Heteroscedastic Tobit Regression

Achim Zeileis
Universität Innsbruck

Abstract

The **htobit** package (https://R-Forge.R-project.org/projects/uibk-rprog-2017/) fits tobit regression models with conditional heteroscedasticy using maximum likelihood estimation. A brief overview of the package is provided, along with some illustrations and a replication of results from the **crch** package.

Keywords: heteroscedastic tobit, regression, R.

1. Introduction

The heteroscedastic tobit model assumes an underlying latent Gaussian variable

$$y_i^* \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

which is only observed if positive and zero otherwise: $y_i = \max(0, y_i^*)$. The latent mean μ_i and scale σ_i (latent standard deviation) are linked to two different linear predictors

$$\begin{aligned} \mu_i &= x_i^\top \beta \\ \log(\sigma_i) &= z_i^\top \gamma \end{aligned}$$

where the regressor vectors x_i and z_i can be set up without restrictions, i.e., they can be identical, overlapping or completely different or just including an intercept, etc.

See also Messner, Mayr, and Zeileis (2016) for a more detailed introduction to this model class as well as a better implementation in the package **crch**. The main purpose of **htobit** is to illustrate how to create such a package *from scratch*.

2. Implementation

As usual in many other regression packages for R (R Core Team 2019), the main model fitting function `htobit()` uses a formula-based interface and returns an (S3) object of class `htobit`:

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

Actually, the `formula` can be a two-part `Formula` (Zeileis and Croissant 2010), specifying separate sets of regressors x_i and z_i for the location and scale submodels, respectively.

Method	Description
`print()`	Simple printed display with coefficients
`summary()`	Standard regression summary; returns `summary.htobit` object (with `print()` method)
`coef()`	Extract coefficients
`vcov()`	Associated covariance matrix
`predict()`	(Different types of) predictions for new data
`fitted()`	Fitted values for observed data
`residuals()`	Extract (different types of) residuals
`terms()`	Extract terms
`model.matrix()`	Extract model matrix (or matrices)
`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 **htobit**.

The underlying workhorse function is `htobit_fit()` which has a matrix interface and returns an unclassed list.

A number of standard S3 methods are provided, see Table 1.

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**), `Boot()` (**car**), etc.

3. Illustration

To illustrate the package's use in practice, a comparison of several homoscedastic and heteroscedastic tobit regression models is applied to data on budget shares of alcohol and tobacco for 2724 Belgian households (taken from Verbeek 2004). The homoscedastic model from Verbeek (2004) can be replicated by:

```
R> data("AlcoholTobacco", package = "htobit")
R> library("htobit")
R> ma <- htobit(alcohol ~ (age + adults) * log(expenditure) + oldkids + youngkids,
+    data = AlcoholTobacco)
R> summary(ma)

Call:
htobit(formula = alcohol ~ (age + adults) * log(expenditure) +
    oldkids + youngkids, data = AlcoholTobacco)

Standardized residuals:
    Min      1Q  Median      3Q     Max
```

```
-1.0698 -0.4407 -0.1364  0.3934  8.3170

Coefficients (location model):
                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -0.1591533  0.0437782  -3.635 0.000278 ***
age                     0.0134938  0.0108824   1.240 0.214989    
adults                  0.0291901  0.0169469   1.722 0.084989 .  
log(expenditure)        0.0126679  0.0032156   3.939 8.17e-05 ***
oldkids                -0.0026408  0.0006049  -4.366 1.27e-05 ***
youngkids              -0.0038789  0.0023835  -1.627 0.103651    
age:log(expenditure)   -0.0008093  0.0008006  -1.011 0.312067    
adults:log(expenditure) -0.0022484  0.0012232  -1.838 0.066051 .  

Coefficients (scale model with log link):
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.71236    0.01533  -242.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood:  4755 on 9 Df
Number of iterations in BFGS optimization: 16
```

This model is now modified in two directions: First, the variables influencing the location parameter are also employed in the scale submodel. Second, because the various coefficients for different numbers of persons in the household do not appear to be very different, a restricted specification for this is used. Using a Wald test for testing linear hypotheses from **car** (Fox and Weisberg 2019) yields

```
R> library("car")
R> linearHypothesis(ma, "oldkids = youngkids")

Linear hypothesis test

Hypothesis:
oldkids - youngkids = 0

Model 1: restricted model
Model 2: alcohol ~ (age + adults) * log(expenditure) + oldkids + youngkids

  Df  Chisq Pr(>Chisq)
1                     
2  1 0.2639     0.6075

R> linearHypothesis(ma, "oldkids = adults")

Linear hypothesis test
```

```
Hypothesis:
- adults  + oldkids = 0

Model 1: restricted model
Model 2: alcohol ~ (age + adults) * log(expenditure) + oldkids + youngkids

  Df  Chisq Pr(>Chisq)
1
2  1 3.4994    0.06139 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Therefore, the following models are considered:

```
R> AlcoholTobacco$persons <- with(AlcoholTobacco, adults + oldkids + youngkids)
R> ma2 <- htobit(alcohol ~ (age + adults) * log(expenditure) + oldkids + youngkids |
+    (age + adults) * log(expenditure) + oldkids + youngkids, data = AlcoholTobacco)
R> ma3 <- htobit(alcohol ~ age + log(expenditure) + persons | age +
+    log(expenditure) + persons, data = AlcoholTobacco)
R> BIC(ma, ma2, ma3)

    df       BIC
ma   9 -9439.553
ma2 16 -9735.109
ma3  8 -9777.154
```

The BIC would choose the most parsimonious model but a likelihood ratio test would prefer the unconstrained person coefficients:

```
R> library("lmtest")
R> lrtest(ma, ma2, ma3)

Likelihood ratio test

Model 1: alcohol ~ (age + adults) * log(expenditure) + oldkids + youngkids
Model 2: alcohol ~ (age + adults) * log(expenditure) + oldkids + youngkids |
    (age + adults) * log(expenditure) + oldkids + youngkids
Model 3: alcohol ~ age + log(expenditure) + persons | age + log(expenditure) +
    persons
  #Df LogLik Df   Chisq Pr(>Chisq)
1   9 4755.4
2  16 4930.8  7 350.925  < 2.2e-16 ***
3   8 4920.2 -8  21.234   0.006551 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

	htobit		crch	
	location	scale	location	scale
(Intercept)	-0.072^{***}	0.176	-0.072^{***}	0.176
	(0.014)	(0.515)	(0.014)	(0.515)
age	0.002^{***}	0.064^{***}	0.002^{***}	0.064^{***}
	(0.000)	(0.013)	(0.000)	(0.013)
log(expenditure)	0.006^{***}	-0.278^{***}	0.006^{***}	-0.278^{***}
	(0.001)	(0.038)	(0.001)	(0.038)
persons	-0.002^{***}	-0.111^{***}	-0.002^{***}	-0.111^{***}
	(0.000)	(0.014)	(0.000)	(0.014)

Significance: $*** \equiv p < 0.001$; $** \equiv p < 0.01$; $* \equiv p < 0.05$

Table 2: Replication of **crch** results using **htobit**.

4. Replication

To assess the reliability of the `htobit()` implementation, it is benchmarked against the `crch()` function of (Messner *et al.* 2016), using the same restricted model as above.

```
R> library("crch")
R> ca3 <- crch(alcohol ~ age + log(expenditure) + persons | age +
+    log(expenditure) + persons, data = AlcoholTobacco, left = 0)
```

Using a model table from **memisc** (Elff 2019) it can be easily seen the results can be replicated using both packages (see Table 2).

```
R> library("memisc")
R> mtable("htobit" = ma3, "crch" = ca3)
```

References

Elff M (2019). *memisc: Tools for Management of Survey Data and the Presentation of Analysis Results*. R package version 0.99.17.2, URL `https://CRAN.R-project.org/package=memisc`.

Fox J, Weisberg S (2019). *An R Companion to Applied Regression*. 3rd edition. Sage, Thousand Oaks. URL `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`.

Messner JW, Mayr GJ, Zeileis A (2016). "Heteroscedastic Censored and Truncated Regression with crch." *The R Journal*, **8**(1), 173–181. `doi:10.32614/RJ-2016-012`.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Verbeek M (2004). *A Guide to Modern Econometrics*. 2nd edition. John Wiley & Sons, Chichester.

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.

Affiliation:

Achim Zeileis
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@R-project.org
URL: https://statmath.wu.ac.at/~zeileis/