Ejemplo 10.1 *Modelos Probit y Logit para la probabilidad de tener carro propio en Holanda*

El archivo NLCAR contiene información para 2,820 hogares utilizada por Cramer (2001) para estudiar los determinantes de la probabilidad de tener carro propio en Holanda. Las variables son las siguientes:

CAR :	Número de carros disponibles en el hogar
	0: El hogar no dispone de carro propio.
	1: El hogar posee carro usado.
	2: El hogar posee carro nuevo.
	3: El hogar posee mas de un carro.
INC :	Ingreso del hogar por adulto equivalente.
$SIZE$:	Tamaño del hogar.
$BUSCAR$:	Variable dummy que denota la presencia (o no) de carro de la empresa.
$URBA$:	Grado de urbanización, medido en escala de 1 (campo) a 6 (ciudad).
$HHAGE$:	Edad de la cabeza del hogar.
$LINC$:	Logaritmo del ingreso.
$CAR01$:	Variable dummy que denota la presencia de carro en el hogar.

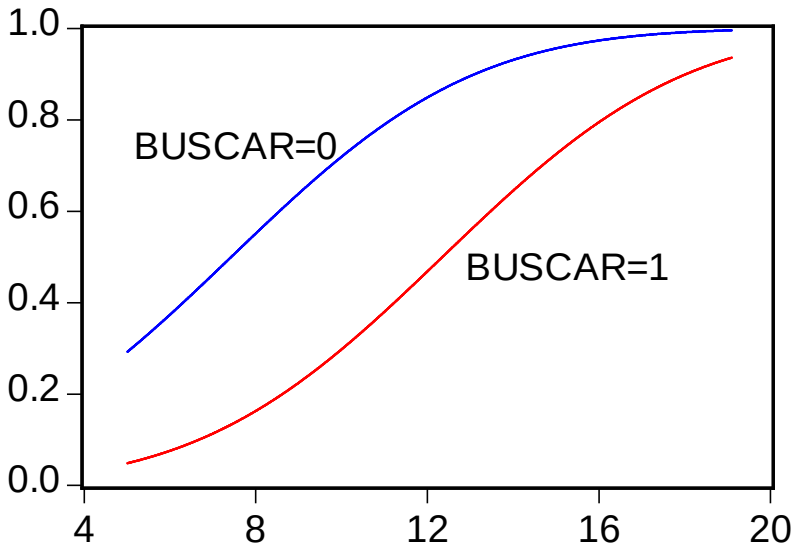
Modelo de regresión lineal (OLS)
Dependent Variable: $CAR01$
Included observations: 2820
White Heteroskedasticity-Consistent Standard Errors and Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.052199	0.174963	-0.298343	0.7655
LINC	0.077026	0.018025	4.273286	0.0000
BUSCAR	-0.418032	0.025732	-16.245420	0.0000
R-squared	0.086485	Mean dependent var		0.641844
Adjusted R-squared	0.085837	S.D. dependent var		0.479543
S.E. of regression	0.458500	Akaike info criterion		1.279352
Sum squared resid	592.1973	Schwarz criterion		1.285676
Log likelihood	-1800.887	F-statistic		133.3469
Durbin-Watson stat	1.911756	Prob(F-statistic)		0.000000

Modelo Probit
Dependent Variable: $CAR01$
Included observations: 2820

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-1.684658	0.526593	-3.19917	0.0014
LINC	0.226523	0.054501	4.15628	0.0000
BUSCAR	-1.111295	0.077643	-14.31280	0.0000
Mean dependent var	0.641844	S.D. dependent var		0.479543
S.E. of regression	0.458496	Akaike info criterion		1.223000
Sum squared resid	592.1860	Schwarz criterion		1.229324
Log likelihood	-1721.430	Hannan-Quinn criter.		1.225282
Restr. log likelihood	-1839.627	Avg. log likelihood		-0.610436
LR statistic (2 df)	236.3926	McFadden R-squared		0.064250
Probability(LR stat)	0.000000			

BUSCAR1=@cnorm(-1.684657762 + 0.2265233798*LINC - 1.111295374)
BUSCAR0=@cnorm(-1.684657762 + 0.2265233798*LINC)



Dependent Variable: $CAR01$
Method: ML - Binary Probit (Quadratic hill climbing)
Included observations: 2820
Prediction Evaluation (success cutoff C = 0.5)

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
P(Dep=1)<=C	247	93	340	0	0	0
P(Dep=1)>C	763	1717	2480	1010	1810	2820
Total	1010	1810	2820	1010	1810	2820
Correct	247	1717	1964	0	1810	1810
% Correct	24.46	94.86	69.65	0	100	64.18
% Incorrect	75.54	5.14	30.35	100	0	35.82
Total Gain*	24.46	-5.14	5.46			
Percent Gain**	24.46	NA	15.25			

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
E(No. of Dep=0)	417.94	592.30	1010.24	361.74	648.26	1010.00
E(No. of Dep=1)	592.06	1217.70	1809.76	648.26	1161.74	1810.00
Total	1010.00	1810.00	2820.00	1010.00	1810.00	2820.00
Correct	417.94	1217.70	1635.65	361.74	1161.74	1523.48
% Correct	41.38	67.28	58.00	35.82	64.18	54.02
% Incorrect	58.62	32.72	42.00	64.18	35.82	45.98
Total Gain*	5.56	3.09	3.98			
Percent Gain**	8.67	8.63	8.65			

* Change in " %Correct"from default (constant probability) specification

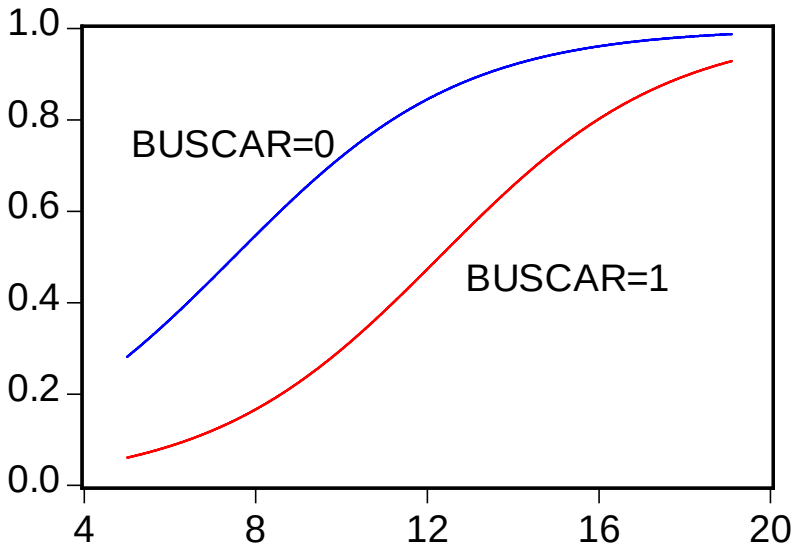
** Percent of incorrect (default) prediction corrected by equation

Modelo Logit
Dependent Variable: *CAR*01
Included observations: 2820

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-2.817963	0.877137	-3.212682	0.0013
LINC	0.376002	0.090903	4.136313	0.0000
BUSCAR	-1.800486	0.129930	-13.857390	0.0000

Mean dependent var	0.641844	S.D. dependent var	0.479543
S.E. of regression	0.458496	Akaike info criterion	1.223008
Sum squared resid	592.1853	Schwarz criterion	1.229332
Log likelihood	-1721.441	Hannan-Quinn criter.	1.225290
Restr. log likelihood	-1839.627	Avg. log likelihood	-0.610440
LR statistic (2 df)	236.3704	McFadden R-squared	0.064244
Probability(LR stat)	0.000000		

BUSCAR0=@logit(-2.817963298 + 0.37600163790*LINC)
BUSCAR1=@logit(-2.817963298 + 0.3760016379*LINC - 1.800485565)



Dependent Variable: $CAR01$
Method: ML - Binary Logit (Quadratic hill climbing)
Included observations: 2820
Prediction Evaluation (success cutoff C = 0.5)

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
P(Dep=1)<=C	247	93	340	0	0	0
P(Dep=1)>C	763	1717	2480	1010	1810	2820
Total	1010	1810	2820	1010	1810	2820
Correct	247	1717	1964	0	1810	1810
% Correct	24.46	94.86	69.65	0.00	100.00	64.18
% Incorrect	75.54	5.14	30.35	100.00	0.00	35.82
Total Gain*	24.46	-5.14	5.46			
Percent Gain**	24.46	NA	15.25			

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
E(No. of Dep=0)	417.82	592.18	1010.00	361.74	648.26	1010.00
E(No. of Dep=1)	592.18	1217.82	1810.00	648.26	1161.74	1810.00
Total	1010.00	1810.00	2820.00	1010.00	1810.00	2820.00
Correct	417.82	1217.82	1635.65	361.74	1161.74	1523.48
% Correct	41.37	67.28	58.00	35.82	64.18	54.02
% Incorrect	58.63	32.72	42.00	64.18	35.82	45.98
Total Gain*	5.55	3.10	3.98			
Percent Gain**	8.65	8.65	8.65			

*Change in " % Correct"from default (constant probability) specification
**Percent of incorrect (default) prediction corrected by equation

$LINC$	Modelo Probit		Modelo Logit	
	$BUSCAR = 0$	$BUSCAR = 1$	$BUSCAR = 0$	$BUSCAR = 1$
5.000	0.29	0.048	0.281	0.061
5.005	0.291	0.048	0.282	0.061
5.010	0.291	0.048	0.282	0.061
5.015	0.292	0.048	0.282	0.061
...	...	...	...	...
7.435	0.5	0.133	0.494	0.139
7.440	0.5	0.133	0.495	0.139
7.445	0.501	0.134	0.495	0.14
7.450	0.501	0.134	0.496	0.14
...	...	...	...	...
12.620	0.88	0.525	0.873	0.532
12.625	0.88	0.525	0.873	0.532
12.630	0.88	0.526	0.873	0.533
12.635	0.88	0.526	0.874	0.533
...	...	...	...	...
15.170	0.96	0.739	0.947	0.748
15.175	0.96	0.739	0.947	0.748
15.180	0.96	0.74	0.947	0.748
15.185	0.96	0.74	0.947	0.749
...	...	...	...	...
19.080	0.996	0.937	0.987	0.928
19.085	0.996	0.937	0.987	0.928
19.090	0.996	0.937	0.987	0.928
19.095	0.996	0.937	0.987	0.928

Ejemplo 10.2 *Modelos Probit y Logit para la probabilidad de participar en la fuerza de trabajo*

El archivo MROZ contiene información para una muestra de 753 mujeres que nos permite estudiar los determinantes de la probabilidad de participar en la fuerza de trabajo. De las 753 mujeres en la muestra, 428 reportan que trabajaron durante el año, mientras que las restantes 325 reportan no haber trabajado. Las variables son las siguientes:

AGE :	Edad
$EDUC$:	Educación
$EXPER$:	Años de experiencia laboral
$NWIFEINC$:	Ingreso del compañero
$KIDSLT6$:	Número de hijos con menos de 6 años
$KIDSGE6$:	Número de hijos con más de 6 años

Variable	OLS		Logit		Probit	
	Coefficient	Std. Error White	Coefficient	Std. Error	Coefficient	Std. Error
$NWIFEINC$	-0.0034	(0.0015)	-0.0213	(0.0091)	-0.0120	(0.0053)
$EDUC$	0.0380	(0.0073)	0.2212	(0.0444)	0.1309	(0.0258)
$EXPER$	0.0395	(0.0058)	0.2059	(0.0323)	0.1233	(0.0188)
$EXPERSQ$	-0.0006	(0.0002)	-0.0032	(0.0010)	-0.0019	(0.0006)
AGE	-0.0161	(0.0024)	-0.0880	(0.0144)	-0.0529	(0.0083)
$KIDSLT6$	-0.2618	(0.0318)	-1.4434	(0.2030)	-0.8683	(0.1161)
$KIDSGE6$	0.0130	(0.0135)	0.0601	(0.0798)	0.0360	(0.0453)
Constant	0.5855	(0.1523)	0.4255	(0.8592)	0.2701	(0.5048)
Obs	753		753		753	
R-squared	0.264		0.220		0.221	
Log likelihood	-423.892		-401.765		-401.302	

Los signos de los coeficientes son los mismos para todos los modelos, y las mismas variables son estadísticamente significativas.

Dependent Variable: $INLF$
Method: ML - Binary Probit (Quadratic hill climbing)
Included observations: 753
Prediction Evaluation (success cutoff C = 0.5)

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
P(Dep=1)<=C	205	80	285	0	0	0
P(Dep=1)>C	120	348	468	325	428	753
Total	325	428	753	325	428	753
Correct	205	348	553	0	428	428
% Correct	63.08	81.31	73.44	0	100	56.84
% Incorrect	36.92	18.69	26.56	100	0	43.16
Total Gain*	63.08	-18.69	16.6			
Percent Gain**	63.08	NA	38.46			

* Change in " % Correct"from default (constant probability) specification

** Percent of incorrect (default) prediction corrected by equation

La principal diferencia entre el modelo de regresión lineal, y los modelos Logit y Probit, es que el primero asume un efecto marginal constante para cada una de las variables independientes del modelo, mientras que los modelos Logit y Probit implican efecto marginal decreciente.

Por ejemplo, en el modelo de regresión lineal un hijo menor de 6 años adicional reduce la probabilidad de participar en la fuerza laboral en 0.262, independientemente no solo del número de hijos que ya tenga la mujer, sino también del nivel (o valores) de las demás variables del modelo. Por el contrario, en el modelo Probit tendríamos:

	$\bar{x}_i$	$\bar{x}_i$	$\bar{x}_i$	β_i	$\beta_i \cdot \bar{x}_i$	$\beta_i \cdot \bar{x}_i$	$\beta_i \cdot \bar{x}_i$
$NWIFEINC$	20.13	20.13	20.13	-0.0120	-0.242	-0.242	-0.242
$EDUC$	12.3	12.3	12.3	0.1309	1.610	1.610	1.610
$EXPER$	10.6	10.6	10.6	0.1233	1.307	1.307	1.307
$EXPERSQ$	112.36	112.36	112.36	-0.0019	-0.212	-0.212	-0.212
AGE	42.5	42.5	42.5	-0.0529	-2.246	-2.246	-2.246
$KIDSLT6$	0	1	2	-0.8683	0	-0.868	-1.737
$KIDSGE6$	1	1	1	0.0360	0.036	0.036	0.036
$Constant$	1	1	1	0.2701	0.270	0.270	0.270
				$\sum \beta_i \cdot \bar{x}_i$	0.523	-0.345	-1.213
				$\Phi\left(\sum \beta_i \cdot \bar{x}_i\right)$	0.700	0.365	0.113
						-0.335	-0.253

Ejemplo 11.1 *Modelo Probit ordenado para la probabilidad del estado de salud*

En este ejemplo utilizamos información de la encuesta de calidad de vida del DANE para Bogotá en 2003. La variable dependiente corresponde a 11.182 observaciones del estado de salud autoreportado por una muestra de individuos, donde $Y_i = 1$ si es muy bueno,$Y_i = 2$ si es bueno,$Y_i = 3$ si es regular, $Y_i = 4$ si es malo. Como variables independientes se incluyen el logaritmo del ingreso del individuo (ly) y su género ($genero = 1$, si es hombre).

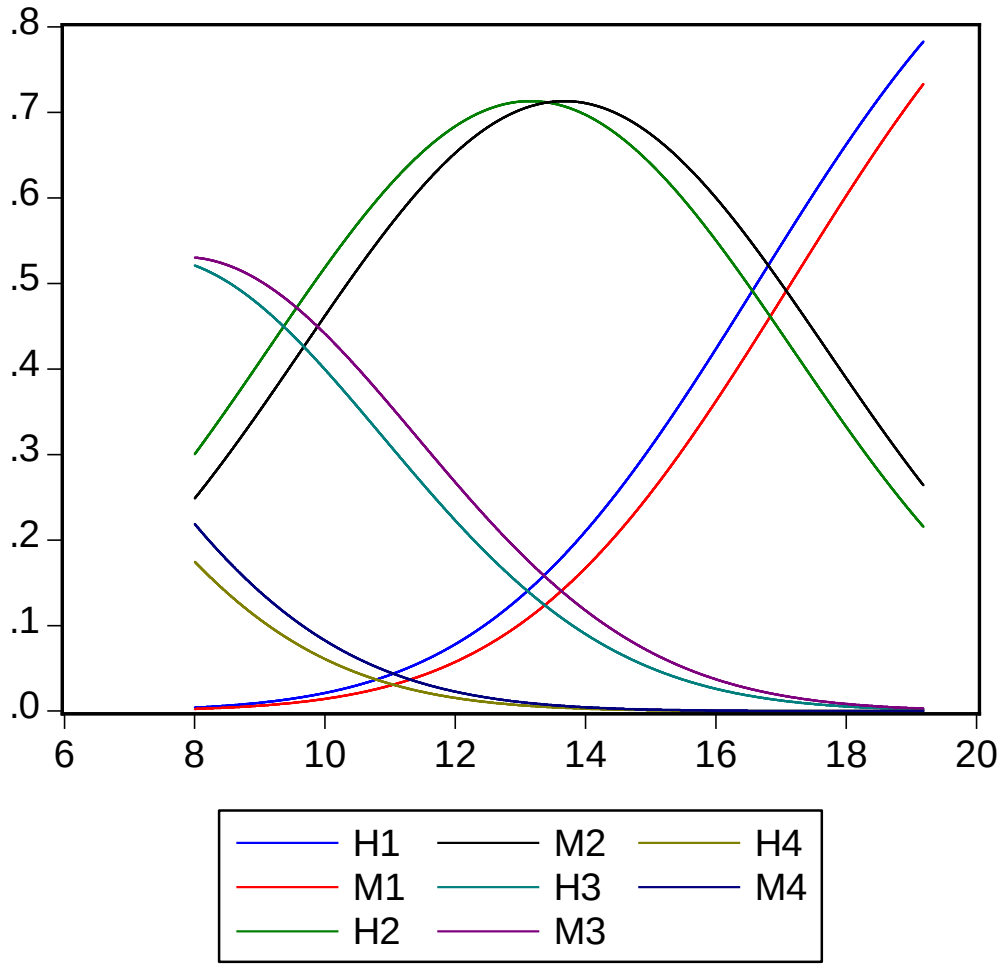
Dependent Variable: Y
ML - Ordered Probit
Sample: 1 11182
Ordered indicator values: 4

Variable	Coefficient	Std. Error	z-Statistic	Prob.
GENERO	-0.1592	0.0224	-7.1214	0.0000
LY	-0.3061	0.0127	-24.1324	0.0000
Limit Points				
LIMIT2	-5.2488	0.1710	-30.6989	0.0000
LIMIT3	-3.1193	0.1668	-18.7037	0.0000
LIMIT4	-1.6718	0.1686	-9.9152	0.0000

Akaike info criterion	1.6654	Schwarz criterion	1.6686
Log likelihood	-9306.0340	Hannan-Quinn criter.	1.6665
Restr. log likelihood	-9646.4600	Avg. log likelihood	-0.8322
LR statistic (2 df)	680.8520	LR index (Pseudo-R2)	0.0353
Probability(LR stat)	0.0000		

Como se puede observar, las variables incluídas son estadisticamente significativas.

A partir de estos resultados, podemos estimar las diferentes probabilidades para diferentes valores de ly, dependiendo del género del individuo.



Ejemplo 11.2 *Modelo Logit multinomial para la relación entre empleo e historia escolar*

El archivo KEANE.RAW contiene información sobre empleo e historia escolar de una muestra de individuos jóvenes para los años 1981 a 1987. En este ejemplo se utiliza la información de 1987.

- La variable dependiente se denomina *status* y está definida como $status = 0$ si el individuo esta estudiando, $status = 1$ si el individuo no está estudiando y tampoco trabajando, y $status = 2$ si el individuo esta trabajando.
- Las variables independientes incluyen *educ*, *exper*, *exper*2, *black* y una constante.

Como categoría base ($J = 0$) se utilizan los individuos que están estudiando, es decir $status = 0$. La muestra es de un total de 1,717 observaciones, de las cuales 99, 332 y 1,286 observaciones están en las categorías 0, 1 y 2, respectivamente.

`Instrucción en STATA: mlogit status educ exper expersq black, basecategory(0)`

Status	Coef.	Std. Err.	z	p.val	[95 % conf.	int.]
1						
educ	-0.674	-0.070	-9.640	0.000	-0.811	-0.537
exper	-0.106	-0.173	-0.610	0.540	-0.446	0.233
expersq	-0.013	-0.025	-0.500	0.620	-0.062	0.037
black	0.813	-0.303	2.690	0.007	0.220	1.406
Constant	10.278	-1.133	9.070	0.000	8.057	12.499
2						
educ	-0.315	-0.065	-4.830	0.000	-0.442	-0.187
exper	0.849	-0.157	5.410	0.000	0.541	1.156
expersq	-0.077	-0.023	-3.370	0.001	-0.122	-0.032
black	0.311	-0.282	1.110	0.269	-0.240	0.863
Constant	5.544	-1.086	5.100	0.000	3.414	7.673
Pseudo R2		0.243				
Log likelihood		-907.857				

(Outcome status==0 is the comparison group).

Las magnitudes de estos coeficientes son difíciles de interpretar. En su lugar podemos calcular efectos parciales, o la diferencia en probabilidad.

Por ejemplo, consideremos dos individuos de raza negra, cada uno con 5 años de experiencia. Un individuo con 16 años de educación tiene una probabilidad de empleo que es 0.042 puntos porcentuales mayor que la de un individuo con 12 años de educación, mientras que la diferencia en probabilidad de estar en $status = 1$ es 0.072 puntos porcentuales mas baja. Por construcción, la diferencia en probabilidad en $status = 0$ es 0.030 mas alta para el individuo con 16 años de educación.

Variable	Status = 1					Status = 2				
	β_i	$\bar{x}_i$	$\bar{x}_i$	$\beta_i \cdot \bar{x}_i$	$\beta_i \cdot \bar{x}_i$	β_i	$\bar{x}_i$	$\bar{x}_i$	$\beta_i \cdot \bar{x}_i$	$\beta_i \cdot \bar{x}_i$
educ	-0.674	16	12	-10.778	-8.084	-0.315	16	12	-5.035	-3.776
exper	-0.106	5	5	-0.531	-0.531	0.849	5	5	4.244	4.244
expersq	-0.013	25	25	-0.313	-0.313	-0.077	25	25	-1.933	-1.933
black	0.813	1	1	0.813	0.813	0.311	1	1	0.311	0.311
Constant	10.278	1	1	10.278	10.278	5.544	1	1	5.544	5.544
$\sum \beta_i \cdot \bar{x}_i$				-0.531	2.163				3.132	4.390

$$\begin{aligned}\text{Prob}\,(s = 2\,|educ = 16\,) &= \frac{\exp(3{,}132)}{1 + \exp(-0{,}531) + \exp(3{,}132)} = \frac{22{,}920}{1 + 0{,}588 + 22{,}920} \\ &= 0{,}935\end{aligned}$$

$$\begin{aligned}\text{Prob}\,(s = 2\,|educ = 12\,) &= \frac{\exp(4{,}390)}{1 + \exp(2{,}163) + \exp(4{,}390)} = \frac{8{,}697}{1 + 8{,}697 + 22{,}920} \\ &= 0{,}893\end{aligned}$$

La diferencia en probabilidad para $status = 2$ es $0{,}935 - 0{,}893 = 0{,}042$.

$$\begin{aligned}\text{Prob}\,(s = 1\,|educ = 16\,) &= \frac{\exp(-0{,}531)}{1 + \exp(-0{,}531) + \exp(3{,}132)} = \frac{0{,}588}{1 + 0{,}588 + 22{,}920} \\ &= 0{,}024\end{aligned}$$

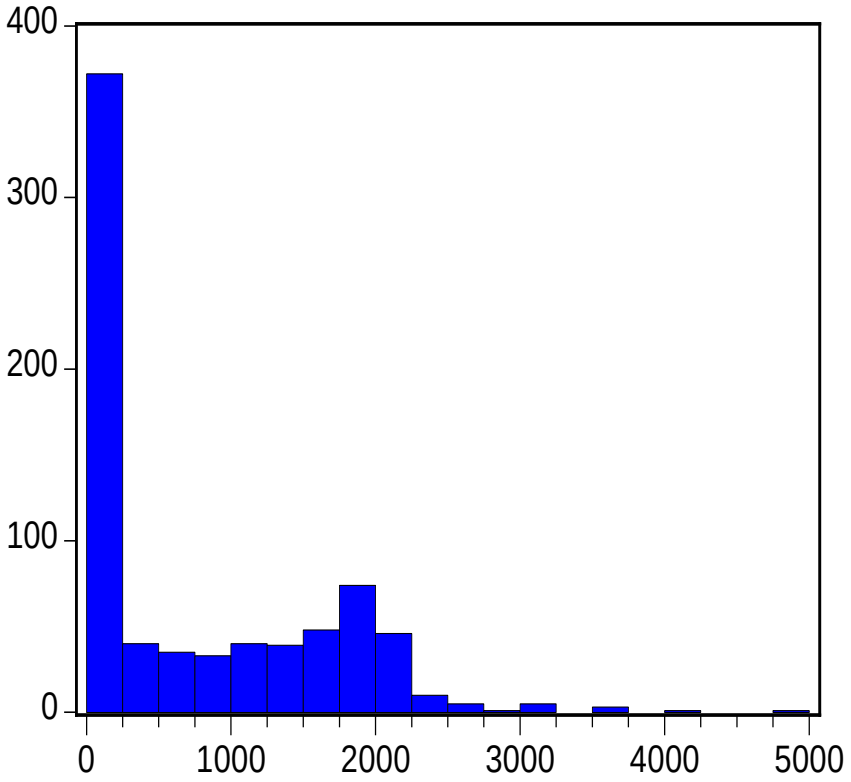
$$\begin{aligned}\text{Prob}\,(s = 1\,|educ = 12\,) &= \frac{\exp(2{,}163)}{1 + \exp(2{,}163) + \exp(4{,}390)} = \frac{8{,}697}{1 + 8{,}697 + 22{,}920} \\ &= 0{,}096\end{aligned}$$

La diferencia en probabilidad para $status = 1$ es $0{,}024 - 0{,}096 = -0{,}072$.
Por defecto la diferencia en probabilidad para $status = 0$ es 0.030.

Ejemplo 12.1 *Modelo Tobit para la probabilidad de horas trabajadas por mujeres casadas*

En este ejemplo utilizamos datos de Mroz (1987) para estimar una ecuación de forma reducida para el número de horas trabajadas en un año por mujeres casadas. La ecuación es una "forma reducida" pues no incluye el salario como variable explicativa (variable que es poco probable que sea exógena).

La muestra consiste de 753 observaciones, 428 de las cuales corresponden a mujeres que trabajaron por fuera del hogar durante el año en cuestión; es decir, 325 mujeres trabajaron cero (0) horas durante el año. Para las mujeres que efectivamente trabajaron, el rango de variación de horas trabajadas fluctúa entre 12 y 4,950.



Series: HOURS	
Sample 1 753	
Observations 753	
Mean	740.5764
Median	288.0000
Maximum	4950.000
Minimum	0.000000
Std. Dev.	871.3142
Skewness	0.922531
Kurtosis	3.193949
Jarque-Bera	107.9888
Probability	0.000000

Los coeficientes estimados del modelo TOBIT tienen el mismo signo de los coeficientes del modelo estimado por OLS, y la significancia estadística de los coeficientes es similar. Debe recordarse que la comparación de la magnitud de los coeficientes estimados en los modelos no resulta informativa.

Variable	OLS		TOBIT	
	Coef	(s.e)	Coef	(s.e)
$NWIFEINC$	-3.447	-2.544	-8.814	-4.459
$EDUC$	28.761	-12.955	80.646	-21.583
$EXPER$	65.673	-9.963	131.564	-17.279
$EXPERSQ$	-0.700	-0.325	-1.864	-0.538
AGE	-30.512	-4.364	-54.405	-7.419
$KIDSLT6$	-442.090	-58.847	-894.022	-111.878
$KIDSGE6$	-32.779	-23.176	-16.218	-38.641
$Constant$	1330.482	-270.785	965.305	-446.436
Log-likelihood			-3819.095	
R-squared	0.266		0.273	
$\hat{\sigma}$	750.179		1122.022	

Para obtener el efecto parcial debemos multiplicar los coeficientes estimados en el modelo TOBIT por el factor de ajuste correspondiente, evaluado en el valor promedio de cada una de las variables independientes, y utilizando el cuadrado de $\overline{EXPER}$ en lugar de usar el promedio de $EXPER_i^2$.

Variable	β_i	$\bar{x}_i$	$\beta_i \cdot \bar{x}_i$
$NWIFEINC$	-8.814	20.129	-177.422
$EDUC$	80.646	12.287	990.881
$EXPER$	131.564	10.631	1398.635
$EXPERSQ$	-1.864	113.014	-210.676
AGE	-54.405	42.538	-2314.272
$KIDSLT6$	-894.022	0.238	-212.523
$KIDSGE6$	-16.218	1.353	-21.947
$Constant$	965.305	1.000	965.305
$\sum \beta_i \cdot \bar{x}_i$			417.981

Condicionando en que el número de horas trabajadas es positivo, para calcular $\partial E\left[y\left|X, y>0\right.\right]/\partial x_j$ el factor de ajuste esta dado por:

$$\frac{\partial E\left[y\left|X, y>0\right.\right]}{\partial x_j} = \beta_j \left\{1 - \lambda\left(X\beta/\hat{\sigma}\right)\left[X\beta/\hat{\sigma} + \lambda\left(X\beta/\hat{\sigma}\right)\right]\right\}.$$

Para calcular esta expresión debemos tener en cuenta que:

$$\begin{aligned} X\beta/\hat{\sigma} &= 417{,}981/1122{,}022 \\ &= 0{,}373 \end{aligned}$$

$$\begin{aligned} \phi\left(X\beta/\hat{\sigma}\right) &= \phi\left(0{,}373\right) \\ &= 0{,}372 \qquad @dnorm\left(\bullet\right) \end{aligned}$$

$$\begin{aligned} \Phi\left(X\beta/\hat{\sigma}\right) &= \Phi\left(0{,}373\right) \\ &= 0{,}645 \qquad @cnorm\left(\bullet\right) \end{aligned}$$

$$\begin{aligned} \lambda\left(X\beta/\hat{\sigma}\right) &= \frac{\phi\left(X\beta/\hat{\sigma}\right)}{\Phi\left(X\beta/\hat{\sigma}\right)} \\ &= \frac{\phi\left(0{,}373\right)}{\Phi\left(0{,}373\right)} \\ &= \frac{0{,}372}{0{,}645} \\ &= 0{,}577 \end{aligned}$$

Por consiguiente,

$$\begin{aligned} \left\{1-\lambda\left(X\beta/\hat{\sigma}\right)\left[X\beta/\hat{\sigma}+\lambda\left(X\beta/\hat{\sigma}\right)\right]\right\} &= \left\{1-0{,}577\left[0{,}373+0{,}577\right]\right\} \\ &= 0{,}452, \end{aligned}$$

que es lo mismo que:

$$\frac{\partial E\left[y\left|X,y>0\right.\right]}{\partial x_j}=\beta_j\times 0{,}452.$$

Por ejemplo, condicionando en que el número de horas trabajadas es positivo, un año adicional de educación (a partir de los valores promedio de las variables) incrementará el valor esperado de horas trabajadas en aproximadamente $80{,}646\times 0{,}452=36{,}452$ horas.

En cuanto al efecto de un hijo adicional de edad menor o igual a 6 años, la caída en horas trabajadas es igual a $-894{,}022\times 0{,}452=-404{,}098$. Debe observarse que esta cifra no tiene sentido para una mujer que trabaja menos de 404,098 horas. En este caso, resulta más conveniente estimar el valor esperado para dos valores diferentes de la variable KIDSLT6, y calcular la diferencia en horas trabajadas.

El factor de ajuste $\Phi\left(X\beta/\hat{\sigma}\right)$ =Prob$(y>0\left|X\right.)$ nos da la probabilidad estimada de observar una respuesta positiva dados los valores de las variables independientes. Si $\Phi\left(X\beta/\hat{\sigma}\right)$ es cercano a uno, entonces es poco probable observar $y_i=0$ cuando $X_i=\bar{X}$. En otras palabras, cuando $\Phi\left(X\beta/\hat{\sigma}\right)$ es cercano a uno no debe haber mucha diferencia entre los resultados de estimar un modelo OLS o un modelo TOBIT.

En el ejemplo de Mroz, el factor de ajuste es $\Phi\left(X\beta/\hat{\sigma}\right)=\Phi\left(0{,}373\right)=0{,}645$, que indica que la probabilidad que una mujer este en la fuerza de trabajo, evaluada en los valores promedio de las variables independientes, es del 0.645.

Por consiguiente, el efecto de cada variable independiente condicionando en $hours$, es decir cuando se tienen en cuenta los individuos que no trabajan ($hours=0$) así como aquellos que trabajan ($hours>0$), es mayor que el efecto cuando únicamente condicionamos en los individuos que trabajan ($hours>0$).

Podemos entonces multiplicar los coeficientes de las variables independientes que son "continuas" para compararlos con los coeficientes que resultan del modelo estimado por OLS. Por ejemplo, el efecto en el modelo TOBIT de un año adicional de educación es $80{,}646 \times 0{,}645 \approx 52{,}017$, que es casi el doble que 28.761.

El R^2 del modelo TOBIT (0.273) corresponde al cuadrado del coeficiente de correlación entre y_i y $\hat{y}_i$.

Ejemplo 13.1 *Modelo Poisson para la probabilidad del número de veces que un individuo es arrestado*

En este ejemplo utilizamos el archivo GROGGER.XLS que contiene la siguiente información para una muestra de 2725 individuos:

$NARR86$	Número de arrestos en 1986
$NFARR86$	Número de arrestos por crímenes graves
$NPARR86$	Número de arrestos por crímenes contra la propiedad
$PCNV$	Proporción de convicciones anteriores
$AVGSEN$	Sentencia promedio (en meses)
$TOTTIME$	Tiempo total pasado en prisión desde los 18 años de edad
$PTIME86$	Tiempo en prisión en 1986
$QEMP86$	Número de trimestres en que el individuo ha estado trabajando en 1986
$INC86$	Ingreso (en $100)
$DURAT$	Tiempo durante el que ha estado desempleado
$BLACK$	Dummy igual a uno si el individuo es de raza negra
$HISPAN$	Dummy igual a uno si el individuo es de origen hispano
$BORN60$	Dummy igual a uno si el individuo nació en 1960

Utilizando esta información formulamos un modelo de regresión Poisson para estudiar los determinantes del número de veces que una persona fue arrestada durante 1986.

Dependent Variable: $NARR86$
Included observations: 2725
Frequencies for dependent variable

			Cumulative	
Value	Count	Percent	Count	Percent
0	1970	72	1970	72.29
1	559	20	2529	92.81
2	121	4	2650	97.25
3	42	1	2692	98.79
4	12	0	2704	99.23
5	13	0	2717	99.71
6	4	0	2721	99.85
7	1	0	2722	99.89
9	1	0	2723	99.93
10	1	0	2724	99.96
12	1	0	2725	100

Dependent Variable: NARR86
Method: ML/QML - Poisson Count (Quadratic hill climbing)
Included observations: 2725

Variable	Coeff.	S.E.	z-Stat.	S.E.*	z-Stat. *
$PCNV$	-0.402	0.085	-4.726	0.105	-3.837
$AVGSEN$	-0.024	0.020	-1.192	0.025	-0.968
$TOTTIME$	0.024	0.015	1.660	0.018	1.348
$PTIME86$	-0.099	0.021	-4.763	0.025	-3.867
$QEMP86$	-0.038	0.029	-1.310	0.036	-1.064
$INC86$	-0.008	0.001	-7.762	0.001	-6.303
$BLACK$	0.661	0.074	8.950	0.091	7.267
$HISPAN$	0.500	0.074	6.761	0.091	5.490
$BORN60$	-0.051	0.064	-0.797	0.079	-0.647
$Constant$	-0.600	0.067	-8.916	0.083	-7.239
R-squared	0.077				
Log likelihood	-2248.761				
Restr. log likelihood	-2441.921				
LR statistic (9 df)	386.320				
Probability(LR stat)	0.000				
Mean dependent var	0.404				
S.D. dependent var	0.859				
$\hat{\sigma}$	1.232				

* indica que los errores estándar, y los correspondientes estadísticos z, han sido corregidos por $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = 1{,}232$. Los errores estándar corregidos son mayores que los no corregidos, y por consiguiente los estadísticos z son menores (en valor absoluto).

Los coeficientes de los modelos OLS y Poisson no son comparables, y tienen una interpretación diferente.

Dependent Variable: $NARR86$
Included observations: 2725

	OLS		Poisson	
Variable	Coeff.	S.E.	Coeff.	S.E. *
$PCNV$	-0.132	0.040	-0.402	0.050
$AVGSEN$	-0.011	0.012	-0.024	0.015
$TOTTIME$	0.012	0.009	0.024	0.012
$PTIME86$	-0.041	0.009	-0.099	0.011
$QEMP86$	-0.051	0.014	-0.038	0.018
$INC86$	-0.001	0.000	-0.008	0.000
$BLACK$	0.327	0.045	0.661	0.056
$HISPAN$	0.194	0.040	0.500	0.049
$BORN60$	-0.022	0.033	-0.051	0.041
$Constant$	0.577	0.038	-0.600	0.047

Por ejemplo, en el modelo estimado por OLS si $\Delta pcnv = 0{,}10$, el número esperado de arrestos disminuye 0.013. Por su parte, en el modelo Poisson si $\Delta pcnv = 0{,}10$, el número esperado de arrestos disminuye 0.0402 (aproximadamente 4 %). En el modelo Poisson el coeficiente asociado a la variable $BLACK$ indica que el número esperado de arrestos para un individuo de raza negra es aproximadamente 66 % mas alto que para un individuo de raza blanca.

Al igual que en el modelo Tobit, el coeficiente de determinación R^2 se calcula como el cuadrado del coeficiente de correlación entre y_i y $\hat{y}_i = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k\right)$.

Utilizando:

$$\Pr\left(y = h \,|x\right) = \frac{\exp^{-\exp(x\beta)} \exp\left(x\beta\right)^h}{h!},$$

se puede calcular la probabilidad para varios valores de h (utilizando los valores promedio de las variables independientes):

Variable	β_i	$\bar{x}_i$	$\beta_i \cdot \bar{x}_i$
$PCNV$	-0.402	0.358	-0.144
$AVGSEN$	-0.024	0.632	-0.015
$TOTTIME$	0.024	0.839	0.021
$PTIME86$	-0.099	0.387	-0.038
$QEMP86$	-0.038	2.309	-0.088
$INC86$	-0.008	54.967	-0.444
$BLACK$	0.661	0.161	0.106
$HISPAN$	0.500	0.218	0.109
$BORN60$	-0.051	0.363	-0.019
$Constant$	-0.600	1.000	-0.600
$\sum \beta_i \cdot \bar{x}_i$			-1.111

A partir de la información de la última línea podemos calcular $\exp\left(-\exp\left(-1{,}111\right)\right) = 0{,}720$ y además $\exp\left(-1{,}111\right) = 0{,}329$. Las probabilidades serían entonces:

$$\begin{aligned}
\Pr\left(y = 0 \,|x\right) &= 0{,}71952 \\
\Pr\left(y = 1 \,|x\right) &= 0{,}23685 \\
\Pr\left(y = 2 \,|x\right) &= 0{,}03898 \\
\Pr\left(y = 3 \,|x\right) &= 0{,}00428 \\
\Pr\left(y = 4 \,|x\right) &= 0{,}00035
\end{aligned}$$

Ejemplo 13.2 *Modelo Poisson para determinar el efecto de la educación sobre la fertilidad de las mujeres en Botswana*

En este ejemplo utilizamos el archivo FERTIL2.XLS con el propósito de investigar el efecto de la educación sobre la fertilidad de las mujeres en Botswana. El archivo contiene, entre otras, la siguiente información para una muestra de 4,358 mujeres:

$CHILDREN$	Número de hijos vivos
$EDUC$	Educación (en años)
AGE	Edad
$EVERMARR$	Dummy que indica si la mujer estuvo alguna vez casada
$URBAN$	Dummy que indica si la mujer vive en un área urbana
$ELECTRIC$	Dummy que indica si el hogar de la mujer tiene acceso a energía eléctrica
TV	Dummy que indica si la mujer tiene televisor en su hogar

Dependent Variable: $CHILDREN$
Included observations: 4358
Frequencies for dependent variable

Value	Count	Percent	Cumulative Count	Cumulative Percent
0	1132	25	1132	25.98
1	905	20	2037	46.74
2	696	15	2733	62.71
3	528	12	3261	74.83
4	392	8	3653	83.82
5	255	5	3908	89.67
6	196	4	4104	94.17
7	134	3	4238	97.25
8	68	1	4306	98.81
9	32	0	4338	99.54
10	13	0	4351	99.84
11	3	0	4354	99.91
12	3	0	4357	99.98
13	1	0	4358	100

Dependent Variable: $CHILDREN$
Method: ML/QML - Poisson Count (Quadratic hill climbing)
Included observations: 4358

Variable	Coeff.	S.E.	z-Stat.	S.E.*	z-Stat. *
$EDUC$	-0.022	0.003	-7.437	0.003	-8.588
AGE	0.337	0.010	33.949	0.009	39.200
AGE^2	-0.004	0.000	-28.331	0.000	-32.779
$EVERMARR$	0.315	0.024	12.875	0.021	14.867
$URBAN$	-0.086	0.022	-3.975	0.019	-4.590
$ELECTRIC$	-0.121	0.039	-3.103	0.034	-3.584
TV	-0.145	0.047	-3.054	0.041	-3.526
$Constant$	-5.375	0.163	-33.001	0.141	-38.108
R-squared	0.597				
Log likelihood	-6497.060				
Restr. log likelihood	-9580.731				
LR statistic (9 df)	6167.343				
Probability(LR stat)	0.000				
Mean dependent var	2.268				
S.D. dependent var	2.222				
$\hat{\sigma}$	0.866				

* indica que los errores estándar, y los correspondientes estadísticos z, han sido corregidos por $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = 0{,}866$. Este ejemplo ilustra el caso en que $\sigma^2 < 1$ (sub-dispersión). Los errores estándar corregidos son menores que los no corregidos, y por consiguiente los estadísticos z son mayores (en valor absoluto).

Los coeficientes estimados por OLS y Poisson no son comparables, y tienen una interpretación diferente.

Dependent Variable: $CHILDREN$
Included observations: 4358

	OLS		Poisson	
Variable	Coeff.	S.E.	Coeff.	S.E.*
$EDUC$	-0.064	0.006	-0.022	0.003
AGE	0.272	0.017	0.337	0.009
AGE^2	-0.002	0.000	-0.004	0.000
$EVERMARR$	0.682	0.052	0.315	0.021
$URBAN$	-0.228	0.046	-0.086	0.019
$ELECTRIC$	-0.262	0.076	-0.121	0.034
TV	-0.250	0.090	-0.145	0.041
$Constant$	-3.394	0.245	-5.375	0.141

Por ejemplo, en el modelo estimado por OLS un año adicional de educación reduce el número esperado de hijos en 0.064. Por su parte, en el modelo Poisson un año adicional de educación disminuye el número esperado de arrestos en aproximadamente 2.2 %. En el modelo Poisson el coeficiente asociado a la variable

TV indica que el número esperado de hijos para una mujer con televisión en su hogar es aproximadamente 14.5 % mas bajo comparado con una mujer que no tiene televisión.

Al igual que en el modelo Tobit, el coeficiente de determinación R^2 se calcula como el cuadrado del coeficiente de correlación entre y_i y $\hat{y}_i = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k\right)$.

Utilizando:

$$\Pr\left(y = h \,|x\right) = \frac{\exp^{-\exp(x\beta)} \exp\left(x\beta\right)^h}{h!},$$

se puede calcular la probabilidad para varios valores de h (utilizando los valores promedio de las variables independientes):

Variable	β_i	$\bar{x}_i$	$\beta_i \cdot \bar{x}_i$
$EDUC$	-0.022	5.856	-0.127
AGE	0.337	27.404	9.244
AGE^2	-0.004	750.996	-3.091
$EVERMARR$	0.315	0.476	0.150
$URBAN$	-0.086	0.517	-0.044
$ELECTRIC$	-0.121	0.140	-0.017
TV	-0.145	0.093	-0.013
$Constant$	-5.375	1.000	-5.375
$\sum \beta_i \cdot \bar{x}_i$			0.727

A partir de la información de la última línea podemos calcular $\exp\left(-\exp\left(0{,}727\right)\right) = 0{,}126$ y además $\exp\left(0{,}727\right) = 2{,}068$. Las probabilidades serían entonces:

$$\begin{aligned}
\Pr\left(y = 0 \,|x\right) &= 0{,}12642 \\
\Pr\left(y = 1 \,|x\right) &= 0{,}26145 \\
\Pr\left(y = 2 \,|x\right) &= 0{,}27036 \\
\Pr\left(y = 3 \,|x\right) &= 0{,}18639 \\
\Pr\left(y = 4 \,|x\right) &= 0{,}09637 \\
\Pr\left(y = 5 \,|x\right) &= 0{,}03986 \\
\Pr\left(y = 6 \,|x\right) &= 0{,}01374 \\
\Pr\left(y = 7 \,|x\right) &= 0{,}00406
\end{aligned}$$