --quiet
printf "Solo el intercepto (q = 1): t de female = %.2f ; t al cuadrado = %.1f\n", $coeff(female)/$stderr(female),
    ↪ ($coeff(female)/$stderr(female))^2
```

Conjunto de restricciones

- 1: b[femeduc] = 0
- 2: b[femexper] = 0
- 3: b[femtenure] = 0

Estadístico de contraste: F(3, 518) = 3,0446, con valor p = 0,0284658

Solo las pendientes (q = 3): F(3,518) = 3,04 valor p = 0,028

Solo el intercepto (q = 1): t de female = -8,09 ; t al cuadrado = 65,4

Lo que debe observar

La orden **chow** devuelve $F(4, 518) = 18,8$ con valor p del orden de 10^{-14} , las cifras de la lección 16: datos como estos serían muy raros si hombres y mujeres compartiesen la ecuación. En la regresión aumentada que imprime, ningún t de **female** ni de sus interacciones supera 2 en valor absoluto, aunque el F conjunto es 18,8: es el caso de la lección 13, un F conjunto alto con t individuales pequeños, por la colinealidad entre la indicadora y sus interacciones que la lección 16 explicó.

Las preguntas una a una sí dicen qué difiere. Solo las pendientes: $F(3, 518) = 3,04$ con valor p 0,028, y el contraste al 5% rechaza que las tres pendientes sean comunes. Solo el intercepto: el t de **female** es $-8,09$ y su cuadrado, 65,4, es el F con una restricción.

Actividad 5 - El contraste de Chow con Boston: ¿la misma ecuación junto al río?

Vuelva al Boston original y repita las tres preguntas con **chas** como indicadora: ¿comparten la ecuación los barrios junto al río y los demás? Si no, ¿difieren las pendientes? ¿Y el intercepto?

El guion vuelve a abrir **boston.csv** porque la Actividad 4 cargó otro fichero, y repite el modelo común. Después, la misma secuencia que con **wage1**: **chow chas --dummy**, el modelo con las ocho columnas con **restrict** sobre las tres interacciones, y el **t** de **chas** al cuadrado.

```
open ../../data/boston.csv --quiet
series lmedv = log(medv)
```

```
ols lmedv const rm lstat dis --quiet
chow chas --dummy
series chasrm = chas*rm
series chaslstat = chas*lstat
series chasdis = chas*dis
ols lmedv const rm lstat dis chas chasrm chaslstat chasdis --quiet
restrict --quiet
    b[chasrm] = 0
    b[chaslstat] = 0
    b[chasdis] = 0
end restrict
printf "Solo las pendientes (q = 3): F(3,%d) = %.2f   valor p = %.3f\n", $df, $test, $pvalue
ols lmedv const rm lstat dis chas --quiet
printf "Solo el intercepto (q = 1): t de chas = %.2f ; t al cuadrado = %.1f\n", $coeff(chas)/$stderr(chas),
↳ ($coeff(chas)/$stderr(chas))^2
```

Regresión aumentada para el contraste de Chow
MCO, usando las observaciones 1-506
Variable dependiente: lmedv

	coeficiente	Desv. típica	Estadístico t	valor p
const	2,75459	0,146450	18,81	5,03e-60 ***
rm	0,126919	0,0194444	6,527	1,65e-10 ***
lstat	-0,0395459	0,00212855	-18,58	6,35e-59 ***
dis	-0,00742667	0,00582157	-1,276	0,2027
chas	0,601964	0,550865	1,093	0,2750
ch_rm	-0,0805410	0,0725791	-1,110	0,2677
ch_lstat	-0,00145377	0,00927645	-0,1567	0,8755
ch_dis	0,0317869	0,0325491	0,9766	0,3292
Media de la vble. dep.	3,034513	D.T. de la vble. dep.	0,408757	
Suma de cuad. residuos	26,02864	D.T. de la regresión	0,228618	
R-cuadrado	0,691518	R-cuadrado corregido	0,687182	
F(7, 498)	159,4795	Valor p (de F)	7,9e-123	
Log-verosimilitud	32,75395	Criterio de Akaike	-49,50791	
Criterio de Schwarz	-15,69561	Crit. de Hannan-Quinn	-36,24678	

Contraste de Chow de diferencia estructural con respecto a chas
F(4, 498) = 4,53957 con valor p 0,0013

Conjunto de restricciones
1: b[chasrm] = 0
2: b[chaslstat] = 0
3: b[chasdis] = 0

Estadístico de contraste: F(3, 498) = 0,897837, con valor p = 0,442109

Solo las pendientes (q = 3): F(3,498) = 0,90 valor p = 0,442
Solo el intercepto (q = 1): t de chas = 3,93 ; t al cuadrado = 15,5

Lo que debe observar

El F conjunto es $F(4, 498) = 4,54$ con valor p 0,0013: datos como estos serían raros si los barrios junto al río y los demás compartiesen la ecuación, y el contraste al 5 % rechaza que la compartan. Como con **wage1**, ningún t de la regresión aumentada supera 2: el de **chas** es 1,09 y los de las interacciones, $-1,11$, $-0,16$ y $0,98$.

Con Boston, las preguntas una a una dan la respuesta contraria a la de **wage1**. Solo las pendientes: $F(3, 498) = 0,90$ con valor p 0,44, y datos como estos no serían raros si las tres pendientes fuesen comunes. Solo el intercepto: el t de **chas** es 3,93 y su cuadrado, 15,5. Con **wage1** el contraste sugería pendientes distintas por sexo; con Boston sugiere la misma ecuación para los dos grupos salvo el intercepto, que es el modelo de la Actividad 2: dos ecuaciones con las mismas pendientes, una 0,160 por encima de la otra en logaritmos. El F conjunto dice que algo difiere; qué difiere lo dicen las preguntas una a una.

Preguntas de interpretación para la clase

1. El coeficiente de **rm** en la Actividad 1 es 0,124. ¿Por qué el porcentaje exacto, 13,2 %, difiere del aproximado, 12,4 %, y cuál de los dos se informa?
2. En la Actividad 2, ¿por qué un barrio junto al río tiene un precio un 17 % mayor y uno sin río un precio un 15 % menor, si el coeficiente es el mismo?
3. ¿Qué compara el coeficiente de **chas** y qué no compara? ¿Qué cambiaría si en lugar de **chas** el fichero trajese la indicadora de los barrios sin río?
4. En la Actividad 3, el coeficiente de **chas** pasa de 0,160 a 0,563 al añadir la interacción. ¿Qué mide en cada modelo y por qué el segundo se estima con tan poca precisión?
5. El contraste de Chow rechaza con **wage1** y con Boston. ¿Por qué la conclusión sobre qué difiere es distinta en los dos casos?
6. En las dos regresiones aumentadas ningún t supera 2 y los dos F conjuntos rechazan. ¿Es una contradicción?
7. Los R^2 de los modelos con **lmedv** y con **medv** de la Actividad 2 son 0,690 y 0,657. ¿Por qué no se pueden comparar, y qué se compara en su lugar?

Para profundizar

- Wooldridge, J. M. (2020), *Introductory Econometrics*, cap. 7, secciones 7.2 a 7.4 (indicadoras, interacciones y el contraste de Chow como F de las interacciones).
- Harrison, D. y Rubinfeld, D. L. (1978), “Hedonic housing prices and the demand for clean air”, *Journal of Environmental Economics and Management* 5, 81–102: el origen de los datos, con **chas** entre los regresores de su ecuación hedónica.
- Manual de Gretl, orden **chow** (las opciones **--dummy** y **--limit-to=lista**, que permite contrastar solo un subconjunto de los regresores).

Código completo de la práctica

```
# -----
# Copyright (C) 2026 Marcos Bujosa
# Licencia: GNU General Public License v3.0 o posterior
# Este código es software libre y puede ser redistribuido y/o modificado bajo los términos de la GPL.
# Ver el archivo LICENSE del repositorio para más detalles.
# -----

# Los dos primeros comandos son necesarios para que Gretl guarde los resultados de la práctica en el
↪ directorio de trabajo
# al ejecutar lo siguiente desde un terminal (use los nombres y ruta que correspondan)
#
#   DIRECTORIO="Nombre_Directorio_trabajo" gretlcli -b ruta/nombre_fichero_de_la_practica.inp
#
# Si esto no le funciona en su sistema, comente las siguientes dos líneas y sitúese en el directorio de
↪ trabajo de gretl
# que corresponda (configure dicho directorio de trabajo desde la ventana principal de Gretl).

string directory = getenv("DIRECTORIO")
set workdir "@directory"

set verbose off

open .././data/boston.csv --quiet
outfile --quiet descripcion.txt
    summary medv rm lstat dis chas --simple
    series lmedv = log(medv)
end outfile

outfile --quiet modeloComun.txt
    ols lmedv const rm lstat dis
    printf "Una habitacion mas por vivienda: %.1f%% aproximado, %.1f%% exacto\n", 100*$coeff(rm),
    ↪ 100*(exp($coeff(rm))-1)
end outfile

outfile --quiet conChas.txt
    ols lmedv const rm lstat dis chas
    printf "Junto al rio respecto de sin rio: %.1f%% exacto (aproximado %.1f%%)\n",
    ↪ 100*(exp($coeff(chas))-1), 100*$coeff(chas)
    printf "Sin rio respecto de junto al rio: %.1f%% exacto\n", 100*(exp(-$coeff(chas))-1)
    ols medv const rm lstat dis chas --quiet
    printf "En niveles: chas = %.2f mil dolares (ee %.2f, t = %.2f)\n", $coeff(chas), $stderr(chas),
    ↪ $coeff(chas)/$stderr(chas)
end outfile

outfile --quiet prevision.txt
    ols lmedv const rm lstat dis chas --quiet
    series prevLog = exp($yhat) * exp($sigma^2/2)
    printf "Factor de correccion exp(s^2/2) = %.4f\n", exp($sigma^2/2)
    printf "Precio medio observado %.2f ; media de exp(yhat) %.2f ; media de la prevision corregida
    ↪ %.2f\n", mean(medv), mean(exp($yhat)), mean(prevLog)
    scalar r2log = corr(medv, prevLog)^2
    ols medv const rm lstat dis chas --quiet
    printf "Correlacion al cuadrado con medv: modelo en logaritmos %.3f ; modelo en niveles %.3f (su R2:
    ↪ %.3f)\n", r2log, corr(medv, $yhat)^2, $rsq
end outfile

outfile --quiet interaccion.txt
    series chasrm = chas*rm
    ols lmedv const rm lstat dis chas chasrm
    printf "Pendiente de rm sin rio: %.4f ; junto al rio: %.4f ; t de la interaccion: %.2f\n", $coeff(rm),
    ↪ $coeff(rm)+$coeff(chasrm), $coeff(chasrm)/$stderr(chasrm)
end outfile

open wage1 --quiet
```

```

outfile --quiet chowWage.txt
    ols lwage const educ exper tenure --quiet
    chow female --dummy
end outfile

outfile --quiet pendientesWage.txt
    series femeduc = female*educ
    series femexper = female*exper
    series femtenure = female*tenure
    ols lwage const educ exper tenure female femeduc femexper femtenure --quiet
    restrict --quiet
        b[femeduc] = 0
        b[femexper] = 0
        b[femtenure] = 0
    end restrict
    printf "Solo las pendientes (q = 3): F(3,%d) = %.2f  valor p = %.3f\n", $df, $test, $pvalue
    ols lwage const educ exper tenure female --quiet
    printf "Solo el intercepto (q = 1): t de female = %.2f ; t al cuadrado = %.1f\n",
        ↪ $coeff(female)/$stderr(female), ($coeff(female)/$stderr(female))^2
end outfile

open ../../data/boston.csv --quiet
series lmedv = log(medv)
outfile --quiet chowBoston.txt
    ols lmedv const rm lstat dis --quiet
    chow chas --dummy
    series chasrm = chas*rm
    series chaslstat = chas*lstat
    series chasdis = chas*dis
    ols lmedv const rm lstat dis chas chasrm chaslstat chasdis --quiet
    restrict --quiet
        b[chasrm] = 0
        b[chaslstat] = 0
        b[chasdis] = 0
    end restrict
    printf "Solo las pendientes (q = 3): F(3,%d) = %.2f  valor p = %.3f\n", $df, $test, $pvalue
    ols lmedv const rm lstat dis chas --quiet
    printf "Solo el intercepto (q = 1): t de chas = %.2f ; t al cuadrado = %.1f\n",
        ↪ $coeff(chas)/$stderr(chas), ($coeff(chas)/$stderr(chas))^2
end outfile

```

Respuestas

1. La lectura aproximada usa $\log(1+p) \approx p$, que vale para p pequeño; 0,124 ya está por encima del 0,1 en que la lección 15 situó el límite. El cálculo exacto deshace el logaritmo: $e^{0,124} - 1 = 0,132$. Se informa el exacto; la aproximación sirve para la lectura rápida.
2. Los dos porcentajes describen el mismo cociente de precios, $e^{0,160} = 1,173$ y su inverso 0,853, calculado sobre bases distintas: el 17% es la diferencia sobre el precio de los barrios sin río, la categoría base, y el 15% la diferencia sobre el precio de los barrios junto al río. Al informar hay que decir respecto de qué grupo se calcula.
3. Compara el precio mediano de dos barrios con el mismo número medio de habitaciones, el mismo porcentaje de población de estatus bajo y la misma distancia a los centros de empleo, uno junto al río y otro no. No compara viviendas, y no dice nada de las variables que no están en el modelo. Con la indicadora contraria el coeficiente sería $-0,160$, la constante subiría en 0,160 y el ajuste sería el mismo: cambia la categoría base, no el modelo.
4. En el modelo sin interacción, 0,160 es la diferencia entre grupos con cualquier valor de **rm**, porque las dos rectas son paralelas. Con la interacción, la diferencia depende de **rm**, $0,563 - 0,062\text{rm}$, y el coeficiente de **chas** es esa diferencia en **rm** igual a cero. Ningún barrio tiene menos de 3,56 habitaciones de media, así que el coeficiente 0,563 es una extrapolación de las dos rectas hasta un punto sin datos, y su error estándar, 0,310, refleja esa falta de datos. Centrar **rm** en su media daría la diferencia en un punto con datos.
5. El F conjunto solo dice que las cuatro restricciones no se cumplen a la vez. Con **wage1**, el F de solo las pendientes rechaza (3,04, valor p 0,028); con Boston no rechaza (0,90, valor p 0,44) y lo que separa a los grupos es el intercepto ($t = 3,93$). Son dos conjuntos de datos distintos, y el contraste conjunto no distingue entre pendientes distintas e interceptos distintos: hay que preguntar una cosa cada vez.
6. No. Cada t pregunta por un coeficiente con los otros tres en el modelo, y la indicadora y sus interacciones comparten casi toda su variación (en **wage1**, la correlación entre **female** y **female*educ** es 0,96). Cada una por separado aporta poco al ajuste, y entre las cuatro aportan mucho: es el caso de la lección 13, un F conjunto alto con t individuales pequeños, explicado por la colinealidad de la lección 14.
7. Miden el ajuste a regresandos distintos, el logaritmo del precio y el precio en miles de dólares, cada uno con su propia variación total. Para comparar hay que medir sobre la misma variable: las previsiones del modelo en logaritmos se pasan a dólares con la corrección $e^{s^2/2}$ y se calcula el cuadrado de su correlación con **medv**, 0,728, frente al R^2 del modelo en niveles, 0,657, que es esa misma medida para ese modelo. La corrección no cambia la correlación, porque multiplicar por una constante no la altera; lo que cambia es la media de las previsiones, que pasa a coincidir con la observada.