oglmx: A Package for Estimation of Ordered Generalized Linear Models.

Nathan Carroll
Universität Regensburg

Abstract

Ordered discrete dependent variable models such as ordered probit and ordered logit are frequently used across the social sciences to study outcomes including health status, happiness, wealth and educational attainment. Unlike in the case of OLS, unaccounted for heteroskedasticity in these models can lead to biased parameter estimates. This paper introduces the **oglmx** package for the R statistical environment that permits estimation of generalized models that allow the user to model the form of the heteroskedasticity.

Keywords: ordered multinomial models, heteroskedasticity, heterogeneous choice model, location-scale model, discrete ordered outcome.

1. Introduction

Ordered discrete dependent variable models are common across the social sciences, examples of outcomes that require such models include categorical measures of health status (Case, Lubotsky, and Paxson 2002), happiness or life-satisfaction (Gerdtham and Johannesson 2001), wealth (Hartog and Oosterbeek 1998) and educational attainment (Dearden, Meghir, and Ferri 2002). Standard models such as ordered probit and ordered logit assume that error variances are constant across observations, or homoskedastic. When using ordinary least squares to estimate a linear relationship in the presence of heteroskedasticity in the error term parameter estimates remain consistent, though standard errors need to be adjusted via a variance-covariance estimator that takes account of the heteroskedasticity. However, in the case of models such as ordered probit and ordered logit failure to account for heteroskedasticity can lead to biased parameter estimates in addition to misspecified standard errors. This paper introduces the **oglmx** package developed for the R statistical environment (R Core Team 2015) that allows the user to model the form of the heteroskedasticity in various ordered discrete dependent variable models.

The simplest discrete dependent variable models, those for the case of binary outcomes, can be estimated using the glm function available in the core distribution of R, while ordered dependent variable models beyond the binary case are included in the **MASS** package (Venables and Ripley 2002) under the polr function. The standard ordered probit and logit models include a normalization of the error variance which implies that the scale of the estimated parameters are of little relevance to the researcher, instead researchers are often interested in the marginal effects of particular variables on the probabilities of each observable value of the dependent variable. R packages returning marginal effects include **erer** (Sun 2014) and

mfx (Fernihough 2014). The **oglmx** package includes a margins function that returns marginal effects (and their standard errors) for all models estimated by the oglmx function. The package includes link functions for probit, logit, cauchit, complementary log-log and log-log while allowing the user to specify the functional form used to model the variance of the standard error. The function is written so that it is sufficiently flexible to allow estimation of interval regression with fixed boundaries across observations in addition to the ordered models provided by function polr.[1] The oglmx package makes use of the **maxLik** (Henningsen and Toomet 2011) package to maximise the likelihood for the user specified model.

The paper is organised as follows: Section 2 describes the models estimated by the oglmx function, section 3 gives an outline of how the core functions of the package work, section 4 provides a working example and section 5 concludes.

2. Ordered Multinomial Models

Ordered multinomial models such as ordered probit and ordered logit are frequently motivated using a latent variable model. Let y denote a random variable taking on one of J possible values $\{0, 1, 2, \ldots, J-1\}$. Importantly there is an order to these J responses. There is an unobserved latent variable y^* that is determined by:

$$y^* = \mathbf{x}\beta + \sigma\epsilon$$

where x is a $1 \times K$ vector of explanatory variables that may or may not contain a constant depending on the particular model to be estimated, β is a $K \times 1$ vector of parameters, ϵ is a mean zero random error term and σ is a parameter that allows the variance of the error term to be shifted up or down. Let $\alpha_1 < \alpha_2 < \ldots < \alpha_{J-1}$ be threshold parameters that determine the observed outcome as follows:

$$\begin{aligned} y &= 0 \qquad \text{if } y^* \leq \alpha_1 \\ y &= 1 \qquad \text{if } \alpha_1 < y^* \leq \alpha_2 \\ &\vdots \\ y &= J-1 \quad \text{if } y^* > \alpha_{J-1}. \end{aligned}$$

Given a distribution function for the error term ϵ, vector of parameters β and the set of threshold parameters we can obtain the probabilities for each of the outcomes.

$$\begin{aligned} P(y=0) &= P(y^* \leq \alpha_1) = P\left(\epsilon \leq \frac{\alpha_1 - \mathbf{x}\beta}{\sigma}\right) \\ P(y=1) &= P(y^* \leq \alpha_2) - P(y^* \leq \alpha_1) = P\left(\epsilon \leq \frac{\alpha_2 - \mathbf{x}\beta}{\sigma}\right) - P\left(\epsilon \leq \frac{\alpha_1 - \mathbf{x}\beta}{\sigma}\right) \\ &\vdots \\ P(y=J-1) &= P(y^* > \alpha_{J-1}) = P\left(\epsilon > \frac{\alpha_{J-1} - \mathbf{x}\beta}{\sigma}\right). \end{aligned}$$

[1] The oglm command in Stata (**?**) offers a similar set of models as the oglmx function but omits interval regression and does not allow flexibility in the function used to model the variance of the error term.

More generally by setting $\alpha_0 = -\infty$ and $\alpha_J = \infty$ we have:

$$\begin{aligned} P(y = j) &= P\left(\epsilon \leq \frac{\alpha_{j+1} - \mathbf{x}\beta}{\sigma}\right) - P\left(\epsilon \leq \frac{\alpha_j - \mathbf{x}\beta}{\sigma}\right) \\ &= F\left(\frac{\alpha_{j+1} - \mathbf{x}\beta}{\sigma}\right) - F\left(\frac{\alpha_j - \mathbf{x}\beta}{\sigma}\right) \end{aligned}$$

where F is the assumed cumulative distribution function (cdf) for the error term ϵ. The various models estimated by the oglmx function vary according to the assumed distribution of the error term, e.g. logistic distribution for ordered logit and standard normal distribution for ordered probit, and the parameters that are known versus those that are estimated, e.g. under interval regression the threshold parameters are known while the constant in the latent variable equation and the variance of the error term are estimated whereas with ordered probit the levels of the constant and the variance of the error term are imposed while the threshold parameters are estimated.

To allow for heteroskedasticity the variance of the error term is permitted to vary by allowing σ to be determined by the following equation:

$$\sigma = g(\mathbf{z}\delta)$$

where $\mathbf{z}$ is a $1 \times L$ vector of variables that explain the level of the variance. As was the case for the vector $\mathbf{x}$, $\mathbf{z}$ may or may not include a constant. δ is an $L \times 1$ vector of parameters. The function $g(.)$ should ideally return a positive value for all observed levels of variables in $\mathbf{z}$, with this in mind the default option of the package is to use the exponential function. There is no restriction regarding the choice of variables in $\mathbf{z}$, it may contain the same variables as $\mathbf{x}$ or be entirely different.

The oglmx function obtains estimates of the parameters of the model by maximising the log-likelihood function, that for a sample consisting of n observations is given by:

$$\mathcal{L}(\beta, \delta, \alpha) = \sum_{i=1}^{n} \sum_{j=0}^{J-1} I(y_i = j) \log\left(F\left(\frac{\alpha_{j+1} - \mathbf{x}_i\beta}{g(\mathbf{z}_i\delta)}\right) - F\left(\frac{\alpha_j - \mathbf{x}_i\beta}{g(\mathbf{z}_i\delta)}\right)\right).$$

where $I(.)$ is the indicator function. Following the usual properties of maximum likelihood estimators the parameter estimates obtained from maximising the likelihood are consistent and asymptotically normal and the asymptotic variance of the estimated parameters can be estimated straightforwardly (Wooldridge 2002). The main body of code in the oglmx function calculates the above log-likelihood, score vector and Hessian matrix given a vector of parameter values. This function is passed as an argument to a Newton-Raphson type algorithm via the **maxLik** package.

Many popular models are included under the above framework as the distribution function $F(.)$ and the parameters that are estimated or imposed are varied. Table 1 lists some of the models and the parameter restrictions imposed in each case.

2.1. Marginal Effects

In many multinomial models the absolute magnitude of estimated regression coefficients is of little relevance, for example for an ordered probit the scale of the coefficients is fixed by fixing

Table 1: Parameter Assumptions for Popular Ordered Dependent Variable Models.

Model	No. Outcomes	Error Distribution	Constant in Mean Equation	Error Variance (σ)	Threshold Parameters
Probit	2	Standard Normal	Included	Set =1	Set =0
Logit	2	Standard Logistic	Included	Set =1	Set =0
Ordered Probit	>2	Standard Normal	Excluded	Set =1	Estimated
Ordered Logit	>2	Standard Logistic	Excluded	Set =1	Estimated
Interval Regression	>2	User Choice	Included	Estimated	Specified

the variance of the error term to unity and the constant equal to zero. Instead researchers are interested in the marginal effect of a variable on the probability of each outcome. In homoskedastic models the signs of regression coefficients are informative of the sign of the marginal effects for outcomes at the extreme of the distribution, but not for intermediate outcomes. In contrast in a heteroskedastic model the sign of a variable's coefficient(s) is on its own uninformative on the sign of any marginal effect when it enters the equation for the variance. For a continuous variable v contained in the kth element of vector $\mathbf{x}$ and the lth element of vector $\mathbf{z}$ the marginal effect of that variable on the probability of outcome j, denoted $ME_j(\mathbf{x}, \mathbf{z})$ occurring is given by:

$$\begin{aligned} ME_j(\mathbf{x}, \mathbf{z}) = \frac{\partial P(y=j)}{\partial v} = &- \frac{\beta_k}{g(\mathbf{z}\delta)} \left(f\left(\frac{\alpha_{j+1} - \mathbf{x}\beta}{g(\mathbf{z}\delta)} \right) - f\left(\frac{\alpha_j - \mathbf{x}\beta}{g(\mathbf{z}\delta)} \right) \right) \\ &- \frac{\delta_l g'(\mathbf{z}\delta)}{g(\mathbf{z}\delta)} \left(\frac{\alpha_{j+1} - \mathbf{x}\beta}{g(\mathbf{z}\delta)} f\left(\frac{\alpha_{j+1} - \mathbf{x}\beta}{g(\mathbf{z}\delta)} \right) - \frac{\alpha_j - \mathbf{x}\beta}{g(\mathbf{z}\delta)} f\left(\frac{\alpha_j - \mathbf{x}\beta}{g(\mathbf{z}\delta)} \right) \right) \end{aligned} \tag{1}$$

where $f(.)$ is the probability density function of error term ϵ. In general the sign of marginal effects depends on the sign of the relevant coefficients and the relative value of the mean of the latent variable ($\mathbf{x}\beta$) and relevant threshold parameters α_j and α_{j+1}. Equation 1 is the correct formula when the variable under consideration is continuous, however for a binary variable it may be preferable to consider the full change of a variable from zero to one rather than a change at the margin. In this case the marginal effect of a binary variable contained in x and/or z is calculated using:

$$ME_j(\mathbf{x}, \mathbf{z}) = F\left(\frac{\alpha_{j+1} - \mathbf{x}_1\beta}{g(\mathbf{z}_1\delta)} \right) - F\left(\frac{\alpha_j - \mathbf{x}_1\beta}{g(\mathbf{z}_1\delta)} \right) - \left(F\left(\frac{\alpha_{j+1} - \mathbf{x}_0\beta}{g(\mathbf{z}_0\delta)} \right) - F\left(\frac{\alpha_j - \mathbf{x}_0\beta}{g(\mathbf{z}_0\delta)} \right) \right) \tag{2}$$

where $\mathbf{x}_1$ and $\mathbf{z}_1$ denote vectors with the variable of interest set equal to one while $\mathbf{x}_0$ and $\mathbf{z}_0$ set the variable equal to zero.

Equations 1 and 2 tell us how to calculate the marginal effect given a particular set of values of the components of the vectors $\mathbf{x}$ and $\mathbf{z}$. Two main methods are used to summarize the marginal effects for a sample of data, the marginal effect at mean (MEM) and the average marginal effect (AME). The MEM for a particular variable calculates the marginal effect supposing that all variables were at the means for the sample, that is:

$$MEM_j = ME_j(\bar{\mathbf{x}}, \bar{\mathbf{z}})$$

where $\bar{\mathbf{x}}$ and $\bar{\mathbf{z}}$ denote vectors of means. The AME calculates the marginal effect for each observed set of variables $\mathbf{x}_i$ and $\mathbf{z}_i$ and averages the marginal effects across the sample, that is:

$$AME_j = \frac{1}{n} \sum_{i=1}^{n} ME_j(\mathbf{x}_i, \mathbf{z}_i).$$

The margins.oglmx function included in the **oglmx** package can calculate either of these two measures of the marginal effect. The marginal effects are a non-linear function of the estimated parameters, an approximation to the standard errors of the marginal effects can be obtained via application of the delta method. The margins.oglmx function calculates the standard errors using the delta method.

3. Package implementation

The current implementation of the package follows the standard for model estimation in R, a user input function oglmx that takes as input a formula for the empirical model being estimated and the data frame where the data is to be sourced from, and a oglmx.fit function called by oglmx that estimates the model. The oglmx.fit function contains further functions that return the analytic log-likelihood, score vector and hessian for a given parameter vector and these are used by the likelihood maximization procedure of the **maxLik** package which given the analytic hessian performs the optimization with a Newton-Raphson algorithm by default.

4. Using the oglmx package

To illustrate the use of the package a random generated dataset will be used. An outcome variable y will take on one of four possible values, without loss of generality we may specify them as $\{-1, 0, 1, 2\}$. There is an underlying latent variable y^*, the value of this variable relative to threshold values $\{-0.5, 0.5, 1.5\}$ determine the observed value of outcome y. The latent variable will be a linear function of two variables and a mean zero random error term, the variance of the random error term will be a function of the same two variables.

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \sigma\epsilon$$

where ϵ will have a standard normal distribution. The variance σ will be given by:

$$\sigma = \exp\left(\delta_0 + \delta_1 x_1 + \delta_2 x_2\right)$$

Variable x_1 will be binary and x_2 will be continuous. Specifically x_1 will take on value 1 with probability 0.75 and x_2 will be drawn from a standard normal distribution.

```
>   set.seed(242)
>   n<-250
>   x1<-rbinom(n,1,0.75) # binary variable
>   x2<-rnorm(n) # continuous variable
>   sampledata<-cbind(rep(1,n),x1,x2)
>   # set true parameter values
>   meanparams<-c(0.5,1,-0.5)
>   varparams<- c(0,0.5,-0.5)
>   # generate latent variable
>   ystar<-sampledata%*%meanparams+rnorm(n)*exp(sampledata%*%varparams)
>   # generate outcome variable
>   threshparams<-c(-0.5,0.5,1.5)
>   outcomes<-c(-1,0,1,2)
>   setvalue<-function(x){
+     locate<-outcomes[1:(length(outcomes)-1)][x<threshparams]
+     if (length(locate)==0){
+       return(outcomes[length(outcomes)])
+     } else {
```

```
+       return(locate[1])
+     }
+   }
>   y<-sapply(ystar,setvalue)
>   sampleframe<-data.frame(y,x1,x2)
```

The parameter values chosen for the example are for the mean equation $\beta_0 = 0.5$, $\beta_1 = 1$ and $\beta_2 = 1$ and for the variance $\delta_0 = 0$, $\delta_1 = 0.5$ and $\delta_2 = -0.5$. Given the parameters and knowing the source distribution of the variables we can calculate the expected marginal effects at means for the two variables. Table 2 displays the implied marginal effects.

Table 2: Implied Marginal Effects at Means for Example Data.

Outcome	x_1	x_2
$y = -1$	-0.017	-0.05
$y = 0$	-0.133	0.08
$y = 1$	-0.154	0.139
$y = 2$	0.304	-0.169

Note that the marginal effect of the variable x_2 is negative for the two extreme outcomes ($y = -1$ and $y = 2$). This possibility cannot be captured in a standard ordered probit model unless including the variable non-linearly in the regression equation, for example by adding the square of the variable, which generates additional complications in calculating marginal effects.

The standard ordered probit can be estimated with the oglmx function with the correct specification of optional arguments, or by using the oprobit.reg function included with the package.

```
>   library("oglmx")
>   results.oprob<-oglmx(y ~ x1 + x2, data=sampleframe, link="probit",
+                        constantMEAN = FALSE, constantSD = FALSE,
+                        delta=0,threshparam = NULL)
>   results.oprob1<-oprobit.reg(y ~ x1 + x2, data=sampleframe)
>   summary(results.oprob)

Ordered Probit Regression
Log-Likelihood: -307.3201
No. Iterations: 4
McFadden's R2: 0.07043651
AIC: 624.6402
    Estimate Std. error t value  Pr(>|t|)
x1  0.979599   0.160485  6.1040 1.034e-09 ***
x2 -0.214214   0.072843 -2.9408  0.003274 **
----- Threshold Parameters -----
                  Estimate Std. error t value  Pr(>|t|)
```

```
Threshold (-1->0) -0.48969    0.14503 -3.3764 0.0007344 ***
Threshold (0->1)   0.23025    0.14050  1.6388 0.1012615
Threshold (1->2)   1.06668    0.15215  7.0109 2.368e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(results.oprob1)

Ordered Probit Regression
Log-Likelihood: -307.3201
No. Iterations: 4
McFadden's R2: 0.07043651
AIC: 624.6402
    Estimate Std. error t value  Pr(>|t|)
x1  0.979599   0.160485  6.1040 1.034e-09 ***
x2 -0.214214   0.072843 -2.9408  0.003274 **
----- Threshold Parameters -----
                  Estimate Std. error t value  Pr(>|t|)
Threshold (-1->0) -0.48969    0.14503 -3.3764 0.0007344 ***
Threshold (0->1)   0.23025    0.14050  1.6388 0.1012615
Threshold (1->2)   1.06668    0.15215  7.0109 2.368e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> margins.oglmx(results.oprob,ascontinuous = TRUE)

Marginal Effects on Pr(Outcome==-1)
   Marg. Eff Std. error t value  Pr(>|t|)
x1 -0.190014   0.035860 -5.2988 1.166e-07 ***
x2  0.041551   0.014685  2.8295  0.004662 **
-------------------------------------
Marginal Effects on Pr(Outcome==0)
   Marg. Eff Std. error t value  Pr(>|t|)
x1 -0.158101   0.034587 -4.5711 4.852e-06 ***
x2  0.034573   0.012631  2.7372  0.006197 **
-------------------------------------
Marginal Effects on Pr(Outcome==1)
    Marg. Eff Std. error t value Pr(>|t|)
x1 -0.0187645  0.0222300 -0.8441   0.3986
x2  0.0041033  0.0050714  0.8091   0.4184
-------------------------------------
Marginal Effects on Pr(Outcome==2)
   Marg. Eff Std. error t value  Pr(>|t|)
x1  0.366880   0.060214  6.0930 1.108e-09 ***
x2 -0.080227   0.027374 -2.9308  0.003381 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In a typical ordered probit model, whether standard or heteroskedastic, the scale of the parameter vectors is not identified. Identifying assumptions are necessary, for the standard ordered probit the constant in the latent variable equation is set equal to zero while the variance of the error term is set equal to one. These assumptions are somewhat arbitrary and may be replaced by alternative assumptions. For example, in the case that the threshold values are meaningful two (or more) of them can be imposed and the intercept and variance of the error term can be estimated which leads to parameter estimates with a meaningful scale. The code below estimates a heteroskedastic probit model, first with a standard assumption of no constant in either the latent variable mean equation or the variance equation and secondly using the fact we know the threshold values so fixing two of them to their true values.

```
>   results.oprobhet<-oglmx(y ~ x1 + x2, ~ x1 + x2, data=sampleframe,
+                           constantMEAN = FALSE, constantSD = FALSE)
>   results.oprobhet1<-oglmx(y ~ x1 + x2, ~ x1 + x2, data=sampleframe,
+                           constantMEAN = TRUE, constantSD = TRUE,
+                           threshparam=c(-0.5,NA,1.5))
>   summary(results.oprobhet)

Heteroskedastic Ordered Probit Regression
Log-Likelihood: -275.2048
No. Iterations: 13
McFadden's R2: 0.167577
AIC: 564.4096
----- Mean Equation ------
   Estimate Std. error t value  Pr(>|t|)
x1  1.50473    0.26265  5.7290 1.010e-08 ***
x2 -0.62815    0.10735 -5.8512 4.879e-09 ***
----- SD Equation ------
    Estimate Std. error t value  Pr(>|t|)
x1  0.669208   0.168019  3.9829 6.807e-05 ***
x2 -0.598387   0.095268 -6.2811 3.362e-10 ***
----- Threshold Parameters -----
                  Estimate Std. error t value  Pr(>|t|)
Threshold (-1->0) -1.06454    0.18111 -5.8780 4.152e-09 ***
Threshold (0->1)   0.04495    0.16252  0.2766    0.7821
Threshold (1->2)   1.43314    0.27329  5.2440 1.572e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>   summary(results.oprobhet1)

Heteroskedastic Ordered Probit Regression
Log-Likelihood: -275.2048
No. Iterations: 9
McFadden's R2: 0.167577
AIC: 564.4096
----- Mean Equation ------
```

```
             Estimate Std. error t value  Pr(>|t|)
(Intercept)  0.352425   0.118452  2.9753  0.002927 **
x1           1.204904   0.129457  9.3074 < 2.2e-16 ***
x2          -0.502989   0.074112 -6.7869 1.146e-11 ***
----- SD Equation ------
             Estimate Std. error t value  Pr(>|t|)
(Intercept) -0.222216   0.137325 -1.6182    0.1056
x1           0.669208   0.168019  3.9829 6.807e-05 ***
x2          -0.598387   0.095268 -6.2811 3.362e-10 ***
----- Threshold Parameters -----
                 Estimate Std. error t value  Pr(>|t|)
Threshold (0->1) 0.388418   0.088539   4.387 1.149e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>   margins.oglmx(results.oprobhet)

Marginal Effects on Pr(Outcome==-1)
   Marg. Eff Std. error t value Pr(>|t|)
x1 -0.055846   0.046467 -1.2018 0.229425
x2 -0.065320   0.022854 -2.8582 0.004261 **
-------------------------------------
Marginal Effects on Pr(Outcome==0)
   Marg. Eff Std. error t value  Pr(>|t|)
x1 -0.260152   0.049548 -5.2505 1.517e-07 ***
x2  0.068614   0.013979  4.9085 9.178e-07 ***
-------------------------------------
Marginal Effects on Pr(Outcome==1)
   Marg. Eff Std. error t value  Pr(>|t|)
x1 -0.128857   0.056905 -2.2644   0.02355 *
x2  0.206568   0.038219  5.4049 6.486e-08 ***
-------------------------------------
Marginal Effects on Pr(Outcome==2)
   Marg. Eff Std. error t value  Pr(>|t|)
x1  0.444855   0.037565  11.842 < 2.2e-16 ***
x2 -0.209862   0.024839  -8.449 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>   margins.oglmx(results.oprobhet1)

Marginal Effects on Pr(Outcome==-1)
   Marg. Eff Std. error t value Pr(>|t|)
x1 -0.055846   0.046467 -1.2018 0.229425
x2 -0.065320   0.022854 -2.8582 0.004261 **
-------------------------------------
Marginal Effects on Pr(Outcome==0)
```

```
   Marg. Eff Std. error t value  Pr(>|t|)
x1 -0.260152   0.049548 -5.2505 1.517e-07 ***
x2  0.068614   0.013979  4.9085 9.178e-07 ***
-------------------------------------
Marginal Effects on Pr(Outcome==1)
   Marg. Eff Std. error t value  Pr(>|t|)
x1 -0.128857   0.056905 -2.2644   0.02355 *
x2  0.206568   0.038219  5.4049 6.486e-08 ***
-------------------------------------
Marginal Effects on Pr(Outcome==2)
   Marg. Eff Std. error t value  Pr(>|t|)
x1  0.444855   0.037565  11.842 < 2.2e-16 ***
x2 -0.209862   0.024839  -8.449 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Considering the results for the parameter estimates, the size of the parameters in the mean equation are different but the signs are the same and the conclusions reached from t-tests of significance are the same. More importantly, the estimated marginal effects from these two procedures are identical. The oglm command in Stata that permits estimation of heteroskedastic ordered models imposes the no-constant assumption used in the first estimation above. Given the results and knowing the threshold values it is possible to transform the results to obtain the correctly scaled parameter estimates, however by allowing a flexible specification the **oglmx** package avoids the need for this transformation.

Using the **lmtest** package we can test whether the inclusion of the variables x_1 and x_2 in the variance equation are jointly significant using a likelihood ratio test.

```
>   library("lmtest")
>   lrtest(results.oprob,results.oprobhet)

Likelihood ratio test

Model 1: y ~ x1 + x2
Model 2: y ~ (x1 + x2 | x1 + x2)
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   5 -307.32
2   7 -275.20  2 64.231  1.128e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Unsurprisingly, we see that by modelling the heteroskedasticity we find a significant improvement in model fit.

5. Summary and future development

This paper introduces the **oglmx** package for the R programming language. In linear models heteroskedasticity in the error term does not affect the consistency of parameter estimates and provided suitable standard error corrections are used inferences drawn from OLS estimates are reliable. For non-linear models such as probit and ordered probit this is no longer the case, heteroskedasticity can lead to substantially biased estimates of marginal effects. The **oglmx** package permits the user to model the heteroskedasticity in order to obtain consistent estimates of marginal effects and reliable statistical tests.

Ai and Norton (2003) identified an error committed by many applied researchers when interpreting results when using non-linear models such as probit and logit and including interaction terms in their model. Further development of the package is expected to lead the margins function to identify functions of variables that enter the equations for the mean and standard deviation equations, i.e. interaction terms and polynomials of variables and estimate the true marginal effects desired by researchers.

Inclusion of fixed effects in non-linear models such as probit, logit and their ordered equivalents can lead to biased estimates due to the incidental parameters problem recognised by Neyman and Scott (1948). Further development of the package will add methods to reduce this bias, for example that suggested by Carro (2007) and adapted to the ordered outcome case with two types of fixed effects in Carro and Traferri (2014).

References

Ai C, Norton EC (2003). "Interaction terms in logit and probit models." *Economics Letters*, **80**(1), 123–129. ISSN 01651765. doi:10.1016/S0165-1765(03)00032-6. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.197.5996{&}rep=rep1{&}type=pdf.

Carro JM (2007). "Estimating dynamic panel data discrete choice models with fixed effects." *Journal of Econometrics*, **140**(2), 503–528. ISSN 03044076. doi:10.1016/j.jeconom.2006.07.023.

Carro JM, Traferri A (2014). "State Dependence and Heterogeneity in Health Using a Bias-Corrected Fixed-Effects Estimator." *Journal of Applied Econometrics*, **29**, 181–207. ISSN 01451707. doi:10.1002/jae.

Case A, Lubotsky D, Paxson C (2002). "Economic Status and Health in Childhood: The Origins of the Gradient." *American Economic Review*, **92**(5), 1308–1334.

Dearden L, Meghir C, Ferri J (2002). "The effect of school quality on educational attainment and wages." *The Review of Economics and Statistics*, **84**(February), 1–20. doi:10.1162/003465302317331883. URL http://dx.doi.org/10.1162/003465302317331883.

Fernihough A (2014). *mfx: Marginal Effects, Odds Ratios and Incidence Rate Ratios for GLMs*. URL http://cran.r-project.org/package=mfx.

Gerdtham UG, Johannesson M (2001). "The relationship between happiness, health, and socio-economic factors: Results based on Swedish microdata." *Journal of Socio-Economics*, **30**(6), 553–557. ISSN 10535357. doi:10.1016/S1053-5357(01)00118-4.

Hartog J, Oosterbeek H (1998). "Health, wealth and happiness: why pursue a higher education?" *Economics of Education Review*, **17**(3), 245–256. ISSN 02727757. doi:10.1016/S0272-7757(97)00064-2.

Henningsen A, Toomet O (2011). "maxLik: A package for maximum likelihood estimation in R." *Computational Statistics*, **26**, 443–458. ISSN 09434062. doi:10.1007/s00180-010-0217-1.

Neyman J, Scott EL (1948). "Consistent Estimates Based on Partially Consistent Observations." *Econometrica*, **16**(1), 1–32.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org/.

Sun C (2014). *erer: Empirical Research in Economics with R*. URL http://cran.r-project.org/package=erer.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Fourth edition. Springer, New York. URL http://www.stats.ox.ac.uk/pub/MASS4.

Wooldridge JM (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press Books. The MIT Press. URL http://ideas.repec.org/b/mtp/titles/0262232197.html.

Affiliation:

Nathan Carroll
Institut für Volkswirtschaftslehre und Ökonometrie
Faculty of Business, Economics and Management Information Systems
Universität Regensburg
Universitätsstrasse 31
93053 Regensburg, Germany
E-mail: nathan.carroll@ur.de