Mixed Frequency Data Sampling Regression Models: the R Package midasr

Virmantas Kvedaras
Vilnius University

Vaidotas Zemlys
Vilnius University

Abstract

The implementation of MIDAS approach in the R package **midasr** is described.

Keywords: MIDAS, specification test.

1. Introduction

In econometric applications it is common to encounter time series which are sampled in different frequencies, e.g. quarterly vs yearly, etc. In order to use differently sampled series in regression analysis one of the series is usually aggregated. It is evident that some of the information is lost during such transformation. One of the solutions to this problem is the mixed data sampling (MIDAS) approach introduced in **?**. It has gained popularity in financial and some macroeconomic applications (see e.g. **?**, and **?**, for a recent overview of various contributions). In the most cases, it is used for the forecasting purposes.

The main idea of MIDAS approach is based on observation that aggregation of high frequency time series is actualy a specific embedding of the high frequency domain to the low frequency domain. Say we have yearly series Y_t and quaterly series x_τ and we want to estimate the model

$$Y_t = f(x_\tau) + \varepsilon_t$$

The usual approach is to aggregate x_τ to a yearly sampling frequency:

$$X_t = \frac{1}{4}(x_{4t} + x_{4t-1} + x_{4t-2} + x_{4t-3}),$$

where we assume that yearly time series are observed at the same time as the fourth quarter of the quarterly time series. Then we can rewrite the model in the following form:

$$Y_t = \alpha + \beta X_t + \varepsilon_t$$

If we substitute the aggregation equation we see that this model is a restricted form of the more general model:

$$Y_t = \alpha + w_1 x_{4t} + w_2 x_{4t-1} + w_3 x_{x4t-2} + w_4 x_{4t-3} + \varepsilon_t$$

We can extend this model for general frequency ratio m and including more lags resulting in so called U-MIDAS model:

$$Y_t = \alpha + \sum_{h=0}^{k} w_h x_{tm-h} + \varepsilon_t.$$

If frequency ratio m is high, such model might not be feasible to estimate, due to lack of degrees of freedom. To solve this problem MIDAS approach suggests restricting the weights:

$$w_h = g(h, \lambda), \quad , h = 0, 1, ..., k$$

where g is some function and λ is a vector of hyper-parameters. In MIDAS literature function g is usually chosen from a fixed set of functions. The important question is then whether this chosen function is the correct one. Recently Kvedaras and Zemlys **?** proposed a test which lets to test hypothesis whether the chosen weight function is appropriate.

Package **midasr** is aimed at applied researcher. It allows to estimate the MIDAS regression model and test its feasibility.

2. Theory

2.1. Simple MIDAS model

Consider a situation where we observe processes $y = \{y_t \in \mathbb{R},\ t = 0, \pm1, \pm2 \dots\}$ and $\boldsymbol{z} = \{\boldsymbol{z}_t \in \mathbb{R}^\ell,\ \ell \in \mathbb{N},\ t = 0, \pm1, \pm2 \dots\}$ are observed at a low frequency, whereas $x = \{x_\tau \in \mathbb{R},\ \tau = 0, \pm1, \pm2 \dots\}$ is observed at a higher frequency with m high-frequency observations available per each single low frequency period t.

The MIDAS regression of y on $\boldsymbol{z}$ and x, assuming that x makes an influence up to k high-frequency lags, has the following representation

$$y_t = \sum_{h=0}^{k} w_h x_{tm-h} + \boldsymbol{z}_t' \gamma + \varepsilon_t, \tag{1}$$

where the coefficients w_h are usually constrained by a functional constraint:

$$w_h = g(\boldsymbol{\lambda}, h), h = 0, ..., k \tag{2}$$

where $\boldsymbol{\lambda}$ is a vector of hyper-parameters.

2.2. Estimation

For the unrestricted model the coefficients β_j and γ can be estimated using OLS. For the restricted model, where weights depend on hyper-parameters we estimate the model parameters by non-linear least squares:

$$(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\lambda}}) = \operatorname{argmin} \sum_{t=[k/m]+1}^{n} \left(y_t - \sum_{h=0}^{k} g(\boldsymbol{\lambda}, h) x_{tm-h} - \boldsymbol{z}_t' \gamma \right)^2 . \tag{3}$$

Suppose that ε_t are i.i.d and independent of x_t and $\boldsymbol{z}_t$. Suppose next that x and $\boldsymbol{z}$ are stationary and ergodic or deterministic with usual regularity conditions. Suppose that g is twice-differentiable continuous function with respect to $\boldsymbol{\lambda}$. Then the estimates are consistent and assymptotically normal.

3. Testing

It is of interest to test whether the restriction (2) holds. Kvedaras and Zemlys **?** have developed the test based on the difference between the estimates of weights of restricted model and the estimates of weights of unrestricted model. Denote by $\hat{\boldsymbol{\theta}}' = (\hat{\boldsymbol{w}}', \hat{\boldsymbol{\gamma}}')$ the OLS estimates of the model (1) and by $\hat{\boldsymbol{f}}' = (g(\hat{\boldsymbol{\lambda}}, 0), ..., g(\hat{\boldsymbol{\lambda}}, k), \hat{\boldsymbol{\gamma}}')$. Then under null hypothesis that $w_h = g(\boldsymbol{\lambda})$:

$$(\hat{\boldsymbol{\theta}}' - \hat{\boldsymbol{f}})\boldsymbol{A}(\hat{\boldsymbol{\theta}} - \boldsymbol{f}) \sim \chi^2(k - r),$$

where $\boldsymbol{A}$ is suitable normalisation matrix and r is the dimension of the vector $\boldsymbol{\lambda}$.

4. Implementation in midasr package

Since estimation of MIDAS model is a NLS problem, it is possible to estimate it using existing R functions, such as `nls` for example. There are however several challenges related to using existing R code:

- The data in the model cannot be in one `data.frame`, since the time series of different frequency are of different lengths.
- MIDAS model requires special transformation of the high frequency time series:

$$\begin{bmatrix} x_1 \\ \vdots \\ x_{Nm} \end{bmatrix} \rightarrow \begin{bmatrix} x_{lm} & x_{lm-1} & \cdots & x_{lm-k} \\ x_{(l+1)m} & x_{(l+1m)-1} & \cdots & x_{(l+1)m-k} \\ \cdots & \cdots & \cdots & \cdots \\ x_{Nm} & x_{Nm-1} & \cdots & x_{Nm-k} \end{bmatrix} \tag{4}$$

  where N is the low-frequency sample size and k is the high-frequency lag. There is no existing R function which performs such operation.

The package **midasr** gives a solution to these challenges. It lets applied researcher to specify the MIDAS model using familiar `formula` interface making the estimation of MIDAS models similar to other existing regression models.

4.1. Embedding of high-frequency time series

High frequency time series are embedded to low-frequency with `fmls` function. This function takes as an arguments the high-frequency time series, the number of high-frequency lags and the frequency ratio and performs the transformation (4):

```
> library(midasr)
> x <- 1:16
> fmls(x,3,4)
```

```
     X.0/m X.1/m X.2/m X.3/m
[1,]     4     3     2     1
[2,]     8     7     6     5
[3,]    12    11    10     9
[4,]    16    15    14    13
```

It is assumed that for each low-frequency observation there are exactly m high-frequency observations. This means that the frequency ratio must divide the length of the time series passed to `fmls`. The result of this division is the number of rows of the resulting matrix.

If $m = 1$ this function behaves exactly as function `embed`, with exception that it padds the resulting matrix with `NA`'s

```
> x <- 1:4
> fmls(x,2,1)
```

```
     X.0/m X.1/m X.2/m
[1,]    NA    NA    NA
[2,]    NA    NA    NA
[3,]     3     2     1
[4,]     4     3     2
```

```
> embed(x,3)
```

```
     [,1] [,2] [,3]
[1,]    3    2    1
[2,]    4    3    2
```

4.2. Estimation of unrestricted MIDAS model

Recall that unrestricted MIDAS model can be estimated with OLS. Function `fmls` enables to use existing R function `lm` for estimating linear models. For illustration purposes we use data from the Okun's law example analyzed in **?** and **?**. MIDAS model is then

$$\Delta \log Y_t = \alpha_0 + \alpha_1 t + \sum_{h=0}^{k} w_h \Delta U_{tm-h} \varepsilon_t,$$

where Y and U denote real GDP and unemployment rate of US, and Δ is the difference operator.

```
> data("USrealgdp")
> data("USunempr")
> y <- diff(log(USrealgdp))
> x <- window(diff(USunempr), start = 1949)
> trend <- 1:length(y)
> lm(y~trend+fmls(x,11,12))
```

```
Call:
lm(formula = y ~ trend + fmls(x, 11, 12))

Coefficients:
          (Intercept)                  trend   fmls(x, 11, 12)X.0/m
            0.0424249             -0.0003034              0.0018017
 fmls(x, 11, 12)X.1/m   fmls(x, 11, 12)X.2/m   fmls(x, 11, 12)X.3/m
           -0.0118845             -0.0037937             -0.0081820
 fmls(x, 11, 12)X.4/m   fmls(x, 11, 12)X.5/m   fmls(x, 11, 12)X.6/m
           -0.0095754             -0.0061851             -0.0231792
 fmls(x, 11, 12)X.7/m   fmls(x, 11, 12)X.8/m   fmls(x, 11, 12)X.9/m
           -0.0252373             -0.0303896             -0.0396104
fmls(x, 11, 12)X.10/m  fmls(x, 11, 12)X.11/m
           -0.0299635             -0.0192491
```

As we see there is no problem of passing different length time series to `lm`. R simply picks the data from the environment it is called from. Sometimes it is necessary to specify data directly. This is possible by using `midas_u` function:

```
> ldata <- data.frame(y=y,trend=trend)
> hdata <- data.frame(x=x)
> midas_u(y~trend+fmls(x,11,12),ldata,hdata)

Call:
lm(formula = y ~ trend + fmls(x, 11, 12), data = ee)

Coefficients:
          (Intercept)                  trend   fmls(x, 11, 12)X.0/m
            0.0424249             -0.0003034              0.0018017
 fmls(x, 11, 12)X.1/m   fmls(x, 11, 12)X.2/m   fmls(x, 11, 12)X.3/m
           -0.0118845             -0.0037937             -0.0081820
 fmls(x, 11, 12)X.4/m   fmls(x, 11, 12)X.5/m   fmls(x, 11, 12)X.6/m
           -0.0095754             -0.0061851             -0.0231792
 fmls(x, 11, 12)X.7/m   fmls(x, 11, 12)X.8/m   fmls(x, 11, 12)X.9/m
           -0.0252373             -0.0303896             -0.0396104
fmls(x, 11, 12)X.10/m  fmls(x, 11, 12)X.11/m
           -0.0299635             -0.0192491
```

As we see `lm` object is returned, so it is possible to do everything what can be done with usual `lm` object.

4.3. Estimation of restricted MIDAS model

The restricted model requires that weights are specified. For our example let us use normalized exponential Almon polynomial weights:

$$w_i = w_i(\beta, \theta_1, ..., \theta_P) = \beta \frac{\exp(\theta_1 i + ... + \theta_P i^P)}{\sum_{l=1}^{n} \exp(\theta_1 l + ... + \theta_P l^P)}$$

To use it in **midasr** we need to specify it in the following way:

```
> nealmon <- function(p,d) {
+   i <- (1:d)/100
+   plc <- poly(i,degree=length(p)-1,raw=TRUE) %*% p[-1]
+   as.vector(p[1] * exp(plc)/sum(exp(plc)))
+ }
```

The first argument must be a vector with all the hyper parameters and the second argument must be the number of the coefficients. Note the scaling of the polynomial, which is done purely for computational reasons.

Having specified the weight function we can proceed to estimating of the corresponding MIDAS model.

```
> midas_r(y~fmls(x,11,12,nealmon),start=list(x=c(0,0,0)))

MIDAS regression model
 model: y ~ fmls(x, 11, 12, nealmon)
(Intercept)          x1          x2          x3
    0.03243    -0.19137    15.62062     1.76868

Function was used for fitting
```

Here we specify that we use weight function `nealmon` for variable `x` and we supply starting values for the optimisation. The result is an object of class `midas_r`, which is a type of fitted model object.

References

Foroni C, Marcellino M, Schumacher C (2011). "U-MIDAS: MIDAS regressions with unrestricted lag polynomials." (2011,35). URL http://ideas.repec.org/p/zbw/bubdp1/201135.html.

Ghysels E, Valkanov PSCR (2004). "The MIDAS touch: Mixed data sampling regression models." URL http://http://rady.ucsd.edu/faculty/directory/valkanov/pub/docs/midas-touch.pdf.

Kvedaras V, Račkauskas A (2010). "Regression Models with Variables of Different Frequencies: The Case of a Fixed Frequency Ratio." *Oxford Bulletin of Economics and Statistics*, **72**(5), 600–620. ISSN 1468-0084. doi:10.1111/j.1468-0084.2010.00585.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0084.2010.00585.x/abstract.

Kvedaras V, Zemlys V (2012). "Testing the functional constraints on parameters in regressions with variables of different frequency." *Economics Letters*, **116**(2), 250 – 254. ISSN 0165-1765. doi:10.1016/j.econlet.2012.03.009. URL http://www.sciencedirect.com/science/article/pii/S0165176512000961.

Sinko A, Sockin M, Ghysels E (2012). "Matlab Toolbox for Mixed Sampling Frequency Data Analysis using MIDAS Regression Models." URL `http://www.unc.edu/%7Eghysels/papers/MIDAS_Usersguide_Version3.pdf`.

Affiliation:

Vaidotas Zemlys
Department of Econometric Analysis
Faculty of Mathematics and Informatics
Vilnius University
Naugarduko g. 24, Vilnius, Lithuania
E-mail:`Vaidotas.Zemlys@mif.vu.lt`
URL: `http://vzemlys.wordpress.com`