Modelos de elección discreta
Aplicaciones en ordenador

Román Salmerón Gómez

Para ilustrar cómo abordar el análisis de **Modelos de elección discreta** con el software econométrico **Gretl** resolveremos el siguiente problema obtenido de Santos y otros [1]:

> *Un banco dispone de una base de datos de antiguos receptores de créditos en la que se recoge información (ver tabla del Cuadro 1) acerca de la devolución del mismo, ingresos, situación laboral y cargas del cliente. Basándose en ella quiere obtener un modelo que le permita conocer con un alto nivel de fiabilidad qué clientes devolverán el crédito y cuáles no. Decide para su construcción emplear la técnica de regresión logística.*

Modelo lineal de probabilidad

Como es sabido, el modelo lineal de probabilidad consiste en estimar por **MCO** el modelo planteado, que en este caso es:

$$credito_t = \beta_1 + \beta_2 ingresos_t + \beta_3 laboral_t + \beta_4 cargas_t + u_t. \tag{1}$$

Los resultados obtenidos son los siguientes:

Modelo 1: MCO, usando las observaciones 1–57
Variable dependiente: credito

	Coeficiente	Desv. Típica	Estadístico t	Valor p
const	−0.0309352	0.0863636	−0.3582	0.7216
ingresos	0.0412042	0.00870463	4.7336	0.0000
laboral	0.337372	0.0515758	6.5413	0.0000
cargas	−0.191641	0.0753203	−2.5443	0.0139

Media de la vble. dep.	0.508772	D.T. de la vble. dep.	0.504367
Suma de cuad. residuos	3.812977	D.T. de la regresión	0.268222
R^2	0.732340	R^2 corregido	0.717190
$F(3, 53)$	48.33754	Valor p (de F)	3.44e–15
Log-verosimilitud	−3.797228	Criterio de Akaike	15.59446
Criterio de Schwarz	23.76666	Hannan–Quinn	18.77045

Sin embargo, este modelo presenta diversos problemas, como es el de proporcionar estimaciones fuera del intervalo [0, 1]. Lo cual no es posible ya que se está analizando la probabilidad de devolución del crédito. A continuación se puede observar como se obtienen estimaciones negativas y superiores a 1:

```
      credito     estimada      residuo
52   0.000000     0.092677    -0.092677
53   0.000000    -0.127806     0.127806
54   0.000000    -0.119565     0.119565
55   1.000000     1.138260    -0.138260
56   1.000000     1.055851    -0.055851
57   0.000000     0.800887    -0.800887
```

Por tanto, no le dedicaremos más tiempo.

Modelo Logit

Para estimar un modelo Logit con Gretl (versión 1.9.13) hay que seguir la ruta *Modelo* → *Variable dependiente limitada* → *Logit* → *Binario*. En tal caso aparecerá una nueva ventana donde hay que:

- Especificar la variable dependiente y las independientes.
- Indicar si se desean obtener los detalles de las iteraciones realizadas en la estimación máximo verosímil.
- Elegir si se desea mostrar las pendientes evaluadas en la media de cada variable o el p-valor asociado al contraste de significación individual de cada coeficiente.

En este caso se obtienen los siguientes resultados:

Modelo 2: Logit, usando las observaciones 1–57
Variable dependiente: credito
Desviaciones típicas basadas en el Hessiano

	Coeficiente	Desv. Típica	z	Valor p
const	−6.60045	2.38599	−2.7663	0.0057
ingresos	0.493865	0.210969	2.3409	0.0192
laboral	3.79399	1.36302	2.7835	0.0054
cargas	−2.21248	1.33000	−1.6635	0.0962

Media de la vble. dep.	0.508772	D.T. de la vble. dep.	0.504367
R^2 de McFadden	0.766499	R^2 corregido	0.665235
Log-verosimilitud	−9.223416	Criterio de Akaike	26.44683
Criterio de Schwarz	34.61904	Hannan–Quinn	29.62283

Número de casos 'correctamente predichos' = 53 (93.0 por ciento)
Contraste de razón de verosimilitudes: $\chi^2(3) = 60.554$ [0.0000]

		Predicho	
		0	1
Observado	0	26	2
	1	2	27

La salida obtenida nos resulta muy familiar, si bien hay algunas opciones nuevas:

- El pseudo R^2 de McFadden.
- El contraste de razón de verosimilitudes (para la significación conjunta).
- El número (porcentaje) de casos "correctamente predichos" y un cuadro que muestra en detalle el acierto del modelo en su predicción mediante la comparación entre los predichos y los observados (el umbral es 0'5).

Una alternativa a la razón de verosimilitudes es el Test de Wald, el cual se encuentra en la ruta *Contrastes* → *Omitir variables* de la ventana de resultados. Eligiendo todos los regresores y seleccionando el contraste comentado se obtiene la siguiente salida:

```
Contraste sobre el Modelo 2:

  Hipótesis nula: los parámetros de regresión son cero para las variables
    ingresos, laboral, cargas
  Estadístico de contraste: F(3, 53) = 3.74381, Valor p 0.0163267
```

Evidentemente deben salir resultados similares en ambos casos.

Modelo Probit

Estimar suponiendo un modelo Probit es un proceso totalmente análogo. En este caso la ruta a seguir es: *Modelo → Variable dependiente limitada → Probit → Binario.*

Modelo 6: Probit, usando las observaciones 1–57
Variable dependiente: credito
Desviaciones típicas basadas en el Hessiano

	Coeficiente	Desv. Típica	z	Valor p
const	−3.73552	1.32075	−2.8283	0.0047
ingresos	0.258077	0.0996376	2.5902	0.0096
laboral	2.19874	0.731338	3.0065	0.0026
cargas	−1.20333	0.691663	−1.7398	0.0819

Media de la vble. dep.	0.508772	D.T. de la vble. dep.	0.504367
R^2 de McFadden	0.767750	R^2 corregido	0.666485
Log-verosimilitud	−9.174030	Criterio de Akaike	26.34806
Criterio de Schwarz	34.52026	Hannan–Quinn	29.52406

Número de casos 'correctamente predichos' = 53 (93.0 por ciento)
Contraste de razón de verosimilitudes: $\chi^2(3) = 60.653$ [0.0000]

		Predicho	
		0	1
Observado	0	26	2
	1	2	27

Contraste de normalidad de los residuos –
Hipótesis nula: el error se distribuye normalmente
Estadístico de contraste: $\chi^2(2) = 17.7544$
con valor p = 0.000139534

Contraste de omisión de variables –
Hipótesis nula: los parámetros son cero para las variables
ingresos
laboral
cargas
Estadístico de contraste: $F(3, 53) = 4.57457$
con valor p = $P(F(3, 53) > 4.57457) = 0.00637471$

Anexo: modelos de elección discreta con el entorno R

Analizar modelos de elección discreta mediante una regresión Logit o Probit se realiza mediante el comando *glm*:

```
glm(funcion, family=binomial("logit/probit"))
```

donde `funcion` es la relación entre variables independientes y la dependiente y en `family` hay que especificar si es un modelo logit o probit.

Guardados los datos del ejercicio con formato *.txt* donde las columnas están separadas por tabulaciones, un posible código a usar para analizar el problema mediante un modelo logit es el siguiente:

```
# leo datos
    datos = read.table("datos.txt", header=T, sep="\t")
    attach(datos)
```

```
# regresión logística
    logreg = glm(credito~ingresos+laboral+cargas, family=binomial("logit"))
    summary(logreg)

# odd-ratio
    exp(logreg$coef)

# intervalos de confianza
    confint(logreg)

# tabla de clasificación
    yajus=fitted.values(logreg)
    tcc(mean(credito),yajus,credito)
```

Adviértase que la función *tcc* calcula ta tasa de aciertos del modelo y ha tenido que ser implementa (ver [2]) con el siguiente código:

```
tcc<-function(corte,yajus,y)
{
    verpos<-table(yajus>corte & y==1)[2]
    falpos<-table(yajus>corte & y==0)[2]
    falneg<-table(yajus<corte & y==1)[2]
    verneg<-table(yajus<corte & y==0)[2]
    tasa<-(verpos+verneg)/(verpos+falpos+falneg+verneg)*100
    tasa
}
```

Los resultados obtenidos son los siguientes:

```
> summary(logreg)

    Call:
    glm(formula = credito ~ ingresos + laboral + cargas, family = binomial("logit"))

    Deviance Residuals:
         Min        1Q    Median        3Q       Max
    -2.51521  -0.10923   0.02721   0.24917   1.21595

    Coefficients:
                Estimate Std. Error z value Pr(>|z|)
    (Intercept)  -6.6004     2.3860  -2.766  0.00567 **
    ingresos      0.4939     0.2110   2.341  0.01923 *
    laboral       3.7940     1.3630   2.784  0.00538 **
    cargas       -2.2125     1.3300  -1.664  0.09620 .
    ---
    Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    (Dispersion parameter for binomial family taken to be 1)

        Null deviance: 79.001  on 56  degrees of freedom
    Residual deviance: 18.447  on 53  degrees of freedom
    AIC: 26.447

    Number of Fisher Scoring iterations: 7
```

```
> exp(logreg$coef)

    (Intercept)    ingresos     laboral      cargas
     0.00135976  1.63863764 44.43337841  0.10942864

> confint(logreg)

    Waiting for profiling to be done...
                        2.5 %     97.5 %
    (Intercept) -12.7646873 -2.9543283
    ingresos      0.1734533  1.0445330
    laboral       1.6724962  7.3297632
    cargas       -5.4247879  0.2102243

> yajus=fitted.values(logreg)
> tcc(mean(credito),yajus,credito)

        TRUE
    92.98246
```

Para ajustar un modelo probit simplemente hay que modificar el código anterior cambiando logit por probit.

Referencias

[1] Santos, J., Muñoz, A., Juez, P. y Cortiñas, P. (2003). *Diseño de encuestas para estudios de mercado. Técnicas de muestreo y análisis multivariante*. Editorial Ramón Areces, S.A. Madrid.

[2] Salazar, A. (2011). *Modelos de respuesta discreta en R y aplicación con datos reales*. Trabajo fin de máster del Máster Oficial en estadística Aplicada de la Universidad de Granada. Dirección web: http://masteres.ugr.es/moea/pages/tfm1011/modelosderespuestadiscretaenryaplicacion.

credito	ingresos	laboral	cargas	credito	ingresos	laboral	cargas
1	4	2	0	1	25	1	1
1	4.5	2	1	0	2	0	1
1	5	2	0	0	2	0	1
1	4	2	0	1	10	2	0
0	2.5	1	1	0	4	1	0
0	2.5	1	1	1	13	2	0
1	5	2	0	1	7.7	2	0
1	6	2	1	0	2.4	1	1
0	3	0	0	1	8	2	0
0	1.6	0	0	1	5.5	1	0
1	7	2	0	0	3.2	0	1
0	4	1	1	0	3.1	1	0
1	7.6	2	0	0	3.6	1	1
1	3	2	0	1	14	2	0
0	1.4	0	1	1	12	1	0
0	1.8	0	1	1	10	2	1
0	4	0	0	0	2.4	0	0
0	2	1	1	0	2.1	0	1
0	6	2	1	0	4	1	1
1	7.2	1	0	1	6.9	2	0
1	15	1	1	0	3.1	0	1
1	10	1	1	1	8.2	1	0
0	1.4	0	1	1	12	2	0
0	4	1	1	0	3	0	0
0	2	0	0	0	2.3	0	1
1	14	1	0	0	2.5	0	1
1	10.3	1	0	1	12	2	0
1	7.5	2	1	1	10	2	0
0	1.4	1	1	0	12	1	0

Cuadro 1: Información acerca de la devolución de créditos