Heteroscedastic Probit Regression

Judith Santer
Universität Innsbruck

Abstract

The **hetprobit** package (https://R-Forge.R-project.org/projects/uibk-rprog-2017/) fits heteroscedastic probit regression models via maximum likelihood. In the following the methodology of heteroscedastic probit models is briefly presented. The implementation of such models in R and the practical use of it using the example of voter turnout data from Nagler (1991) are shown. Finally, a replication of the results of the **glmx** package is presented.

Keywords: heteroscedastic probit, regression, R.

1. Introduction

In standard probit regression models the probability of a success, i.e. $P(Y_i = 1)$, is modelled as

$$P(Y_i = 1) = \pi_i = \Phi(x_i^\top \beta)$$

where $\Phi(\cdot)$ is the cdf of a standard normal distribution.
The general assumption that the variance of the error term is constant and due to identifiability set to one is relaxed in the heteroscedastic probit model. The variance can vary systematically and is now modelled as a multiplicative function of regressor variables, i.e.

$$\sigma_i = \exp(z_i^\top \gamma).$$

The probability of success is now represented by

$$\pi_i = \Phi\left(\frac{x_i^\top \beta}{\exp(z_i^\top \gamma)}\right).$$

Note that the scale model is only identified without intercept.

For a detailed discussion of heteroscedastic probit models see e.g. Harvey (1976), Alvarez and Brehm (1995), Keele and Park (2006) and Freeman, Keele, Park, Salzman, and Weickert (2015).
Sections 2 and 3 show the implementation of such models in R (R Core Team 2017) and the practical use of it using the example of voter turnout data from Nagler (1991). Finally, a replication of the results of the **glmx** (Zeileis, Koenker, and Doebler 2015) package is presented.

2. Implementation

As usual in many other regression packages for R, the main model fitting function `hetprobit()` uses a formula-based interface and returns an (S3) object of class `hetprobit`:

```
hetprobit(formula, data, subset, na.action,
  model = TRUE, y = TRUE, x = FALSE,
  control = hetprobit_control(...), ...)
```

Actually, the `formula` can be a two-part `Formula` (Zeileis and Croissant 2010), specifying sets of regressors x_i and z_i for the mean and scale submodels, respectively. The specification of formula `y ~ x1 + x2` is the short version of `y ~ x1 + x2 | x1 + x2` with exactly the same set of regressors used in the mean and scale equation. Different sets of regressors, e.g. `y ~ x1 + x2 | z1`, `y ~ x1 + x2 | z1 + x2` and `y ~ x1 + x2 | 1` are also possible. The last specification assumes a constant scale (`~ 1`), i.e. in this setting a homoscedastic probit model would be estimated.
By default the model frame (`model = TRUE`) and the response (`y = TRUE`) are returned whereas the model matrix is not (`x = FALSE`).

The underlying workhorse function is `hetprobit_fit()` which has a matrix interface and returns an unclassed list with e.g. mean and scale coefficients, fitted values, raw residuals. In order to estimate the coefficients via maximum likelihood the `optim()` function is used. If the starting values are not set by the user, the coefficients estimates returned by `glm()` with `family = binomial(link = "probit")` are used for the mean equation. The starting values for the coefficients in the scale model are set to zero. Remember that there is no intercept in the scale model.
By default analytical gradients together with the `"BFGS"`-method are employed and the hessian is approximated numerically.

Additionally, numerous standard S3 methods are provided (see Table 1). As usual fitted means of the observed response variable can be extracted by the generic function `fitted()`.

Due to these methods a number of useful utilities work automatically, e.g., `AIC()`, `BIC()`, `coeftest()` (**lmtest**), `lrtest()` (**lmtest**), `waldtest()` (**lmtest**), `linearHypothesis()` (**car**), `mtable()` (**memisc**), etc.

3. Illustration

This section is devoted to present the functionality of the package using data on voter turnout of the U.S. presidential elections in 1984. The data has first been analyzed by Nagler (1991) to see whether registration laws and education have an influence on the propensity to vote. Beyond these two effects further controls were included (see Table 2). In 1994 Nagler fitted his skewed logit model to the data and Altman and McDonald (2003) replicated this study with focus on numerical accuracy. The data and further materials needed for the replication are available in their paper supplements. [1]

[1] In the original work of Nagler (1991) 98,860 persons were interviewed, in the study of Altman and McDonald (2003) only 98,857 observations could be replicated. Thus, the model output of `mn` is slightly different to the published results of Nagler (1991).

Method	Description
`print()`	Simple printed display with coefficients
`summary()`	Standard regression summary; returns `summary.hetprobit` object (with `print()` method)
`coef()`	Extract coefficients
`vcov()`	Associated covariance matrix
`predict()`	(Different types of) predictions for new data
`residuals()`	Extract (different types of) residuals
`terms()`	Extract terms
`model.matrix()`	Extract model matrix (or matrices)
`update()`	Update and re-fit a model
`nobs()`	Extract number of observations
`logLik()`	Extract fitted log-likelihood
`bread()`	Extract bread for **sandwich** covariance
`estfun()`	Extract estimating functions (= gradient contributions) for **sandwich** covariances
`getSummary()`	Extract summary statistics for `mtable()`

Table 1: S3 methods provided in **hetprobit**.

Variable	Description	Mean/ % of 'yes'
`vote`	Did the respondent vote?	67 %
`education`	Years of education of the respondent	5.3
`age`	Age of the respondent	43.8
`south`	Is the respondent from the South?	22 %
`govelection`	Were gubernatorial elections held?	18 %
`closing`	How many days before the election has the registration been closed?	24.7

Table 2: Variables in the `VoterTurnout` dataset.

```
R> data("VoterTurnout", package = "hetprobit")
R> library("hetprobit")
R> mn <- glm(vote ~ age + I(age^2) + south + govelection +
+    (education + I(education^2)) * closing,
+    data = VoterTurnout, family = binomial(link = "probit"))
R> summary(mn)


Call:
glm(formula = vote ~ age + I(age^2) + south + govelection + (education +
    I(education^2)) * closing, family = binomial(link = "probit"),
    data = VoterTurnout)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.834  -1.080   0.603   0.873   2.236
```

```
Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -2.7443830  0.1074452  -25.54  < 2e-16 ***
age                     0.0695702  0.0013162   52.86  < 2e-16 ***
I(age^2)               -0.0005047  0.0000136  -37.17  < 2e-16 ***
south                  -0.1116227  0.0104396  -10.69  < 2e-16 ***
govelection             0.0043185  0.0113522    0.38    0.704
education               0.2647146  0.0417268    6.34  2.2e-10 ***
I(education^2)          0.0050968  0.0041839    1.22    0.223
closing                 0.0011137  0.0037293    0.30    0.765
education:closing      -0.0032780  0.0015114   -2.17    0.030 *
I(education^2):closing  0.0002829  0.0001521    1.86    0.063 .
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 125375  on 98856  degrees of freedom
Residual deviance: 110744  on 98847  degrees of freedom
AIC: 110764

Number of Fisher Scoring iterations: 4
```

The same estimates could have been obtained by using the package's `hetprobit()` function, but with a trade-off in efficiency compared to `glm()`.

```
R> mn1 <- hetprobit(vote ~ age + I(age^2) + south + govelection +
+    (education + I(education^2)) * closing | 1,
+    data = VoterTurnout)
```

In a next step the replicated homoscedastic model will be modified in such a way that all regressors (including interactions) in the mean model are also part of the scale submodel (full model `m1`). Additionally, a reduced model without interaction effects will be fitted (model `m2`).

```
R> m1 <- hetprobit(vote ~ age + I(age^2) + south + govelection +
+    (education + I(education^2)) * closing |
+    age + I(age^2) + south + govelection +
+    (education + I(education^2)) * closing,
+    data = VoterTurnout)
R> summary(m1)

Heteroscedastic probit model

Call:
hetprobit(formula = vote ~ age + I(age^2) + south + govelection +
    (education + I(education^2)) * closing | age + I(age^2) +
```

```
    south + govelection + (education + I(education^2)) * closing,
    data = VoterTurnout)

Standardized residuals:
   Min     1Q Median     3Q    Max
-4.863 -0.875  0.434  0.667  3.564

Coefficients (binomial model with probit link):
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.828516   0.481262   -3.80  0.00015 ***
age                    0.061776   0.015678    3.94 0.000081 ***
I(age^2)              -0.000437   0.000113   -3.87  0.00011 ***
south                 -0.086041   0.022420   -3.84  0.00012 ***
govelection           -0.007977   0.011533   -0.69  0.48911
education             -0.212899   0.073907   -2.88  0.00397 **
I(education^2)         0.068554   0.018688    3.67  0.00024 ***
closing               -0.003237   0.003378   -0.96  0.33797
education:closing      0.000495   0.001871    0.26  0.79119
I(education^2):closing -0.000298   0.000261   -1.14  0.25446

Latent scale model coefficients (with log link):
                       Estimate Std. Error z value Pr(>|z|)
age                   -0.0251456  0.0028671   -8.77  < 2e-16 ***
I(age^2)               0.0002434  0.0000297    8.19  2.6e-16 ***
south                  0.0482924  0.0196845    2.45    0.014 *
govelection           -0.0393607  0.0204857   -1.92    0.055 .
education              0.0078116  0.0845681    0.09    0.926
I(education^2)         0.0176976  0.0072757    2.43    0.015 *
closing                0.0029284  0.0088071    0.33    0.740
education:closing     -0.0018285  0.0030853   -0.59    0.553
I(education^2):closing  0.0001353  0.0002673    0.51    0.613
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -5.51e+04 on 19 Df
Number of iterations in BFGS optimization: 38

R> m2 <- update(m1, . ~ . - (education + I(education^2)):closing  |
+    . - (education + I(education^2)):closing)
R> summary(m2)

Heteroscedastic probit model

Call:
hetprobit(formula = vote ~ age + I(age^2) + south + govelection +
    education + I(education^2) + closing | age + I(age^2) +
    south + govelection + education + I(education^2) + closing,
```

```
    data = VoterTurnout)

Standardized residuals:
   Min     1Q Median     3Q    Max
-4.691 -0.873  0.433  0.665  3.372

Coefficients (binomial model with probit link):
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.5257682  0.2062887   -7.40  1.4e-13 ***
age             0.0532561  0.0073816    7.21  5.4e-13 ***
I(age^2)       -0.0003736  0.0000559   -6.68  2.3e-11 ***
south          -0.0746493  0.0123502   -6.04  1.5e-09 ***
govelection    -0.0083568  0.0100220   -0.83      0.4
education      -0.1673403  0.0277078   -6.04  1.5e-09 ***
I(education^2)  0.0520946  0.0069924    7.45  9.3e-14 ***
closing        -0.0055299  0.0007122   -7.76  8.2e-15 ***

Latent scale model coefficients (with log link):
                 Estimate Std. Error z value Pr(>|z|)
age            -0.0244315  0.0029053   -8.41  < 2e-16 ***
I(age^2)        0.0002420  0.0000305    7.94  1.9e-15 ***
south           0.0461163  0.0196665    2.34    0.019 *
govelection    -0.0458717  0.0205164   -2.24    0.025 *
education      -0.0732671  0.0355107   -2.06    0.039 *
I(education^2)  0.0236485  0.0030141    7.85  4.3e-15 ***
closing         0.0006819  0.0008211    0.83    0.406
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -5.51e+04 on 15 Df
Number of iterations in BFGS optimization: 49
```

Comparing these three models by means of information criteria the homoscedastic probit model (`mn`) replicated at the beginning would be the least preferred one. The BIC that penalizes complex models more strongly than the AIC is in favor of the reduced heteroscedastic model `m2`.

```
R> AIC(mn, m1, m2)

   df    AIC
mn 10 110764
m1 19 110221
m2 15 110238

R> BIC(mn, m1, m2)

   df    BIC
mn 10 110859
```

```
m1 19 110401
m2 15 110380
```

A likelihood ratio test on the nested models `m1` and `m2` would prefer the full heteroscedastic model `m1` over the reduced one.

```
R> library("lmtest")
R> lrtest(mn, m1, m2)

Likelihood ratio test

Model 1: vote ~ age + I(age^2) + south + govelection + (education + I(education^2)) *
    closing
Model 2: vote ~ age + I(age^2) + south + govelection + (education + I(education^2)) *
    closing | age + I(age^2) + south + govelection + (education +
    I(education^2)) * closing
Model 3: vote ~ age + I(age^2) + south + govelection + education + I(education^2) +
    closing | age + I(age^2) + south + govelection + education +
    I(education^2) + closing
  #Df LogLik Df Chisq Pr(>Chisq)
1  10 -55372
2  19 -55091  9 560.9    < 2e-16 ***
3  15 -55104 -4  24.8   0.000055 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

4. Replication

This package is a somewhat simpler reimplementation of the function `hetglm()` from the package **glmx**. In case of an implementation not limited to the probit link for the mean and log link for the scale equation **glmx** offers more flexibility. In particular, `hetglm` offers analytical Hessian and flexible link functions for the mean and scale submodel among further features.
For illustration purposes a sparser heteroscedastic model than in Section 3 is used. Therefore, the polynomials of `age` and `education` as well as the interaction between `education` and `closing` are removed from the model equation.

```
R> library("glmx")
R> m0 <- hetglm(vote ~ age + south + govelection + education +
+    I(education^2) + closing |
+    age + south + govelection + education + I(education^2) + closing,
+    data = VoterTurnout, method = "BFGS", hessian = TRUE)

Call:
hetglm(formula = vote ~ age + south + govelection + education +
```

```
    closing | age + south + govelection + education + closing,
    data = VoterTurnout, method = "BFGS", hessian = TRUE)

Deviance residuals:
   Min     1Q Median     3Q    Max
-2.571 -1.085  0.619  0.851  2.935

Coefficients (binomial model with probit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.626438   0.222880  -20.76  < 2e-16 ***
age          0.065286   0.002959   22.06  < 2e-16 ***
south       -0.146467   0.022745   -6.44  1.2e-10 ***
govelection -0.027736   0.024274   -1.14     0.25
education    0.634912   0.029734   21.35  < 2e-16 ***
closing     -0.012997   0.000956  -13.60  < 2e-16 ***

Latent scale model coefficients (with log link):
             Estimate Std. Error z value Pr(>|z|)
age          0.016815   0.000473   35.55  < 2e-16 ***
south        0.069190   0.019713    3.51  0.00045 ***
govelection -0.044501   0.020308   -2.19  0.02843 *
education   -0.009016   0.005155   -1.75  0.08030 .
closing      0.001334   0.000790    1.69  0.09130 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -5.55e+04 on 11 Df
LR test for homoskedasticity: 1.23e+03 on 5 Df, p-value: <2e-16
Dispersion: 1
Number of iterations in BFGS optimization: 24
```

In the `hetglm()` call for the benchmark model `m0` two arguments needed to be switched from the defaults: The method has to be changed to `method = "BFGS"` (rather than `"nlminb"`) and the `hessian`-argument is set to `TRUE` in order to derive the hessian numerically.

```
R> m <- hetprobit(vote ~ age + south + govelection + education + closing |
+    age + south + govelection + education + closing,
+    data = VoterTurnout)
```

Using a model table from **memisc** (Elff 2016) the replicated estimation results (model `m`) can be easily embedded in a TEX-file (see Table 3). However, the method `getSummary()` is not (yet) supported in the **glmx** package.

```
R> library("memisc")
R> toLatex(mtable(m, summary.stats = c("Log-likelihood", "AIC", "BIC", "N")))
```

	mean	scale
(Intercept)	−4.618***	
	(0.223)	
age	0.065***	0.017***
	(0.003)	(0.000)
south	−0.147***	0.069***
	(0.023)	(0.020)
govelection	−0.027	−0.045*
	(0.024)	(0.020)
education	0.634***	−0.009
	(0.030)	(0.005)
closing	−0.013***	0.001
	(0.001)	(0.001)
Log-likelihood	−55479.7	
AIC	110981.4	
BIC	111085.9	
N	98857	

Table 3: Replication of **glmx** results using **hetprobit**.

References

Altman M, McDonald MP (2003). "Replication with Attention to Numerical Accuracy." *Political Analysis*, **11**(3), 302–307. doi:10.1093/pan/mpg016.

Alvarez RM, Brehm J (1995). "American Ambivalence Towards Abortion Policy: Development of a Heteroskedastic Probit Model of Competing Values." *American Journal of Political Science*, **39**(4), 1055–1082. doi:10.2307/2111669.

Elff M (2016). ***memisc**: Tools for Management of Survey Data and the Presentation of Analysis Results*. R package version 0.99.8, URL https://CRAN.R-project.org/package=memisc.

Freeman E, Keele L, Park D, Salzman J, Weickert B (2015). "The Plateau Problem in the Heteroskedastic Probit Model." *ArXiv e-prints*. URL https://arxiv.org/abs/1508.03262.

Harvey AC (1976). "Estimating Regression Models with Multiplicative Heteroscedasticity." *Econometrica*, **44**(3), 461–465. doi:10.2307/1913974.

Keele L, Park D (2006). *Ambivalent about Ambivalence: A Re-examination of Heteroskedastic Probit Models*. Unpublished Manuscript, Penn State University.

Nagler J (1991). "The Effect of Registration Laws and Education on U.S. Voter Turnout." *The American Political Science Review*, **85**(4), 1393–1405. doi:10.2307/1963952.

Nagler J (1994). "Scobit: An Alternative Estimator to Logit and Probit." *American Journal of Political Science*, **38**(1), 230–255. doi:10.2307/2111343.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Zeileis A, Croissant Y (2010). "Extended Model Formulas in R: Multiple Parts and Multiple Responses." *Journal of Statistical Software*, **34**(1), 1–13. doi:10.18637/jss.v034.i01.

Zeileis A, Koenker R, Doebler P (2015). ***glmx**: Generalized Linear Models Extended*. R package version 0.1-1, URL https://cran.r-project.org/package=glmx.

Affiliation:

Judith Santer
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: Judith.Santer@uibk.ac.at