UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2012, Vol. 5, No. 2

Modelo *Logit* y *Probit*: un caso de aplicación

Logit and Probit Models: an application

Orlando Moscote Flórez[a]
orlandomoscote@usantotomas.edu.co

William Arley Rincón[b]
williamrincon@usantotomas.edu.co

Resumen

Los modelos de regresión con respuesta cualitativa son modelos en los cuales la variable dependiente puede ser de naturaleza cualitativa, mientras que las variables independientes pueden ser cualitativas o cuantitativas o una mezcla de las dos; por ejemplo, si se está estudiando la relación entre ingresos y el poseer o no una vivienda, la respuesta solo puede tomar dos valores (si posee una vivienda o no la posee), la variable independiente puede ser los ingresos familiares, el estrato social de esa familia y la cantidad de personas en el hogar, entre otras. Los modelos de respuesta cualitativa no tienen que restringirse simplemente a respuestas de sí o no; la variable respuesta puede tomar más de dos valores, ser tricotómica o politómica, también se establecen modelos en los que la variable dependiente es de carácter ordinal o de carácter nominal, en donde no hay preestablecido ningún tipo de orden.

En el presente artículo se analiza el modelo lineal de probabilidad (MLP) y el modelo *logit* como alternativa al MLP, con la idea de subsanar algunos de los problemas que presenta dicho modelo. Los modelos mencionados son ilustrados utilizando una muestra de datos seleccionados del archivo de datos "seguimiento a usuarios línea base", la cual es parte de las bases de datos del estudio que se desarrolló en el marco del convenio interadministrativo 012 en el año 2009, entre la UAES (Unidad Administrativa Especial de Servicios Públicos) y la Universidad Distrital Francisco José de Caldas, convenio en el cual participó uno de los autores de este artículo.

***Palabras clave*:** respuesta cualitativa, modelo lineal de probabilidad, modelo *logit*.

[a]Docente, Facultad de Estadística, Universidad Santo Tomás.
[b]Docente, Facultad de Estadística, Universidad Santo Tomás.

Abstract

Regression models are qualitative response regression models in which the dependent variable can be of qualitative nature, while the independent variables can be qualitative or quantitative or a mixture of both, for example if you are studying the relationship between incomes and possessing or not a home, the answer can only take two values (if you own a home or if you do not); the independent variable may be the family incomes, the social status of the family and the number of people in the home, among others. Qualitative response models do not need to be restricted to a simply yes or no, the variable response can take more than two values, for example it can be trichotomous or polytomous, also, there are some models in which the dependent variable is ordinal or nominal in nature where there is no a predetermined order.

This article reviews the linear probability model (LPM) and the logit model as an alternative to the LPM; what is more, as an idea of solving some of the problems with the first mentioned model. The two models mentioned above are illustrated using a sample selected from the data base "baseline monitoring users", which is a part of the study database that was developed under the administrative agreement 012 in 2009, between The UAES (Special Administrative Unit of Public Services) and Francisco José de Caldas University , from this administrative agreement one of the authors of this article was involved in.

***Key words*:** qualitative response, linear probability model, logit model.

1. Modelos de regresión de respuesta cualitativa

Los modelos de regresión con respuesta cualitativa son modelos de regresión en los cuales la variable dependiente puede ser de naturaleza cualitativa, mientras que las variables independientes pueden ser cualitativas o cuantitativas, o una mezcla de las dos; por ejemplo, si se está estudiando la relación entre ingresos y el pagar o no impuesto de renta, la respuesta o regresada solo puede tomar dos valores (si paga impuesto de renta o no paga dicho impuesto); otros ejemplos en que la regresada es cualitativa son si la familia posee o no vivienda propia, se aprueba o pierde un curso, padece determinada enfermedad o no la padece. La variable cualitativa en estos tipos de modelos no tiene que restringirse simplemente a respuestas de sí o no, la variable respuesta puede tomar más de dos valores, ser tricotómica o politómica, también se establecen modelos en lo que la variable dependiente es de carácter ordinal o de carácter nominal, en donde no hay preestablecido ningun tipo de orden. En este artículo se analizan algunos de los modelos en donde la variable dependiente es de carácter binario o dicotómica (sí o no). Hay cuatro métodos para crear un modelo de probabilidad para una variable de respuesta binaría (Green 2001):

- El modelo lineal de probabilidad (MLP).

- El modelo *logit*.
- El modelo *probit*.
- El modelo *tobit*.

1.1. Modelo lineal de probabilidad

En un modelo en donde Y es cuantitativa, el objetivo consiste en estimar su valor esperado o media esperada, dados los valores de las regresoras. En los modelos donde Y es cualitativa (dicotómica), el objetivo es encontrar la probabilidad de que un acontecimiento suceda, como por ejemplo poseer una vivenda propia, pagar impuesto de renta, padecer una determinada enfermedad, votar por el candidato del partido M, etcétera. Los modelos de regresión con respuesta cualitativa a menudo se conocen como modelos de probabilidad. Un modelo lineal de probabilidad puede ser escrito de la siguiente manera:

$$Y_i = ß_1 + ß_2 X_i + u_i$$

Donde X_i son por ejemplo los ingresos de la persona i; y la variable regresada Y toma el valor $Y_i = 1$, si la persona paga impuesto de renta; y $Y_i = 0$ si no se paga tal impuesto, el modelo anterior donde la variable regresada es binaria o dicotómica recibe el nombre de modelo lineal de probabilidad (MLP). Esto obedece a que la esperanza condicional de Y_i dado X_i, $E(Y_i \mid X_i)$ puede interpretarse como la probabilidad condicional de que el suceso tenga lugar dado X_i ; es decir $P(Y_i = 1 \mid X_i)$. Así, por ejemplo $E(Y_i \mid X_i)$, da la probabilidad de que una persona pague impuesto de renta y perciba unos ingresos por una cantidad X_i. La variable respuesta Y_i toma los valores 0 o 1. Ahora si π_i es la probabilidad de $Y_i = 1$ (es decir el suceso ocurre) y $(1-\pi_i)$ es la probabilidad de $Y_i = 0$ (el suceso no ocurre) la variable Y_i es una variable aleatoria Bernoulli cuya distribución de probabilidad es (Hosmer & Lemeshow 2000):

Y_i	Probabilidad
1	$P(y_i = 1) = \pi_i$
0	$P(y_i = 0) = 1 - \pi_i$

Como se parte del supuesto de que $E(\varepsilon_i) = 0$, entonces el valor esperado de la variable dependiente será:

$$E(Y_i) = 1 \cdot (\pi_i) + 0 \cdot (1 - \pi_i) = \pi_i$$

Lo que implica que:

$$E(Y_i \mid X_i) = E(\mathbf{x}_i^{'}\beta) = \pi_i$$

Esto significa que la respuesta esperada es la probabilidad de que la variable dependiente tome el valor de 1, es decir la esperanza condicional del modelo en realidad se interpreta como la probabilidad condicional de Y_i . Como la probabilida π_i debe encontrarse entre 0 y 1, tenemos la restricción: $0 \leqslant (E(Y_i \mid X_i)) \leqslant 1$. Es decir, la esperanza condicional o probabilidad condicional debe encontrarse entre 0 y 1.

En el modelo con variable dependiente dicotómica se presentan algunas dificultades teóricas, como:

1. Al ser la respuesta binaria, entonces los términos de error ε_i solamente pueden tomar dos valores, que son:

$$\varepsilon_i = 1 - \mathbf{x}_i^{'}\beta \;\text{ si } y_i = 1$$
$$\varepsilon_i = -\mathbf{x}_i^{'}\beta \text{ si } \;y_i = 0$$

   Por lo tanto, los errores en el MLP no tienen distribución normal. El incumplimiento del supuesto de normalidad en los errores en este caso no impide que se realicen las estimaciones puntuales vía MCO, pues sabemos que estas estimaciones aún permanecen insesgadas. Además, puede demostrarse con el uso del teorema central del límite que, conforme el tamaño de la muestra aumenta indefinidamente, los estimadores MCO, tienden a tener una distribución normal (Hosmer & Lemeshow 2000).

2. La varianza del error no es constante, puesto que:

$$\begin{aligned}\sigma^2_{\varepsilon_i} &= E\left[\varepsilon_i - E(\varepsilon_i)\right]^2 \\ &= (1-\pi_i)^2\pi_i + (0-\pi_i)^2(1-\pi_i) \\ &= \pi_i(1-\pi_i)\end{aligned}$$

   como $\pi_i = E(Y_i \mid X_i) = ß_1 + ß_2 X_i$, lo que implica que la varianza del término de error es función de los valores de X, y por tanto no es homocedástica. En presencia de heterocedasticidad, los estimadores MCO, aunque insesgados no son eficientes; es decir no tienen varianza mínima. Como la varianza de los errores ε_i depende de $E(Y_i \mid X_i)$, una forma de resolver el problema de heterocedasticidad es transformar el modelo $Y_i = ß_1 + ß_2 X_i + u_i$ dividiendo ambos lados entre:

$$\sqrt{E(Y_i \mid X_i)[1 - E(Y_i \mid X_i)]} = \sqrt{\pi_i(1-\pi_i)} = \sqrt{w_i}$$

   Luego de realizar esta transformación, el término de error es homocedástico, con lo cual podemos calcular el modelo mediante mínimos cuadrados ponderados (MCP), donde w_i son las ponderaciones.

   Teóricamente lo que se acaba de mencionar es correcto, pero en la práctica se desconoce la verdadera $E(Y_i \mid X_i)$ y por lo tanto se desconoce la ponderación w_i. Para calcularlas se puede utilizar el siguiente procedimiento:

a) Se efectúa la regresión $E(Y_i \mid X_i) = ß_1 + ß_2 X_i$ por MCO, sin tener en cuenta el problema de heterocedasticidad para obtener $\widehat{Y}_i$ el valor estimado de la verdadera $E(Y_i \mid X_i)$. Luego se calcula la ponderación estimada $\widehat{w_i} = \widehat{Y}_i(1 - \widehat{Y}_i)$.

b) Con el $\widehat{w_i}$ realizar la transformación de los datos dividiendo el modelo inicial por $\sqrt{w_i}$ y estimar la ecuación trasformada mediante el método de MCP.

3. Una restricción importante del modelo es que se debe cumplir que :

$$0 \leqslant E(Y_i \mid X_i) = \pi_i \leqslant 1$$

Esta restricción genera problemas en la selección del modelo adecuado, puesto que se requiere ajustar un modelo para los cuales los valores estimados para la respuesta se encuentren entre 0 y 1. Aunque *a priori* esto es verdadero, no hay garantía de que necesariamente los estimadores $E(Y_i \mid X_i)$ cumplan esta restricción, el problema ocurre porque MCO no tiene en cuenta la condición $0 \leqslant E(Y_i) \leqslant 1$. Las dos formas de establecer si $\widehat{Y}_i$ se encuentra entre 0 y 1, son: estimar el modelo lineal de probabilidad mediante el método de MCO y determinar si el $\widehat{Y}_i$ se encuentra entre 0 y 1. Realizando aproximaciones cuando los valores son menores que 0 se supone que $\widehat{Y}_i$ es 0; y si son mayores que 1 se supone que $\widehat{Y}_i$ es 1. El segundo es utilizar un método que garantice que las probabilidades estimadas se encuentren con seguridad entre 0 y 1.

1.2. El modelo *logit* o logístico

La restricción $0 \leqslant (E(Y_i \mid X_i)) \leqslant 1$ genera problemas en la selección del modelo adecuado, puesto que se requiere ajustar un modelo para el cual los valores estimados para la respuesta se encuentren entre 0 y 1.

Muchas funciones han sido propuestas, pero una función monótonamente creciente o (decreciente), en forma de S (o de S invertida), tal como la función logística, es la que se suele utilizar con más frecuencia, entre otras, por las siguientes razones:

1. Desde el punto de vista matemático, es una función extremedamente flexible y fácil de utilizar.
2. Tiene una interpretación relativamente sencilla.
3. La evidencia empiríca ha demostrado que este modelo es adecuado en la mayoría de los casos en los cuales la respuesta es binaría.

El modelo logístico tiene la forma:

$$E(y) = \frac{e^{\mathbf{x}'\beta}}{1 + e^{\mathbf{x}'\beta}} \tag{1}$$

Donde $\mathbf{x}$ es el vector de variables explicatorias y β es el vector de parámetros, que también puede expresarse como:

$$E(y) = \frac{1}{1 + e^{-\mathbf{x}'\beta}} \tag{2}$$

o sea:

$$\pi_i = \frac{1}{1 + e^{-\mathbf{x}'\beta}}$$

que es equivalente a:

$$1 - \pi_i = \frac{1}{1 + e^{\mathbf{x}'\beta}}$$

Con lo cual se tiene que:

$$\frac{\pi_i}{1 - \pi_i} = \frac{1 + e^{\mathbf{x}'\beta}}{1 + e^{-\mathbf{x}'\beta}} = e^{\mathbf{x}'\beta} \tag{3}$$

A esta transformación se le conoce como transformación *logit* de la probabilidad π_i y la relación $\frac{\pi_i}{1 - \pi_i}$ una razón de probabilidades o ventaja (*odds ratio*).

Si se toma el logaritmo natural, se obtiene:

$$Ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}'\beta \tag{4}$$

Con lo cual se tiene que el logaritmo de la razón de probabilidades es lineal, tanto en las variables como en los parámetros. La estimación de estos puede realizarse mediante el método de máxima verosimilitud (Green 2001).

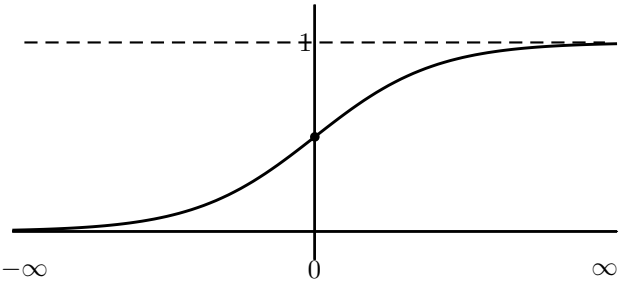

Figura 1: *Modelo logit. Fuente: Elaboración propia.*

1.3. Características del modelo *logit*

1. A pesar de que el modelo transformado es lineal en las variables, las probabilidades no son lineales.
2. El modelo *logit* supone que el logaritmo de la razón de probabilidades está linealmente relacionado con las variables explicatorias.
3. En el modelo *logit* los coeficientes de regresión expresan el cambio en el logaritmo de las probabilidades, cuando una de las variables explicatorias cambia en una unidad, permaneciendo constantes las demás (Gujarati 2010).

1.4. Estimación de los parámetros vía máxima verosimilitud

La forma general del modelo *logit* se puede expresar como:

$$y_i = E(y_i) + \varepsilon_i \tag{5}$$

donde las observaciones y_i son variables aleatorias independientes Bernoulli, con valores esperados:

$$\begin{aligned} E(y_i) &= \pi_i \\ &= \frac{e^{\mathbf{x}'\beta}}{1 + e^{\mathbf{x}'\beta}} \end{aligned}$$

Como cada observación sigue una distribución Bernoulli, su distribución será:

$$f_i(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, i = 1, 2, , 3, \ldots, n$$

Como las observaciones son independientes, la función de verosimilitud será:

$$\begin{aligned} L(y_1, y_2, \ldots, y_n, \beta) &= \prod_{i=1}^{n} f_i(y_i) \\ &= \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \end{aligned}$$

Al tomar logaritmo a la función de verosimilitud:

$$\begin{aligned} LnL(y_1, y_2, \ldots, y_n, \beta) &= Ln \prod_{i=1}^{n} f_i(y_i) \\ &= \sum_{i=1}^{n} \left[y_i Ln \left(\frac{\pi_i}{1 - \pi_i} \right) \right] + \sum_{i=1}^{n} Ln(1 - \pi_i) \end{aligned}$$

Como:

$$1-\pi_i = \frac{1}{1+e^{\,\mathbf{x}'\beta}} \text{ y } Ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}'\beta,$$

El logaritmo de la verosimilitud se puede expresar para el modelo de regresión logística:

$$LnL(\mathbf{y},\beta) = \sum_{i=1}^{n} y_i \mathbf{x}_i'\beta - \sum_{i=1}^{n} Ln[1+e^{\mathbf{x}_i'\beta}]$$

En muchas aplicaciones del modelo se dispone de informacion repetidas para cada uno de los valores de las variables. Sea y_i la cantidad de 1 observados para la *i-ésima* observación y n_i la cantidad de ensayos en cada observación, entonces el logaritmo de la verosimilitud se puede presentar:

$$Ln\, L(\mathbf{y},\beta) = \sum_{i=1}^{n} y_i\pi_i + \sum_{i=1}^{n} n_i Ln\,(1-\pi_i) - \sum_{i=1}^{n} y_i\, Ln\,(1-\pi_i)$$

Los estimadores de máxima verosimilitud se pueden obtener mediante un algoritmo de mínimos cuadrados iterativamente re ponderados.

Si $\hat{\beta}$ es el estimador obtenido, mediante el método iterativo y siendo ciertas las hipótesis del modelo, se puede demostrar que en forma asintótica:

$$E(\hat{\beta}) = \beta \quad \text{y} \quad V(\hat{\beta}) = (\mathbf{X}'V^{-1}\mathbf{X})^{-1}$$

El valor estimado del predictor lineal es $\hat{\eta}_i = \mathbf{x}_i\hat{\beta}$, y el valor esperado del modelo de regresión logístico, se suele expresar:

$$\begin{aligned}\hat{y}_i = \hat{\pi}_i &= \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}} \\ &= \frac{e^{(\mathbf{x}_i'\hat{\beta})}}{1+e^{(\mathbf{x}_i'\hat{\beta})}} \\ &= \frac{1}{1+e^{(-\mathbf{x}_i'\hat{\beta})}}\end{aligned}$$

2. Aplicación

Para ilustrar la metodología antes mencionada, se toma como ejemplo una muestra de los datos obtenidos luego de la aplicación de la encuesta a usuarios (línea base) del servicio de Aseo en la ciudad de Bogotá D.C., dentro del marco del convenio administrativo 012; se analizan los resultados a una pregunta con respuesta binaria, la cual es : ¿realiza separación de material reciclable en la fuente?, con dos opciones

de respuesta sí y no, en función de: la cantidad de residuo sólido que estima genera a la semana; de la cantidad de material que recicla a la semana y en función del estrato social y del tipo de usuario: 1(pequeño generador); 2 (mediano generador) y 3 (gran generador).

El modelo *logit* planteado es :

$$\pi_i = \frac{1}{1 + e^{-\mathbf{x}'\beta}} \tag{6}$$

en donde las variables son:

$x1$: Tipo de usuario: 1(pequeño generador); 2(mediano generador) y 3(gran generador).

$x2$: Autoriza recibir información acerca del proceso de reciclaje: 0 (no); 1 (sí).

$x3$: Cantidad de material que recicla a la semana.

$x4$: Cantidad de residuo sólido que estima genera a la semana.

$x5$: Estrato social.

y: Realiza separación de material reciclable en la fuente.

Los resultados obtenidos pueden observarse a continuación:

```
   Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.41366    0.63169  -2.238  0.02523 *
x1            0.13466    0.38462   0.350  0.72626
x2            0.48341    0.27966   1.729  0.08389
x3          237.23325   77.79066   3.050  0.00229 **
x4           -0.62633    0.06956  -9.005  < 2e-16 ***
x5            0.17895    0.11909   1.503  0.13295
```

El modelo obtenido es de la forma (6) donde :

$$X'\beta = -1.4136 + 0.1346\,\texttt{x1} + 0.4834\,\texttt{x2} + 237.2332\,\texttt{x3} - 0.6263\,\texttt{x4} + 0.1789\,\texttt{x5}$$

Se puede observar que el valor del coeficiente 0,13466 significa que, dejando todas las demás variables constantes, un cambio de usuario por ejemplo de usuario 1 a usuario 2, aumentaría la probabilidad de separar material reciclable en la fuente en 0,13 veces. El valor 0,48 significa que las personas que están de acuerdo en recibir información sobre el proceso de reciclaje aumentan en 0,48 veces la probabilidad de separar efectivamente material reciclable en la fuente, dejando las demás variables constantes. El valor 237,23325 significa que un aumento en 1K de la cantidad de material que recicla a la semana, aumenta la probabilidad de la variable respuesta en 237,23 veces la probabilidad de separar material en la fuente, manteniendo constantes las demás variables. Por último, un aumento en 1K de la cantidad de

residuo sólido que se genera a la semana produce un incremento de 0,17 veces la probabilidad de separar material reciclable en la fuente.

3. Conclusiones

1. El modelo *logit* calculado para este conjunto de datos mostró que dos de las variables utilizadas pueden ser consideradas no significativas a un nivel del 10 %, estas variables son: Tipo de usuario y estrato social.

2. Al correr el modelo sin las variables que resultan no significativas, se obtienen los siguientes resultados.

```
 Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.63587    0.26408  -2.408  0.01605 *
x2            0.41679    0.27467   1.517  0.12916
x3          231.32387   79.79663   2.899  0.00374 **
x4           -0.61056    0.06802  -8.977  < 2e-16 ***
```

   Se puede ver que las variables x3 y x4 siguen siendo significativas, aunque la variable x2 no, a partir de lo cual se puede concluir que las dos variables más importantes en la decisión de reciclar material en la fuente están relacionadas con la cantidad de residuo sólido que se produce a la semana y la cantidad de este material que es separado por semana.

3. El modelo *logit* tiene la ventaja sobre los modelos lineales de probabilidad, que las probabilidades calculadas siempre están entre cero y uno, con lo cual se evita el tener que hacer aproximaciones a 0,01 cuando las probabilidades son negativas, o a0,99 cuando son mayores a 1.

4. La medida usual de bondad de ajuste, el coeficiente de determinación, no es apropiado en el caso del modelo *logit*, pero existen otras medidas útiles como el criterio de clasificación o el criterio de Hosmer-Lemeshow.

5. Aunque el modelo *logit* es lineal en las variables explicatorias, las probabilidades en sí mismas no lo son, lo cual contrasta con el modelo lineal de probabilidad, en donde las probabilidades aumentan linealmente con las variables independientes.

Recibido: 7 de febrero de 2012
Aceptado: 1 de agosto de 2012

Referencias

Green, W. (2001), *Análisis Econométrico*, Prentice Hall.

Gujarati, D. (2010), *Econometría*, McGraw Hill.

Hosmer, D. W. & Lemeshow, S. (2000), *Applied Logistic Regression*, Jonh Wiley.