

Maestría en Econometría - UTDT

Examen Final - Microeconometría I

PRIMERA PARTE: Logit, Probit, Tobit y Heckman.

Hacer los ejercicios 1 a 8 del capítulo 11 del libro “The Practice of Econometrics” de Ernst R. Berndt.

Ejercicio 1: Inspecting Mroz’s 1975 Panel Study of Income Dynamics Data.

El propósito de este ejercicio es ayudarlo a familiarizarse con las características destacadas de la serie de datos en el archivo de datos MROZ. Mientras explora este conjunto de datos, calculará medias aritméticas, desviaciones estándar y valores mínimos y valores máximos para las variables en toda la muestra y para los datos ordenados en varios subgrupos. También construirá y guardará dos variables que se utilizarán en los ejercicios posteriores de este capítulo.

(a) Para comprobar si los datos son los mismos que los empleados por Mroz [1987, Tabla III, pág. 769], calcular e imprimir las medias aritméticas y desviaciones estándar de cada una de las 19 variables en los datos del archivo MROZ, utilizando la muestra completa de 753 observaciones. Los resultados que usted obtener de dicho cálculo debe ser igual (para cada variable nombrada, seguido de paréntesis que encierra primero la media y, luego, la desviación estándar): LFP (0.56839, 0.49563), WHRS (740.57636, 871.31422), KL6 (0.23772, 0.52396), K618 (1.35325, 1.31987), WA (42.53785, 8.07257), WE (12.28685, 2.28025), WW (2.37457, 3.24183), RPWG (1.84973, 2.41989), HHRS (2267.27092, 595.56665), HA (45.12085, 8.05879), HE (12.49137, 3.02080), HW (7.48218, 4.23056), FAMINC (23080.59495, 12190.20203), MTR (0.67886, 0.08350), WMED (9.25100, 3.36747), WFED (8.80876, 3.57229), UN (8.62351, 3.11493), CIT (0.64276, 0.47950) y AX (10,63081, 8,06913). ¿Tus resultados coinciden con los de Mroz? (Debido a las diferentes condiciones de redondeo convenciones, el programa de software de su computadora puede generar valores ligeramente diferentes. Pero las medias y las desviaciones estándar deberían ser muy cercanas a las reportadas aquí.) También calcule los valores mínimo y máximo para cada una de estas variables; esta es una práctica, particularmente, útil, ya que, al hacer esto, a menudo, se pueden detectar errores en la codificación de datos. ¿Alguno sus valores mínimo-máximo son “sospechosos”? ¿Por qué o por qué no?

En la tabla 1, se presentan estadísticas descriptivas (media, desviación estándar, mínimo y máximo) de las 19 variables de la base de datos MROZ. Por un lado, se puede observar que todas ellas coinciden con los de Mroz (1987). Por otra parte, se observa que ninguno de los valores mínimo-máximo son “sospechosos”, ya que todos se encuentran en valores razonables.

Tabla 1. Estadísticas descriptivas de todas las variables.

Variable	Obs.	Media	Desvío estándar	Mínimo	Máximo
lfp	753	0,568	0,496	0	1
whrs	753	740,576	871,314	0	4.950
kl6	753	0,238	0,524	0	3
k618	753	1,353	1,320	0	8
wa	753	42,538	8,073	30	60
we	753	12,287	2,280	5	17
ww	753	2,375	3,242	0	25
rpwg	753	1,850	2,420	0	9,98
hhrs	753	2.267,271	595,567	175	5.010
ha	753	45,121	8,059	30	60
he	753	12,491	3,021	3	17
hw	753	7,482	4,231	0,412	40,509
faminc	753	23.080,59	12.190,2	1500	96.000
mtr	753	0,679	0,083	0,442	0,942
wmed	753	9,251	3,367	0	17
wfed	753	8,809	3,572	0	17
un	753	8,624	3,115	3	14
cit	753	0,643	0,48	0	1
ax	753	10,631	8,069	0	45

Fuente: Elaboración propia en base a Mroz (1987).

(b) Ahora, compare la muestra de mujeres trabajadoras (las primeras 428 observaciones en el archivo de datos de MROZ) con las que no trabajaban por un salario en 1975 (las 325 observaciones finales). Calcule las medias aritméticas y las desviaciones estándar para cada una de las 19 variables en estas dos submuestras, luego imprímalas y compárelas. Mroz señala que las medias aritméticas en las muestras activas y no activas son bastante similares para variables como WA, WE, K618, HA, HE y HHRS. ¿Consideras que éste también es el caso? Sin embargo, las medias en las muestras que trabajan y no trabajan tienden a diferir más para variables como KL6 y HW. ¿En qué se diferencian y qué podría implicar esto con respecto a los salarios de reserva de las mujeres en estas dos muestras? ¿Por qué? En la medida en que la experiencia previa de las mujeres en el mercado laboral (AX) refleje sus preferencias o gustos por el trabajo en el mercado, se podría esperar que las medias de AX difieran entre las muestras que trabajan y las que no trabajan. ¿Es éste el caso? Interpreta esta diferencia. ¿Existen otras diferencias en las medias de las variables de las dos submuestras que podrían afectar la participación en la fuerza laboral o las horas trabajadas? Si es así, coméntelos.

En las tablas 2 y 3, se presentan estadísticas descriptivas (media, desviación estándar, mínimo y máximo) de las 19 variables de la base de datos MROZ para la muestra de las mujeres que trabajaron y la de las que no trabajaron, respectivamente. Por un lado, se puede observar que, efectivamente, las medias aritméticas en las muestras activas y no activas son bastante similares para las variables wa, we, k618, ha, he y hhrrs. Sin embargo, las medias de las variables kl6 y hw difieren entre las dos muestras, siendo menor para ambas variables en el caso de la muestra de las mujeres que no trabajaron. Dicho en otras palabras, las mujeres que no participan en el mercado laboral tienen, en promedio, más niños menores de 6 años en el hogar y el promedio de las ganancias medias por hora del

esposo es mayor, por lo cual se puede pensar que los salarios de reserva de estas mujeres son mayores a los de las mujeres que trabajaron porque el costo de oportunidad de las primeras es mayor. Por otra parte, las medias de la variable ax difieren entre las dos muestras, siendo menor en el caso de la muestra de las mujeres que no trabajaron, reflejando, eventualmente, sus menores preferencias por el trabajo en el mercado laboral.

Tabla 2. Estadísticas descriptivas de todas las variables (observaciones con $lfp=1$).

Variable	Obs.	Media	Desvío estándar	Mínimo	Máximo
lfp	428	1	0	1	1
$whrs$	428	1.302,93	776,274	12	4.950
$kl6$	428	0,140	0,392	0	2
$k618$	428	1,350	1,316	0	8
wa	428	41,972	7,721	30	60
we	428	12,659	2,285	5	17
ww	428	4,178	3,310	0,128	25
$rpwg$	428	3,186	2,440	0	9,980
$hhrs$	428	2.233,465	582,909	175	5.010
ha	428	44,610	7,950	30	60
he	428	12,612	3,0315	4	17
hw	428	7,226	3,571	0,513	26,578
$faminc$	428	24.130,42	11.671,26	2.400	91.044
mtr	428	0,668	0,077	0,442	0,942
$wmed$	428	9,516	3,308	0	17
$wfed$	428	8,988	3,523	0	17
un	428	8,546	3,033	3	14
cit	428	0,640	0,481	0	1
ax	428	13,037	8,056	0	38

Fuente: Elaboración propia en base a Mroz (1987).

Tabla 3. Estadísticas descriptivas de todas las variables (observaciones con $lfp = 0$).

Variable	Obs.	Media	Desvío estándar	Mínimo	Máximo
lfp	325	0	0	0	0
whrs	325	0	0	0	0
kl6	325	0,366	0,637	0	3
k618	325	1,357	1,327	0	7
wa	325	43,283	8,468	30	60
we	325	11,797	2,182	5	17
ww	325	0	0	0	0
rpwg	325	0,09	0,532	0	4,8
hhrs	325	2.311,791	609,882	640	4.640
ha	325	45,794	8,163	30	60
he	325	12,332	2,999	3	17
hw	325	7,819	4,953	0,412	40,509
faminc	325	21.698,05	12.728,15	1.500	96.000
mtr	325	0,693	0,09	0,442	0,942
wmed	325	8,902	3,418	0	17
wfed	325	8,572	3,628	0	17
un	325	8,726	3,221	3	14
cit	325	0,646	0,479	0	1
ax	325	7,462	6,919	0	45

Fuente: Elaboración propia en base a Mroz (1987).

(c) En el modelo de oferta de trabajo estimado por Mroz, se supone que, al tomar sus decisiones sobre la oferta de trabajo, la esposa toma como dados todos los ingresos no laborales del hogar más los ingresos laborales de su marido. Mroz a esta suma la llama ingreso de propiedad de la esposa y la calcula como ingreso familiar total menos el ingreso laboral obtenido por la esposa. Para la muestra completa de 753 observaciones, calcule esta variable de ingreso de propiedad (llamada, por ejemplo, PRIN) como $PRIN = FAMINC - (WHRS * WW)$. También calcule e imprima su media y desviación estándar; estos deberían ser iguales a 20.129 y 11.635, respectivamente. Guarde PRIN para utilizarla en ejercicios posteriores de este capítulo.

En la tabla 4, se presentan estadísticas descriptivas de la variable *prin* construida.

Tabla 4. Estadísticas descriptivas de la variable *prin*.

Variable	Obs.	Media	Desvío estándar	Mínimo	Máximo
prin	753	20.128,96	11.634,8	-29,057	96.000

Fuente: Elaboración propia en base a Mroz (1987).

(d) Una de las variables que, a menudo, se emplea en los análisis empíricos de la participación en la fuerza laboral es la tasa salarial. Sin embargo, como se enfatizó anteriormente en este capítulo, la tasa salarial, generalmente, no se observa para las mujeres que no están trabajando. Algunos analistas han intentado abordar este problema (aunque de manera insatisfactoria; consultar la Sección 11.3.B.1) estimando una

ecuación de determinación de salarios utilizando datos sobre los trabajadores únicamente y, luego, utilizando las estimaciones de los parámetros resultantes y las características de la muestra no trabajadora para construir salarios ajustados o previstos para cada uno de los no trabajadores.

*Restringir su muestra a trabajadores (las primeras 428 observaciones), tome el logaritmo natural de la variable de tasa salarial de la esposa WW y calcule esta variable transformada logarítmicamente LWW. Calcular e imprimir la media y desviación estándar de LWW para esta muestra. Luego, para toda la muestra de 753 observaciones, construya el cuadrado de la variable experiencia de la esposa y llamarla AX2, es decir, generar $AX2 = AX * AX$ (y, para más adelante, el cuadrado de la edad de la esposa, $WA2 = WA * WA$). A continuación, siguiendo la literatura sobre capital humano sobre la determinación de salarios que se resume en el capítulo 5 de este libro y utilizando sólo las 428 observaciones del trabajo muestra, estime por MCO una ecuación típica de determinación de salarios, en la cual LWW se regresa en un término constante, WA, WE, CIT, AX y AX2. ¿Tiene sentido esta ecuación? ¿Por qué o por qué no? Luego, usa las estimaciones de parámetros de esta ecuación y los valores de las variables WA, WE, CIT, AX y AX2 para las 325 mujeres de la muestra que no trabajan para generar el salario logarítmico previsto o ajustado para los no trabajadores. Llama a esta variable log-salario ajustada para los que no son trabajadores FLWW. Calcular la media aritmética y la desviación estándar de la variable FLWW para los no trabajadores, y compárelos con los de LWW de la muestra de trabajo. ¿La diferencia en las medias es sustancial? ¿Cómo interpreta este resultado? Finalmente, para toda la muestra de 753 observaciones y para su uso en ejercicios posteriores en los puntos de este capítulo, genere una variable llamada LWW1 para la cual las primeras 428 observaciones (la muestra de trabajo) $LWW1 = LWW$ y para las cuales las últimas 325 observaciones (la muestra que no trabaja) $LWW1 = FLWW$, desde arriba. Tenga en cuenta que su variable LWW1 construida debe incluir ya sea el salario real o el previsto para cada individuo de la muestra. Para asegurarse de haber calculado correctamente la serie de datos LWW1, calcular e imprimir su media y desviación estándar; ellos deberían iguales a 1,10432 y 0,58268, respectivamente. Guarde la serie de datos LWW1 para utilizar en ejercicios posteriores de este capítulo.*

En la tabla 5, se presenta la estimación por MCO de una ecuación de salarios, en la cual *lww* se regresa en un término constante, *wa*, *we*, *cit*, *ax* y *ax2*. En la tabla 6, se presentan estadísticas descriptivas de las variables *flww* (sólo para la muestra de las mujeres que no trabajaron) y *lww* (sólo para la muestra de las mujeres que trabajaron). Se puede observar que la diferencia en las medias es importante, siendo alrededor de 22% mayor para el caso de las mujeres que trabajaron.

Tabla 5. *Estimación por MCO de ecuación de salarios.*

Source	SS	df	MS	Number of obs	=	428
				F(5, 422)	=	15.85
Model	35.3048763	5	7.06097526	Prob > F	=	0.0000
Residual	188.022565	422	.445551101	R-squared	=	0.1581
				Adj R-squared	=	0.1481
Total	223.327441	427	.523015084	Root MSE	=	.6675

lww	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
wa	-.0001945	.0048945	-0.04	0.968	-.00098152	.0094263
we	.105676	.0143701	7.35	0.000	.0774302	.1339218
cit	.0545581	.0686923	0.79	0.428	-.0804637	.1895799
ax	.0410581	.0132119	3.11	0.002	.0150888	.0670275
ax2	-.0007945	.0004006	-1.98	0.048	-.0015819	-7.14e-06
_cons	-.5231308	.2782635	-1.88	0.061	-1.070086	.0238243

Fuente: Elaboración propia en base a Mroz (1987).

Tabla 6. *Estadísticas descriptivas (variables flww y lww).*

Variable	Obs.	Media	Desvío estándar	Mínimo	Máximo
flww	325	0,975	0,292	-0,006	1,758
lww	428	1,19	0,723	-2,054	3,219

Fuente: Elaboración propia.

Ejercicio 2: Estimating the Hours Worked Equation Using Procedure I.

El propósito de este ejercicio es presentarle un procedimiento común, pero lamentablemente inapropiado, para estimar la ecuación de horas trabajadas. Específicamente, en este ejercicio, se implementa el Procedimiento I, en el que se estima mediante MCO una ecuación de horas trabajadas utilizando la muestra completa de 753 observaciones y salarios previstos para los no trabajadores y se fijan las horas trabajadas para los no trabajadores en cero. Esta ecuación, a menudo, se denomina ecuación de regresión normal truncada. También se calculan las respuestas implícitas a los cambios en la tasa de consumo y en el ingreso de la propiedad, tanto en nivel como en forma de elasticidad.

(a) *Inspeccione la variable WHRS (horas trabajadas de la esposa) y verifique que, siempre que la variable LFP (participación en la fuerza laboral) sea igual a cero, el valor de WHRS también sea cero. Luego, utilizando el procedimiento de estimación MCO y el conjunto completo de 753 observaciones en el archivo de datos MROZ, haga una regresión WHRS en un término constante y en las variables KL6, K618, WA, WE, LWWI (construido en la parte (d) del Ejercicio 1) y PRIN (construido en la parte (c) del Ejercicio 1). ¿Los signos de los parámetros estimados concuerdan con su intuición? ¿Por qué o por qué no? ¿Cuál es el valor de R^2 ? ¿Por qué este valor podría ser tan bajo cuando se emplea el Procedimiento I?*

En la tabla 7, se presenta la estimación por MCO (no condicional, al no restringir la muestra a las mujeres que trabajaron) de una ecuación de horas trabajadas, en la cual *whrs* se regresa en un término constante, *kl6*, *k618*, *wa*, *we*, *lww1* y *prin*. Por un lado, se puede observar que los signos de todos los parámetros estimados concuerdan con lo intuitivo, ya que es esperable que la presencia de niños en el hogar, la edad y el ingreso de propiedad disminuyan la cantidad de horas trabajadas, mientras que la educación y el salario la aumenten.

Por otra parte, se observa que el R^2 es igual a 0,1225, el cual puede ser relativamente bajo porque, cuando se emplea el Procedimiento I, las horas trabajadas de las mujeres que no trabajaron se fijan en cero, lo cual puede llevar a una pérdida de información y a una reducción en la variabilidad explicada por el modelo, lo que resulta en un R^2 más bajo en comparación con otros métodos de estimación que no imponen esta fijación.

Tabla 7. Estimación por MCO (no condicional) de ecuación de horas trabajadas.

Source	SS	df	MS	Number of obs	=	753
Model	69918508.6	6	11653084.8	F(6, 746)	=	17.35
Residual	500991215	746	671569.994	Prob > F	=	0.0000
Total	570909724	752	759188.463	R-squared	=	0.1225
				Adj R-squared	=	0.1154
				Root MSE	=	819.49

whrs	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
kl6	-497.7525	63.93751	-7.78	0.000	-623.2714	-372.2337
k618	-79.16349	24.97937	-3.17	0.002	-128.2017	-30.12527
wa	-20.31185	4.535243	-4.48	0.000	-29.2152	-11.40849
we	43.65517	15.42005	2.83	0.005	13.38332	73.92703
lww1	111.0774	57.20935	1.94	0.053	-1.233081	223.3879
prin	-.0104968	.0027022	-3.88	0.000	-.0158016	-.005192
_cons	1383.091	287.1968	4.82	0.000	819.2806	1946.901

Fuente: Elaboración propia en base a Mroz (1987).

(b) Utilizando las estimaciones de parámetros MCO anteriores y las fórmulas de elasticidad debajo de la ecuación (11.48), evaluadas en los mismos puntos que se indican al final de la tabla 11.2, calcula la elasticidad de las horas trabajadas con respecto a los salarios y con respecto al ingreso de la propiedad. ¿La elasticidad salarial es compensada o no compensada? ¿Por qué? Luego, siguiendo los procedimientos de Mroz, que también se analizan en la ecuación (11.48), calcule la respuesta implícita de las horas trabajadas ante un cambio de 1 dólar en el salario, evaluada en el mismo punto. ¿Cómo se compara esta estimación con las presentadas en la tabla 11.2? Finalmente, calcule la respuesta implícita de las horas trabajadas ante un aumento de \$1.000 en el ingreso de la propiedad. Comente cómo se compara esta estimación con las presentadas en la Tabla 11.2.

Por un lado, la elasticidad de las horas trabajadas con respecto a los salarios y con respecto al ingreso de la propiedad, evaluadas en los puntos mencionados, son 0,074 y -0,007, respectivamente. Por otro lado, la respuesta implícita de las horas trabajadas ante un cambio de 1 dólar en el salario, evaluada en el mismo punto anterior, es 24,684. Por último, la respuesta implícita de las horas trabajadas ante un aumento de \$1.000 en el ingreso de la propiedad es -10,497. Estas dos últimas estimaciones son relativamente pequeñas en valor absoluto respecto a las presentadas en la tabla 11.2 mencionada.

(c) Aunque este Procedimiento I es simple y fácil de implementar, tiene varios defectos graves. ¿Cuáles son las principales deficiencias?

Las principales deficiencias de este Procedimiento I son:

- **Sesgo de selección:** Al excluir a los no trabajadores de la muestra y fijar sus horas trabajadas en cero, se introduce un sesgo de selección en la estimación. Esto se debe a que los no trabajadores pueden tener características diferentes a los

trabajadores y, al ignorarlos, se pierde información importante sobre los determinantes de las horas trabajadas.

- Endogeneidad: Existe la posibilidad de que la participación en la fuerza laboral (LFP) esté, endógenamente, relacionada con las horas trabajadas y otras variables explicativas en el modelo. Al no abordar, adecuadamente, la endogeneidad, los coeficientes estimados pueden estar sesgados y ser inconsistentes.
- Pérdida de eficiencia: Fijar las horas trabajadas de los no trabajadores en cero reduce la variabilidad en los datos y puede llevar a una pérdida de eficiencia en la estimación de los parámetros del modelo. Esto significa que los estimadores MCO pueden no ser los más eficientes en presencia de este tipo de truncamiento en los datos.
- No considera la elección discreta: Al fijar las horas trabajadas de los no trabajadores en cero, se está asumiendo implícitamente que esta decisión es exógena y que los individuos no tienen preferencias sobre trabajar o no trabajar. Sin embargo, en la realidad, la elección de participar en la fuerza laboral y la cantidad de horas trabajadas son decisiones discretas que pueden depender de factores individuales y contextuales.

Ejercicio 3: Comparing OLS, Probit and Logit Estimates of the Labor Force Participation Decision.

El objetivo de este ejercicio es ayudarlo a adquirir experiencia con técnicas simples de estimación de variables dependientes limitadas, calculando y, luego, comparando estimaciones MCO, Probit y Logit de una ecuación típica de participación en la fuerza laboral. La comparación numérica de estos diversos estimadores se basa, en gran medida, en el trabajo de Takeshi Amemiya [1981], quien ha derivado relaciones entre ellos.

(a) Los libros de texto de econometría, típicamente, señalan que, si la variable dependiente en una ecuación es una variable ficticia dicotómica y si se estima una ecuación mediante MCO en la que esta variable dependiente está relacionada, linealmente, con un término de intercepción, varios regresores y un término de error estocástico, entonces, la ecuación resultante (a menudo, llamada modelo de probabilidad lineal) adolece de, al menos, dos defectos: (1) los valores ajustados no están confinados al intervalo 0-1, por lo que su interpretación como probabilidades es inapropiado; y (2) los residuos de dicha ecuación son heterocedásticos. Tenga en cuenta que, en nuestro contexto, la LFP es una variable dependiente dicotómica.

Utilizando los procedimientos de estimación MCO y el archivo de datos MROZ para las 753 observaciones, estime los parámetros de un modelo de probabilidad lineal en el que LFP se relaciona, linealmente, con un término de intersección, la variable LWW1 construida en la parte (d) del Ejercicio 1, KL6, K618, WA, WE, UN, CIT, el ingreso de propiedad de la esposa PRIN (construido en la parte (c) del Ejercicio 1) y un término de error estocástico. ¿Tienen sentido los signos de estos parámetros estimados por MCO? ¿Por qué o por qué no? Comente sobre la conveniencia de utilizar los errores estándar estimados por MCO para realizar pruebas de significación estadística. Luego, recupere e imprima los valores ajustados de este modelo de probabilidad lineal estimado. ¿Para cuántas observaciones los valores ajustados son negativos? ¿Para cuántos son mayores que 1? ¿Por qué esto complica la interpretación de este modelo? ¿Qué es R^2 en este modelo? ¿Tiene alguna interpretación útil? ¿Por qué o por qué no?

En la tabla 8, se presenta la estimación por MCO de una ecuación de participación en la fuerza laboral, en la cual *lfp* se regresa en un término constante, *lww1*, *kl6*, *k618*, *wa*, *we*, *un*, *cit* y *prin*. Por un lado, se puede observar que los signos de todos los parámetros estimados por MCO tienen sentido, ya que es esperable que la presencia de niños en el hogar, la edad, la tasa de desempleo del país y el ingreso de propiedad disminuyan la participación en la fuerza laboral, mientras que el salario y la educación la aumenten; el único signo que, eventualmente, se esperaría distinto es el de la *dummy* indicativa de vivir en una gran ciudad, el cual se esperaría sea positivo.

Por otra parte, no es conveniente utilizar los errores estándar estimados por MCO para realizar pruebas de significación estadística, ya que los errores del modelo son heterocedásticos y, en presencia de heterocedasticidad, estos errores estándar estimados pueden estar sesgados y conducir a pruebas de significación incorrectas.

Por último, los valores ajustados son negativos para 7 observaciones y son mayores que 1 para 9 observaciones, lo cual complica la interpretación de este modelo porque no es posible hablar de probabilidades negativos ni mayores a 1.

Tabla 8. Estimación por MCO de ecuación de participación en la fuerza laboral.

Source	SS	df	MS	Number of obs	=	753
Model	28.9458948	8	3.61823684	F(8, 744)	=	17.28
Residual	155.781861	744	.209384222	Prob > F	=	0.0000
				R-squared	=	0.1567
				Adj R-squared	=	0.1476
Total	184.727756	752	.245648611	Root MSE	=	.45759

lfp	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
lww1	.0932754	.0319824	2.92	0.004	.0304888	.1560619
kl6	-.2906401	.0357228	-8.14	0.000	-.3607696	-.2205106
k618	-.0083879	.0139698	-0.60	0.548	-.0358129	.019037
wa	-.0116091	.0025492	-4.55	0.000	-.0166136	-.0066046
we	.0420989	.0086508	4.87	0.000	.025116	.0590817
un	-.003487	.0054787	-0.64	0.525	-.0142426	.0072686
cit	-.004477	.0367137	-0.12	0.903	-.0765518	.0675978
prin	-6.77e-06	1.54e-06	-4.40	0.000	-9.79e-06	-3.75e-06
_cons	.6922918	.162686	4.26	0.000	.3729135	1.01167

Fuente: Elaboración propia en base a Mroz (1987).

(b) *Un posible procedimiento de estimación más apropiado cuando la variable dependiente es dicotómica se basa en el supuesto de que la distribución acumulativa de las perturbaciones estocásticas es la logística; el estimador de máxima verosimilitud resultante suele denominarse logit.*

Con LFP como variable dependiente y con un término constante, LWWI (del inciso (d) del Ejercicio 1), KL6, K618, WA, WE, UN, CIT y PRIN (del inciso (c) del Ejercicio 1) como variables explicativas, utilice la muestra completa de 753 observaciones en el archivo de datos MROZ y estime los parámetros basándose en un procedimiento logit de máxima verosimilitud. ¿Tienen sentido los signos de estos parámetros logit estimados? ¿Por qué o por qué no? ¿Cuál de los parámetros logit estimados es significativamente diferente de cero? Interpretar. ¿El algoritmo computacional logit no lineal de su programa de computadora alcanza la convergencia rápidamente, después de sólo unas pocas iteraciones (digamos, menos de cinco)? Algunos programas informáticos proporcionan resultados de “bondad de ajuste” para el modelo logit estimado, como una medida pseudo-R² o una medida que indica qué porcentaje de las predicciones son “correctas”. Consulte los resultados y el manual de su computadora para conocer la interpretación de dichas medidas de bondad de ajuste. Finalmente, compare sus estimaciones logit con las estimaciones de MCO o del modelo de probabilidad lineal del inciso (a). En particular, siguiendo a Takeshi Amemiya [1981], cada una de las estimaciones de los parámetros de pendiente MCO debería ser aproximadamente igual a 0,25 veces la estimación del parámetro de pendiente logit correspondiente. ¿Qué tan bien le va al trabajo de aproximación de Amemiya en esta muestra? Además, Amemiya muestra que cada uno de los términos del intercepto de MCO y de intersección de la variable ficticia de MCO debe ser, aproximadamente, igual a 0,25 veces la estimación del parámetro logit correspondiente, más 0,5. ¿Se corresponden bien sus términos

intercepto MCO y logit y la intersección de la variable ficticia con la aproximación de Amemiya?

En la tabla 9, se presenta la estimación Logit de una ecuación de participación en la fuerza laboral, en la cual *lfp* se regresa en un término constante, *lww1*, *kl6*, *k618*, *wa*, *we*, *un*, *cit* y *prin*. Por un lado, se puede observar que los signos de estos parámetros logit estimados tienen sentido, al igual que sucedía en la estimación por MCO; la diferencia se encuentra en que, ahora, el signo de la *dummy* indicativa de vivir en una gran ciudad es positivo. Por otro lado, los parámetros de *lww1*, *kl6*, *wa*, *we* y *prin* son los parámetros logit estimados que son significativamente diferentes de cero. Por último, el algoritmo computacional logit no lineal que usa Stata (por *default*, el algoritmo de Newton-Raphson) alcanza la convergencia rápidamente, después de sólo 4 iteraciones.

Tabla 9. Estimación Logit de ecuación de participación en la fuerza laboral.

Logistic regression				Number of obs = 753		
				LR chi2(8) = 130.80		
				Prob > chi2 = 0.0000		
Log likelihood = -449.47564				Pseudo R2 = 0.1270		

	lfp	Coefficient	Std. err.	z	P> z	[95% conf. interval]

	lw1	.4644233	.1572362	2.95	0.003	.1562461 .7726006
	kl6	-1.469201	.1985189	-7.40	0.000	-1.858291 -1.080112
	k618	-.0512143	.0684263	-0.75	0.454	-.1853275 .0828989
	wa	-.0582569	.0129069	-4.51	0.000	-.083554 -.0329599
	we	.2119735	.0441121	4.81	0.000	.1255153 .2984317
	un	-.0185696	.0264618	-0.70	0.483	-.0704337 .0332945
	cit	.0127482	.1781669	0.07	0.943	-.3364525 .361949
	prin	-.0000353	8.11e-06	-4.36	0.000	-.0000512 -.0000194
	_cons	.9510144	.8042517	1.18	0.237	-.6252899 2.527319

Fuente: Elaboración propia en base a Mroz (1987).

Además, como se muestra en la tabla 10, luego de estimar, es posible clasificar a las estimaciones y obtener la sensibilidad y la especificidad del modelo, como así también el porcentaje de observaciones correctamente clasificadas (67,73%, en este caso). Finalmente, se puede observar que la aproximación de Amemiya (1981) se corresponde, relativamente, bien para esta muestra (ver cálculos en *do-file*).

Tabla 10. Clasificación post estimación Logit.

Logistic model for lfp

Classified	True		Total
	D	~D	
+	346	161	507
-	82	164	246
Total	428	325	753

Classified + if predicted $\Pr(D) \geq .5$

True D defined as lfp != 0

Sensitivity	$\Pr(+ D)$	80.84%
Specificity	$\Pr(- \sim D)$	50.46%
Positive predictive value	$\Pr(D +)$	68.24%
Negative predictive value	$\Pr(\sim D -)$	66.67%
False + rate for true ~D	$\Pr(+ \sim D)$	49.54%
False - rate for true D	$\Pr(- D)$	19.16%
False + rate for classified +	$\Pr(\sim D +)$	31.76%
False - rate for classified -	$\Pr(D -)$	33.33%
Correctly classified		67.73%

Fuente: Elaboración propia en base a Mroz (1987).

Tabla 11. Estimación por MCO versus Estimación Logit.

	(1) OLS	(2) Logit
main		
lw1	0.0933*** (0.0320)	0.464*** (0.157)
kl6	-0.291*** (0.0357)	-1.469*** (0.199)
k618	-0.00839 (0.0140)	-0.0512 (0.0684)
wa	-0.0116*** (0.00255)	-0.0583*** (0.0129)
we	0.0421*** (0.00865)	0.212*** (0.0441)
un	-0.00349 (0.00548)	-0.0186 (0.0265)
cit	-0.00448 (0.0367)	0.0127 (0.178)
prin	-0.00000677*** (0.00000154)	-0.0000353*** (0.00000811)
_cons	0.692*** (0.163)	0.951 (0.804)
N	753	753
R-sq	0.157	
pseudo R-sq		0.127

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Fuente: Elaboración propia en base a Mroz (1987).

(c) Otro procedimiento de estimación común que se utiliza en la estimación de modelos de variables dependientes dicotómicas se basa en el supuesto de que la distribución acumulativa de las perturbaciones es normal; esto suele denominarse modelo probit. Las funciones de probabilidad basadas en el modelo probit se analizan en la Sección 11.3.B.1 de este capítulo. Con LFP como variable dependiente y con un término constante, LWWI (del inciso (d) del Ejercicio 1), KL6, K618, WA, WE, UN, CIT y PRIN (del inciso (c) del Ejercicio 1) como variables explicativas, utilice la muestra completa de 753 observaciones en el archivo de datos MROZ y estime los parámetros basándose en un procedimiento probit de máxima verosimilitud. ¿Tienen sentido los signos de estos parámetros probit estimados? ¿Por qué? ¿Cuál de los parámetros probit estimados es significativamente diferente de cero? ¿Por qué? ¿El algoritmo computacional probit de su programa de computadora alcanza la convergencia rápidamente, después de sólo unas pocas iteraciones? ¿Es la convergencia más o menos rápida que con el modelo logit? Como en el inciso (b), interprete cualquier medida de bondad de ajuste que proporcione su programa de software.

En la tabla 12, se presenta la estimación Probit de una ecuación de participación en la fuerza laboral, en la cual *lfp* se regresa en un término constante, *lww1*, *kl6*, *k618*, *wa*, *we*, *un*, *cit* y *prin*. Por un lado, se puede observar que los signos de estos parámetros logit estimados tienen sentido (al igual que sucedía en la estimación Logit). Por otro lado, los parámetros de *lww1*, *kl6*, *wa*, *we* y *prin* son los parámetros probit estimados que son significativamente diferentes de cero (al igual que sucedía en la estimación Logit). Por último, el algoritmo computacional logit no lineal que usa Stata (por *default*, el algoritmo de Newton-Raphson) alcanza la convergencia rápidamente, después de sólo 4 iteraciones (al igual que sucedía en la estimación Logit).

Tabla 12. Estimación Probit de ecuación de participación en la fuerza laboral.

Probit regression					Number of obs = 753		
					LR chi2(8) = 130.95		
					Prob > chi2 = 0.0000		
Log likelihood = -449.39696					Pseudo R2 = 0.1272		

	lfp	Coefficient	Std. err.	z	P> z	[95% conf. interval]	

	lww1	.2820424	.092771	3.04	0.002	.1002145	.4638703
	kl6	-.8808229	.1146558	-7.68	0.000	-1.105544	-.6561017
	k618	-.0297267	.0407748	-0.73	0.466	-.1096439	.0501904
	wa	-.0349665	.0076782	-4.55	0.000	-.0500155	-.0199175
	we	.1277126	.0259654	4.92	0.000	.0768214	.1786039
	un	-.0110548	.0159776	-0.69	0.489	-.0423704	.0202607
	cit	.0100105	.1076264	0.09	0.926	-.2009333	.2209543
	prin	-.0000212	4.71e-06	-4.51	0.000	-.0000304	-.000012
	_cons	.5678982	.4820863	1.18	0.239	-.3769737	1.51277

Fuente: Elaboración propia en base a Mroz (1987).

Además, como se muestra en la tabla 13, luego de estimar, es posible clasificar a las estimaciones y obtener la sensibilidad y la especificidad del modelo, como así también el porcentaje de observaciones correctamente clasificadas (67,6%, en este caso, menor que en la estimación Logit).

Tabla 13. Clasificación post estimación Probit.

Probit model for lfp

		----- True -----		
Classified		D	~D	Total
+		349	165	514
-		79	160	239
Total		428	325	753

Classified + if predicted Pr(D) >= .5

True D defined as lfp != 0

Sensitivity	Pr(+ D)	81.54%
Specificity	Pr(- ~D)	49.23%
Positive predictive value	Pr(D +)	67.90%
Negative predictive value	Pr(~D -)	66.95%
False + rate for true ~D	Pr(+ ~D)	50.77%
False - rate for true D	Pr(- D)	18.46%
False + rate for classified +	Pr(~D +)	32.10%
False - rate for classified -	Pr(D -)	33.05%
Correctly classified		67.60%

Fuente: Elaboración propia en base a Mroz (1987).

(d) Debido a que la distribución normal acumulativa y la distribución logística están muy cerca entre sí, en la mayoría de los casos los modelos estimados logit y probit serán bastante similares. ¿Son similares las probabilidades log-likelihoods maximizadas de la muestra en sus modelos logit y probit estimados? ¿Cuál es mayor? ¿Qué pasa con los signos y la significancia estadística de los parámetros estimados? ¿Son similares? El efecto estimado de un cambio en un regresor sobre la probabilidad de participar en la fuerza laboral, $\frac{\partial P}{\partial X}$, es igual a $P * (1 - P) * \hat{\beta}_{Li}$ en el modelo logit y $f(P) * \hat{\beta}_{Pi}$ en el modelo probit, donde P es la probabilidad de LFP, $\hat{\beta}_{Li}$ y $\hat{\beta}_{Pi}$ son los coeficientes logit y probit estimados, respectivamente, en la i -ésima variable explicativa, y $f(P)$ es la función normal acumulativa correspondiente a P . Evalúe estas derivadas estimadas para los modelos logit y probit, utilizando el LFPR muestral de $\frac{425}{753} = 0,568$ como estimación de P y observando que $f(P) = f(0,568) = 0,393$. En esta media muestral, ¿son similares los efectos estimados para los modelos logit y probit? ¿Qué sucede si evalúa estos efectos en la cola de la distribución, como en $P = 0,9$ donde $f(P) = 0,175$?

Por un lado, las probabilidades log-likelihoods maximizadas de la muestra en los modelos Logit y Probit son 0,568 y 0,571, respectivamente, por lo que es mayor en el caso del Probit. Por otro lado, los signos y la significancia estadística de los parámetros estimados entre ambos modelos son similares (ver tabla 14). Por último, en la media muestral, los efectos estimados para los modelos Logit y Probit son similares, no así si se evalúa estos efectos en la cola de la distribución, en donde los efectos estimados son, en valor absoluto, mayores para el modelo Probit (ver cálculos en *do-file*).

Tabla 14. *Estimación Logit versus Estimación Probit.*

	(1) Logit	(2) Probit
lfp		
lwvl	0.464*** (0.157)	0.282*** (0.0928)
kl6	-1.469*** (0.199)	-0.881*** (0.115)
k618	-0.0512 (0.0684)	-0.0297 (0.0408)
wa	-0.0583*** (0.0129)	-0.0350*** (0.00768)
we	0.212*** (0.0441)	0.128*** (0.0260)
un	-0.0186 (0.0265)	-0.0111 (0.0160)
cit	0.0127 (0.178)	0.0100 (0.108)
prin	-0.0000353*** (0.00000811)	-0.0000212*** (0.00000471)
_cons	0.951 (0.804)	0.568 (0.482)
N	753	753
pseudo R-sq	0.127	0.127

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Fuente: Elaboración propia en base a Mroz (1987).

(e) Dado que la distribución logística tiene una variación de $\frac{\pi^2}{3}$, mientras que la variación del modelo probit generalmente se normaliza a la unidad, una forma de comparar las estimaciones logit y probit es multiplicar cada una de las estimaciones logit por $\frac{\sqrt{3}}{\pi} = \frac{1.73205}{3.14159} \cong 0.5513$ y, luego, comparar estos parámetros logit transformados con las estimaciones probit reales. Amemiya [1981] sostiene, sin embargo, que surge una mejor aproximación si se multiplican las estimaciones de los parámetros logit por 0,625 y, luego, se comparan estos parámetros logit transformados con las estimaciones probit. ¿Cuál de estas dos aproximaciones transforma mejor sus estimaciones logit en estimaciones probit comparables? ¿Por qué?

De las dos aproximaciones mencionadas, la de Amemiya (1981) transforma mejor las estimaciones Logit en estimaciones Probit comparables (ver cálculos en *do-file*), ya que la distancia entre ambas estimaciones es menor considerando esta aproximación.

(f) En este ejercicio, hemos comparado modelos de probabilidad lineal MCO, estimaciones logit y probit de una ecuación de participación en la fuerza laboral y nos hemos centrado en las relaciones numéricas entre los parámetros estimados. Sin

embargo, como se señaló en la Sección 11.3.B.1, existen serios problemas estadísticos con cada uno de los procedimientos MCO, logit y probit particulares que se emplearon en este ejercicio. ¿Cuáles son estos problemas?

Los problemas estadísticos asociados con los procedimientos MCO, Logit y Probit particulares que se emplearon en este ejercicio pueden incluir:

- Supuestos no cumplidos: Los modelos MCO, Logit y Probit tienen supuestos diferentes. MCO asume errores con distribución normal, mientras que Logit y Probit asumen errores con distribución logística y normal, respectivamente. Si estos supuestos no se cumplen, las estimaciones pueden estar sesgadas y las pruebas de hipótesis basadas en los errores estándar pueden no ser válidas.
- Endogeneidad: Ocurre cuando una variable explicativa está correlacionada con el término de error. En los modelos MCO, esto puede llevar a estimaciones sesgadas e inconsistentes; en los modelos Logit y Probit, puede afectar la interpretación de los coeficientes.
- Heterocedasticidad: Ocurre cuando la variabilidad de los errores no es constante a lo largo de los valores de las variables explicativas. En los modelos MCO, esto puede conducir a estimaciones ineficientes y errores estándar sesgados; en los modelos Logit y Probit, puede afectar la precisión de las estimaciones y las pruebas de hipótesis.
- Autocorrelación: Ocurre cuando hay correlación entre los errores en diferentes observaciones. En los modelos MCO, esto puede resultar en estimaciones ineficientes y errores estándar sesgados; en los modelos logit y probit, puede afectar la precisión de las estimaciones y conducir a pruebas de hipótesis incorrectas.

En resumen, los modelos MCO, Logit y Probit pueden enfrentar una variedad de problemas estadísticos que pueden afectar la validez de las estimaciones y la interpretación de los resultados.

Ejercicio 4: Relating the Tobit and Conditional OLS Estimates.

El propósito de este ejercicio es involucrarlo en la estimación e interpretación de un modelo de oferta laboral basado en el Procedimiento Tobit III y enriquecer su comprensión de cómo este modelo Tobit se relaciona con el marco de MCO condicional del Procedimiento II. También tendrá la oportunidad de implementar, empíricamente, los resultados analíticos de Goldberger [1981], Greene [1981] y McDonald y Moffitt [1980], que se analizaron en la Sección 11.3.B.1.

(a) Primero, estime un modelo MCO condicional de oferta laboral (Procedimiento II). En particular, restringiendo su muestra a las mujeres que trabajaron por un salario en 1975 (las primeras 428 observaciones en el archivo de datos MROZ), ejecute una regresión de WHRS en un término constante, KL6, K618, WA, WE, PRIN y LWW1 (consulte la parte (d) del Ejercicio 1 para una discusión sobre LWW1 y PRIN). Compare sus resultados con los obtenidos por Mroz, reproducidos en la ecuación (11.49). Sus estimaciones de error estándar pueden diferir de las reportadas por Mroz, ya que sus estimaciones se ajustan por heterocedasticidad utilizando el procedimiento de estimación robusta de Halbert White [1980]. Si el programa de su computadora lo permite, calcule también los errores estándar robustos de White; sus resultados deberían ser muy parecidos a los informados por Mroz.

En la tabla 15, se presenta la estimación por MCO (condicional, al restringir la muestra a las mujeres que trabajaron) de una ecuación de horas trabajadas, en la cual whrs se regresa en un término constante, kl6, k618, wa, we, prin y lww. Se puede observar que los resultados son muy parecidos a los informados por Mroz, reproducidos en la ecuación (11.49).

Tabla 15. Estimación por MCO (condicional) de ecuación de horas trabajadas.

Linear regression				Number of obs	=	428
				F(6, 421)	=	3.93
				Prob > F	=	0.0008
				R-squared	=	0.0670
				Root MSE	=	755.16

		Robust				
whrs		Coefficient	std. err.	t	P> t	[95% conf. interval]
-----+-----						
kl6		-342.5048	131.7733	-2.60	0.010	-601.5205 -83.48919
k618		-115.0205	29.50866	-3.90	0.000	-173.0232 -57.01786
wa		-7.729976	5.849662	-1.32	0.187	-19.22816 3.768206
we		-14.44486	18.21292	-0.79	0.428	-50.24445 21.35473
prin		-.0042458	.0032235	-1.32	0.189	-.0105821 .0020904
lww1		-17.40781	81.37728	-0.21	0.831	-177.3642 142.5486
_cons		2114.697	350.3186	6.04	0.000	1426.106 2803.289

Fuente: Elaboración propia en base a Mroz (1987).

(b) En la Sección 11.3.B.1, se señaló que estas estimaciones MCO condicionales son estimaciones sesgadas de los parámetros de la ecuación (11.33). ¿Por qué son sesgadas? Siguiendo a Goldberger y Greene, calcule estimaciones consistentes de cada uno de los parámetros de la oferta laboral utilizando el ajuste de la LFP. En particular, calcule la

proporción muestral de observaciones para las cuales WHRS es positivo; en el archivo de datos MROZ, esto es $\frac{428}{753} = 0,568$. Luego, divida cada una de las estimaciones de los parámetros MCO condicionales de la parte (a) por esta proporción. Según Goldberger y Greene, estas estimaciones MCO condicionales transformadas son estimaciones consistentes de los parámetros la oferta de trabajo en la ecuación (11.33). ¿Estas estimaciones transformadas son consistentes? ¿Qué puedes decir sobre la significación estadística de estos parámetros transformados? ¿Por qué?

Las estimaciones MCO condicionales estimadas son estimaciones sesgadas (hacia abajo) de los parámetros de la ecuación (11.33) porque la variable dependiente sólo se corresponde con las horas trabajadas de las mujeres que trabajaron. Siendo así y siguiendo a Goldberger y Greene, se calculan estimaciones consistentes de cada uno de los parámetros de la oferta laboral utilizando el ajuste de la *lfp* (ver cálculos en *do-file*). Cabe notar que no se puede decir mucho sobre la significación estadística de estos parámetros transformados porque no se pueden realizar pruebas de significación con este procedimiento, ya que los errores estándar estimados basados en MCO son inconsistentes.

(c) Ahora, obtenga estimaciones consistentes utilizando el método de estimación Tobit (Procedimiento III). En particular, utilizando el procedimiento Tobit de máxima verosimilitud (la función de verosimilitud muestral correspondiente a la ecuación (11.33) es la ecuación (11.32) con J_i , ahora, igual a $X_i\beta + u_{Hi}$) y la misma forma funcional que en el inciso (a), calcule las estimaciones de los parámetros Tobit y los errores estándar asintóticos. ¿Cómo se comparan estas estimaciones con las aproximaciones de Goldberger y Greene del inciso (b)? ¿Qué puede decir sobre la significancia estadística de estos parámetros Tobit?

En la tabla 16, se presenta la estimación Tobit de una ecuación de horas trabajadas, en la cual *whrs* se regresa en un término constante, *kl6*, *k618*, *wa*, *we*, *prin* y *lww1*. Se puede observar que todos los parámetros son estadísticamente significativos.

Tabla 16. Estimación Tobit de ecuación de horas trabajadas.

Tobit regression				Number of obs		=	753
				Uncensored		=	428
Limits: Lower =		0	Left-censored		=	325	
Upper =		+inf	Right-censored		=	0	
				LR chi2(6)		=	127.54
				Prob > chi2		=	0.0000
Log likelihood = -3891.1202				Pseudo R2		=	0.0161

whrs		Coefficient	Std. err.	t	P> t	[95% conf. interval]	

kl6		-1045.188	125.0131	-8.36	0.000	-1290.606	-799.7688
k618		-100.3541	42.28567	-2.37	0.018	-183.367	-17.34117
wa		-36.5092	7.640147	-4.78	0.000	-51.50792	-21.51049
we		104.9165	25.74198	4.08	0.000	54.38121	155.4517
prin		-.0222349	.0048942	-4.54	0.000	-.0318429	-.012627
lw1		199.3924	88.75162	2.25	0.025	25.16015	373.6247
_cons		1172.019	477.9667	2.45	0.014	233.7009	2110.336

var(e.whrs)		1584438	118914.7			1367375	1835959

Fuente: Elaboración propia en base a Mroz (1987).

(d) Como se analizó en la Sección 11.3.B.1, McDonald y Momt han demostrado (ver ecuación (11.35)) que el efecto total del cambio en un regresor sobre las horas esperadas trabajadas en el modelo Tobit se puede descomponer en dos partes: cambio en las horas trabajadas para quienes ya trabajan ponderado por la probabilidad de trabajar más el cambio en la probabilidad de trabajar ponderado por el valor esperado de las horas trabajadas para quienes trabajan. Utilizando la proporción muestral de quienes trabajan ($\frac{428}{753} = 0,568$) como estimación de $F(z)$, las estimaciones del parámetro Tobit del inciso (c) y los datos evaluados con medias muestrales, calcule A como se describe en la ecuación (11.35). Nota: Para $F(z) = 0,568$, $z = 0,175$ y $f(z) = 0,393$. Del cambio total en las horas trabajadas debido a un cambio de 1% en el salario, ¿qué cantidad resulta de los cambios en las horas trabajadas de aquellos que ya están trabajando? ¿qué cantidad proviene de los nuevos ingresantes a la fuerza laboral? ¿Cuál es la proporción del efecto total sobre las horas trabajadas de un cambio en cualquiera de las variables que se deriva de aquellas mujeres que ya están trabajando?

Por un lado, del cambio total en las horas trabajadas debido a un cambio de 1% en el salario, la cantidad que resulta de los cambios en las horas trabajadas de aquellos que ya están trabajando es 45,632 y la cantidad que proviene de los nuevos ingresantes a la fuerza laboral es 31,255. Por otro lado, la proporción del efecto total sobre las horas trabajadas de un cambio en cualquiera de las variables que se deriva de aquellas mujeres que ya están trabajando es 0,403.

(e) La función de probabilidad (11.32) para el modelo Tobit indica que se puede considerar que el Tobit combina un modelo Probit de participación en la fuerza laboral con un modelo de regresión estándar de horas trabajadas para quienes trabajan (ver Sección 11.3.B.1). Por lo tanto, se podría concluir que, si la muestra se limitara a quienes trabajan, las estimaciones Tobit serían, numéricamente, equivalentes a las estimaciones

MCO. ¿Sería correcta tal conclusión? ¿Por qué o por qué no? Verifique su intuición numéricamente, utilizando datos en el archivo de datos MROZ, estimando un modelo Tobit, restringiendo la muestra a las primeras 428 observaciones (las trabajadoras) y comparando estas estimaciones Tobit con las estimaciones de MCO del inciso (a).

En la tabla 17, se presenta la estimación Tobit de una ecuación de horas trabajadas, en la cual *whrs* se regresa en un término constante, *kl6*, *k618*, *wa*, *we*, *prin* y *lww*, restringiendo la muestra a las mujeres que trabajaron. Se puede observar que estas estimaciones de los parámetros son, numéricamente, equivalentes a las estimaciones de los parámetros por MCO.

Tabla 17. Estimación Tobit de ecuación de horas trabajadas (observaciones con *lpf*= 1).

Tobit regression				Number of obs	=	428
				Uncensored	=	428
Limits: Lower = 0				Left-censored	=	0
Upper = +inf				Right-censored	=	0
				LR chi2(6)	=	29.66
				Prob > chi2	=	0.0000
Log likelihood = -3440.103				Pseudo R2	=	0.0043

whrs		Coefficient	Std. err.	t	P> t	[95% conf. interval]

kl6		-342.5048	99.18476	-3.45	0.001	-537.4625 -147.5471
k618		-115.0205	30.57611	-3.76	0.000	-175.1209 -54.92008
wa		-7.729976	5.484046	-1.41	0.159	-18.50942 3.049472
we		-14.44486	17.82039	-0.81	0.418	-49.47264 20.58292
prin		-.0042458	.0036258	-1.17	0.242	-.0113727 .0028811
lww		-17.40781	53.77026	-0.32	0.746	-123.0987 88.2831
_cons		2114.697	337.3377	6.27	0.000	1451.626 2777.769

var(e.whrs)		560940.7	38345.12			490413.8 641610.2

Fuente: Elaboración propia en base a Mroz (1987).

(f) *En este ejercicio, hemos relacionado, numéricamente, las estimaciones Tobit y MCO condicionales en un modelo de oferta laboral. Sin embargo, como se señaló en la Sección 11.3.B.1, existen serios problemas estadísticos con los procedimientos Tobit y OLS condicional que se emplearon en este ejercicio. ¿Cuáles son estos problemas?*

Los problemas estadísticos asociados con los procedimientos Tobit y MCO condicional que se emplearon en este ejercicio pueden incluir:

- Violación de la normalidad: Tanto el modelo Tobit como el MCO condicional asumen que los errores tienen una distribución normal. Si esta suposición no se cumple, las estimaciones pueden estar sesgadas y las pruebas de hipótesis basadas en los errores estándar pueden no ser válidas.
- Heterocedasticidad: Puede afectar la eficiencia de los estimadores en ambos modelos, Tobit y MCO condicional, lo cual puede conducir a estimaciones ineficientes y errores estándar sesgados.

- Autocorrelación: Puede surgir en ambos modelos, Tobit y MCO condicional, lo cual puede resultar en estimaciones ineficientes y errores estándar sesgados, lo que afecta la precisión de las pruebas de hipótesis.
- Sesgo de selección: El modelo Tobit y el MCO condicional asumen que todas las observaciones tienen la misma probabilidad de ser censuradas o truncadas. Sin embargo, en la práctica, esto puede no ser cierto y puede haber un sesgo de selección si las observaciones censuradas o truncadas tienen características diferentes de las observaciones no censuradas o truncadas.

En resumen, tanto el modelo Tobit como el MCO condicional pueden enfrentar una variedad de problemas estadísticos que pueden afectar la validez de las estimaciones y la interpretación de los resultados.

Ejercicio 5: Identifying Parameters in a Reduced Form Estimation.

El propósito de este ejercicio es explorar cuestiones de identificación de los parámetros estructurales cuando el método de forma reducida del Procedimiento IV se implementa empíricamente. En esencia, este procedimiento implica, primero, estimar una ecuación de salario ecuación usando sólo una muestra de trabajadores, luego usando la muestra completa de observaciones con las horas de los no trabajadores fijadas en cero para estimar una ecuación de horas trabajadas de forma reducida derivada de una relación de salario de reserva y, finalmente, identificando los parámetros estructurales de la ecuación del salario de reserva utilizando la forma reducida y las estimaciones de los parámetros de la ecuación salarial. Como veremos, la ecuación del salario de reserva está subidentificada, apenas identificada o sobreidentificada dependiendo de si cero, uno o más regresores en la ecuación del salario están excluidos de la ecuación del salario de reserva.

(a) *Obtenga, del Ejercicio 1, parte (d), las estimaciones de los parámetros de la ecuación de determinación de salarios de MCO en la que se hizo la regresión de $LWW1$ sobre una constante, WA , WE , AX , $WA2$ y CIT y la muestra se limitó a aquellos que trabajaron (los primeros 428 observaciones). Esto corresponde a la ecuación, con i subíndices eliminados por simplicidad:*

$$lww = g_0 + g_1wa + g_2wa2 + g_3we + g_4cit + g_5ax + e_w, \quad (11.61)$$

donde las g son estimaciones de parámetros MCO y e_w es el residuo MCO.

En la tabla 18, se presenta la estimación por MCO (condicional, al restringir la muestra a las mujeres que trabajaron) de una ecuación de salarios, en la cual $lww1$ se regresa en un término constante, wa , $wa2$, we , cit y ax .

Tabla 18. *Estimación por MCO (condicional) de ecuación de salarios.*

Source	SS	df	MS	Number of obs	=	428
Model	34.3816151	5	6.87632302	F(5, 422)	=	15.36
Residual	188.945826	422	.447738924	Prob > F	=	0.0000
				R-squared	=	0.1540
				Adj R-squared	=	0.1439
Total	223.327441	427	.523015084	Root MSE	=	.66913

lww1	Coefficient	Std. err.	t	P> t	[95% conf. interval]
wa	.0605336	.0461306	1.31	0.190	-.0301408 .151208
wa2	-.0007322	.0005379	-1.36	0.174	-.0017894 .0003251
we	.1074434	.0143904	7.47	0.000	.0791577 .1357291
cit	.0615677	.0687288	0.90	0.371	-.0735258 .1966612
ax	.0171097	.004619	3.70	0.000	.0080306 .0261889
_cons	-1.639786	.9901658	-1.66	0.098	-3.586057 .3064851

Fuente: Elaboración propia en base a Mroz (1987).

(b) *A continuación, especifique tres ecuaciones de salario de reserva estructural alternativas y, luego, utilizando una transformación de la relación de proporcionalidad de Heckman (11.42), escriba sus correspondientes representaciones en forma reducida.*

Específicamente, tres posibles representaciones estructurales para el logaritmo de la ecuación del salario de reserva (LWR), análoga a la ecuación (11.41), son:

Apenas identificada:

$$lwr = a_0 + a_1wa + a_2wa2 + a_3we + a_4cit + a_5kl6 + a_6k618 + a_7prin + a_8un + \epsilon \quad (11.62)$$

Sobreidentificada:

$$lwr = b_0 + b_1wa + b_2wa2 + b_3we + b_4kl6 + b_5k618 + b_6prin + b_7un + \epsilon \quad (11.63)$$

Subidentificada:

$$lwr = c_0 + c_1wa + c_2wa2 + c_3we + c_4cit + c_5ax + c_6kl6 + c_7k618 + c_8prin + c_9un + \epsilon \quad (11.64)$$

donde PRIN es la variable de ingreso de la propiedad construida en el inciso (c) del Ejercicio 1 y ϵ es una perturbación aleatoria. Nótese que, en las Ecs. (11.62), (11.63) y (11.64), hay una (AX), dos (AX y CIT) y cero variables, respectivamente, incluidas en la ecuación LWW (11.61) pero excluidas de la ecuación LWR. Para cada una de estas tres ecuaciones de salario de reserva, emplee, primero, una transformación logarítmica de la relación de proporcionalidad de Heckman (11.42):

$$whrs_i = d(lww_i - lwr_i) \quad \text{si } lww_i > lwr_i \quad (11.65a)$$

$$whrs_i = 0 \quad \text{si } lww_i \leq lwr_i \quad (11.65b)$$

Luego, sustituir, en las Ecs. (11.65), para LWW, la ecuación de determinación de salario estimado por MCO (11.61) y, para LWR, una de las especificaciones de LWR en las Ecs. (11.62)-(11.64). Una vez hecho esto, debería haber obtenido tres ecuaciones alternativas en forma reducida para las horas trabajadas, cada una en términos de los parámetros estructurales subyacentes, las variables exógenas y las perturbaciones.

A continuación, se especifican tres ecuaciones de salario de reserva estructural alternativas:

$$lwr1 = a_0 + a_1wa + a_2wa2 + a_3we + a_4cit + a_5kl6 + a_6k618 + a_7prin + a_8un + \epsilon \quad (1)$$

$$lwr2 = b_0 + b_1wa + b_2wa2 + b_3we + b_4kl6 + b_5k618 + b_6prin + b_7un + \epsilon \quad (2)$$

$$lwr3 = c_0 + c_1wa + c_2wa2 + c_3we + c_4cit + c_5ax + c_6kl6 + c_7k618 + c_8prin + c_9un + \epsilon \quad (3)$$

Utilizando la transformación de la relación de proporcionalidad de Heckman, sus correspondientes representaciones en forma reducida son:

$$whrs1 = d(g_0 - a_0) + d(g_1 - a_1)wa + d(g_2 - a_2)wa2 + d(g_3 - a_3)we + d(g_4 - a_4)cit + dg_5ax - da_5kl6 - da_6k618 - da_7prin - da_8un + (e_w - \epsilon) \quad (4)$$

$$whrs2 = d(g_0 - b_0) + d(g_1 - b_1)wa + d(g_2 - b_2)wa2 + d(g_3 - b_3)we + dg_4cit + dg_5ax - db_4kl6 - db_5k618 - db_6prin - db_7un + (e_w - \epsilon) \quad (5)$$

$$whrs3 = d(g_0 - c_0) + d(g_1 - c_1)wa + d(g_2 - c_2)wa2 + d(g_3 - c_3)we + d(g_4 - c_4)cit + d(g_5 - c_5)ax - dc_6kl6 - dc_7k618 - dc_8prin - dc_9un + (e_w - \epsilon) \quad (6)$$

(c) Utilizando la muestra completa de 753 observaciones en el archivo de datos MROZ y estableciendo las horas de los no trabajadores en cero, estime por MCO la forma reducida de la ecuación correspondiente a la ecuación apenas identificada (11.62) derivado en la parte (b), utilizando las estimaciones g de la ecuación (11.61) para resolver para los parámetros a en la ecuación (11.62). Este procedimiento, a menudo, se llama mínimos cuadrados indirectos. ¿Estas estimaciones estructurales de los parámetros a tienen sentido? ¿Por qué o por qué no? ¿Qué pasa con su estimación de d ? A continuación, emplee un procedimiento de estimación más directo. Específicamente, usando su ecuación en forma reducida correspondiente a la ecuación (11.62), construir variables regresoras transformadas como el producto de los regresores originales en la ecuación (11.61) y sus coeficientes g estimados y, luego, estimar por MCO la ecuación resultante con estas variables transformadas como regresores. Verifica que obtienes las mismas estimaciones estructurales de los parámetros a y d como lo hizo con el procedimiento de mínimos cuadrados indirectos. ¿Las estimaciones del error estándar de este procedimiento directo son apropiadas para hacer inferencias? ¿Por qué o por qué no?

No se verifica que se obtienen las mismas estimaciones estructurales de los parámetros a y d mediante el procedimiento de mínimos cuadrados indirectos y el procedimiento más directo (ver cálculos en *do-file*).

(d) Para comprender las complicaciones que surgen cuando se sobreidentifica la ecuación del salario de reserva, primero, demuestre que, en la ecuación de forma reducida correspondiente a la ecuación (11.63) derivada en el inciso (b), se debe imponer una restricción de parámetro para que se obtenga una estimación única de d . Utilizando la muestra completa de 753 observaciones en el archivo de datos MROZ y estableciendo las horas de los no trabajadores en cero, imponga esta restricción de parámetro, estime mediante MCO restringido la ecuación de forma reducida y sobreidentificada y resuelva los parámetros b en la ecuación (11.63) y para d . ¿Tienen sentido estas estimaciones estructurales de los parámetros b y d ? ¿Por qué o por qué no? Derive, interprete y, luego, implemente, empíricamente, una prueba para el modelo sobreidentificado (11.63) como un caso especial del modelo apenas identificado (11.62).

En la ecuación de forma reducida correspondiente a la ecuación (11.63) derivada en el inciso (b), se debe imponer una restricción de parámetros para que se obtenga una estimación única de d . Esto se debe a que se tiene que coeficiente de cit es igual a dg_4 y que el coeficiente de ax es igual dg_5 .

(e) Demuestre que, si estimara mediante MCO la ecuación en forma reducida correspondiente a la ecuación de salario de reserva subidentificada (11.64) derivada en el inciso (b), no hay manera de obtener estimaciones únicas de los parámetros estructurales c y d . No obstante, si estimaste esta ecuación en forma reducida mediante MCO, ¿cuál sería el R^2 en relación con el de la estimación en forma reducida apenas identificada del inciso (c)? ¿Por qué ocurre esto?

Si se estimara mediante MCO la ecuación en forma reducida correspondiente a la ecuación de salario de reserva subidentificada (11.64), no hay manera de obtener estimaciones únicas de los parámetros estructurales c y d , ya que, en este caso, no es posible obtener una estimación única de d haciendo uso de las estimaciones g .

(f) Existe un grave inconveniente con la estimación del Procedimiento IV que ha realizado en este ejercicio. ¿Cuál es este problema?

El inconveniente principal con la estimación del Procedimiento IV que se ha realizado en este ejercicio es cuando la ecuación del salario de reserva está subidentificada. La subidentificación ocurre cuando hay más parámetros desconocidos que restricciones en el modelo, lo que significa que no es posible estimar todos los parámetros estructurales de manera única.

En el contexto específico del Procedimiento IV descrito, la ecuación del salario de reserva se deriva de una relación entre el salario y las horas trabajadas, asumiendo que las horas no trabajadas tienen un valor de salario de reserva de cero. Sin embargo, esta suposición puede no ser válida en todos los casos, lo que conduce a la subidentificación de los parámetros estructurales.

La subidentificación puede provocar estimaciones sesgadas o inconsistentes de los parámetros del modelo, lo que hace que las conclusiones basadas en estas estimaciones sean poco confiables. Para abordar este problema, se necesitarían estrategias alternativas, como la inclusión de variables instrumentales adicionales o el uso de diferentes especificaciones de modelos que permitan una identificación adecuada de los parámetros.

Ejercicio 6: Implementing the Heckit Generalized Tobit Estimator.

El propósito del Ejercicio 6 es que usted implemente e interprete el Procedimiento VIII de múltiples etapas de Heckit. Esto se logra replicando los resultados reportados por Mroz [1987]. También comparará el procedimiento de selectividad de muestras de Heckit con un método basado en OLS debido a Olsen [1980].

*(a) En la primera etapa del procedimiento de Heckit, se estima una ecuación probit de LFP y se recupera de esta ecuación estimada la inversa del ratio de Mills λ . Mroz, primero, genera una serie de transformaciones polinómicas de las variables de edad, educación y experiencia de la esposa para usarlas como variables explicativas en la ecuación LFP. Para la muestra completa de 753 observaciones en el archivo de datos MROZ, siguiendo a Mroz, genere $AX2 = AX * AX$, $WA2 = WA * WA$, $WE2 = WE * WE$, $WA3 = WA2 * WA$, $WE3 = WE2 * WE$, $WAW = WA * WE$, $WA2WE = WA2 * WE$ y $WAW2 = WA * WE2$. Con esta muestra, estime por máxima verosimilitud un modelo probit en el que la LFP es la variable dependiente y las variables explicativas incluyen un término constante, $KL6$, $K618$, WA , WE , $WA2$, $WE2$, WAW , $WA3$, $WE3$, $WA2WE$, $WAW2$, $WFED$, $WMED$, UN , CIT y $PRIN$ (esta última variable se calculó en la parte (c) del Ejercicio 1). A partir de este modelo probit estimado, calcule la inversa del ratio de Mills para cada observación, guarde esta variable y llámela $INVR1$. (Algunos programas de computadora ofrecen este cálculo como un comando opcional; para otros, debe calcularse mediante fuerza bruta, usando la ecuación (11.37) y valores de la distribución normal). Ahora, rehaga la estimación probit, esta vez sumando las variables de experiencia AX y $AX2$ y llame a los valores de la inversa del ratio de Mills correspondientes $INVR2$. Comente la significancia estadística de los parámetros estimados en estas dos ecuaciones probit. Tenga en cuenta que la variable salario $LWW1$ está excluida como variable explicativa en este modelo probit. Dado que la teoría económica sugiere que la LFP se ve afectada por la tasa salarial, ¿por qué se excluye esta variable $LWW1$? Sin embargo, ¿en qué sentido podría incluirse indirectamente?*

En las tablas 19 y 20, se presentan las estimaciones Probit de una ecuación de participación en la fuerza laboral, en la cual lfp se regresa en un término constante, $kl6$, $k618$, wa , we , $wa2$, $we2$, waw , $wa3$, $we3$, $wa2we$, $waw2$, $wfed$, $wmed$, un , cit y $prin$ (ax y $ax2$, en la tabla 20). Se puede observar que, en la primera estimación, sólo son estadísticamente significativas las variables $kl6$ y $prin$, mientras que, en la segunda estimación, también lo son ax y $ax2$.

Dado que la teoría económica sugiere que la lfp se ve afectada por la tasa salarial, la variable $lww1$ se excluye porque el propósito de la primera etapa del procedimiento Heckit es modelar la participación en la fuerza laboral sin tener en cuenta el salario. Este enfoque se utiliza para evitar problemas de endogeneidad y sesgo de selección en la estimación del modelo. Sin embargo, indirectamente, el salario puede influir en la participación en la fuerza laboral a través de su efecto en otras variables incluidas en el modelo, como el nivel educativo (we , $we2$, $we3$) o la experiencia laboral (ax , $ax2$). Si el salario afecta estas variables explicativas, su impacto en la participación en la fuerza laboral puede ser capturado, indirectamente, a través de ellas. Es importante considerar la validez de esta suposición y realizar pruebas de especificación adecuadas para evaluar si la inclusión indirecta del salario a través de otras variables es apropiada en el modelo.

Tabla 19. *Estimación Probit de ecuación de participación en la fuerza laboral (sin variables de experiencia).*

Probit regression
Log likelihood = -450.47746
Number of obs = 753
LR chi2(16) = 128.79
Prob > chi2 = 0.0000
Pseudo R2 = 0.1251

lfp	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
kl6	-.8579509	.1188431	-7.22	0.000	-1.090879	-.6250226
k618	-.0603097	.0439087	-1.37	0.170	-.1463692	.0257497
wa	.7521026	.76984	0.98	0.329	-.7567562	2.260961
we	1.09907	1.922726	0.57	0.568	-2.669404	4.867544
wa2	-.0151855	.0145537	-1.04	0.297	-.0437101	.0133392
we2	-.0636591	.1056755	-0.60	0.547	-.2707794	.1434611
wawe	-.0154558	.0446633	-0.35	0.729	-.1029943	.0720827
wa3	.000107	.0001011	1.06	0.290	-.0000911	.0003051
we3	.0012697	.0023487	0.54	0.589	-.0033336	.005873
wa2we	-2.10e-06	.0003764	-0.01	0.996	-.0007399	.0007357
wawe2	.0006687	.0008759	0.76	0.445	-.0010481	.0023855
wfed	-.0125127	.017534	-0.71	0.475	-.0468787	.0218532
wmed	.0026455	.018479	0.14	0.886	-.0335726	.0388636
un	-.0102966	.0161046	-0.64	0.523	-.041861	.0212678
cit	.0414417	.10854	0.38	0.703	-.1712928	.2541763
prin	-.000022	4.77e-06	-4.62	0.000	-.0000313	-.0000127
_cons	-14.06639	15.98708	-0.88	0.379	-45.4005	17.26772

Fuente: Elaboración propia en base a Mroz (1987).

Tabla 20. *Estimación Probit de ecuación de participación en la fuerza laboral (con variables de experiencia).*

Probit regression
Log likelihood = -398.23069
Number of obs = 753
LR chi2(18) = 233.29
Prob > chi2 = 0.0000
Pseudo R2 = 0.2265

lfp	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
kl6	-.8504688	.1233302	-6.90	0.000	-1.092192	-.608746
k618	.0236102	.0474965	0.50	0.619	-.0694813	.1167017
wa	.4938567	.8053263	0.61	0.540	-1.084554	2.072267
we	.9622103	2.002132	0.48	0.631	-2.961897	4.886317
wa2	-.0115551	.0154034	-0.75	0.453	-.0417453	.018635
we2	-.0723266	.111544	-0.65	0.517	-.2909488	.1462956
wawe	-.0055927	.046395	-0.12	0.904	-.0965252	.0853398
wa3	.0000949	.000108	0.88	0.379	-.0001167	.0003066
we3	.0015105	.0024888	0.61	0.544	-.0033676	.0063885
wa2we	-.00012	.0003973	-0.30	0.763	-.0008988	.0006587
wawe2	.000672	.0009136	0.74	0.462	-.0011187	.0024626
wfed	-.0042132	.0184251	-0.23	0.819	-.0403257	.0318993
wmed	.0103231	.0193464	0.53	0.594	-.0275951	.0482413
un	-.0149435	.017083	-0.87	0.382	-.0484256	.0185386
cit	.0231706	.1152617	0.20	0.841	-.2027382	.2490795
prin	-.0000129	5.03e-06	-2.57	0.010	-.0000228	-3.08e-06
ax	.1241355	.0193738	6.41	0.000	.0861635	.1621075
ax2	-.0019205	.0006282	-3.06	0.002	-.0031518	-.0006892
_cons	-10.12486	16.53431	-0.61	0.540	-42.53151	22.28179

Fuente: Elaboración propia en base a Mroz (1987).

(b) Luego, restringiendo su muestra a aquellos que trabajan por un salario (las primeras 428 observaciones en MROZ), estime mediante MCO una ecuación de determinación de salario que permita la selectividad de la muestra y compare los resultados con una ecuación que no tenga en cuenta la selectividad de la muestra. En particular, siguiendo a Mroz, sea LWW una función lineal de un término constante, KL6, K618, WA, WE, WA2, WE2, WAVE, WA3, WE3, WA2WE, WAVE2, WMED, WFED, UN, CIT y PRIN. Llame a este conjunto de variables explicativas “Conjunto A”. Estime esta ecuación mediante MCO (si su software lo permite, emplee el procedimiento de error estándar robusto de White [1980]) y comente los signos y la significancia estadística de los parámetros estimados. Rehaga esta estimación de MCO, agregando al Conjunto A las variables de experiencia AX y AX2. Interpretar cualquier cambio en los resultados. Luego, con las mismas 428 observaciones, estime mediante MCO dos ecuaciones de determinación de salarios que permitan la selectividad de la muestra: primero, una ecuación de determinación de salarios LWW con las variables del Conjunto A y la variable de la inversa del ratio de Mills INVR1 (del inciso (a)) incluidas como regresores y, segundo, una ecuación de determinación de salarios LWW con las variables del Conjunto A incluidas, pero con las medidas de experiencia AX, AX2 y la variable INVR2 correspondiente también incluidas. Comente sobre la sensibilidad de los parámetros estimados a la inclusión de las variables de experiencia y al ajuste de selectividad de la muestra. ¿Es significativa la selectividad de la muestra? (Se puede utilizar la teoría de la distribución de muestras grandes y el método de error estándar robusto de White [1980] para realizar inferencias estadísticas).

En las tablas 21 y 22, se presentan las estimaciones por MCO de una ecuación de salarios, en la cual *lww* se regresa en un término constante, *kl6*, *k618*, *wa*, *we*, *wa2*, *we2*, *wawe*, *wa3*, *we3*, *wa2we*, *wawe2*, *wfed*, *wmed*, *un*, *cit* y *prin* (*ax* y *ax2*, en la tabla 22), restringiendo la muestra a las mujeres que trabajaron. En las tablas 23 y 24, se presentan las mismas estimaciones pero incorporando las variables de la inversa del ratio de Mills *invr1* e *invr2*, respectivamente. Se puede observar que, en ambas estimaciones, la selectividad de la muestra no es estadísticamente significativa.

Tabla 21. Estimación por MCO de ecuación de salarios (sin variables de experiencia y sin ajuste de selectividad).

Linear regression	Number of obs	=	428
	F(16, 411)	=	8.53
	Prob > F	=	0.0000
	R-squared	=	0.1772
	Root MSE	=	.66863

lwv	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]
kl6	-.1266363	.1119148	-1.13	0.258	-.3466331 .0933605
k618	-.0464093	.0302085	-1.54	0.125	-.1057917 .012973
wa	.1113025	.4408207	0.25	0.801	-.755242 .9778469
we	-.8514843	1.147582	-0.74	0.459	-3.107346 1.404378
wa2	-.0060282	.0093899	-0.64	0.521	-.0244865 .0124301
we2	.0119921	.0675097	0.18	0.859	-.1207153 .1446994
wawe	.0301834	.0240799	1.25	0.211	-.0171519 .0775186
wa3	.0000348	.0000707	0.49	0.623	-.0001042 .0001738
we3	.0017814	.0015375	1.16	0.247	-.0012409 .0048038
wa2we	.0000746	.0002031	0.37	0.714	-.0003247 .0004738
wawe2	-.0015388	.0005256	-2.93	0.004	-.0025719 -.0005057
wfed	-.0168407	.0111106	-1.52	0.130	-.0386724 .0049909
wmed	-.0090201	.0124633	-0.72	0.470	-.03352 .0154797
un	-.0015374	.0095299	-0.16	0.872	-.0202708 .017196
cit	.0789842	.0680702	1.16	0.247	-.0548249 .2127933
prin	1.86e-06	2.88e-06	0.65	0.518	-3.79e-06 7.52e-06
_cons	2.668241	8.681064	0.31	0.759	-14.39658 19.73307

Fuente: Elaboración propia en base a Mroz (1987).

Tabla 22. Estimación por MCO de ecuación de salarios (con variables de experiencia y sin ajuste de selectividad).

Linear regression	Number of obs	=	428
	F(18, 409)	=	8.10
	Prob > F	=	0.0000
	R-squared	=	0.2100
	Root MSE	=	.65679

lwv	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]
kl6	-.1055143	.1061322	-0.99	0.321	-.3141469 .1031183
k618	-.0171133	.0314647	-0.54	0.587	-.078966 .0447395
wa	.021614	.428565	0.05	0.960	-.820851 .864079
we	-.9660166	1.109505	-0.87	0.384	-3.147061 1.215028
wa2	-.0047736	.0092678	-0.52	0.607	-.0229921 .0134449
we2	.0141661	.0666777	0.21	0.832	-.1169076 .1452399
wawe	.0332155	.0228326	1.45	0.147	-.0116684 .0780994
wa3	.0000274	.0000702	0.39	0.696	-.0001106 .0001655
we3	.0018592	.0015278	1.22	0.224	-.0011441 .0048625
wa2we	.0000709	.0001932	0.37	0.714	-.0003089 .0004506
wawe2	-.0016391	.0005077	-3.23	0.001	-.0026371 -.000641
wfed	-.0129667	.0108686	-1.19	0.234	-.0343321 .0083986
wmed	-.0104555	.0121778	-0.86	0.391	-.0343944 .0134834
un	-.002731	.0093985	-0.29	0.772	-.0212064 .0157445
cit	.0698902	.0682422	1.02	0.306	-.0642589 .2040394
prin	4.61e-06	2.92e-06	1.58	0.116	-1.14e-06 .0000104
ax	.0398315	.0163393	2.44	0.015	.0077121 .0719509
ax2	-.0007028	.0004542	-1.55	0.123	-.0015957 .0001901
_cons	4.379425	8.295024	0.53	0.598	-11.92678 20.68563

Fuente: Elaboración propia en base a Mroz (1987).

Tabla 23. Estimación por MCO de ecuación de salarios (sin variables de experiencia y con ajuste de selectividad).

Linear regression	Number of obs	=	428
	F(17, 410)	=	8.52
	Prob > F	=	0.0000
	R-squared	=	0.1787
	Root MSE	=	.66886

lwv	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]
kl6	.3671854	.5145311	0.71	0.476	-.6442628 1.378634
k618	-.0125544	.045542	-0.28	0.783	-.1020794 .0769705
wa	-.306002	.6007274	-0.51	0.611	-1.486892 .874888
we	-1.52232	1.36434	-1.12	0.265	-4.204294 1.159654
wa2	.0025341	.0124389	0.20	0.839	-.021918 .0269861
we2	.0547208	.0819805	0.67	0.505	-.1064338 .2158753
wave	.0380183	.0256188	1.48	0.139	-.0123423 .0883789
wa3	-.0000279	.0000922	-0.30	0.762	-.0002091 .0001534
we3	.0008792	.0018235	0.48	0.630	-.0027055 .0044638
wa2we	.0000991	.0002036	0.49	0.627	-.0003011 .0004993
wave2	-.0019567	.0006902	-2.83	0.005	-.0033135 -.0005998
wfed	-.0094517	.0125412	-0.75	0.451	-.0341047 .0152013
wmed	-.0096553	.012468	-0.77	0.439	-.0341645 .014854
un	.0043358	.0113179	0.38	0.702	-.0179125 .0265841
cit	.0537633	.0692056	0.78	0.438	-.0822787 .1898054
prin	.0000146	.0000126	1.16	0.246	-.0000101 .0000394
invr1	.8599341	.8561958	1.00	0.316	-.8231473 2.543015
_cons	9.916658	11.44322	0.87	0.387	-12.57804 32.41135

Fuente: Elaboración propia en base a Mroz (1987).

Tabla 24. Estimación por MCO de ecuación de salarios (con variables de experiencia y con ajuste de selectividad).

Linear regression	Number of obs	=	428
	F(19, 408)	=	8.53
	Prob > F	=	0.0000
	R-squared	=	0.2112
	Root MSE	=	.6571

lwv	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]
kl6	.1576996	.5160366	0.31	0.760	-.8567227 1.172122
k618	-.0245922	.0341721	-0.72	0.472	-.0917676 .0425831
wa	-.1272606	.4629312	-0.27	0.784	-1.037289 .7827674
we	-1.271284	1.245965	-1.02	0.308	-3.720596 1.178028
wa2	-.0011548	.0100698	-0.11	0.909	-.02095 .0186404
we2	.0391827	.0794751	0.49	0.622	-.117049 .1954145
wave	.0341165	.0229362	1.49	0.138	-.0109714 .0792044
wa3	-3.37e-06	.0000784	-0.04	0.966	-.0001575 .0001507
we3	.0013101	.0017714	0.74	0.460	-.0021721 .0047923
wa2we	.0001192	.000208	0.57	0.567	-.0002898 .0005281
wave2	-.0018528	.0006669	-2.78	0.006	-.0031638 -.0005417
wfed	-.0116572	.0112614	-1.04	0.301	-.0337948 .0104804
wmed	-.013424	.0130021	-1.03	0.302	-.0389835 .0121356
un	.0021935	.0137871	0.16	0.874	-.0249092 .0292962
cit	.0636221	.068735	0.93	0.355	-.0714968 .1987411
prin	8.52e-06	7.59e-06	1.12	0.262	-6.40e-06 .0000234
ax	.0019818	.0744025	0.03	0.979	-.1442783 .1482418
ax2	-.0001309	.0012035	-0.11	0.913	-.0024968 .002235
invr2	.4435288	.8038702	0.55	0.581	-1.136716 2.023773
_cons	7.073821	9.126592	0.78	0.439	-10.86719 25.01483

Fuente: Elaboración propia en base a Mroz (1987).

(c) Finalmente, restringiendo su muestra a aquellos que trabajan por un salario (las primeras 428 observaciones en MROZ), use los valores ajustados de la ecuación de determinación de salarios en la parte (b) como instrumentos en una estimación de mínimos cuadrados en dos etapas ajustada por selectividad de la muestra de la ecuación de horas trabajadas. Específicamente, siguiendo a Mroz, para una comparación de caso base, primero, emplee la estimación de la variable instrumental (IV) (con el procedimiento de error estándar robusto de White, si está disponible en su software) de una ecuación de horas trabajadas en la que WHRS es la variable dependiente y las variables explicativas incluyen un término constante, KL6, K618, WA, WE, LWW y PRIN; Llame a este conjunto de variables explicativas “Conjunto B”. En esta estimación IV o 2SLS, trate LWW como una variable endógena y utilice las variables del Conjunto A definidas en la parte (b) para formar instrumentos. ¿Cómo se comparan sus resultados con los reportados por Mroz, reproducidos en la ecuación (11.51)? (Nota: Las estimaciones del error estándar de Mroz emplean el procedimiento de error estándar robusto de White. Si su software no permite esto, sus estimaciones del error estándar MCO diferirán, ligeramente, de las de la ecuación (11.51)). A continuación, permita la selectividad de la muestra, pero excluya las variables de experiencia. Específicamente, usando las mismas variables del Conjunto A más la variable de la inversa del ratio de Mills INVR1 para formar el instrumento para LWW, estime por IV o 2SLS una ecuación de horas trabajadas con las variables del Conjunto B como regresores (usando métodos de error estándar robustos si es posible), pero con INVR1 agregado como regresor. ¿Cómo se comparan sus resultados con los reportados por Mroz, dados en la ecuación (11.53)? ¿Es significativa la selectividad de la muestra? ¿Por qué o por qué no? (Nota: Las estimaciones del error estándar de Mroz se basan en una fórmula derivada de su Apéndice; sus estimaciones diferirán, ligeramente, de las del procedimiento de error estándar robusto de White). Luego, estime un modelo mediante 2SLS en el que se incluyen las variables de experiencia AX y AX2 junto con las variables del Conjunto A en la ecuación de determinación de salarios de la primera etapa, pero en la que no se tiene en cuenta la selectividad de la muestra y sólo las variables del Conjunto B son regresoras. Compare sus resultados con los de Mroz, reproducidos en la ecuación (11.50). Finalmente, incluya las variables de experiencia AX y AX2, las variables del Conjunto A y la inversa del ratio de Mills INVR2 como variables para formar el instrumento para LWW y, luego, estime mediante 2SLS la ecuación de horas trabajadas incluyendo como regresores las variables del Conjunto B y el INVR2. Sus resultados deben concordar, estrechamente, con los informados por Mroz, reproducidos en la ecuación (11.52). ¿Es importante la selectividad de la muestra cuando las variables de experiencia se tratan como exógenas? ¿Por qué o por qué no?

En la tabla 25, se presenta la estimación por MC2E de una ecuación de horas trabajadas, en la cual *whrs* se regresa en un término constante, *kl6*, *k618*, *wa*, *we*, *lww* (instrumentada mediante *kl6*, *k618*, *wa*, *we*, *wa2*, *we2*, *wawe*, *wa3*, *we3*, *wa2we*, *wawe2*, *wfed*, *wmed*, *un*, *cit* y *prin*) y *prin*, restringiendo la muestra a las mujeres que trabajaron. Se puede observar que los resultados son muy parecidos a los reportados por Mroz, reproducidos en la ecuación (11.51). En la tabla 26, se presenta la misma estimación anterior, pero incorporando la variable de la inversa del ratio de Mills *invr1*, tanto en la primera como en la segunda etapa de la estimación. Se puede observar que los resultados NO se parecen a los reportados por Mroz, reproducidos en la ecuación (11.53), y que la selectividad de la muestra no es estadísticamente significativa.

Tabla 25. Estimación por MC2E de ecuación de horas trabajadas (sin ajuste de selectividad y sin variables de experiencia en la primera etapa).

Linear regression	Number of obs	=	428
	F(6, 421)	=	3.89
	Prob > F	=	0.0008
	R-squared	=	0.0668
	Root MSE	=	755.22

	whrs	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]
kl6		-337.1825	132.6914	-2.54	0.011	-598.0026 -76.36234
k618		-112.2935	30.55218	-3.68	0.000	-172.3474 -52.23972
wa		-7.853428	5.836681	-1.35	0.179	-19.32609 3.619238
we		-21.03373	30.50365	-0.69	0.491	-80.99215 38.92469
prin		-.004448	.0033124	-1.34	0.180	-.010959 .0020629
lww_hat1		45.73968	222.167	0.21	0.837	-390.9551 482.4345
_cons		2127.531	355.3102	5.99	0.000	1429.128 2825.934

Fuente: Elaboración propia en base a Mroz (1987).

Tabla 26. Estimación por MC2E de ecuación de horas trabajadas (con ajuste de selectividad y sin variables de experiencia en la primera etapa).

Linear regression	Number of obs	=	428
	F(7, 420)	=	3.59
	Prob > F	=	0.0009
	R-squared	=	0.0672
	Root MSE	=	755.94

	whrs	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]
kl6		-248.8672	274.8461	-0.91	0.366	-789.1125 291.3781
k618		-109.3943	31.59931	-3.46	0.001	-171.5068 -47.28181
wa		-4.466792	10.34004	-0.43	0.666	-24.79146 15.85788
we		-37.51752	48.13674	-0.78	0.436	-132.1365 57.10142
prin		-.0022882	.0063867	-0.36	0.720	-.014842 .0102657
lww_hat3		26.56777	228.2598	0.12	0.907	-422.1062 475.2418
invr1		149.9619	366.4688	0.41	0.683	-570.3796 870.3034
_cons		1996.001	494.3158	4.04	0.000	1024.36 2967.642

Fuente: Elaboración propia en base a Mroz (1987).

En la tabla 27, se presenta la misma estimación que en la tabla 25, pero incorporando las variables de experiencia (ax y $ax2$) en la primera etapa de la estimación. Se puede observar que los resultados son muy parecidos a los reportados por Mroz, reproducidos en la ecuación (11.50). En la tabla 28, se presenta la misma estimación anterior, pero incorporando la variable de la inversa del ratio de Mills $invr2$, tanto en la primera como en la segunda etapa de la estimación. Se puede observar que los resultados NO se parecen a los reportados por Mroz, reportados en la ecuación (11.52) y que la selectividad de la muestra es estadísticamente significativa, por lo que pasa a ser importante cuando las variables de experiencia se tratan como exógenas.

Tabla 27. Estimación por MC2E de ecuación de horas trabajadas (sin ajuste de selectividad y con variables de experiencia en la primera etapa).

Linear regression	Number of obs	=	428
	F(6, 421)	=	7.45
	Prob > F	=	0.0000
	R-squared	=	0.0976
	Root MSE	=	742.65

	whrs	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]	
kl6		-284.3733	132.6096	-2.14	0.033	-545.0326	-23.71387
k618		-85.23599	30.13556	-2.83	0.005	-144.4709	-26.00109
wa		-9.078339	5.708645	-1.59	0.113	-20.29934	2.142658
we		-86.40976	25.99426	-3.32	0.001	-137.5045	-35.31505
prin		-.0064547	.0032506	-1.99	0.048	-.0128441	-.0000653
lww_hat2		672.3016	180.645	3.72	0.000	317.2231	1027.38
_cons		2254.871	347.0666	6.50	0.000	1572.672	2937.07

Fuente: Elaboración propia en base a Mroz (1987).

Tabla 28. Estimación por MC2E de ecuación de horas trabajadas (con ajuste de selectividad y con variables de experiencia en la primera etapa).

Linear regression	Number of obs	=	428
	F(7, 420)	=	8.93
	Prob > F	=	0.0000
	R-squared	=	0.1296
	Root MSE	=	730.24

	whrs	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]	
kl6		40.71379	163.0809	0.25	0.803	-279.8426	361.2702
k618		-92.2295	30.60343	-3.01	0.003	-152.3845	-32.07453
wa		3.509587	6.458685	0.54	0.587	-9.185788	16.20496
we		-87.186	25.53279	-3.41	0.001	-137.374	-36.99802
prin		.0038183	.0041052	0.93	0.353	-.004251	.0118877
lww_hat4		50.50553	210.3807	0.24	0.810	-363.0247	464.0358
invr2		555.0556	143.8739	3.86	0.000	272.253	837.8582
_cons		1552.281	398.2029	3.90	0.000	769.5618	2335

Fuente: Elaboración propia en base a Mroz (1987).

(d) Interprete sus hallazgos en la parte (c), comentando, en particular, la importancia de permitir la selectividad de la muestra y cómo esta sensibilidad se ve afectada por el supuesto de exogeneidad de las variables de experiencia. Mroz concluye que, con sus especificaciones preferidas, el efecto salario no compensado para su muestra es pequeño, al igual que el efecto ingreso estimado. ¿Estás de acuerdo? ¿Por qué o por qué no? (Es posible que desee comparar los resultados de aquí con los de otros estudios, presentados en la Tabla 11.2.).

La selectividad de la muestra se refiere al sesgo que puede surgir cuando la muestra observada no es representativa de la población objetivo debido a ciertos criterios

de selección (por ejemplo, solo mujeres que trabajan por salario). Los resultados muestran que:

- Sin variables de experiencia: La corrección por selectividad (*invr1*) no es significativa, sugiriendo que, en este caso, la selectividad de la muestra no afecta de manera importante las estimaciones del modelo.
- Con variables de experiencia: La corrección por selectividad (*invr2*) es significativa, sugiriendo que, en este caso, la selectividad de la muestra afecta en la determinación de las horas trabajadas cuando se consideran las variables de experiencia como exógenas. Esto resalta la importancia de ajustar por selectividad en modelos que incluyen estas variables, ya que su omisión puede llevar a estimaciones sesgadas.

Mroz concluye que, con sus especificaciones preferidas, el efecto salario no compensado para su muestra es pequeño, al igual que el efecto ingreso estimado. En nuestro caso, por un lado, los coeficientes asociados a las variables *lww_hat* son pequeños y, en la mayoría de los casos, no significativos y, por otro lado, los coeficientes de *prin* también son pequeños y, en la mayoría de los casos, no significativos. Por lo tanto, estoy de acuerdo con la conclusión de Mroz.

(e) Ahora, compare el procedimiento de la inversa del ratio de Mills de Heckit con el método del modelo de probabilidad lineal basado en MCO propuesto por Olsen [1980]. En particular, utilizando un modelo con o sin las variables de experiencia AX y AX2, siga el procedimiento de Olsen descrito en la Sección 11.3.B.1 debajo de la ecuación (11.37) y utilizar como regresor en las ecuaciones de determinación de la tasa salarial y de horas trabajadas, en lugar de la inversa del ratio de Mills, la probabilidad ajustada menos 1 de un modelo de probabilidad lineal estimado por MCO. Compare estos resultados con los obtenidos con el procedimiento de Heckit. ¿Difieren sustancialmente?

En las tablas 29 y 30, se presentan las estimaciones por MCO (Olsen, 1980) de una ecuación de salarios, en la cual *lww* se regresa en un término constante, *kl6*, *k618*, *wa*, *we*, *wa2*, *we2*, *wawe*, *wa3*, *we3*, *wa2we*, *wawe2*, *wfed*, *wmed*, *un*, *cit* y *prin* (*ax* y *ax2*, en la tabla 30) y, además, incorporando las variables de la probabilidad ajustada menos 1 *lfp_ols1* y *lfp_ols2*, respectivamente, restringiendo la muestra a las mujeres que trabajaron. Se puede observar que los resultados no difieren sustancialmente con los obtenidos con el procedimiento de Heckit.

Tabla 29. Estimación por MCO de ecuación de salarios mediante método Olsen (1980) (sin variables de experiencia).

Linear regression	Number of obs	=	428
	F(16, 411)	=	8.53
	Prob > F	=	0.0000
	R-squared	=	0.1772
	Root MSE	=	.66863

	lww	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]	
kl6		-.1266363	.1119148	-1.13	0.258	-.3466331	.0933605
k618		-.0464093	.0302085	-1.54	0.125	-.1057917	.012973
wa		.1113025	.4408207	0.25	0.801	-.755242	.9778469
we		-.8514843	1.147582	-0.74	0.459	-3.107346	1.404378
wa2		-.0060282	.0093899	-0.64	0.521	-.0244865	.0124301
we2		.0119921	.0675097	0.18	0.859	-.1207153	.1446994
wawe		.0301834	.0240799	1.25	0.211	-.0171519	.0775186
wa3		.0000348	.0000707	0.49	0.623	-.0001042	.0001738
we3		.0017814	.0015375	1.16	0.247	-.0012409	.0048038
wa2we		.0000746	.0002031	0.37	0.714	-.0003247	.0004738
wawe2		-.0015388	.0005256	-2.93	0.004	-.0025719	-.0005057
wfed		-.0168407	.011106	-1.52	0.130	-.0386724	.0049909
wmed		-.0090201	.0124633	-0.72	0.470	-.03352	.0154797
un		-.0015374	.0095299	-0.16	0.872	-.0202708	.017196
cit		.0789842	.0680702	1.16	0.247	-.0548249	.2127933
prin		1.86e-06	2.88e-06	0.65	0.518	-3.79e-06	7.52e-06
lfp_ols1		0	(omitted)				
_cons		2.668241	8.681064	0.31	0.759	-14.39658	19.73307

Fuente: Elaboración propia en base a Mroz (1987).

Tabla 30. Estimación por MCO de ecuación de salarios mediante método Olsen (1980) (con variables de experiencia).

Linear regression	Number of obs	=	428
	F(18, 409)	=	8.10
	Prob > F	=	0.0000
	R-squared	=	0.2100
	Root MSE	=	.65679

	lww	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]	
kl6		-.1055143	.1061322	-0.99	0.321	-.3141469	.1031183
k618		-.0171133	.0314647	-0.54	0.587	-.078966	.0447395
wa		.021614	.428565	0.05	0.960	-.820851	.864079
we		-.9660166	1.109505	-0.87	0.384	-3.147061	1.215028
wa2		-.0047736	.0092678	-0.52	0.607	-.0229921	.0134449
we2		.0141661	.0666777	0.21	0.832	-.1169076	.1452399
wawe		.0332155	.0228326	1.45	0.147	-.0116684	.0780994
wa3		.0000274	.0000702	0.39	0.696	-.0001106	.0001655
we3		.0018592	.0015278	1.22	0.224	-.0011441	.0048625
wa2we		.0000709	.0001932	0.37	0.714	-.0003089	.0004506
wawe2		-.0016391	.0005077	-3.23	0.001	-.0026371	-.000641
wfed		-.0129667	.0108686	-1.19	0.234	-.0343321	.0083986
wmed		-.0104555	.0121778	-0.86	0.391	-.0343944	.0134834
un		-.002731	.0093985	-0.29	0.772	-.0212064	.0157445
cit		.0698902	.0682422	1.02	0.306	-.0642589	.2040394
prin		4.61e-06	2.92e-06	1.58	0.116	-1.14e-06	.0000104
ax		.0398315	.0163393	2.44	0.015	.0077121	.0719509
ax2		-.0007028	.0004542	-1.55	0.123	-.0015957	.0001901
lfp_ols2		0	(omitted)				
_cons		4.379425	8.295024	0.53	0.598	-11.92678	20.68563

Fuente: Elaboración propia en base a Mroz (1987).

En las tablas 31 y 32, se presentan la estimaciones por MCO (Olsen, 1980) de una ecuación de horas trabajadas, en la cual *whrs* se regresa en un término constante, *kl6*, *k618*, *wa*, *we*, *lww* y *prin* (*ax* y *ax2*, en la tabla 32) y, además, incorporando las variables de la probabilidad ajustada menos 1 *lfp_ols1* y *lfp_ols2*, respectivamente, restringiendo la muestra a las mujeres que trabajaron. Se puede observar que los resultados no difieren sustancialmente con los obtenidos con el procedimiento de Heckit.

Tabla 31. Estimación por MCO de ecuación de horas trabajadas mediante método Olsen (1980) (sin variables de experiencia).

Linear regression	Number of obs	=	428
	F(7, 420)	=	3.94
	Prob > F	=	0.0003
	R-squared	=	0.0693
	Root MSE	=	755.13

	whrs	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]	
kl6		-69.06906	312.0502	-0.22	0.825	-682.4437	544.3056
k618		-103.0086	32.26865	-3.19	0.002	-166.4368	-39.58048
wa		2.480754	11.53572	0.22	0.830	-20.19417	25.15568
we		-66.99559	53.58872	-1.25	0.212	-172.3311	38.3399
prin		.0017307	.00668	0.26	0.796	-.0113998	.0148612
lww_hat5		5.772107	220.8881	0.03	0.979	-428.4119	439.9561
lfp_ols1		902.7641	847.5784	1.07	0.287	-763.26	2568.788
_cons		2485.489	473.3136	5.25	0.000	1555.131	3415.848

Fuente: Elaboración propia en base a Mroz (1987).

Tabla 32. Estimación por MCO de ecuación de horas trabajadas mediante método Olsen (1980) (con variables de experiencia).

Linear regression	Number of obs	=	428
	F(7, 420)	=	9.29
	Prob > F	=	0.0000
	R-squared	=	0.1360
	Root MSE	=	727.57

	whrs	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]	
kl6		38.56339	159.7224	0.24	0.809	-275.3914	352.5182
k618		-90.36106	30.69175	-2.94	0.003	-150.6896	-30.03248
wa		3.143651	6.363656	0.49	0.622	-9.364932	15.65223
we		-83.95151	25.90671	-3.24	0.001	-134.8745	-33.02854
prin		.0035319	.0039723	0.89	0.374	-.0042761	.01134
lww_hat6		88.95557	209.3017	0.43	0.671	-322.4538	500.365
lfp_ols2		1260.458	314.9426	4.00	0.000	641.3978	1879.518
_cons		2574.491	346.5177	7.43	0.000	1893.366	3255.616

Fuente: Elaboración propia en base a Mroz (1987).

(f) ¿En qué sentido el procedimiento de estimación de este ejercicio es un estimador Tobit generalizado?

El procedimiento de estimación de este ejercicio se asemeja a un estimador Tobit generalizado en el sentido de que combina un modelo Probit con una corrección por sesgo de selección. Aunque el enfoque estándar de Tobit se utiliza, típicamente, para manejar datos censurados en el contexto de variables dependientes continuas, el enfoque generalizado de Tobit se puede aplicar a problemas de selección en cualquier tipo de modelo, incluidos los modelos Probit para variables binarias.

El procedimiento Heckit, que se menciona en el ejercicio, es una extensión del estimador Tobit que se utiliza, específicamente, en el contexto de modelos de selección muestral. En este procedimiento, se estima un modelo Probit en la primera etapa para modelar la probabilidad de selección (por ejemplo, la participación en la fuerza laboral) y, luego, se utiliza la inversa del ratio de Mills calculado a partir de esta ecuación para corregir el sesgo de selección en la segunda etapa del modelo.

De manera similar al Tobit estándar, donde se corrige el sesgo de selección generado por la censura en la variable dependiente, el procedimiento Heckit corrige el sesgo de selección causado por la omisión de ciertos individuos en la muestra debido a la no observación de ciertas características.

En resumen, el procedimiento de estimación de este ejercicio es un estimador Tobit generalizado en el sentido de que aborda el problema de selección muestral al combinar un modelo Probit con una corrección por sesgo de selección en la segunda etapa del modelo.

Ejercicio 7: Incorporating Income Taxes Into a Model of Labor Supply.

El propósito de este ejercicio es permitirle evaluar los impactos de la incorporación de impuestos sobre el ingreso en sus estimaciones de la capacidad de respuesta de la oferta laboral a los salarios y los ingresos. Esto implicará crear variables de ingreso virtual y tasa salarial después de impuestos que sean consistentes con una especificación de restricción presupuestaria linealizada (LBC), estimando usando el procedimiento Tobit generalizado de Heckit y comparando los resultados con los reportados por Mroz [1987].

(a) *Nuestra primera tarea es crear varias variables relacionadas con los impuestos. Para toda la muestra de 753 observaciones en el archivo de datos MROZ, primero, cree y guarde una variable de ingreso de propiedad virtual definida como $VPRIN = (1 - MTR) * PRIN$, donde $PRIN$ se creó en la parte (c) del Ejercicio 1 y MTR es el margen variable de tasa impositiva proporcionada en el archivo de datos MROZ. A continuación, genere y guarde $LTAX = \text{LOG}(1 - MTR)$ y el logaritmo de una variable salarial después de impuestos como $LTWW = LTAX + LWW1$, donde $LWW1$ es la variable salarial creada en el Ejercicio 1.*

(Ver cálculos en *do-file*).

(b) *Para implementar el procedimiento de múltiples etapas de Heckit que permite la selectividad de la muestra, inicialmente, debemos estimar una ecuación probit LFP para toda la muestra de 753 observaciones y calcular, a partir de esta estimación, la inversa del ratio de Mills de Heckman. Siguiendo a Mroz, primero, genere las transformaciones polinómicas de las variables edad, educación y experiencia de la esposa, que se utilizarán como variables explicativas en la ecuación de LFP. En particular, para toda la muestra de 753 observaciones en el archivo de datos MROZ, genere y guarde $WA2 = WA * WA$, $WE2 = WE * WE$, $WA3 = WA2 * WA$, $WE3 = WE2 * WE$, $WAVE = WA * WE$, $WA2WE = WA2 * WE$ y $WAVE2 = WA * WE2$. Con esta muestra, estime por máxima verosimilitud un modelo probit en el que la LFP es la variable dependiente y las variables explicativas incluyen un término constante, $KL6$, $K618$, WA , WE , $WA2$, $WE2$, $WAVE$, $WA3$, $WE3$, $WA2WE$, $WAVE2$, $WFED$, $WMED$, UN , CIT y $PRIN$ (esta última variable se calculó en la parte (c) del Ejercicio 1). A partir de este modelo probit estimado, calcule la inversa del ratio de Mills para cada observación, guarde esta variable y llámela $INVR$. (Algunos programas de computadora ofrecen este cálculo como un comando opcional; para otros, debe calcularse mediante fuerza bruta usando la ecuación (11.37) y valores de la distribución normal).*

(Ver cálculos en *do-file*).

(c) *Ahora, estime un modelo con impuestos incluídos, similar al presentado en la ecuación (11.55), por 2SLS. En concreto, utilizando como variables exógenas un término constante, $KL6$, $K618$, WA , WE , $WA2$, $WE2$, $WAVE$, $WA3$, $WE3$, $WA2WE$, $WAVE2$, $WMED$, $WFED$, UN , CIT , $PRIN$ y la variable inversa del ratio de Mills $INVR$ del inciso (b), forme un instrumento para el logaritmo de la variable de tasa salarial después de impuestos $LTWW$ (creada en el inciso (a)). Con estos instrumentos para $LTWW$, estime*

mediante 2SLS un modelo en el que WHRS sea una función lineal de un término constante, LTWW, LTAX, KL6, K618, WA, WE, PRIN, VPRIN, INVR y un término de perturbación aleatoria. Si su software lo permite, obtenga las estimaciones robustas del error estándar de White. Comente el signo y la magnitud de los coeficientes estimados de salario, ingreso de la propiedad e impuestos. Usando la información proporcionada debajo de la ecuación (11.55) y su matriz de varianza-covarianza estimada de los coeficientes estimados, formule y pruebe la hipótesis de que las mujeres toman en cuenta, de manera óptima, los impuestos sobre el ingreso al tomar decisiones sobre la oferta laboral. Luego, formule y pruebe la hipótesis de que las mujeres ignoran por completo los impuestos sobre el ingreso cuando toman decisiones sobre la oferta laboral. Interprete los resultados de su prueba. ¿Concuerdan sus conclusiones con las de Mroz y Rosen [1976]? ¿Por qué o por qué no?

En la tabla 33, se presenta la estimación por MC2E de una ecuación de horas trabajadas, en la cual *whrs* se regresa en un término constante, *ltww* (instrumentada mediante *kl6*, *k618*, *wa*, *we*, *wa2*, *we2*, *wawe*, *wa3*, *we3*, *wa2we*, *wawe2*, *wfed*, *wmed*, *un*, *cit* y *prin*), *ltax*, *kl6*, *k618*, *wa*, *we*, *prin*, *vprin* e *invr* (inversa del ratio de Mills). Se puede observar que la variable *ltww* no es estadísticamente significativa, mientras que las variables *ltax*, *prin* y *vprin* sí lo son.

Tabla 33. Estimación por MC2E de ecuación de horas trabajadas (sin variables de experiencia).

Linear regression	Number of obs	=	753
	F(9, 743)	=	19.56
	Prob > F	=	0.0000
	R-squared	=	0.2751
	Root MSE	=	746.32

	whrs	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]
ltww_hat1		-84.97777	289.7	-0.29	0.769	-653.7059 483.7503
ltax		1375.743	214.4775	6.41	0.000	954.6887 1796.797
kl6		-98.85902	129.4066	-0.76	0.445	-352.9051 155.187
k618		15.41402	34.40144	0.45	0.654	-52.12157 82.94961
wa		-.7300246	6.28139	-0.12	0.908	-13.06141 11.60136
we		-26.84565	41.38082	-0.65	0.517	-108.0829 54.39161
prin		-.0880061	.022341	-3.94	0.000	-.131865 -.0441472
vprin		.1143824	.0300593	3.81	0.000	.0553711 .1733938
invr		518.5601	251.0986	2.07	0.039	25.6129 1011.507
_cons		3155.132	836.0154	3.77	0.000	1513.898 4796.365

Fuente: Elaboración propia en base a Mroz (1987).

Al probar la hipótesis nula de que las mujeres toman en cuenta, de manera óptima, los impuestos sobre el ingreso al tomar decisiones sobre la oferta laboral, se encuentra que ésta se rechaza (ver test en *do-file*). También, al probar la hipótesis nula de que las mujeres ignoran por completo los impuestos sobre el ingreso cuando toman decisiones sobre la oferta laboral, se encuentra que ésta se rechaza (ver test en *do-file*). Estos resultados concuerdan con las conclusiones de Rosen, pero no así con las de Mroz, que encuentra que ninguna de estas dos hipótesis nula puede rechazarse.

(d) *Experimente con otras dos especificaciones plausibles cualesquiera para las ecuaciones probit, determinación de salarios y/o horas trabajadas con impuestos incluidos, y comente sobre la sensibilidad de sus resultados a los cambios en las especificaciones. ¿Parecen sólidos los resultados de Mroz sobre la falta de importancia de los impuestos sobre el ingreso?*

En la tabla 34, se presenta la misma estimación anterior, pero incorporando las variables de experiencia (*ax* y *ax2*), tanto en la primera como en la segunda etapa de la estimación. Se puede observar que los resultados no varían mucho ante el cambio de especificación.

Sí, parecen sólidos los resultados de Mroz sobre la falta de importancia de los impuestos sobre el ingreso.

Tabla 34. *Estimación por MC2E de ecuación de horas trabajadas (con variables de experiencia).*

Linear regression

Number of obs = 753

F(11, 741) = 32.64

Prob > F = 0.0000

R-squared = 0.3661

Root MSE = 698.87

		Robust					
	whrs	Coefficient	std. err.	t	P> t	[95% conf. interval]	
ltww_hat2		-134.8892	258.2309	-0.52	0.602	-641.8405	372.0622
ltax		1159.754	197.8351	5.86	0.000	771.3694	1548.138
kl6		-114.8352	98.13238	-1.17	0.242	-307.4858	77.81541
k618		15.88057	30.18826	0.53	0.599	-43.38414	75.14527
wa		-7.524676	6.36223	-1.18	0.237	-20.01482	4.965467
we		-21.45393	31.13736	-0.69	0.491	-82.58187	39.67402
prin		-.0649567	.0193054	-3.36	0.001	-.1028566	-.0270569
vprin		.0853239	.026976	3.16	0.002	.0323653	.1382825
ax		29.64786	16.11428	1.84	0.066	-1.98721	61.28293
ax2		-.3491333	.3805019	-0.92	0.359	-1.096123	.3978567
invr_new		450.7461	202.7755	2.22	0.027	52.66327	848.829
_cons		2661.194	763.8603	3.48	0.001	1161.606	4160.782

Fuente: Elaboración propia en base a Mroz (1987).

Ejercicio 8: Specifying and Estimating an Extended Tobit Model.

En este ejercicio, especificará y estimará un modelo Tobit extendido de ecuaciones simultáneas en el que se estiman los parámetros de la tasa salarial, el salario de reserva (e, implícitamente, las horas de trabajo) y las ecuaciones de participación en la fuerza laboral estimadas, simultáneamente, mediante el uso del método de máxima verosimilitud de información completa (FIML), con restricciones de parámetros de ecuaciones cruzadas impuestas. Como se señaló en la Sección 11.3.B.1, Killingsworth [1983] ha llamado a este Procedimiento VI.

(a) Comenzamos especificando la ecuación salarial para los trabajadores únicamente, donde se especifica que LWW es una función lineal de un término constante, WA , $WA2$, WE , CIT , AX y un término de perturbación aleatoria ϵ_{Wi} , como en las ecuaciones (11.38) y (11.61). A continuación, especificamos la forma funcional de la ecuación del salario de reserva derivada de la función de tasa marginal de sustitución. En particular, supongamos que el logaritmo de la ecuación del salario de reserva sea LWR y que tenga la forma funcional como en la ecuación sobreidentificada (11.63), pero denotemos el término de perturbación en la ecuación (11.63) por ϵ_{Ri} , como en la ecuación (11.41). Finalmente, siguiendo a Heckman [1974b], especifique que las horas trabajadas, $WHRS$, son una proporción d de la diferencia entre LWW y LWR si $LWW_i > LWR_i$, y es cero si $LWW_i \leq LWR_i$, de manera análoga a las Ecs. (11.65a) y (11.65b). Finalmente, sustituya, analíticamente, en las ecuaciones de horas trabajadas (11.65a) y (11.65b), sus especificaciones por las ecuaciones LWW y LWR , y observe que, en la ecuación (11.65a), el término de perturbación, ahora denotado $d_{\epsilon_{Di}}$, es igual a $d_{\epsilon_{Di}} \equiv d(\epsilon_{Wi} - \epsilon_{Ri})$, que es similar a la ecuación (11.45a) pero con d reemplazando a b .

Se tienen las siguientes ecuaciones:

$$lww_i = g_0 + g_1 wa_i + g_2 wa2_i + g_3 we_i + g_4 cit_i + g_5 ax_i + \epsilon_{Wi} \quad (11.61)$$

$$lwr_i = b_0 + b_1 wa_i + b_2 wa2_i + b_3 we_i + b_4 k618_i + b_6 prin_i + b_7 un_i + \epsilon_{Ri} \quad (11.63)$$

$$whrs_i = d(lww_i - lwr_i) \quad \text{si } lww_i > lwr_i \quad (11.65a)$$

$$whrs_i = 0 \quad \text{si } lww_i \leq lwr_i \quad (11.65b)$$

Reemplazando (11.61) y (11.63) en (11.65a), se tiene:

$$whrs_i = d[(g_0 + g_1 wa_i + g_2 wa2_i + g_3 we_i + g_4 cit_i + g_5 ax_i + \epsilon_{Wi}) - (b_0 + b_1 wa_i + b_2 wa2_i + b_3 we_i + b_4 k618_i + b_6 prin_i + b_7 un_i + \epsilon_{Ri})]$$

$$whrs_i = d(g_0 + g_1 wa_i + g_2 wa2_i + g_3 we_i + g_4 cit_i + g_5 ax_i + \epsilon_{Wi} - b_0 - b_1 wa_i - b_2 wa2_i - b_3 we_i - b_4 k618_i - b_6 prin_i - b_7 un_i - \epsilon_{Ri})$$

$$whrs_i = d[(g_0 - b_0) + (g_1 - b_1) wa_i + (g_2 - b_2) wa2_i + (g_3 - b_3) we_i + g_4 cit_i + g_5 ax_i - b_4 k618_i - b_6 prin_i - b_7 un_i + (\epsilon_{Wi} - \epsilon_{Ri})]$$

$$whrs_i = d(g_0 - b_0) + d(g_1 - b_1) wa_i + d(g_2 - b_2) wa2_i + d(g_3 - b_3) we_i + d g_4 cit_i + d g_5 ax_i - d b_4 k618_i - d b_6 prin_i - d b_7 un_i + d(\epsilon_{Wi} - \epsilon_{Ri}).$$

$$whrs_i = d(g_0 - b_0) + d(g_1 - b_1) wa_i + d(g_2 - b_2) wa2_i + d(g_3 - b_3) we_i + d g_4 cit_i + d g_5 ax_i - d b_4 k618_i - d b_6 prin_i - d b_7 un_i + d_{\epsilon_{Di}} \quad \text{si } lww_i > lwr_i.$$

$$whrs_i = 0 \quad \text{si } lww_i \leq lwr_i.$$

(b) Ahora, supongamos que ϵ_{Wi} y ϵ_{Ri} tienen distribución normal conjunta con varianzas σ_W^2 y σ_R^2 y covarianza σ_{WR} . Esto implica que, en la ecuación de horas trabajadas que construyó en el inciso (a), el término de perturbación ϵ_{Di} se distribuye normalmente con varianza $\sigma_D^2 = \sigma_W^2 + \sigma_R^2 - 2\sigma_{WR}$ y se distribuye normalmente conjuntamente con ϵ_{Wi} , con covarianza $\sigma_{DW} = \sigma_W^2 - \sigma_{WR}$. Dados estos supuestos distributivos, escriba la función de verosimilitud para toda la muestra de 753 observaciones, análoga a la ecuación (11.47), donde d , ahora, reemplaza a b y los otros términos son como se definen en la ecuación (11.47).

(c) Utilizando el software de computadora apropiado y las 753 observaciones del archivo de datos MROZ, así como la función de verosimilitud construida en el inciso (b), estime por máxima verosimilitud los parámetros que aparecen en el modelo Tobit extendido que consiste en las ecuaciones de salario y horas trabajadas. Mroz informa que, cuando estimó un modelo similar a éste, obtuvo un gran coeficiente positivo y estadísticamente significativo en LWW, pero un coeficiente negativo y marginalmente significativo en PRIN. ¿Son sus resultados, cualitativamente, similares a los de Mroz?

(d) Finalmente, siguiendo a Mroz, especifique un modelo alternativo en el que la variable de experiencia AX se excluya de la ecuación de determinación de la tasa salarial. Como en los incisos (a) y (b), escriba la función de verosimilitud correspondiente y, luego, estime los parámetros utilizando el método FIML. Compare sus resultados con los obtenidos en el inciso (c). Mroz descubrió que, cuando se omitía la variable experiencia AX de la ecuación de determinación de salarios, el coeficiente FIML estimado sobre LWW en la ecuación de horas trabajadas era positivo pero mucho menor que cuando se incluía AX (y estadísticamente insignificamente diferente de cero). Sin embargo, el coeficiente estimado eficiente en PRIN fue negativo y estadísticamente significativo, pero menor en valor absoluto que cuando la variable AX se incluyó en la ecuación (11.61). ¿Sus resultados coinciden con los de Mroz? ¿Por qué o por qué no?

(e) El modelo especificado y estimado en este ejercicio se basa, implícitamente, en un supuesto importante sobre la continuidad de la relación de horas trabajadas. ¿Cuál es este supuesto y bajo qué consideraciones prácticas podría violarse? ¿Qué procedimientos de estimación están disponibles que no requieren este fuerte supuesto?

El supuesto importante sobre la continuidad de la relación de horas trabajadas implica que la variable horas trabajadas tiene un rango continuo de valores, es decir,

puede tomar cualquier valor dentro de un cierto intervalo. Este supuesto es fundamental para modelos como el Tobit, que asumen que la variable dependiente (en este caso, las horas trabajadas) está censurada en la parte inferior (es decir, hay una cantidad mínima de horas que un individuo puede trabajar), pero puede tomar valores continuos por encima de este límite.

Sin embargo, en la práctica, este supuesto podría violarse si hay discontinuidades en la relación de horas trabajadas. Por ejemplo, podría haber una restricción legal o contractual que impida a los trabajadores trabajar menos de un cierto número de horas, lo que crearía una discontinuidad en la relación de horas trabajadas en el límite inferior. También podría haber restricciones prácticas, como la disponibilidad de horas de trabajo por parte del empleador, que podrían hacer que la relación de horas trabajadas no sea continua.

Cuando este supuesto se viola, las estimaciones basadas en el modelo Tobit pueden estar sesgadas o ser inconsistentes. En tales casos, sería más apropiado utilizar otros métodos de estimación que no requieran el supuesto de continuidad de la relación de horas trabajadas. Algunos de estos métodos incluyen:

- Modelos de regresión truncada: En lugar de asumir que la variable dependiente está censurada en la parte inferior y es continua por encima de ese límite, los modelos de regresión truncada permiten que la variable dependiente esté truncada en ambas direcciones, es decir, puede tener valores observados sólo dentro de ciertos intervalos.
- Modelos de selección muestral: Estos modelos abordan, directamente, el problema de la selección muestral, que surge cuando la muestra observada no es representativa de la población de interés debido a la censura o al truncamiento de la variable dependiente.
- Modelos de conteo: Si las horas trabajadas se registran como conteos discretos (por ejemplo, el número de horas trabajadas en un período de tiempo determinado), los modelos de conteo como el modelo de Poisson o el modelo de conteo inflado con ceros podrían ser más apropiados.

SEGUNDA PARTE: Análisis de Supervivencia.

La base de datos utilizada para los ejercicios sale de un ensayo controlado aleatorizado (RCT) de tratamientos para perros diagnosticados con linfosarcoma en varias clínicas veterinarias.

El estudio involucró 7 clínicas y 300 perros que, aleatoriamente, fueron asignados a tratamientos de radioterapia, quimioterapia, ambos o ninguno. En este ensayo clínico, se incluyeron perros con edades hasta 15 años.

La variable de interés es el tiempo medido en días (meses) desde el diagnóstico hasta la muerte debido a la linfosarcoma. Se considera que los perros que murieron de otras causas o no volvieron a las entrevistas de seguimiento son eventos censurados.

Las variables del dataset están descritas en la próxima tabla. La base de datos para trabajar es: linfosarcoma.dta.

Variable	Descripción	Rango
id	No. de identificación de cada perro	1 – 300
clinic	No. de identificación de cada clínica	1 – 7
age	Edad del perro al momento del diagnóstico	1,4 – 14,4
rad	Radiación	0 =no, 1 = yes
chemo	Quimioterapia	0 =no, 1 =yes
died	muerte debido a linfosarcoma	0 = censored, 1 = died
days	tiempo (en días) desde el diagnóstico a la muerte (o censura)	7 – 1363
months	tiempo (en meses) desde el diagnóstico a la muerte (o censura)	0,45 – 45,5

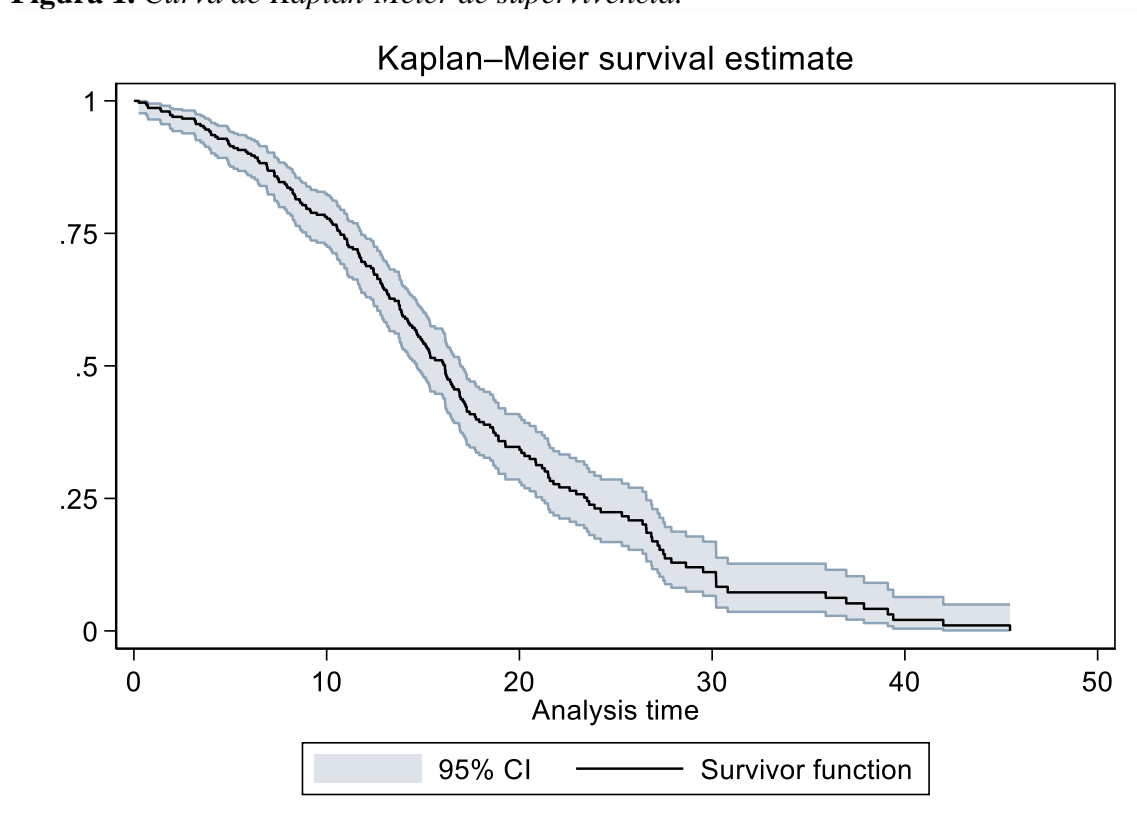
1. Introducción al Análisis de Supervivencia.

Ejercicio 1.

Generar una tabla de Kaplan-Meier para los datos mensuales. ¿Qué proporción de perros sobreviven 1 año, 2 años y 5 años? Generar la curva de supervivencia e incluir intervalos de confianza alrededor de dicha curva.

Se genera la tabla de Kaplan-Meier para los datos mensuales (ver cálculos en *do-file*). En ella, se puede observar que la proporción de perros que sobreviven 1 año, 2 años y 5 años, es 0,688, 0,224 y 0, respectivamente. En la figura 1, se presenta la curva de Kaplan-Meier de supervivencia, con intervalos de confianza al 95%.

Figura 1. Curva de Kaplan-Meier de supervivencia.



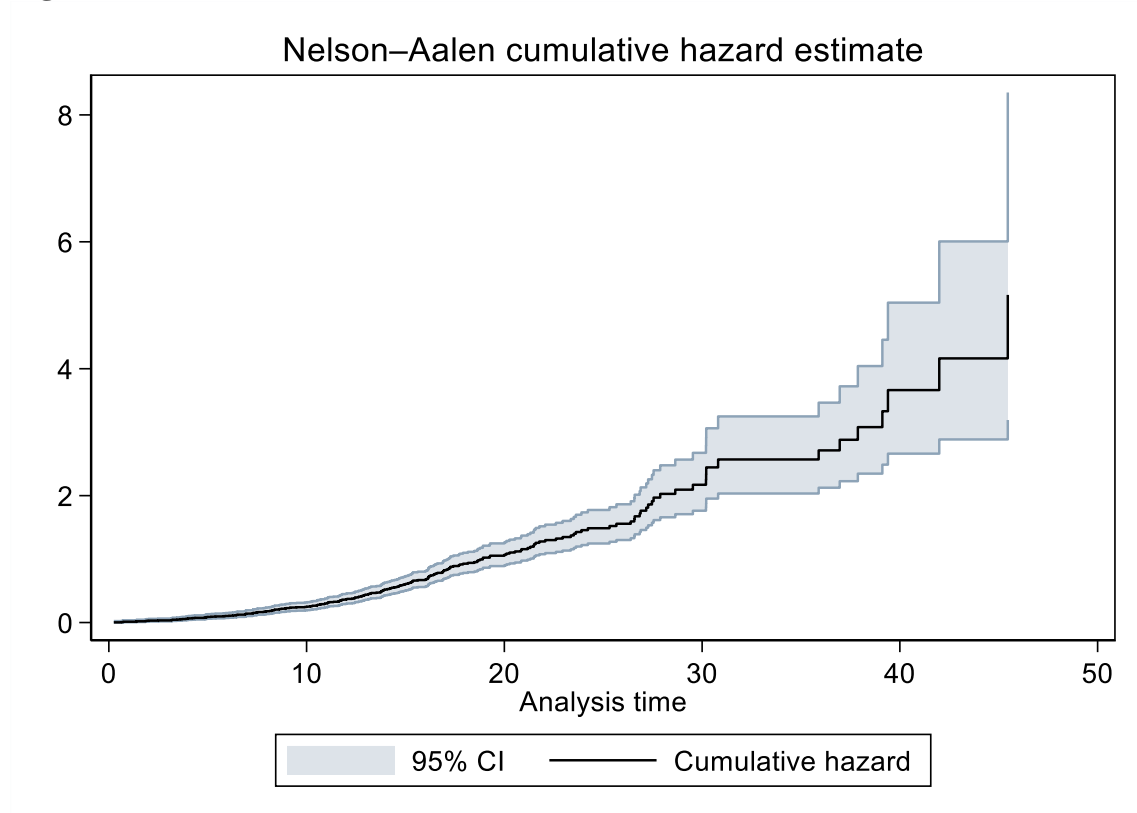
Fuente: Elaboración propia.

Ejercicio 2.

Generar la curva de Nelson-Aalen de hazard acumulada. ¿En qué punto la función de hazard acumulada llega a 1? Generar un gráfico con el hazard suavizada. Interpretar el gráfico de la función de hazard.

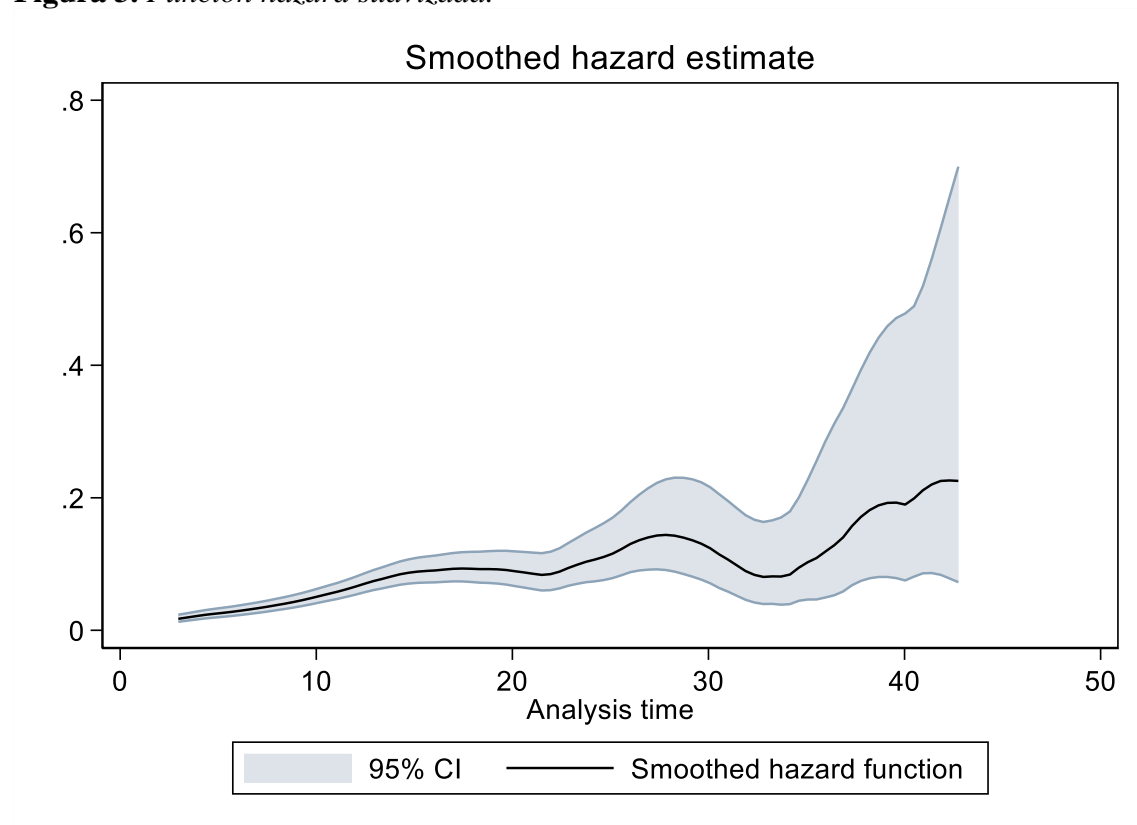
En la figura 2, se presenta la curva Nelson-Aalen de *hazard* acumulada. La función de *hazard* acumulada llega a 1 a los 18,91 meses.

Figura 2. Curva de Nelson-Aalen de hazard acumulada.



Fuente: Elaboración propia.

En la figura 3, se presenta la función *hazard* suavizada. Esta función permite visualizar cómo cambia la tasa de riesgo de ocurrencia de un evento a lo largo del tiempo; representa la probabilidad instantánea de que ocurra un evento en un momento dado, dado que el sujeto de estudio ha sobrevivido hasta ese momento; en otras palabras, es una medida de la tasa de riesgo en un momento específico del tiempo.

Figura 3. *Función hazard suavizada.*

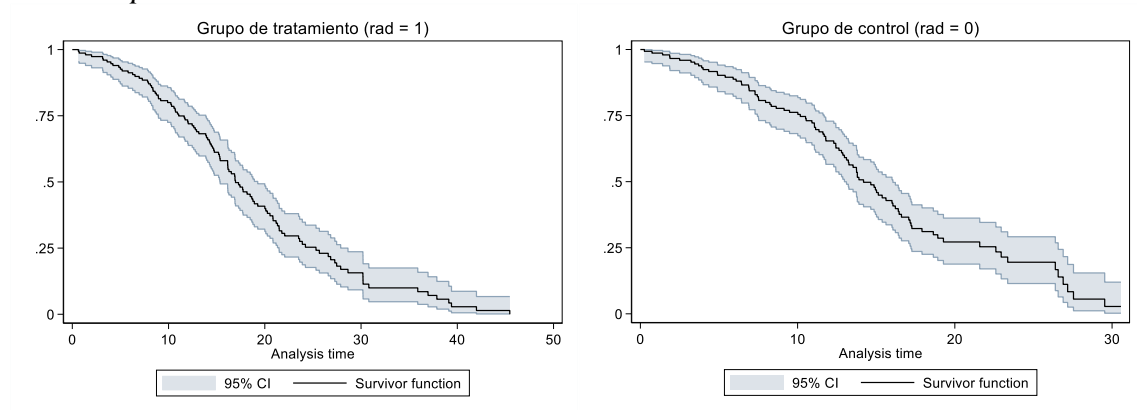
Fuente: Elaboración propia.

Ejercicio 3.

Generar la curva de Kaplan-Meier de supervivencia para evaluar los efectos de radiación en las expectativas de supervivencia de los perros de la muestra. Generar un gráfico de los tiempos de supervivencia para el grupo de tratamiento y para el grupo de control. Repetir el ejercicio para cuando se aplica quimioterapia.

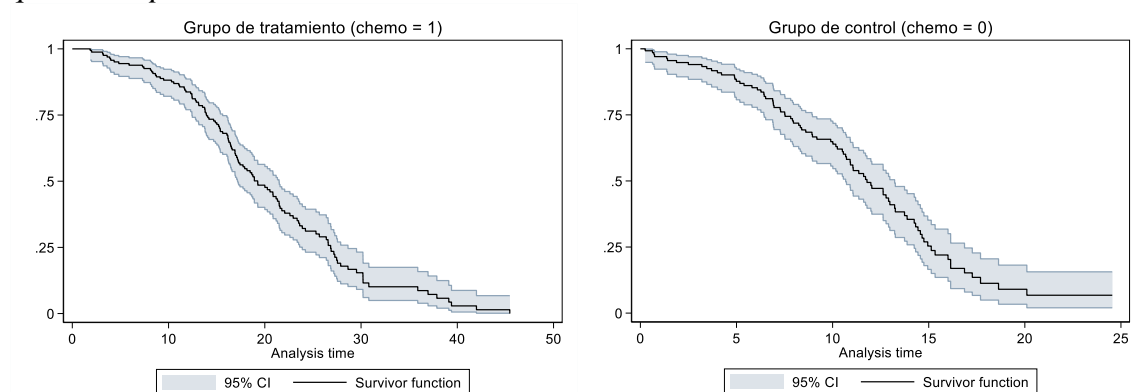
En las figuras 4 y 5, se presentan las curvas de Kaplan-Meier de supervivencia para grupo de tratamiento y grupo de control para los tratamientos de radioterapia y quimioterapia, respectivamente. En ambos casos, se puede observar que la curva correspondiente al grupo de tratamiento se encuentra más elevada que la correspondiente al grupo de control, indicando que, para los grupos de tratamiento, para cada momento del tiempo, hay una mayor proporción de perros que sobreviven.

Figura 4. Curvas de Kaplan-Meier de supervivencia para evaluar los efectos de la radioterapia.



Fuente: Elaboración propia.

Figura 5. Curvas de Kaplan-Meier de supervivencia para evaluar los efectos de la quimioterapia.



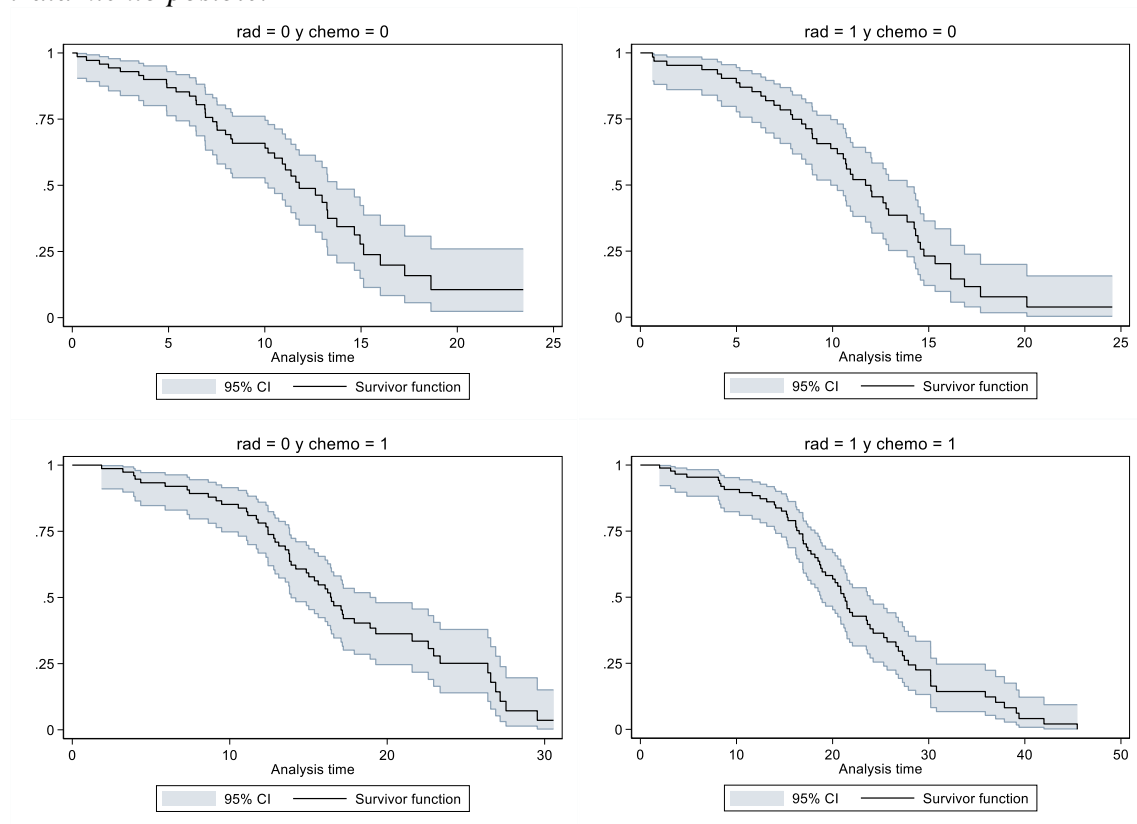
Fuente: Elaboración propia.

Ejercicio 4.

Crear una nueva variable que represente cada una de las cuatro combinaciones de tratamiento posibles. Generar un gráfico mostrando cada una de las funciones de supervivencia.

En la figura 6, se presenta la curva de Kaplan-Meier de supervivencia para cada de las cuatro combinaciones de tratamiento posibles. Se puede observar que la curva correspondiente al grupo que recibe los dos tratamientos ($rad=1$ y $chemo=1$) se encuentra más elevada que las otras tres curvas, indicando que, para este grupo, para cada momento del tiempo, hay una mayor proporción de perros que sobreviven.

Figura 6. Curva de Kaplan-Meier de supervivencia para cada combinación de tratamiento posible.



Fuente: Elaboración propia.

Ejercicio 5.

Usar un test de Wilcoxon (revisar la literatura) para determinar si hay diferencias estadísticamente significativas entre las cuatro combinaciones de tratamiento.

Se realizan dos test Wilcoxon por separado, uno para la variable *rad* y otro para la variable *chemo*¹. En ambos casos, con un nivel de significancia del 1%, estos datos aportan evidencia suficiente para indicar que los grupos analizados (por un lado, *rad*= 0 versus *rad*= 1 y, por otro lado, *chemo*= 0 y *chemo*= 1) son estadísticamente diferentes.

¹ No encontré un test de Wilcoxon que me pudiera analizar una variable con más de dos categorías.

2. Estimando un Modelo de Cox (Cox Proportional Hazard Model).**Ejercicio 1.**

Estimar un modelo de Cox con *chemo* y *rad* como predictores. ¿Qué efecto tienen los tratamientos en las probabilidades de supervivencia? Usar el tiempo definido en días.

En la tabla 1 (tabla 2), se presentan los coeficientes (*hazard ratios*) de un modelo de Cox con *age*, *rad* y *chemo* como predictores. Se puede observar que ambos tratamientos tienen un efecto positivo sobre la probabilidad de supervivencia, ya que, como ambos coeficientes estimados son negativos, los perros que reciben los tratamientos tienen un riesgo relativo más bajo de experimentar el estado “muerte” en comparación con los perros que no reciben los tratamientos. En relación a esto, en la tabla 2, se observa que los *hazard ratio* de estas variables son menores a 1 (lo cual es consistente con coeficientes estimados negativos) y, dado que el *hazard* es una medida inversa de la supervivencia, un *hazard ratio* menor 1 implica una mayor (menor) supervivencia para la categoría correspondiente a *dummy*= 1 (*dummy*= 0), siendo, en este caso, *dummy* igual a *rad* o a *chemo*.

Tabla 1. Modelo de Cox 1 (coeficientes).

Cox regression with Breslow method for ties

No. of subjects =	300	Number of obs =	300
No. of failures =	207		
Time at risk =	127,744		
Log likelihood =	-931.6115	LR chi2(3) =	72.93
		Prob > chi2 =	0.0000

_t	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
age	.1045706	.0273839	3.82	0.000	.0508991	.1582422
rad	-.2917368	.1466947	-1.99	0.047	-.5792531	-.0042204
chemo	-1.233281	.1649521	-7.48	0.000	-1.556581	-.9099809

Fuente: Elaboración propia.

Tabla 2. Modelo de Cox 1 (hazard ratios).

Cox regression with Breslow method for ties

No. of subjects =	300	Number of obs =	300
No. of failures =	207		
Time at risk =	127,744		
Log likelihood =	-931.6115	LR chi2(3) =	72.93
		Prob > chi2 =	0.0000

_t	Haz. ratio	Std. err.	z	P> z	[95% conf. interval]	
age	1.110234	.0304026	3.82	0.000	1.052217	1.17145
rad	.7469651	.1095758	-1.99	0.047	.5603167	.9957885
chemo	.2913351	.0480563	-7.48	0.000	.2108557	.4025319

Fuente: Elaboración propia.

Ejercicio 2.

El estudio fue realizado en 7 clínicas y podría ser una confounding variable. Agregar la variable clinic al modelo y determinar si agregar esta variable tiene algún efecto en los coeficientes de chemo y rad.

En la tabla 3 (tabla 4), se presentan los coeficientes (*hazard ratios*) de un modelo de Cox con *age*, *rad* y *chemo* como predictores, pero, ahora, agregando a la variable *clinic*. Se puede observar que agregar esta variable no tiene un efecto muy relevante en los coeficientes de *chemo* y *rad*, ya que estos varían mínimamente, al igual que su significatividad estadística (la cual empeora un poco para el caso de la variable *rad*).

Tabla 3. Modelo de Cox 2 (coeficientes).

Cox regression with Breslow method for ties

No. of subjects =	300	Number of obs =	300
No. of failures =	207		
Time at risk =	127,744		
Log likelihood =	-925.3136	LR chi2(9) =	85.52
		Prob > chi2 =	0.0000

_t	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
age	.1156871	.0280631	4.12	0.000	.0606844	.1706899
rad	-.2796114	.148758	-1.88	0.060	-.5711717	.0119488
chemo	-1.254296	.1675252	-7.49	0.000	-1.582639	-.9259523
_Iclinic_2	-.4824995	.2881326	-1.67	0.094	-1.047229	.0822301
_Iclinic_3	-.0506161	.2554809	-0.20	0.843	-.5513494	.4501171
_Iclinic_4	-.2359797	.2552263	-0.92	0.355	-.7362141	.2642547
_Iclinic_5	-.0951635	.2496705	-0.38	0.703	-.5845086	.3941817
_Iclinic_6	-.3706164	.267898	-1.38	0.167	-.8956869	.1544541
_Iclinic_7	-.7563722	.2692629	-2.81	0.005	-1.284118	-.2286267

Fuente: Elaboración propia.

Tabla 4. Modelo de Cox 2 (hazard ratios).

Cox regression with Breslow method for ties

No. of subjects =	300	Number of obs =	300
No. of failures =	207		
Time at risk =	127,744		
Log likelihood =	-925.3136	LR chi2(9) =	85.52
		Prob > chi2 =	0.0000

_t	Haz. ratio	Std. err.	z	P> z	[95% conf. interval]	
age	1.122645	.0315049	4.12	0.000	1.062564	1.186123
rad	.7560775	.1124726	-1.88	0.060	.5648632	1.012021
chemo	.2852767	.047791	-7.49	0.000	.2054322	.396154
_Iclinic_2	.6172387	.1778466	-1.67	0.094	.3509087	1.085706
_Iclinic_3	.9506435	.2428712	-0.20	0.843	.5761718	1.568496
_Iclinic_4	.7897967	.2015769	-0.92	0.355	.4789236	1.30246
_Iclinic_5	.9092243	.2270065	-0.38	0.703	.5573797	1.48317
_Iclinic_6	.6903087	.1849323	-1.38	0.167	.408327	1.167021
_Iclinic_7	.4693661	.1263829	-2.81	0.005	.2768947	.7956255

Fuente: Elaboración propia.

Ejercicio 3.

*Utilizar tanto un test de Wald como un test de cociente de verosimilitud para evaluar la significatividad de la variable *clinic* como predictor. Utilizar el criterio de información de Akaike para determinar si el modelo que mejor ajusta los datos es el que incorpora la variable *clinic* o el que no la incorpora.*

Se pretende evaluar la significatividad de la variable *clinic* como predictor. Por un lado, utilizando un test de Wald, con un nivel de significancia del 5%, la variable *clinic* no es estadísticamente significativa. Por otro lado, utilizando un test de cociente de verosimilitud, con un nivel de significancia del 5%, se rechaza la hipótesis nula de que el modelo más simple (el que no incorpora la variable *clinic*) es el que mejor ajusta a los datos. Por último, al utilizar el criterio de información de Akaike para determinar el modelo que mejor ajusta a los datos, resulta que el modelo que incorpora la variable *clinic* es el que tiene un menor AIC (1.868,627 versus 1.869,223), por lo que se concluye que este modelo es el que mejor ajusta a los datos.

Ejercicio 4.

¿Hay alguna evidencia de la interacción entre *chemo* y *rad*? ¿Cuál es el efecto de recibir ambos tratamientos sobre la probabilidad de morir?

En la tabla 5 (tabla 6), se presentan los coeficientes (*hazard ratios*) de un modelo de Cox con *age*, *rad*, *chemo* y *clinic* como predictores, pero, ahora, agregando a la interacción entre *rad* y *chemo*. Se puede observar que el efecto de recibir ambos tratamientos sobre la probabilidad de morir es negativo, es decir, la interacción de ambos tratamientos tiene un efecto positivo sobre la probabilidad de supervivencia. Además, es notable destacar que, ahora, cuando sólo se recibe el tratamiento de radioterapia, el efecto sobre la probabilidad de morir es positivo (anteriormente, cuando no incorporábamos la interacción, era negativo), aunque no es estadísticamente significativo.

Tabla 5. Modelo de Cox 3 (coeficientes).

Cox regression with Breslow method for ties

No. of subjects =	300	Number of obs =	300
No. of failures =	207		
Time at risk =	127,744		
Log likelihood =	-920.29014	LR chi2(10) =	95.57
		Prob > chi2 =	0.0000

_t	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
age	.1252957	.028208	4.44	0.000	.0700091	.1805824
rad#chemo						
no#yes	-.8118311	.218913	-3.71	0.000	-1.240893	-.3827696
yes#no	.2554841	.2232517	1.14	0.252	-.1820812	.6930493
yes#yes	-1.51176	.2268242	-6.66	0.000	-1.956327	-1.067193
_Iclinic_2	-.5079383	.2887136	-1.76	0.079	-1.073807	.0579301
_Iclinic_3	-.1381408	.2576148	-0.54	0.592	-.6430565	.3667748
_Iclinic_4	-.3703841	.2603957	-1.42	0.155	-.8807502	.1399821
_Iclinic_5	-.1969114	.252787	-0.78	0.436	-.6923647	.298542
_Iclinic_6	-.4020562	.2680656	-1.50	0.134	-.927455	.1233427
_Iclinic_7	-.9039418	.2735173	-3.30	0.001	-1.440026	-.3678576

Fuente: Elaboración propia.

Tabla 6. Modelo de Cox 3 (hazard ratios).

Cox regression with Breslow method for ties

No. of subjects = 300
 No. of failures = 207
 Time at risk = 127,744
 Log likelihood = -920.29014
 Number of obs = 300
 LR chi2(10) = 95.57
 Prob > chi2 = 0.0000

_t	Haz. ratio	Std. err.	z	P> z	[95% conf. interval]	
age	1.133484	.0319733	4.44	0.000	1.072518	1.197915
rad#chemo						
no#yes	.4440442	.097207	-3.71	0.000	.289126	.68197
yes#no	1.291086	.2882372	1.14	0.252	.8335337	1.999804
yes#yes	.2205216	.0500196	-6.66	0.000	.1413767	.3439728
_Iclinic_2	.6017349	.1737291	-1.76	0.079	.3417053	1.059641
_Iclinic_3	.870976	.2243763	-0.54	0.592	.5256832	1.443073
_Iclinic_4	.6904691	.1797952	-1.42	0.155	.4144719	1.150253
_Iclinic_5	.8212634	.2076047	-0.78	0.436	.5003914	1.347892
_Iclinic_6	.6689432	.1793206	-1.50	0.134	.3955591	1.131272
_Iclinic_7	.4049702	.1107664	-3.30	0.001	.2369216	.6922157

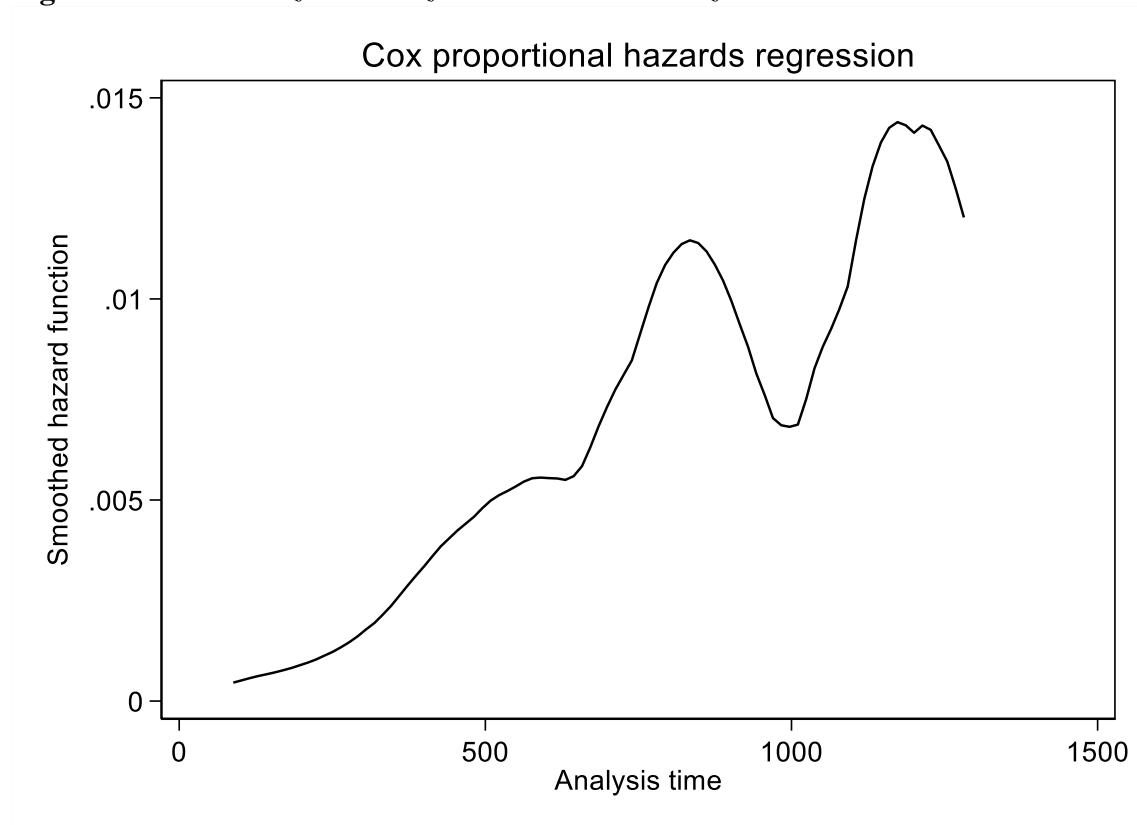
Fuente: Elaboración propia.

Ejercicio 5.

¿Cuál es la forma de la hazard en el baseline hazard?

En la figura 7, se presenta la función *hazard* en el *baseline hazard* (considerando el modelo de Cox 3).

Figura 7. Función hazard suavizada en el baseline hazard.



Fuente: Elaboración propia.

Ejercicio 6.

Realizar un análisis estratificado por clínica (*strata(clinic)*) de *chemo*, *rad* y su interacción. Comparar este modelo con el que incluye la variable *clinic* como efecto fijo. ¿Cómo difieren estos modelos? ¿Cuál tiene la mayor log-verosimilitud? ¿Son los coeficientes para las variables de tratamiento similares? Explicar, en sus propias palabras, cuáles son los efectos de cada uno de los tratamientos sobre la probabilidad de que un perro sobreviva. Estimar un modelo que determine si el efecto de quimioterapia varía de una clínica a la otra.

En la tabla 7 (tabla 8), se presentan los coeficientes (*hazard ratios*) de un modelo de Cox estratificado por clínica con *age*, *rad*, *chemo* y su interacción como predictores. Al comparar este modelo con el que incluye la variable *clinic* como efecto fijo, se puede observar que el modelo aquí estimado, por un lado, tiene un mayor efecto absoluto de *chemo* y de la interacción y, por otro lado, tiene la mayor log-verosimilitud (menos negativa).

Los efectos de cada uno de los tratamientos son:

- *rad*= 0 y *chemo*= 1: efecto positivo sobre la probabilidad de que un perro sobreviva de recibir el tratamiento de quimioterapia pero no de radioterapia.
- *rad*= 1 y *chemo*= 0: efecto negativo (pero no significativo) sobre la probabilidad de que un perro sobreviva de recibir el tratamiento de radioterapia pero no de quimioterapia.
- *rad*= 1 y *chemo*= 1: efecto positivo sobre la probabilidad de que un perro sobreviva de recibir ambos tratamientos.

Tabla 7. Modelo de Cox 4 (coeficientes).

Stratified Cox regression with Breslow method for ties
Strata variable: clinic

No. of subjects =	300	Number of obs =	300
No. of failures =	207		
Time at risk =	127,744		
Log likelihood =	-541.46723	LR chi2(4) =	77.11
		Prob > chi2 =	0.0000

_t	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
age	.110702	.0290864	3.81	0.000	.0536938	.1677102
rad#chemo						
no#yes	-.8747791	.2255114	-3.88	0.000	-1.316773	-.4327849
yes#no	.1865514	.2313938	0.81	0.420	-.266972	.6400749
yes#yes	-1.508527	.2349946	-6.42	0.000	-1.969108	-1.047947

Fuente: Elaboración propia.

Tabla 8. Modelo de Cox 4 (hazard ratios).

Stratified Cox regression with Breslow method for ties
Strata variable: clinic

No. of subjects =	300	Number of obs =	300
No. of failures =	207		
Time at risk =	127,744		
Log likelihood =	-541.46723	LR chi2(4) =	77.11
		Prob > chi2 =	0.0000

_t	Haz. ratio	Std. err.	z	P> z	[95% conf. interval]	
age	1.117062	.0324913	3.81	0.000	1.055161	1.182594
rad#chemo						
no#yes	.4169541	.0940279	-3.88	0.000	.2679987	.6487
yes#no	1.205087	.2788496	0.81	0.420	.7656945	1.896623
yes#yes	.2212355	.0519891	-6.42	0.000	.1395813	.3506571

Fuente: Elaboración propia.

En la tabla 9 (tabla 10), se presentan los coeficientes (*hazard ratios*) de un modelo de Cox con *age*, *rad*, *chemo*, su interacción y la interacción entre *chemo* y *clinic* como predictores. Se puede observar que la gran mayoría de las interacciones no son estadísticamente significativas y, para los casos en que sí lo son, no hay grandes diferencias en sus efectos, por lo que concluiría que el efecto de quimioterapia no varía de una clínica a otra.

Tabla 9. Modelo de Cox 5 (coeficientes).

Cox regression with Breslow method for ties

No. of subjects = 300
 No. of failures = 207
 Time at risk = 127,744
 Log likelihood = -918.59407
 Number of obs = 300
 LR chi2(16) = 98.96
 Prob > chi2 = 0.0000

_t	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
age	.1256619	.0294777	4.26	0.000	.0678867	.1834371
rad#chemo						
no#yes	-1.769954	.4163616	-4.25	0.000	-2.586008	-.9539003
yes#no	.3499103	.236425	1.48	0.139	-.1134742	.8132948
yes#yes	-2.450626	.430462	-5.69	0.000	-3.294316	-1.606936
chemo#clinic						
no#2	-.4072162	.4230318	-0.96	0.336	-1.236343	.4219109
no#3	-.3877265	.4126055	-0.94	0.347	-1.196418	.4209654
no#4	-.6833391	.4369149	-1.56	0.118	-1.539677	.1729984
no#5	-.3144385	.4064653	-0.77	0.439	-1.111096	.4822188
no#6	-.2561675	.4422568	-0.58	0.562	-1.122975	.6106399
no#7	-1.121181	.4751976	-2.36	0.018	-2.052551	-.1898104
yes#1	.8234076	.3352236	2.46	0.014	.1663813	1.480434
yes#2	.1344455	.4162758	0.32	0.747	-.6814401	.950331
yes#3	.8558539	.3416664	2.50	0.012	.1862	1.525508
yes#4	.641404	.3314385	1.94	0.053	-.0082036	1.291012
yes#5	.6625776	.3404581	1.95	0.052	-.0047081	1.329863
yes#6	.3410004	.3441839	0.99	0.322	-.3335876	1.015588
yes#7	0	(omitted)				

Fuente: Elaboración propia.

Tabla 10. Modelo de Cox 5 (hazard ratios).

Cox regression with Breslow method for ties

No. of subjects =	300	Number of obs =	300
No. of failures =	207		
Time at risk =	127,744		
Log likelihood =	-918.59407	LR chi2(16) =	98.96
		Prob > chi2 =	0.0000

_t	Haz. ratio	Std. err.	z	P> z	[95% conf. interval]	
age	1.133899	.0334247	4.26	0.000	1.070244	1.201339
rad#chemo						
no#yes	.1703408	.0709234	-4.25	0.000	.0753201	.3852356
yes#no	1.41894	.335473	1.48	0.139	.8927272	2.255327
yes#yes	.0862395	.0371228	-5.69	0.000	.0370934	.2005009
chemo#clinic						
no#2	.6655003	.2815278	-0.96	0.336	.2904443	1.524873
no#3	.6785979	.2799932	-0.94	0.347	.3022749	1.523432
no#4	.5049281	.2206107	-1.56	0.118	.2144504	1.188864
no#5	.7301987	.2968005	-0.77	0.439	.329198	1.619664
no#6	.7740123	.3423122	-0.58	0.562	.3253106	1.841609
no#7	.3258948	.1548644	-2.36	0.018	.128407	.8271159
yes#1	2.27825	.7637232	2.46	0.014	1.181023	4.394852
yes#2	1.143902	.4761788	0.32	0.747	.5058879	2.586566
yes#3	2.353383	.804072	2.50	0.012	1.204663	4.597478
yes#4	1.899145	.62945	1.94	0.053	.99183	3.636463
yes#5	1.939786	.6604159	1.95	0.052	.995303	3.780527
yes#6	1.406354	.4840443	0.99	0.322	.7163491	2.760987
yes#7	1	(omitted)				

Fuente: Elaboración propia.

Ejercicio 7.

Estimar un modelo que use como variable explicativa una variable categórica que represente los cuatro grupos de tratamiento (es decir, cada uno de los dos tratamientos y su interacción) y efectos fijos por clínica y evaluar el supuesto de proportional hazard utilizando un gráfico de la función hazard acumulada. ¿Qué se puede concluir?

En la tabla 11 (tabla 12), se presentan los coeficientes (*hazard ratios*) de un modelo de Cox con *age*, *terapy* (considera las cuatro combinaciones de tratamientos posibles) y *clinic* como predictores. Como era de esperar, los resultados son, exactamente, los mismos que los obtenidos en el modelo de Cox 3 (Ejercicio 4).

Tabla 11. Modelo de Cox 6 (coeficientes).

Cox regression with Breslow method for ties

No. of subjects =	300	Number of obs =	300
No. of failures =	207		
Time at risk =	127,744		
Log likelihood =	-920.29014	LR chi2(10) =	95.57
		Prob > chi2 =	0.0000

_t	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
age	.1252957	.028208	4.44	0.000	.0700091	.1805824
_Iterapy_2	-.8118311	.218913	-3.71	0.000	-1.240893	-.3827696
_Iterapy_3	.2554841	.2232517	1.14	0.252	-.1820812	.6930493
_Iterapy_4	-1.51176	.2268242	-6.66	0.000	-1.956327	-1.067193
_Iclinic_2	-.5079383	.2887136	-1.76	0.079	-1.073807	.0579301
_Iclinic_3	-.1381408	.2576148	-0.54	0.592	-.6430565	.3667748
_Iclinic_4	-.3703841	.2603957	-1.42	0.155	-.8807502	.1399821
_Iclinic_5	-.1969114	.252787	-0.78	0.436	-.6923647	.298542
_Iclinic_6	-.4020562	.2680656	-1.50	0.134	-.927455	.1233427
_Iclinic_7	-.9039418	.2735173	-3.30	0.001	-1.440026	-.3678576

Fuente: Elaboración propia.

Cox regression with Breslow method for ties

	_t	Haz. ratio	Std. err.	z	P> z	[95% conf. interval]	
	age	1.133484	.0319733	4.44	0.000	1.072518	1.197915
	_Iterapy_2	.4440442	.097207	-3.71	0.000	.289126	.68197
	_Iterapy_3	1.291086	.2882372	1.14	0.252	.8335337	1.999804
	_Iterapy_4	.2205216	.0500196	-6.66	0.000	.1413767	.3439728
	_Iclinic_2	.6017349	.1737291	-1.76	0.079	.3417053	1.059641
	_Iclinic_3	.870976	.2243763	-0.54	0.592	.5256832	1.443073
	_Iclinic_4	.6904691	.1797952	-1.42	0.155	.4144719	1.150253
	_Iclinic_5	.8212634	.2076047	-0.78	0.436	.5003914	1.347892
	_Iclinic_6	.6689432	.1793206	-1.50	0.134	.3955591	1.131272
	_Iclinic_7	.4049702	.1107664	-3.30	0.001	.2369216	.6922157

En la figura 8, se presenta la curva Nelson-Aalen de *hazard* acumulada para la variable *therapy*. Se puede concluir que se cumple el supuesto de *proportional hazard*, es decir, que la relación entre la variable explicativa y el riesgo de ocurrencia del evento de interés es constante a lo largo del tiempo, ya que las curvas son, aproximadamente, paralelas a lo largo del tiempo.

The figure is a line plot titled "Nelson-Aalen cumulative hazard estimates". The x-axis is labeled "Analysis time" and ranges from 0 to 1500 with major ticks every 500 units. The y-axis ranges from 0.00 to 5.00 with major ticks every 1.00 unit. There are four data series, each represented by a step function of a different color: blue for "therapy = 1", green for "therapy = 3", maroon for "therapy = 2", and orange for "therapy = 4". All curves start at (0, 0). The blue curve (therapy 1) rises to about 2.1 by time 600 and then plateaus. The green curve (therapy 3) rises to about 2.9 by time 700 and then plateaus. The maroon curve (therapy 2) rises to about 3.0 by time 900 and then plateaus. The orange curve (therapy 4) rises more gradually, reaching about 4.5 by time 1350.

Analysis time	therapy = 1	therapy = 2	therapy = 3	therapy = 4
0	0.00	0.00	0.00	0.00
250	0.40	0.20	0.40	0.10
500	1.60	0.80	1.80	0.30
750	2.10	1.40	2.90	0.90
1000	2.10	1.90	2.90	1.90
1250	2.10	2.90	2.90	3.50
1350	2.10	2.90	2.90	4.50

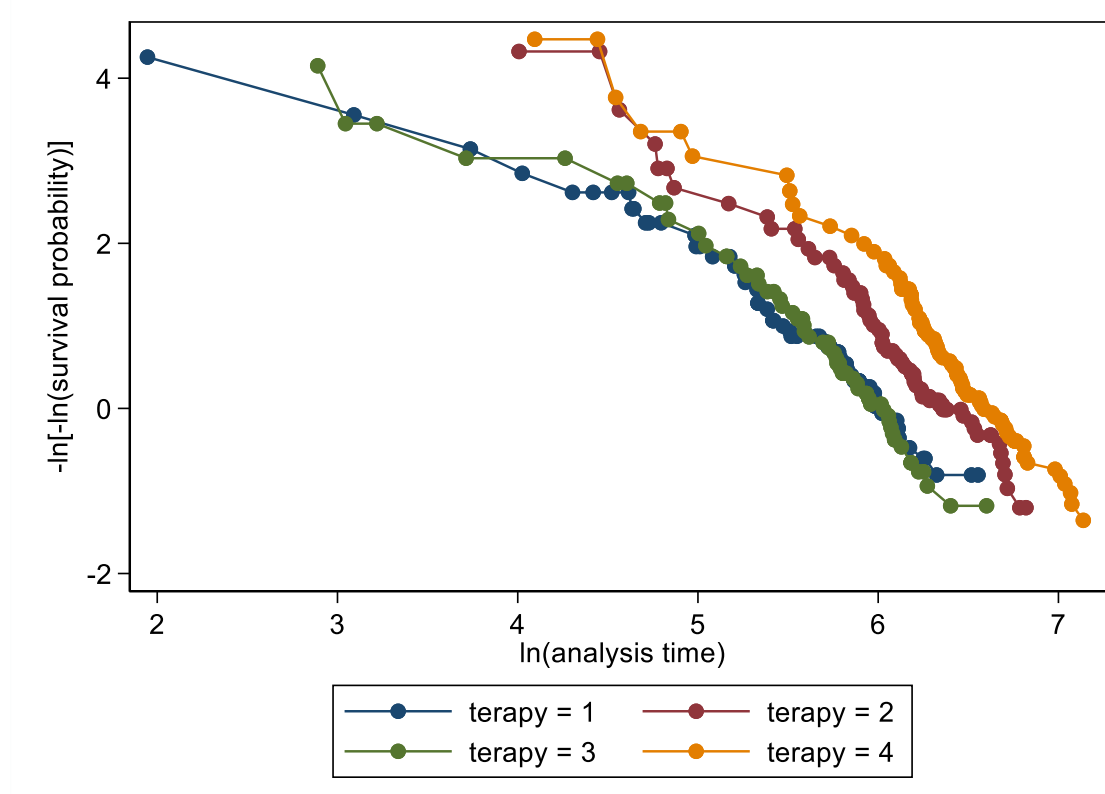
Fuente: Elaboración propia.

Ejercicio 8.

Evaluar también el supuesto de proportional hazard utilizando gráficos de residuos Schoenfeld escalados y un test estadístico de proporcionalidad basado en los residuos de Schoenfeld (ayuda: ver schoenfeld y scaledsch en Stata).

En la figura 9 y en la tabla 13, se evalúa el supuesto de *proportional hazard* utilizando gráficos de residuos Schoenfeld escalados y un test estadístico de proporcionalidad basado en los residuos de Schoenfeld, respectivamente. En el primer caso, se puede observar que las curvas son, relativamente, paralelas, por lo que no se viola el supuesto de *proportional hazard*. En el segundo caso, no se rechaza la hipótesis nula de *proportional hazard*. Por lo tanto, se puede concluir que se cumple el supuesto de *proportional hazard*.

Figura 9. Gráfico de residuos Schoenfeld escalados.



Fuente: Elaboración propia.

Tabla 13. Test estadístico de proporcionalidad basado en los residuos de Schoenfeld.

Test of proportional-hazards assumption

Time function: Analysis time

	chi2	df	Prob>chi2
Global test	4.66	10	0.9128

Fuente: Elaboración propia.

3. Modelos Paramétricos.

En este ejercicio, se tienen que calcular estimaciones para diferentes tipos de modelos de supervivencia paramétricos. Las variables explicativas de los modelos serán: la edad del diagnóstico, su cuadrado y los tratamientos rad y chemo.

Ejercicio 1.

Ajustar modelos paramétricos con el tiempo de supervivencia modelado mediante: una distribución exponencial, una Weibull y una distribución lognormal.

En las tablas 14, 15 y 16, se presentan las estimaciones de modelos paramétricos con el tiempo de supervivencia modelado mediante una distribución exponencial, una Weibull y una lognormal, respectivamente.

Tabla 14. Modelo paramétrico (distribución exponencial).

Exponential PH regression

No. of subjects =	300	Number of obs =	300
No. of failures =	207		
Time at risk =	127,744		
		Wald chi2(5) =	41.98
Log pseudolikelihood =	-315.62373	Prob > chi2 =	0.0000

		Coefficient	Robust std. err.	z	P> z	[95% conf. interval]	
age		-.0337868	.0900661	-0.38	0.708	-.210313	.1427394
age2		.0066531	.0058958	1.13	0.259	-.0049024	.0182086
rad		.1583477	.159347	0.99	0.320	-.1539666	.4706621
chemo		-.2735383	.1530009	-1.79	0.074	-.5734146	.026338
rad_chemo		-.349043	.1979503	-1.76	0.078	-.7370185	.0389325
_cons		-6.345979	.3360271	-18.89	0.000	-7.00458	-5.687378

Fuente: Elaboración propia.

Tabla 15. Modelo paramétrico (distribución Weibull).

Weibull PH regression

No. of subjects = 300
 No. of failures = 207
 Time at risk = 127,744
 Log pseudolikelihood = -243.48756
 Number of obs = 300
 Wald chi2(5) = 78.17
 Prob > chi2 = 0.0000

		Robust				
_t	Coefficient	std. err.	z	P> z	[95% conf. interval]	
age	-.0437837	.134874	-0.32	0.745	-.308132	.2205645
age2	.0098644	.0090807	1.09	0.277	-.0079335	.0276623
rad	.1379884	.2168381	0.64	0.525	-.2870064	.5629832
chemo	-.7562293	.2126099	-3.56	0.000	-1.172937	-.3395216
rad_chemo	-.6779556	.2786234	-2.43	0.015	-1.224047	-.1318637
_cons	-13.78937	1.018176	-13.54	0.000	-15.78496	-11.79379
/ln_p	.8068914	.0672935	11.99	0.000	.6749986	.9387842
p	2.240931	.1508			1.96403	2.556871
1/p	.4462431	.0300292			.3911031	.5091571

Fuente: Elaboración propia.

Tabla 16. Modelo paramétrico (distribución lognormal).

Lognormal AFT regression

No. of subjects = 300
 No. of failures = 207
 Time at risk = 127,744
 Log pseudolikelihood = -280.49695
 Number of obs = 300
 Wald chi2(5) = 64.04
 Prob > chi2 = 0.0000

		Robust				
_t	Coefficient	std. err.	z	P> z	[95% conf. interval]	
age	.0665496	.079323	0.84	0.401	-.0889206	.2220198
age2	-.0091573	.0054042	-1.69	0.090	-.0197494	.0014347
rad	-.0304744	.1575335	-0.19	0.847	-.3392344	.2782855
chemo	.4473685	.1372041	3.26	0.001	.1784533	.7162837
rad_chemo	.2797488	.1875355	1.49	0.136	-.087814	.6473115
_cons	5.811232	.2803129	20.73	0.000	5.261829	6.360635
/lnsigma	-.3324743	.0923369	-3.60	0.000	-.5134513	-.1514973
sigma	.7171471	.0662191			.5984267	.8594202

Fuente: Elaboración propia.

Ejercicio 2.

Interpretar los coeficientes estimados cuidadosamente.

Seguidamente, se presenta una breve descripción de cómo se interpretan estos coeficientes estimados en cada distribución:

➤ Distribución exponencial:

En un modelo con distribución exponencial, el coeficiente estimado para una covariable continua representa el cambio en el logaritmo del *hazard* (tasa de riesgo) por un incremento unitario en la covariable, manteniendo todas las demás covariables constantes. Por ejemplo, el coeficiente estimado de la covariable *age*, para un perro con 5 años de edad en el momento del diagnóstico de linfoma, es 0,033 ($= -0,034 + 2 * 0,007 * 5$), que significa que un incremento en un año de edad en el momento del diagnóstico está asociado con un incremento del 3,3% en el *hazard* (tasa de riesgo), manteniendo todas las demás covariables constantes.

En el caso de las covariables discretas (*rad*, *chemo* y su interacción), haciendo uso de los coeficientes estimados ($e^{\hat{\beta}_k} - 1$), se llega a cuál es la brecha en la tasa de riesgo entre la *categoría_k*, $k = 1, 2, 3$, y la *categoría₀*, manteniendo todas las demás covariables constantes. Por ejemplo, el coeficiente estimado para la variable *rad* es 0,158, que implica que la brecha en la tasa de riesgo entre los perros que recibieron sólo el tratamiento de radioterapia (*rad*= 1 y *chemo*= 0) y los que no recibieron ningún tratamiento es igual a 0,172 ($= e^{0,158} - 1$), manteniendo todas las demás covariables constantes. Los respectivos valores para la *categoría₂* y la *categoría₃* son negativos e iguales a -0,239 ($= e^{-0,274} - 1$) y -0,371 ($= e^{0,158-0,274-0,349} - 1$), respectivamente.

➤ Distribución Weibull:

En un modelo con distribución Weibull, la interpretación del coeficiente estimado es similar a la con distribución exponencial. Sin embargo, debido a la flexibilidad adicional de la distribución Weibull, el efecto de una covariable puede ser no lineal en términos de la tasa de riesgo. Por lo tanto, la interpretación del coeficiente en la distribución Weibull sigue siendo un cambio en el logaritmo del *hazard* por un incremento unitario en la covariable, manteniendo todas las demás covariables constantes, pero la forma exacta del cambio puede variar dependiendo de los parámetros de forma y escala de la distribución Weibull.

➤ Distribución lognormal:

En un modelo con distribución lognormal, el coeficiente estimado para una covariable continua representa el cambio en la mediana (o la media, dependiendo de la parametrización) del tiempo hasta el evento por un incremento unitario en la covariable, manteniendo todas las demás covariables constantes. Por ejemplo, el coeficiente estimado de la covariable *age*, para un perro con 5 años de edad en el momento del diagnóstico de linfoma, es -0,025 ($= 0,067 + 2 (-0,009) * 5$), significa que un incremento en un año de edad en el momento del diagnóstico está asociado con un decremento del 2,5% en la mediana del tiempo hasta el evento, manteniendo todas las demás covariables constantes.

En el caso de las covariables discretas (*rad*, *chemo* y su interacción), haciendo uso de los coeficientes estimados ($e^{\hat{\beta}_k} - 1$), se llega a cuál es la brecha en la mediana del tiempo hasta el evento entre la *categoría_k*, $k = 1, 2, 3$, y la *categoría₀*, manteniendo todas las demás covariables constantes. Por ejemplo, el coeficiente estimado de la variable *rad* es -0,03, que implica que la brecha en la mediana del tiempo hasta el evento entre los perros que recibieron sólo el tratamiento de radioterapia (*rad*= 1 y *chemo*= 0) y los que no recibieron ningún tratamiento es igual a -0,03 ($= e^{-0,03} - 1$), manteniendo todas las demás covariables constantes. Los respectivos valores para la *categoría₂* y la *categoría₃* son positivos e iguales a 0,564 ($= e^{0,447} - 1$) y 1,01 ($= e^{-0,03+0,447+0,28} - 1$), respectivamente.

Ejercicio 3.

Comparar cuál de los tres modelos ajusta mejor a los datos.

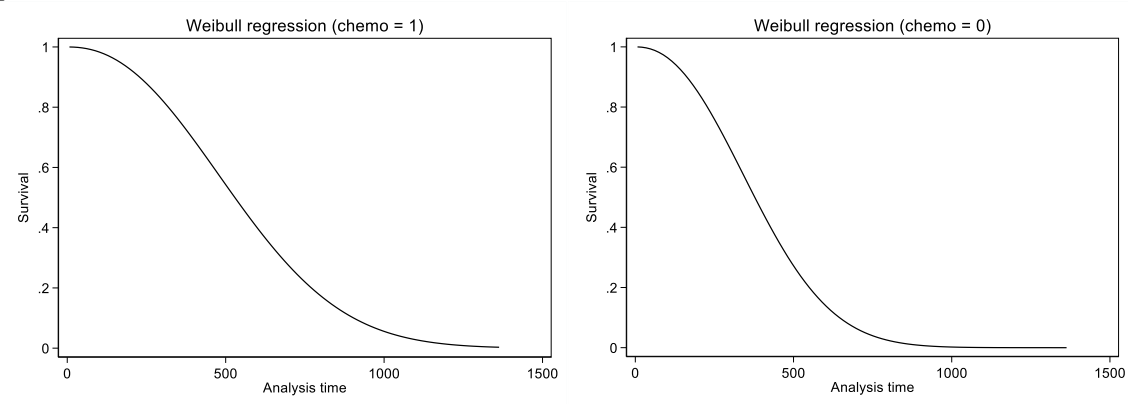
Los AIC (BIC) de los modelos que usan distribución exponencial, Weibull y lognormal son 643,247, 500,975 y 574,994 (665,47, 526,902 y 600,92), respectivamente, por lo que se concluye que el modelo con distribución Weibull es el que ajusta mejor a los datos.

Ejercicio 4.

Calcular la curva de supervivencia para cada grupo de quimioterapia para los modelos que usan distribuciones de Weibull y lognormal.

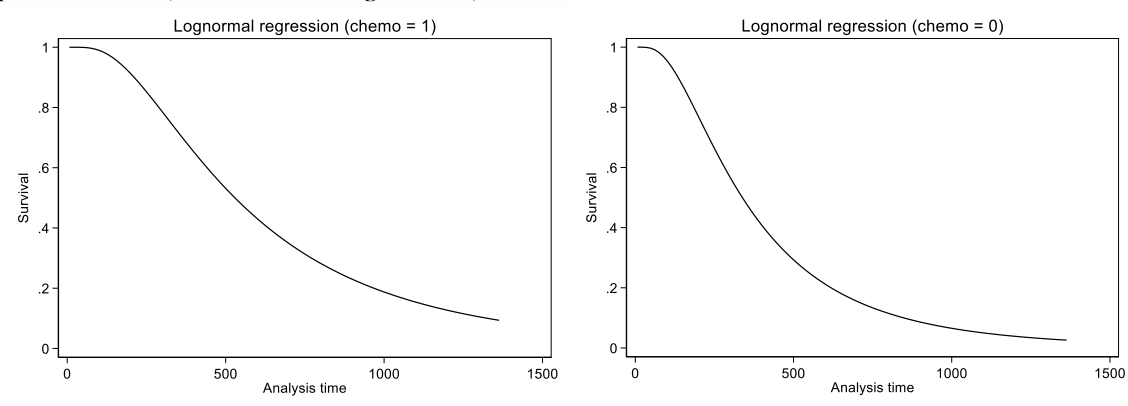
En las figuras 10 y 11, se presentan las curvas de supervivencia para grupo de tratamiento y grupo de control para el tratamiento de quimioterapia en los modelos que usan distribución Weibull y lognormal, respectivamente. En ambos modelos, se puede observar que la curva correspondiente al grupo de tratamiento se encuentra más elevada que la correspondiente al grupo de control, indicando que, para los grupos de tratamiento, para cada momento del tiempo, hay una mayor proporción de perros que sobreviven.

Figura 10. *Curvas de supervivencia para tratamiento de quimioterapia en modelo paramétrico (distribución Weibull).*



Fuente: Elaboración propia.

Figura 11. *Curvas de supervivencia para tratamiento de quimioterapia en modelo paramétrico (distribución lognormal).*



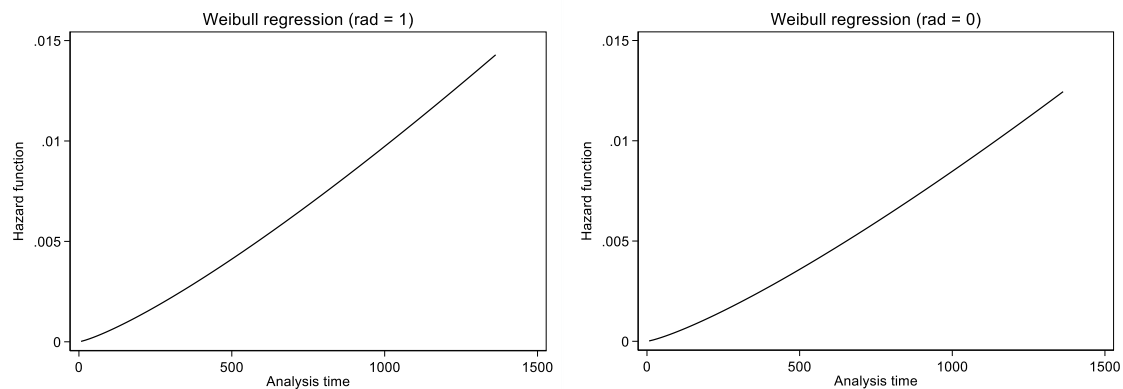
Fuente: Elaboración propia.

Ejercicio 5.

Dibujar la hazard en la línea de base para cada grupo de radiación para los modelos que usan distribuciones de Weibull y lognormal.

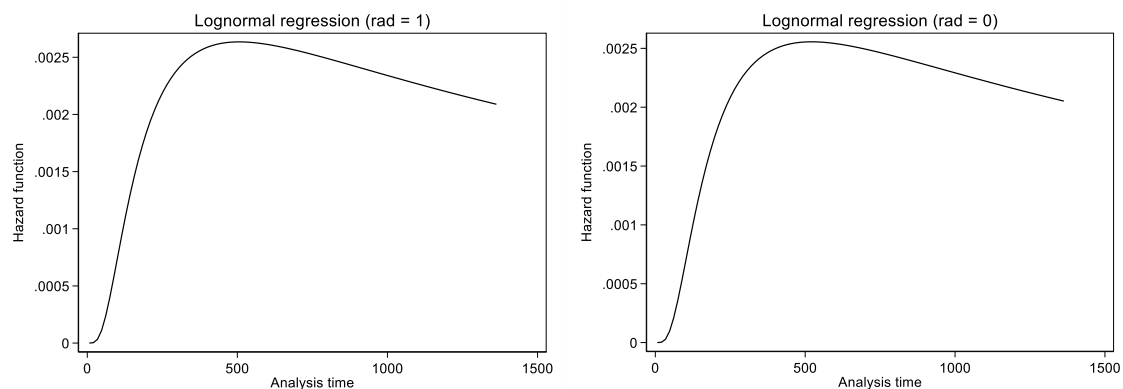
En las figuras 12 y 13, se presentan las funciones *hazard* para grupo de tratamiento y grupo de control para el tratamiento de radiación en los modelos que usan distribución Weibull y lognormal, respectivamente. Por un lado, al comparar grupos, no pareciera haber grandes diferencias entre grupo de tratamiento y grupo de control, indicando una tasa de riesgo semejante a lo largo del tiempo. Por otro lado, al comparar ambos modelos, mientras que la función *hazard* en el modelo que usa distribución lognormal no crece en todo el período (empieza a decrecer alrededor del día 500), la función *hazard* en el modelo que usa distribución Weibull crece en todo el período.

Figura 12. Funciones *hazard* para tratamiento de radiación en modelo paramétrico (distribución Weibull).



Fuente: Elaboración propia.

Figura 13. Funciones *hazard* para tratamiento de radiación en modelo paramétrico (distribución lognormal).



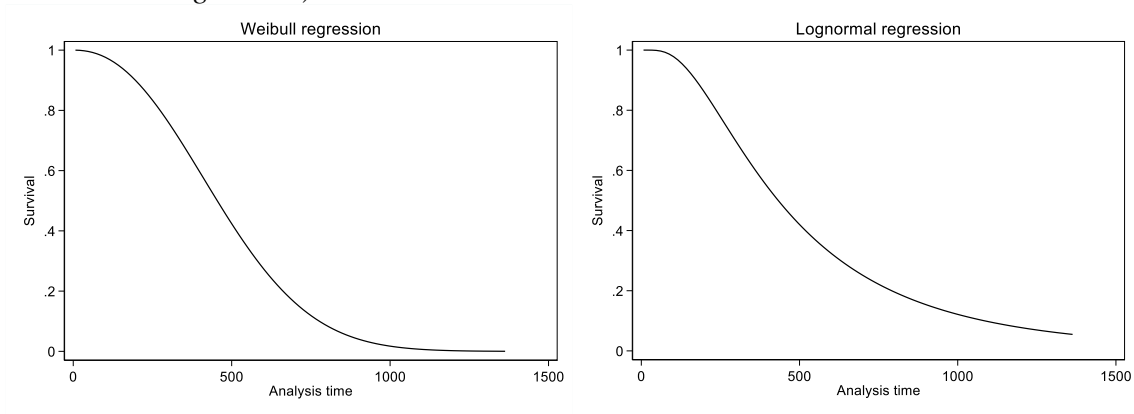
Fuente: Elaboración propia.

Ejercicio 6.

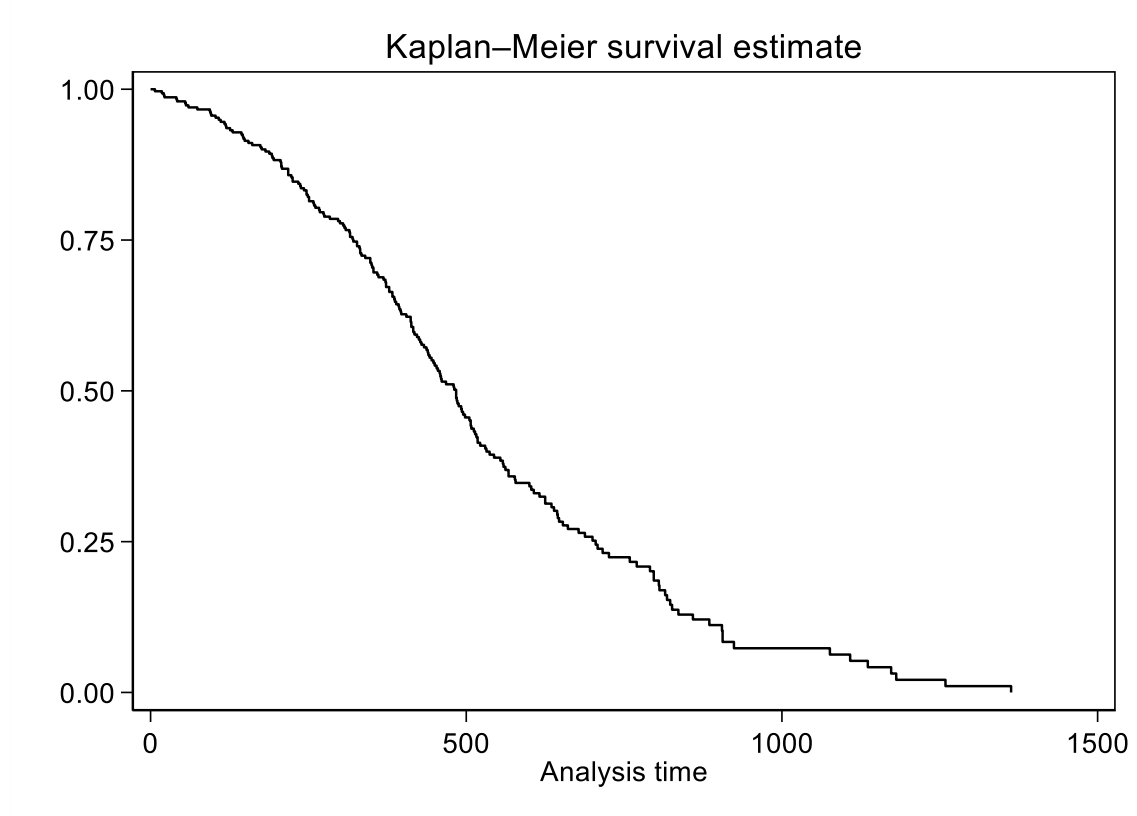
Evaluar el ajuste de los modelos que usan distribuciones de Weibull y lognormal.

En la figura 14, se presentan las curvas de supervivencia de los modelos que usan distribución Weibull y lognormal, respectivamente, mientras que, en la figura 15, se presenta la curva de Kaplan-Meier de supervivencia. Siguiendo este criterio, se puede observar que el modelo que usa distribución Weibull ajusta mejor a los datos, ya que su curva de supervivencia se asemeja más a la curva de Kaplan-Meier que lo que se asemeja la curva de supervivencia del modelo que usa distribución log-normal.

Figura 14. *Curvas de supervivencia en modelos paramétricos (distribución Weibull y distribución lognormal).*



Fuente: Elaboración propia.

Figura 15. *Curva de Kaplan-Meier de supervivencia.*

Fuente: Elaboración propia.

En las tablas 17 y 18, se presentan los criterios de información de los modelos que usan distribución Weibull y lognormal, respectivamente. Siguiendo este criterio, se puede observar que el modelo que usa distribución Weibull ajusta mejor a los datos, ya que tanto AIC como BIC son menores a los del modelo que usa distribución lognormal.

Tabla 17. *Criterios de información en modelo paramétrico (distribución Weibull).*

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	300	-278.6239	-243.4876	7	500.9751	526.9016

Fuente: Elaboración propia.

Tabla 18. *Criterios de información en modelo paramétrico (distribución lognormal).*

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	300	-309.5887	-280.4969	7	574.9939	600.9204

Fuente: Elaboración propia.

Ejercicio 7.

Resumir las principales conclusiones del análisis.

Las principales conclusiones del análisis son:

- Para el grupo de tratamiento de quimioterapia, para cada momento del tiempo, hay una mayor proporción de perros que sobreviven.
- El modelo que usa distribución Weibull es que el ajusta mejor a los datos.

Referencias.

Berndt, E. R. (1991). The Practice of Econometrics: Classic and Contemporary. Addison-Wesley Publishing Company.

Mroz, T. A. (1987). The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions. *Econometrica*, 55(4), 765-799. Recuperado de <https://www.jstor.org/stable/1911029>