UCD CENTRE FOR ECONOMIC RESEARCH

WORKING PAPER SERIES

2011

Simple Logit and Probit Marginal Effects in R

Alan Fernihough, University College Dublin

WP11/22

October 2011

UCD SCHOOL OF ECONOMICS
UNIVERSITY COLLEGE DUBLIN
BELFIELD DUBLIN 4

Simple Logit and Probit Marginal Effects in R

Alan Fernihough*

Abstract

This paper outlines a simple routine to calculate the marginal effects of logit and probit regressions using the popular statistical software package R. I compare results obtained using this procedure with those produced using Stata. An extension of this routine to the generalized linear mixed effects regression is also presented.

1 Introduction

A common approach in empirical economic research is to model binary variables using a generalized linear model with a binomial distribution. The advantage of this approach is that it restricts predictions of the dependent variable to values between zero and one, unlike ordinary least squares regression (OLS). One difference of this approach is that the estimated coefficients are not marginal effects, as in OLS, but multiplicative effects. Fortunately, transforming these coefficients into marginal effects is a reasonably straightforward procedure.

In recent years, the open-source statistical program R has exploded in popularity. The primary attraction of R is the extensive repository of packages which have been contributed by researchers across a huge range of disciplines. Currently, the R package repository features 3,369 packages. Surprisingly, to my knowledge there is no general function which easily computes marginal effects from all potential binary dependent models similar to the `mfx` command as in Stata.[1]

The aim of this paper is to present a quick solution to this problem, which is easy to implement.

*Fernihough is a graduate student at the UCD School of Economics; he acknowledges financial support from the Irish Research Council for the Humanities and Social Sciences. Thanks to Keith O'Hara for some comments. All errors and omissions are mine. **Email: alan.fernihough@gmail.com**.

[1] Some support is offered in both the 'tonymisc' and 'erer' packages.

2 Binary Dependent Variables

Let $\mathrm{E}(y_i|\mathbf{x_i})$ represent the expected value of a dependent variable y_i given a vector of explanatory variables $\mathbf{x_i}$, for an observation unit i. In the case where $\mathbf{y}$ is a linear function of $(\mathbf{x_1}, \ldots, \mathbf{x_j}) = \mathbf{X}$ and $\mathbf{y}$ is a continuous variable the following model with j regressors can be estimated via ordinary least squares:

$$\mathbf{y} = \mathbf{X}'\beta \tag{1}$$

or

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x_1} \ldots + \beta_j\mathbf{x_j} \tag{2}$$

so the additive vector of predicted coefficients can be obtained from the usual computation $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. From (1) and (2) it is straightforward to see that the marginal effect of the variable $\mathbf{x_k}$, where $k \in \{1, \ldots, j\}$, on the dependent variable is $\partial\mathbf{y}/\partial\mathbf{x_k} = \beta_k$. In other words, a unit increase in the variable $\mathbf{x_k}$ increases the variable $\mathbf{y}$ by β_k units.

The standard approach to modeling dichotomous/binary variables (so $\mathbf{y} \in \{0, 1\}$) is to estimate a generalized linear model under the assumption that $\mathbf{y}$ follows some form of Bernoulli distribution. In econometrics, researchers will either use the logistic (logit) or standard normal cumulative (probit) distributions. Thus, the expected value of the dependent variable becomes:

$$\mathbf{y} = G(\mathbf{X}'\beta) \tag{3}$$

where G is the specified binomial distribution. Since $\mathbf{y} \in$ (0,1), the predicted value for observation i (y_i) represents the conditional probability that y_i is one, or $\Pr(y_i = 1)$. In the case of the logistic regression the generalized linear model can be specified:

$$\Pr(y_i = 1) = \mathrm{logit}^{-1}(\beta_0 + \beta_1\mathbf{x_1} + \ldots + \beta_j\mathbf{x_j}) \tag{4}$$

where the inverse logit function is used here to map the linear predictions into probabilities. From (3), we see that the marginal effects must be calculated using the the chain rule so:

$$\frac{\partial\mathbf{y}}{\partial\mathbf{x_k}} = \beta_k \times \frac{dG}{d\mathbf{X}'\beta} \tag{5}$$

so the marginal effect of variable $\mathbf{x_k}$ depends on the derivative: $dG/d\mathbf{X}'\beta$, which is either a logistic or normal probability density function, depending on the choice of G.

As outlined in Kleiber & Zeileis (2008), there are two main approaches to calculating marginal effects from binary dependent variable models. The first uses the average of the sample marginal effects, while the other uses average marginal effects. The average of the sample marginal effects is calculated as follows:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x_k}} = \beta_k \times \frac{\sum_{i=0}^{n} g(\mathbf{X}'\hat{\beta})}{n} \tag{6}$$

where there are n observations in the dataset and g is the probability density function for either the normal or logistic distribution. In essence, one can calculate the marginal effect for a each variable by using the estimated coefficient (corresponding to the inner-part of the chain rule) multiplied by the average value of all appropriately transformed predicted values.

The second approach calculates the marginal effect for $\mathbf{x}_k$ by taking predicted probability calculated when all regressors are held at their mean value from the same formulation with the exception of adding one unit to $\mathbf{x}_k$. The derivation of this marginal effect is captured by the following:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x_k}} = G(\hat{\beta}_0+\hat{\beta}_1\bar{\mathbf{x_1}}+\ldots+\hat{\beta}_k(\bar{\mathbf{x_k}}+1)+\ldots+\hat{\beta}_j\bar{\mathbf{x_j}})-G(\hat{\beta}_0+\hat{\beta}_1\bar{\mathbf{x_1}}+\ldots+\hat{\beta}_k(\bar{\mathbf{x_k}})+\ldots+\hat{\beta}_j\bar{\mathbf{x_j}}) \tag{7}$$

where the marginal effect for a variable is computed by subtracting the conditional predicted probability when all variables are held at their mean values from the same conditional predicted probability, except with the variable of interest increased by one-unit ($\bar{\mathbf{x_k}} + 1$).

3 Simple Functions of Logit and Probit Marginal Effects in R

Section 2 specified two methods by which marginal effects for either a logit of probit regression can be calculated. In this section, I outline a basic user-written R-function which calculates the average of the sample marginal effects, as in equation (6), and their associated standard errors. Dealing with binary/dummy or factor variables adds complexity in calculating the average marginal effects of equation (7). Given the objective of this paper, I do not present a function which calculates average marginal effects. It is noteworthy that the marginal effects produced by other statistical software programs, such as `Stata`, calculate average marginal effects by default. However, there is no reason to believe that the marginal effects produced by one method are superior to the other. Similarly, the standard errors produced by the following `R` function are via simulation – which captures uncertainty in both the regression coefficients and the probability density function. Alternatively, one could compute standard errors using the delta method, as in `Stata`. Once again the difference between the two approaches is minimal and since both methods are "approximations" there is little reason to believe one is more robust than the other.

A function which calculates the average of the sample marginal effects for either a probit or logit model in R is displayed below. The default number of simulations from which the standard errors are calculated is 1,000. However, the user can change this number using the second argument.

```
mfx <- function(x,sims=1000){
      set.seed(1984)
      pdf <- ifelse(as.character(x$call)[3]=="binomial(link = \"probit\")",
                                mean(dnorm(predict(x, type = "link"))),
                                mean(dlogis(predict(x, type = "link"))))
      pdfsd <- ifelse(as.character(x$call)[3]=="binomial(link = \"probit\")",
                                     sd(dnorm(predict(x, type = "link"))),
                                     sd(dlogis(predict(x, type = "link"))))
      marginal.effects <- pdf*coef(x)
      sim <- matrix(rep(NA,sims*length(coef(x))), nrow=sims)
      for(i in 1:length(coef(x))){
           sim[,i] <- rnorm(sims,coef(x)[i],diag(vcov(x)^0.5)[i])
           }
      pdfsim <- rnorm(sims,pdf,pdfsd)
      sim.se <- pdfsim*sim
      res <- cbind(marginal.effects,sd(sim.se))
      colnames(res)[2] <- "standard.error"
      ifelse(names(x$coefficients[1])=="(Intercept)",
               return(res[2:nrow(res),]),return(res))
}
```

4 Comparison with Other Software

To demonstrate how the function above works and also the similarities between this function and the mfx command in Stata, I perform a basic analysis. To complete this exercise, I use data from the car package in R. These data comprise of individual level information on income, education, gender, age, and language for 3,987 individuals. Creating a binary dependent variable called h.wage – signaling whether an individual earns a 'high wage' – I estimate the probability that an individual is in the 'high wage' cohort conditional on their age, education, a dummy variable taking the value one where the individual is a male and two dummy variables to indicate languages spoken other than English.

The code below displays the necessary R syntax and output as if displayed in the R console. The equivalent Stata output is also displayed. For example, the estimated marginal effects for education, i.e. the increase in the probability of being in the high wage category for a one year increase in education, are 4.2%, 4.2%, 4.6% and 4.5% for the probit and logit models estimated using R and Stata respectively. Clearly these values are very alike. The marginal effects for the other regressors and their standard errors are very similar in each of the four other specifications.

```
> setwd("C:\\Users\\Alan\\Documents\\My Dropbox\\marginaleffects")
> library(car)
> data(SLID)
> dat1 <- na.omit(SLID)
> dat1$h.wage <- ifelse(dat1$wages>20,1,0)
> p1 <- glm(h.wage ~ education+age+sex+language,data=dat1, family = binomial(link = "probit"))
> mfx(p1)
               marginal.effects standard.error
education           0.042139646    0.018387281
age                 0.009601096    0.004189945
sexMale             0.134059218    0.058575244
languageFrench     -0.015701475    0.028678160
languageOther      -0.009959757    0.020615175
> l1 <- glm(h.wage ~ education+age+sex+language,data=dat1, family = binomial(link = "logit"))
> mfx(l1)
               marginal.effects standard.error
education            0.04243174    0.020708933
age                  0.00947604    0.004632232
sexMale              0.13366522    0.065230162
languageFrench      -0.01573230    0.029935596
languageOther       -0.01063887    0.021267302

##################################################################
Stata Output
******************************************************************
Marginal effects after probit
      y  = Pr(hwage) (predict)
         =  .19364718
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
educat~n |   .0460106      .00226   20.39   0.000   .041589  .050432    13.337
     age |   .0104831      .00058   18.22   0.000   .009355  .011611   37.0981
 _Isex_2*|   .1460536       .0131   11.15   0.000   .120373  .171735   .498119
_Ilang~2*|   -.016739      .02595   -0.64   0.519  -.067607  .034129   .064961
_Ilang~3*|  -.0107334      .01992   -0.54   0.590  -.049771  .028305   .121395
------------------------------------------------------------------------------
Marginal effects after logit
      y  = Pr(hwage) (predict)
         =  .18771273
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
educat~n |   .0447072      .00221   20.22   0.000   .040373  .049041    13.337
     age |   .0099842      .00055   18.07   0.000   .008901  .011067   37.0981
 _Isex_2*|   .1413666      .01291   10.95   0.000   .116064  .166669   .498119
```

```
_Ilang~2*|  -.0160889      .02536   -0.63   0.526    -.0658  .033622   .064961
_Ilang~3*|  -.0110152      .01918   -0.57   0.566   -.04861   .02658   .121395
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

5 Generalized Linear Mixed Effects Model

The following output contains a function which calculates marginal effects of the fixed effects of a generalized linear mixed effects model. The output of this function applied to the data used in the previous section is also displayed. This model is estimated using the `lme4` package (Bates, 2010).

```
> library(lme4)
> glmermfx <- function(x,nsims=1000){
+   set.seed(1984)
+   pdf <- mean(dlogis(-log((1-fitted(x))/fitted(x))))
+   pdfsd <- sd(dlogis(-log((1-fitted(x))/fitted(x))))
+   marginal.effects <- pdf*fixef(x)
+   sim <- matrix(rep(NA,nsims*length(fixef(x))), nrow=nsims)
+   for(i in 1:length(fixef(x))){
+     sim[,i] <- rnorm(nsims,fixef(x)[i],diag(vcov(x)^0.5)[i])
+     }
+   pdfsim <- rnorm(nsims,pdf,pdfsd)
+   sim.se <- pdfsim*sim
+   res <- cbind(marginal.effects,sd(sim.se))
+   colnames(res)[2] <- "standard.error"
+   ifelse(names(fixef(x))[1]=="(Intercept)",
              return(res[2:nrow(res),]),return(res))
+ }
> glme1 <- lmer(h.wage ~ education+age+sex+(1|language),
     family = binomial(link = logit),data=dat1)
> glmermfx(glme1)
          marginal.effects standard.error
education      0.042502043    0.021134699
age            0.009457282    0.004751698
sexMale        0.133529363    0.067310950
```

References

[1] Christian Kleiber and Achim Zeileis, *Applied Econometrics with R.* Springer, 2008.

[2] Douglas M. Bates, *lme4: Mixed-effects modeling with R.* Springer, 2010.

UCD CENTRE FOR ECONOMIC RESEARCH – RECENT WORKING PAPERS

WP10/37 Alan Fernihough: "Malthusian Dynamics in a Diverging Europe: Northern Italy 1650-1881" November 2010
WP10/38 Cormac Ó Gráda: "The Last Major Irish Bank Failure: Lessons for Today?" November 2010
WP10/39 Kevin Denny and Veruska Oppedisano: "Class Size Effects: Evidence Using a New Estimation Technique" December 2010
WP10/40 Robert Gillanders and Karl Whelan: "Open For Business? Institutions, Business Environment and Economic Development" December 2010
WP10/41 Karl Whelan: "EU Economic Governance: Less Might Work Better Than More" December 2010
WP11/01 Svetlana Batrakova: 'Flip Side of the Pollution Haven: Do Export Destinations Matter?' January 2011
WP11/02 Olivier Bargain, Mathias Dolls, Dirk Neumann, Andreas Peichl and Sebastian Siegloch: 'Tax-Benefit Systems in Europe and the US: Between Equity and Efficiency' January 2011
WP11/03 Cormac Ó Gráda: 'Great Leap into Famine' January 2011
WP11/04 Alpaslan Akay, Olivier Bargain, and Klaus F Zimmermann: 'Relative Concerns of Rural-to-Urban Migrants in China' January 2011
WP11/05 Matthew T Cole: 'Distorted Trade Barriers' February 2011
WP11/06 Michael Breen and Robert Gillanders: 'Corruption, Institutions and Regulation' March 2011
WP11/07 Olivier Bargain and Olivier Donni: 'Optimal Commodity Taxation and Redistribution within Households' March 2011
WP11/08 Kevin Denny: 'Civic Returns to Education: its Effect on Homophobia' April 2011
WP11/09 Karl Whelan: 'Ireland's Sovereign Debt Crisis' May 2011
WP11/10 Morgan Kelly and Cormac Ó Gráda: 'The Preventive Check in Medieval and Pre-industrial England' May 2011
WP11/11 Paul J Devereux and Wen Fan: 'Earnings Returns to the British Education Expansion' June 2011
WP11/12 Cormac Ó Gráda: 'Five Crises' June 2011
WP11/13 Alan Fernihough: 'Human Capital and the Quantity-Quality Trade-Off during the Demographic Transition: New Evidence from Ireland' July 2011
WP11/14 Olivier Bargain, Kristian Orsini and Andreas Peichl: 'Labor Supply Elasticities in Europe and the US' July 2011
WP11/15 Christian Bauer, Ronald B Davies and Andreas Haufler: 'Economic Integration and the Optimal Corporate Tax Structure with Heterogeneous Firms' August 2011
WP11/16 Robert Gillanders: 'The Effects of Foreign Aid in Sub-Saharan Africa' August 2011
WP11/17 Morgan Kelly: 'A Note on the Size Distribution of Irish Mortgages' August 2011
WP11/18 Vincent Hogan, Patrick Massey and Shane Massey: 'Late Conversion: The Impact of Professionalism on European Rugby Union' September 2011
WP11/19 Wen Fan: 'Estimating the Return to College in Britain Using Regression and Propensity Score Matching' September 2011
WP11/20 Ronald B Davies and Amélie Guillin: 'How Far Away is an Intangible? Services FDI and Distance' September 2011
WP11/21 Bruce Blonigen and Matthew T Cole: 'Optimal Tariffs with FDI: The Evidence' September 2011

UCD Centre for Economic Research Email economics@ucd.ie