

Técnicas de Análisis de Datos de Elección Discreta

**Sub-Gerencia de Investigación
GPR**

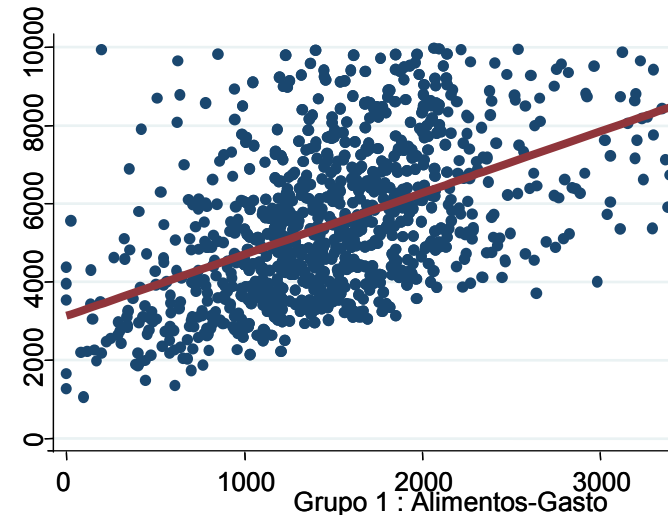
Viernes, 07 de abril de 2006

Contenido

- ✓ Introducción
- ✓ Modelos de variable dependiente binaria
 - Probit
 - Logit
- ✓ Modelos de variable dependiente de elección múltiple
 - Probit ordenado
 - Logit multinomial

Introducción (1)

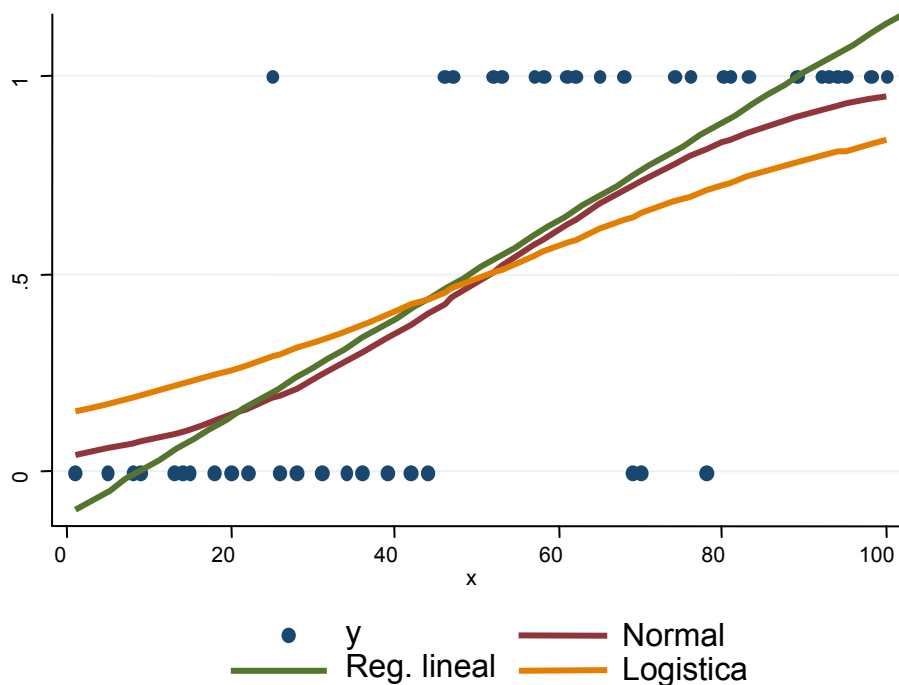
- ✓ Las estimaciones lineales clásicas se usan para identificar asociaciones estadísticas entre variables (dependiente vs. exógenas): $y_i = x_i\beta + e_i$
- ✓ Estas estimaciones funcionan correctamente cuando se hacen sobre variables dependientes continuas:
 - Ingreso del hogar
 - Gasto en telefonía
 - Minutos consumidos
- ✓ Para esto se asumen supuestos sobre la forma del error (homocedasticidad, normalidad)



Introducción (2)

- ✓ Sin embargo, puede ser necesario trabajar con variables dependiente discretas:
 - Acceso a telefonía fija / móvil
 - Calificación sobre el servicio de las empresas (bueno, regular, etc.)
 - Plan tarifario escogido
 - Estrategia de telecomunicación del hogar (fijo, móvil, ambos)
- ✓ En estos casos, los modelos lineales clásicos presentan problemas:
 - $x_i\hat{\beta} \notin [0,1]$
 - En $y_i = x_i\beta + e_i$, e_i tendría una distribución no normal.
 - Heterocedasticidad en el error, producto de la forma de su **varianza**
$$\begin{aligned} \text{Var}[e_i] &= E[y_i - E(y_i)]^2 = E[y_i - x_i\beta]^2 \\ &= E[y_i^2 - 2(y_i)(x_i\beta) + (x_i\beta)^2] \text{ como } y_i = y_i \\ &= x_i\beta - 2(x_i\beta)^2 + (x_i\beta)^2 \\ &= (x_i\beta)(1 - x_i\beta) \end{aligned}$$

Modelos de variable dependiente binaria



- ✓ Por este motivo se recurre a funciones que permiten caracterizar mejor la distribución de la variable dependiente.
- ✓ En el caso de variables dependientes dicotómicas una forma de caracterizar a la variable dependiente es:

$$P(y = 1|x) = G(x\beta) \equiv p(x)$$

- ✓ Este tipo de modelos, debido a que x no afecta directamente a $P(y=1)$, sino a través del índice $x\beta$, son conocidos como **modelos de índices**, donde

$$x\beta = \beta_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

- ✓ Si $G(x\beta)$ se define como la función de dist. normal acum.: PROBIT.
- ✓ Si $G(x\beta)$ se define como la función de dist. logística acum.: LOGIT.

Modelos de variable dependiente binaria (2)

- ✓ Las variables discretas dicotómicas también pueden ser modeladas como realizaciones dependientes de otra variable no observable (*latente*).
- ✓ En este caso, se asume que la variable no observada debe traspasar un umbral para que la variable dependiente tome el valor de 1:

$$\begin{aligned}\Pr[y = 1|x] &= \Pr[y^* > 0] \\ &= \Pr[x' \beta + u > 0]\end{aligned}$$

- ✓ Si se asume que el error está distribuido simétricamente alrededor de cero, entonces podrá replantearse la última expresión como:

$$\Pr[x' \beta + u > 0] = \Pr[-u < x' \beta] = F(x' \beta)$$

- ✓ Donde nuevamente, dependiendo de la forma de la distribución que se le asigne al error, se tratará de un modelo LOGIT o PROBIT.

Modelos de variable dependiente binaria (3)

- ✓ Finalmente, también es posible modelar las variables discretas como resultado de modelos de utilidad aleatoria, donde se asume que el valor observado de la variable discreta representa que dicha alternativa es la que mayor utilidad ofrece al agente.
- ✓ En este caso, se podrían modelar las alternativas como:

$$U_0 = V_0 + \varepsilon_0 \quad \text{y} \quad U_1 = V_1 + \varepsilon_1$$

donde los V representan componentes determinísticos y los ε representan componentes estocásticos (shocks idiosincrásicos).

- ✓ En este caso:
$$\begin{aligned} \Pr[y = 1] &= \Pr[U_1 > U_0] = \Pr[V_1 + \varepsilon_1 > V_0 + \varepsilon_0] \\ &= \Pr[\varepsilon_0 - \varepsilon_1 < V_1 - V_0] \\ &= F(V_1 - V_0) \end{aligned}$$
- ✓ Dependiendo de la forma funcional que se asuma para la diferencia de los errores se puede llegar a los conocidos modelos probit y logit:
 - Si se asume que los errores son normales, su diferencia es normal, y se estaría en el modelo *probit*.
 - Si se asume que los errores son independientes con distribución “*valor extremo tipo 1*”, entonces la diferencia tendría una distribución logística (*logit*).

Modelo probit (1)

- ✓ En cualquiera de las especificaciones planteadas, el modelo probit representa:

$$p = \Phi(x' \beta) = \int_{-\infty}^{x' \beta} \phi(z) dz$$

- ✓ donde $\Phi(\cdot)$ es la distribución acumulada normal estándar.
- ✓ Los efectos marginales, a diferencia de las regresiones lineales, no son los parámetros, sino una función de los mismos:

$$\frac{\partial p}{\partial x_j} = \phi(x' \beta) \beta_j$$

- ✓ Las estimaciones se realizan usando la metodología de máxima verosimilitud, y son fácilmente manejables usando distintos paquetes econométricos (Stata, EVIEWS, SPSS, etc.)

Modelo probit: estimación en Stata

```
. probit movil mieperho ingre anho_est uso_cab j_hombre j_anho_est p_gas_elec
```

```
Iteration 0: log likelihood = -1219.1252
```

```
Iteration 4: log likelihood = -1054.3311
```

```
Probit regression
```

```
Number of obs = 1968
```

```
LR chi2(7) = 329.59
```

```
Prob > chi2 = 0.0000
```

```
Log likelihood = -1054.3311
```

```
Pseudo R2 = 0.1352
```

```
-----
```

movil	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
mieperho	.1000729	.015791	6.34	0.000	.069123	.1310227
ingre	.0000285	9.16e-06	3.11	0.002	.0000106	.0000465
anho_est	.0779782	.0131391	5.93	0.000	.052226	.1037304
uso_cab	.5926038	.12709	4.66	0.000	.3435119	.8416956
j_hombre	-.1439711	.0759859	-1.89	0.058	-.2929006	.0049585
j_anho_est	.0575237	.0104097	5.53	0.000	.0371212	.0779263
p_gas_elec	.4324145	.0959651	4.51	0.000	.2443264	.6205026
_cons	-2.675154	.1598882	-16.73	0.000	-2.988529	-2.361779

Modelo probit: estimación en Stata (2)

```
. dprobit movil mieperho ingre anho_est uso_cab j_hombre j_anho_est p_gas_elec
```

```
Probit regression, reporting marginal effects          Number of obs =   1968
                                                    LR chi2(7)      = 329.59
                                                    Prob > chi2    = 0.0000
Log likelihood = -1054.3311                          Pseudo R2      = 0.1352
```

movil	dF/dx	Std. Err.	z	P> z	x-bar	[95% C.I.]
mieperho	.0336254	.005277	6.34	0.000	4.41108	.023283	.043968	
ingre	9.59e-06	3.09e-06	3.11	0.002	1928.16	3.5e-06	.000016	
anho_est	.0262014	.0043971	5.93	0.000	8.59146	.017583	.03482	
uso_cab	.1991203	.0426183	4.66	0.000	.181727	.11559	.282651	
j_hombre*	-.0493232	.0265011	-1.89	0.058	.744411	-.101264	.002618	
j_anho~t	.0193285	.0034869	5.53	0.000	9.94461	.012494	.026163	
p_gas_~c*	.1326205	.0261905	4.51	0.000	.809959	.081288	.183953	

obs. P	.3104675							
pred. P	.2789491 (at x-bar)							

```
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| correspond to the test of the underlying coefficient being 0
```

Modelo probit: estimación en Stata (3)

Probit model for movil

Classified	True		Total
	D	~D	
+	202	131	333
-	409	1226	1635
Total	611	1357	1968

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

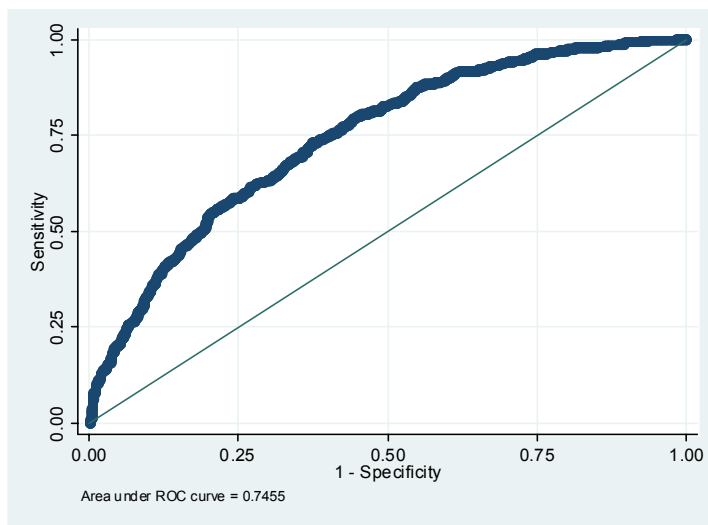
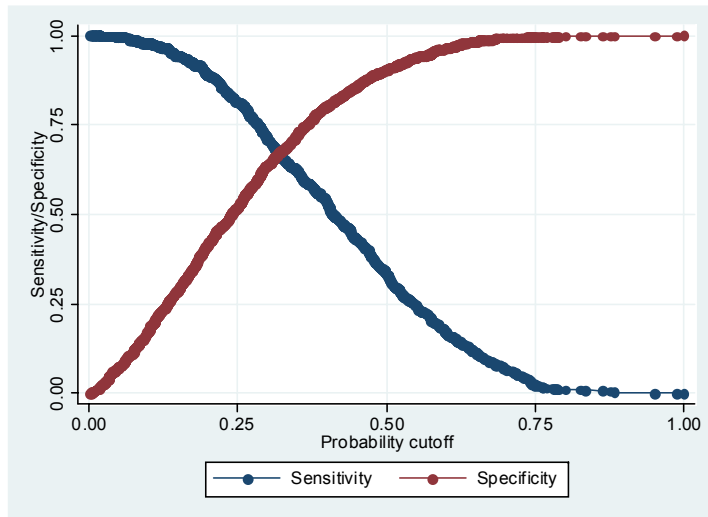
True D defined as movil != 0

Sensitivity	Pr(+ D)	33.06%
Specificity	Pr(- ~D)	90.35%
Positive predictive value	Pr(D +)	60.66%
Negative predictive value	Pr(~D -)	74.98%
False + rate for true ~D	Pr(+ ~D)	9.65%
False - rate for true D	Pr(- D)	66.94%
False + rate for classified +	Pr(~D +)	39.34%
False - rate for classified -	Pr(D -)	25.02%
Correctly classified		72.56%

✓ Los valores predichos pueden ser una medida de bondad del modelo, pero en variables concentradas en un valor, es mejor no considerarlos.

✓ Los porcentajes corresponden a los porcentajes verticales y horizontales del cuadro.

Modelo probit: estimación en Stata (4)



- ✓ Una mejor medida del ajuste es la curva ROC (receiver operating characteristics):
 - Fracción de $y=1$ predichos correctamente (*sensitivity*) contra la fracción de $y=0$ valorados incorrectamente ($1 - \textit{specificity}$), para cada valor de corte.
- ✓ Idealmente, en el primer gráfico el cruce de las curvas debería estar en un parte alta del cuadro.
- ✓ En el segundo gráfico, el área bajo la curva ROC debería acercarse lo más posible a 1.

Modelo logit (1)

- ✓ Igualmente, en cualquiera de las especificaciones planteadas, el modelo logit representa:

$$p = \Lambda(x' \beta) = \frac{e^{x' \beta}}{1 + e^{x' \beta}} = \frac{1}{1 + e^{-x' \beta}}$$

- ✓ donde $\Lambda(\bullet)$ es la distribución acumulada logística.
- ✓ Los efectos marginales tienen la siguiente forma:

$$\frac{\partial p}{\partial x_j} = \Lambda(x' \beta)[1 - \Lambda(x' \beta)]\beta_j$$

- ✓ Las estimaciones se realizan también usando la metodología de máxima verosimilitud, y al igual que los modelos probit son fácilmente manejables usando programas como Stata, EVIEWS o SPSS

Modelo logit: estimación en Stata (1)

```
. logit movil mieperho ingre anho_est uso_cab j_hombre j_anho_est p_gas_elec
```

```
Iteration 0: log likelihood = -1219.1252
```

```
Iteration 4: log likelihood = -1053.0395
```

Logistic regression

Number of obs = 1968

LR chi2(7) = 332.17

Prob > chi2 = 0.0000

Log likelihood = -1053.0395

Pseudo R2 = 0.1362

movil	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
mieperho	.1740915	.0270953	6.43	0.000	.1209857	.2271974
ingre	.0000827	.0000258	3.21	0.001	.0000322	.0001332
anho_est	.1220599	.022496	5.43	0.000	.0779685	.1661513
uso_cab	.9862105	.2126703	4.64	0.000	.5693844	1.403037
j_hombre	-.2405713	.1285968	-1.87	0.061	-.4926165	.0114738
j_anho_est	.0969034	.0178906	5.42	0.000	.0618386	.1319683
p_gas_elec	.7535263	.1730207	4.36	0.000	.414412	1.092641
_cons	-4.523344	.2855136	-15.84	0.000	-5.08294	-3.963747

Modelo logit: estimación en Stata (2)

```
. dlogit2 movil mieperho ingre anho_est uso_cab j_hombre j_anho_est p_gas_elec
```

Marginal effects from logit

Number of obs = 1968

chi2(7) = 283.58

Prob > chi2 = 0.0000

Log Likelihood = -1053.0395

Pseudo R2 = 0.1362

movil	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
mieperho	.0346503	.0053571	6.47	0.000	.0241505	.04515
ingre	.0000165	5.17e-06	3.18	0.001	6.32e-06	.0000266
anho_est	.0242942	.0044411	5.47	0.000	.0155898	.0329985
uso_cab	.1962902	.042186	4.65	0.000	.1136071	.2789733
j_hombre	-.0478821	.0255869	-1.87	0.061	-.0980314	.0022673
j_anho_est	.0192872	.003531	5.46	0.000	.0123665	.0262078
p_gas_elec	.1499779	.0339222	4.42	0.000	.0834917	.2164642
_cons	-.9003025	.0493696	-18.24	0.000	-.9970652	-.8035399

Modelo logit: estimación en Stata (3)

```
. estat class
```

```
Logistic model for movil
```

Classified	True		Total
	D	~D	
+	209	134	343
-	402	1223	1625
Total	611	1357	1968

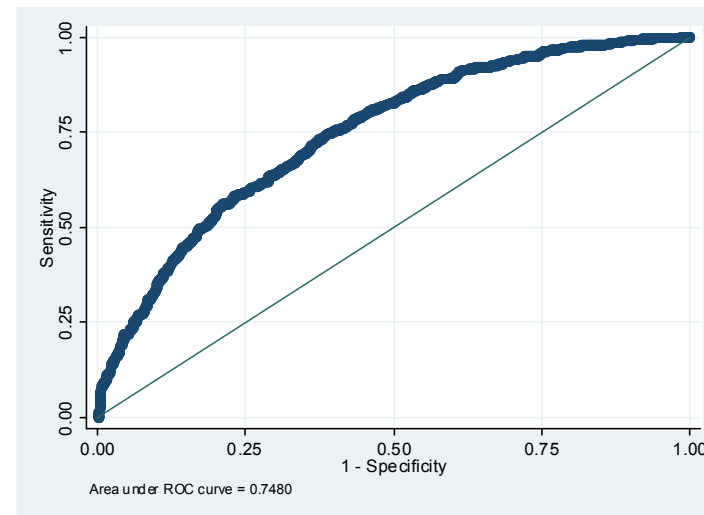
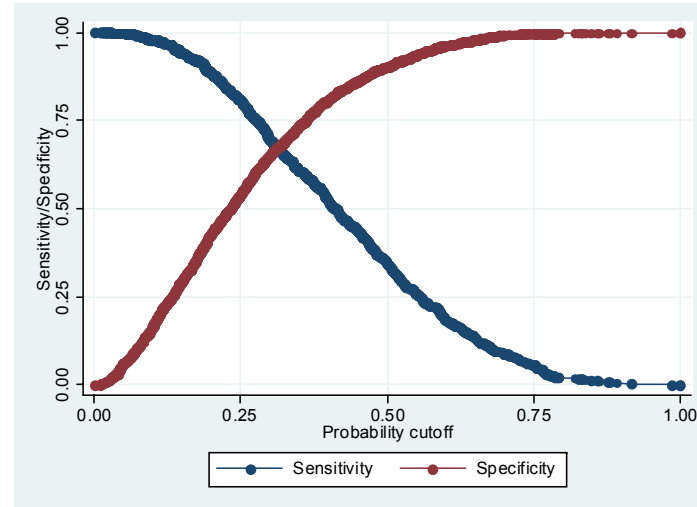
```
Classified + if predicted Pr(D) >= .5
```

```
True D defined as movil != 0
```

Sensitivity	Pr(+ D)	34.21%
Specificity	Pr(- ~D)	90.13%
Positive predictive value	Pr(D +)	60.93%
Negative predictive value	Pr(~D -)	75.26%

False + rate for true ~D	Pr(+ ~D)	9.87%
False - rate for true D	Pr(- D)	65.79%
False + rate for classified +	Pr(~D +)	39.07%
False - rate for classified -	Pr(D -)	24.74%

```
Correctly classified 72.76%
```



Probit vs. Logit

- ✓ Pequeñas diferencias en probabilidades predichas (mayores en las colas de la distribución).
- ✓ Parámetros estimados difieren, como consecuencia de las distintas distribuciones que se asumen.
- ✓ Se puede asumir cierta correspondencia:

$$\hat{\beta}_{Logit} \cong 4\hat{\beta}_{MCO}$$

$$\hat{\beta}_{Probit} \cong 2.5\hat{\beta}_{MCO}$$

$$\hat{\beta}_{Logit} \cong 1.6\hat{\beta}_{Probit}$$

- ✓ Es posible hacer comparaciones basadas en el logaritmo del ratio de verosimilitud, siempre que ambos modelos tengan la misma cantidad de parámetros.
- ✓ Sin embargo, por lo general, los valores de los logaritmos de los ratios suelen ser muy cercanos, lo que implica poca ganancia al pasar de un modelo a otro.

Modelos de variable dependiente de elección múltiple

- ✓ Cuando se trabaja con variables dependientes con más de dos categorías, los modelos binarios resultan insuficientes.
- ✓ En estos casos, debe diferenciarse si la variable dependiente corresponde a:
 - un ordenamiento natural (p.ej. bueno, regular, malo)
 - respuestas no ordenadas (p.ej. plan tarifario escogido).
- ✓ En el caso de ordenamientos naturales, la forma más común de abordarlos son los modelos probit ordenados.
- ✓ Para el caso de respuestas no ordenadas, se suele trabajar con modelos logit multinomiales.

Modelos probit ordenados

- ✓ Los modelos probit suponen variables discretas que toman valores de acuerdo a la siguiente especificación:

$$y_i = 0 \quad \text{si} \quad y_i^* < \gamma_1$$

$$y_i = 1 \quad \text{si} \quad \gamma_1 \leq y_i^* < \gamma_2$$

$$y_i = 2 \quad \text{si} \quad \gamma_2 < y_i^*$$

- ✓ En este caso, los parámetros del modelo son los β y γ .
- ✓ Los γ representan los umbrales que determinan el valor de y_i para el valor alcanzado por y_i^* .
- ✓ Por tanto, la probabilidad de cada alternativa es:

$$\Pr[y_i = 0] = \Pr[y_i^* < \gamma_1] = \Pr[X_i \beta + u_i < \gamma_1] = \Phi(\gamma_1 - X_i \beta)$$

$$\Pr[y_i = 1] = \Pr[\gamma_1 \leq y_i^* < \gamma_2] = \Pr[\gamma_1 < X_i \beta + u_i < \gamma_2] = \Phi(\gamma_2 - X_i \beta) - \Phi(\gamma_1 - X_i \beta)$$

$$\Pr[y_i = 2] = \Pr[y_i^* \geq \gamma_2] = \Pr[X_i \beta + u_i \geq \gamma_2] = \Phi(X_i \beta - \gamma_2)$$

- ✓ Las tres probabilidades se integran en una única expresión que se estima por el método de Máxima Verosimilitud

Probit ordenado: estimación en Stata (1)

```
Ordered probit regression                Number of obs   =       18668
                                         LR chi2(8)      =       193.05
                                         Prob > chi2     =       0.0000
Log likelihood = -16475.481             Pseudo R2       =       0.0058
```

percep_hogar	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
ingre	9.93e-06	5.14e-06	1.93	0.054	-1.50e-07	.00002
anho_est	-.0214823	.002982	-7.20	0.000	-.0273269	-.0156376
tup	-.0663465	.0198465	-3.34	0.001	-.105245	-.027448
uso_cab	.330814	.0450017	7.35	0.000	.2426122	.4190158
pared_lad	-.0558312	.0202208	-2.76	0.006	-.0954633	-.0161991
j_edad	.0010595	.0005769	1.84	0.066	-.0000713	.0021902
j_hombre	-.0450704	.0216848	-2.08	0.038	-.0875718	-.0025691
transf_nac	-.154807	.0183677	-8.43	0.000	-.190807	-.118807
/cut1	-1.662974	.0408116			-1.742964	-1.582985
/cut2	.4259358	.0392084			.3490887	.5027829
/cut3	1.90544	.0433183			1.820538	1.990342

Probit ordenado: estimación en Stata (2)

```
. mfx, predict(p outcome(1))
Marginal effects after oprobit
      y = Pr(percep_hogar==1) (predict, p outcome(1)) = .65781435
-----+-----
variable |          dy/dx      Std. Err.      z    P>|z|    [   95% C.I.   ]      X
-----+-----
      ingre | -1.99e-06          .00000      -1.93   0.054   -4.0e-06   3.2e-08   1030.77
     anho_est |  .0043159          .00006       7.14   0.000   .003131   .005501   6.32673
          tup* |  .0130025          .00038       3.42   0.001   .005558   .020447   .316317
        uso_cab | -.0664627          .00914      -7.27   0.000  -.084372  -.048554   .118399
-----+-----
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

```
. mfx, predict(p outcome(2))
Marginal effects after oprobit
      y = Pr(percep_hogar==2) (predict, p outcome(2)) = .25500659
-----+-----
variable |          dy/dx      Std. Err.      z    P>|z|    [   95% C.I.   ]      X
-----+-----
      ingre |  2.85e-06          .00000       1.93   0.054  -4.4e-08   5.7e-06   1030.77
     anho_est | -.0061686          .00086      -7.18   0.000  -.007853  -.004484   6.32673
          tup* | -.0189602          .00565      -3.36   0.001  -.030028  -.007893   .316317
        uso_cab |  .0949926          .01296       7.33   0.000   .069598   .120388   .118399
-----+-----
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Modelos logit multinomiales (1)

- ✓ A diferencia de los modelos ordenados, en que la elección depende una única función índice, en los modelos de respuesta no ordenada, este supuesto carece de sentido.
- ✓ El enfoque más simple utilizado es el del modelo logit multinomial.
- ✓ Este modelo, que se plantea para $J+1$ opciones posibles corresponde a la siguiente especificación:

$$\Pr[y_i = 0] = \frac{1}{1 + \sum_{j=1}^J \exp(X_i \beta^j)}$$

$$\Pr[y_i = l] = \frac{\exp(X_i \beta^l)}{1 + \sum_{j=1}^J \exp(X_i \beta^j)} \quad \text{para } l = 1, \dots, J.$$

- ✓ En esta especificación, existe un conjunto de parámetros diferente para cada una de las alternativas posibles.

Modelos logit multinomiales (2)

- ✓ Una propiedad importante de los modelos logit multinomial es:

$$\frac{\Pr[y_i = l]}{\Pr[y_i = j]} = \frac{\exp(X_i \beta^l)}{\exp(X_i \beta^j)} = \exp(X_i (\beta^l - \beta^j))$$

- ✓ Esta propiedad permite reducir a un modelo logit binomial la elección entre dos categorías específicas, condicionado a que la elección fue realizada entre dichas variables
- ✓ En *Stata*, la estimación del logit multinomial se realiza por el método de máxima verosimilitud, y se estiman los diferenciales de los parámetros.

Logit multinomial: estimación en Stata (1)

```
. mlogit var1 mieperho ingre tup if dominio==8
```

```
Multinomial logistic regression
```

```
Number of obs = 2208
```

```
LR chi2(9) = 456.36
```

```
Prob > chi2 = 0.0000
```

```
Log likelihood = -2642.5051
```

```
Pseudo R2 = 0.0795
```

		var1	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
1	mieperho		.257311	.030051	8.56	0.000	.1984122	.3162098
	ingre		.000807	.0000628	12.85	0.000	.0006839	.0009301
	tup		-.0087858	.1127649	-0.08	0.938	-.2298009	.2122294
	_cons		-2.498002	.1915335	-13.04	0.000	-2.8734	-2.122603
2	mieperho		.0802862	.0421376	1.91	0.057	-.0023019	.1628744
	ingre		.0005242	.0000799	6.56	0.000	.0003677	.0006808
	tup		.7618867	.1496397	5.09	0.000	.4685983	1.055175
	_cons		-2.595982	.2577029	-10.07	0.000	-3.10107	-2.090894
3	mieperho		.3715936	.0327326	11.35	0.000	.3074389	.4357483
	ingre		.0009472	.0000644	14.71	0.000	.000821	.0010735
	tup		-.0371722	.130377	-0.29	0.776	-.2927064	.218362
	_cons		-3.748492	.2165727	-17.31	0.000	-4.172967	-3.324017

```
(var1==0 is the base outcome)
```


Logit multinomial: estimación en Stata (2)

```
. mfx compute, predict(outcome(1))
Marginal effects after mlogit
      y = Pr(var1==1) (predict, outcome(1))
      = .35306993
```

variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]	X
mieperho	.0270817	.00536	5.05	0.000	.01657 .037594	4.53487
ingre	.0000903	.00001	10.86	0.000	.000074 .000107	1802.84
tup*	-.0330029	.02192	-1.51	0.132	-.075957 .009952	.404891

(*) dy/dx is for discrete change of dummy variable from 0 to 1

```
. mfx compute, predict(outcome(2))
Marginal effects after mlogit
      y = Pr(var1==2) (predict, outcome(2))
      = .11771168
```

variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]	X
mieperho	-.011809	.00389	-3.03	0.002	-.019437 -.004181	4.53487
ingre	-3.19e-06	.00001	-0.49	0.622	-.000016 9.5e-06	1802.84
tup*	.0857562	.01571	5.46	0.000	.054965 .116547	.404891

(*) dy/dx is for discrete change of dummy variable from 0 to 1

Técnicas de Análisis de Datos de Elección Discreta

**Sub-Gerencia de Investigación
GPR**

Viernes, 07 de abril de 2006