# How to estimate a static or dynamic linear regression model if the time series involved have missing values in between valid values

Herman J. Bierens
Pennsylvania State University

December 11, 2006

## 1 Introduction

In principle EasyReg only allows missing values at the beginning and the end of a time series. However, there is a way to estimate a linear regression model using time series with missing values in between valid values. I will explain that in first instance for the case of two time series, $y_t$ and $x_t$, with names $Y$ and $X$, respectively, contained in a data file in space delimited EasyReg text format (the former default format), i.e.,

$$\begin{array}{ll} 2 & -88888 \\ Y & \\ X & \\ y_1 & x_1 \\ \vdots & \vdots \\ y_T & x_T \end{array} \tag{1}$$

where 2 is the number of variables, $-88888$ is the missing value code, for example, separated by at least one space, $Y$ and $X$ are the variable names, and below the last variable name $X$ the data matrix, with $T$ the number of observations. For each observation $t$ the data entries $y_t$ and $x_t$ are separated by at least one space. It is recommended **not** to use the EasyReg default missing value code $(-99999.99)$.

In the last section I will explain the case where $x_t$ is a vector time series.

# 2 Static models

Suppose that you only want to estimate a static linear regression model: $y_t = \alpha + \beta y_t + u_t$, where $u_t$ is the error term. Then the solution is simple: Import the data as cross-section data, and run this regression, because for cross-section data missing values are allowed anywhere.

EasyReg automatically add the observation numbers to the data, as variable "Observation". Therefore, if want to estimate the model with a time trend: $y_t = \alpha + \beta x_t + \gamma.t + u_t$, use the variable "Observation" as $t$.

# 3 Dynamic models based on two time series

Now suppose that you want to estimate a dynamic regression model, for example,

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_1 x_t + \beta_2 x_{t-1} + u_t.$$

Then proceed as follows.
(1) Import the data file in Wordpad, and change the missing value code in the first record to zero:

$$\begin{array}{cc} 2 & 0 \\ Y & \\ X & \\ y_1 & x_1 \\ \vdots & \vdots \\ y_T & x_T \end{array}$$

However, do not change the missing value codes $-88888$ in the data matrix. In this way we fool EasyReg to believe that there are no missing values.
(2) Import the modified data file in EasyReg, as **time series** data.
(3) Open Menu → Input → Transform variables, click the $Dummy = I(x = a)$ button, enter $a = -88888$, and push the enter key. In the next window, select the variable $Y$ only, click "Selection OK", and in the following window click "O.K.". Then the dummy variable $I(Y = -88888)$ will be generated and added to the data. This variable takes the value 1 if $y_t = -88888$, and 0 if $y_t \neq -88888$. The function $I(.)$ is known as the indicator function: $I(true) = 1$, $I(false) = 0$.
(4) Repeat step (3) for variable $X$, which generates the dummy variable $I(X = -88888)$.

(5) Open Menu → Input → Transform variables, click "Time series transformations, and make the lagged variables $y_{t-1}$, $I(y_{t-1} = -88888)$, $x_{t-1}$, $I(x_{t-1} = -88888)$. The variable names involved are

$$\text{LAG1}\left[Y\right],\ \text{LAG1}\left[I(Y = -88888)\right],\ \text{LAG1}\left[X\right],\ \text{LAG1}\left[I(X = -88888)\right],$$

respectively. In general, for each lagged variable $Z$ $(= X \text{ or } Y)$ you have to lag the dummy variable $I(Z = -88888)$ in the same way.
(6) Open Menu → Input → Transform variables, click "Linear combination of variables", make the linear combination

$$\begin{aligned} I(Y &= -88888) + \text{LAG1}\left[I(Y = -88888)\right] + I(X = -88888) \\ &\quad + \text{LAG1}\left[I(X = -88888)\right] \end{aligned}$$

and give it a name, for example "Dummy Missing Value". In general you have to make this linear combination of dummy variables of the type $I(Z = -88888)$ and their lags for all variables $Z$ and their lags in your model. The dummy variable "Dummy Missing Value" is now equal to 1 if one of the variables in your model is a missing value $(= -88888)$, and 0 if not.
(7) Open Menu → Data analysis → Data table, select the variables $Y$, $\text{LAG1}\left[Y\right]$, $X$, $\text{LAG1}\left[X\right]$, and "Dummy Missing Value", and then select "View the data table as a CSV file via Excel". Then the selected variables will be written to file TMP.CSV in the EASYREG.DAT folder. If you have Excel installed on your computed, and have instructed EasyReg where EXCEL.EXE is located, the Excel file TMP.CSV will be opened, otherwise nothing happens.
(8) Import EASYREG.DAT\TMP.CSV in EasyReg, as **cross-section** data, adopt the default missing value code $(-99999.99)$ and overwrite the previous data, or start EasyReg in a new folder.
(9) Open Menu → Input → Transform variables, click "1 if x=a, missing value if not", enter $a = 0$, and push the enter key. In the next window select the variable "Dummy Missing Value", and create the new variable "1 if Dummy Missing Value=0, missing value if not".
(10) Open Menu → Input → Transform variables, click "Multiplicative combination of variables", select the variables $Y$ and "1 if Dummy Missing Value=0, missing value if not", and enter the powers 1 and 1. Then the new variable

$$Y_* = Y \times 1 \text{ if Dummy Missing Value=0, missing value if not}$$

is created. This variable $Y_*$ is for observation $t$ a missing value if one or more of the variables $Y$, LAG1$[Y]$, $X$, and/or LAG1$[X]$ have a missing value for observation $t$, and $Y_* = Y$ if not.
(11) Finally, regress $Y_*$ on LAG1$[Y]$, $X$ and LAG1$[X]$, with an intercept if $\alpha_0 \neq 0$, and you are done.

# 4 What to do if the data file is an Excel file in CSV format?

If your data file (1) is an Excel CSV file, and Windows uses a dot (.) as decimal delimiter, the CSV file will have the structure

$$\begin{array}{l} Y, X \\ y_1, x_1 \\ y_2, x_2 \\ \vdots \\ y_T, x_T \end{array}$$

where missing values in the data matrix are represented by empty spaces. For example, if $y_t$ is a missing value but $x_t$ is not, record $t+1$ in the CSV file looks like "$, x_t$", whereas if $x_t$ is a missing value but $y_t$ is not then record $t+1$ looks like "$y_t,$". If both are missing values then record $t+1$ is empty.

If Windows uses a comma (, ) as decimal delimiter, the CSV file takes the form

$$\begin{array}{l} Y; X \\ y_1; x_1 \\ y_2; x_2 \\ \vdots \\ y_T; x_T \end{array}$$

where the entries are now separated by a semi-colon (; ) rather than a comma.

Instead of steps (1) and (2),
(1) import the CSV file in Excel or Wordpad and fill the empty entries with missing value code $-88888$, and
(2) import the modified CSV file in EasyReg as time series data, with missing value code 0 or the default missing value code $-99999.99$.

Then continue with step (3).

# 5 General dynamic models

Suppose your data takes the form of a vector time series process $z_t = (y_t, x_t')'$, where $x_t = (x_{1,t}, ..., x_{k,t})' \in \mathbb{R}^k$, with variable names $Y$ and $X = (X_1, ..., X_k)'$, so that the variable names of $z_t$ are $Z = (Y, X_1, ..., X_k)' = (Z_0, Z_1, ..., Z_k)'$. Again, assume that your data file is in space delimited EasyReg text format:

$$\begin{array}{llll} k+1 & -88888 & & \\ Z_0 & & & \\ Z_1 & & & \\ \vdots & & & \\ Z_k & & & \\ z_{0,1} & z_{1,1} & \cdots & z_{k,1} \\ \vdots & \vdots & \ddots & \vdots \\ z_{0,T} & z_{1,t} & \cdots & z_{k,T} \end{array}$$

with missing value code $-88888$.

Suppose you want to estimate the dynamic regression model

$$y_t = \alpha_0 + \sum_{i=1}^{p_0} \alpha_i y_{t-i} + \sum_{j=1}^{k} \sum_{i=q_j}^{p_j} \beta_{i,j} x_{j,t-i} + u_t$$

where $0 \leq q_j \leq p_j$ for $j = 1, ..., k$, and $0 = q_0 \leq p_0$.
(1) Import the data file in Wordpad, and change the missing value code in the first record to zero:

$$\begin{array}{llll} k+1 & 0 & & \\ Z_0 & & & \\ Z_1 & & & \\ \vdots & & & \\ Z_k & & & \\ z_{0,1} & z_{1,1} & \cdots & z_{k,1} \\ \vdots & \vdots & \ddots & \vdots \\ z_{0,T} & z_{1,t} & \cdots & z_{k,T} \end{array}$$

(2) Import the modified data file in EasyReg, as **time series** data.
(3) For $j = 0, 1, ..., k$, create the dummy variables $I(Z_j = -88888)$.

(4) For $j = 0, 1, ..., k$, and $i = q_j, ..., p_j$, create the lagged variables

$$\text{LAG}i\left[Z_j\right],\ \text{LAG}i\left[I(Z_j = -88888)\right]$$

Note that if $i = 0$ then

$$\text{LAG}i\left[Z_j\right] = Z_j,\ \text{LAG}i\left[I(Z_j = -88888)\right] = I(Z_j = -88888)$$

which of course don't need to be created, because you already have them or done that.
(5) Make the dummy variable

$$\text{Dummy missing value } = \sum_{j=0}^{k} \sum_{i=q_j}^{p_j} \text{LAG}i\left[I(Z_j = -88888)\right]$$

where again $\text{LAG}i\left[I(Z_j = -88888)\right] = I(Z_j = -88888)$ if $i = 0$.
(6) Write the variables $Z_j$, $\text{LAG}i\left[Z_j\right]$, $\text{LAG}i\left[I(Z_j = -88888)\right]$, $j = 0, 1, ..., k$, $i = q_j, ..., p_j$, and "Dummy missing value" to file EASYREG.DAT\TMP.CSV.
(7) Import the Excel file EASYREG.DAT\TMP.CSV in EasyReg, as **cross section data**, adopt the default missing value code ($-99999.99$) and overwrite the previous data, or start EasyReg in a new folder.
(8) Make the variable "1 if Dummy Missing Value=0, missing value if not".
(9) Make the variable $Y_* = Y \times 1$ if Dummy Missing Value=0, missing value if not.
(10) Regress $Y_*$ on $\text{LAG}i\left[Z_0\right]$ for $i = 1, ..., p_0$ and $\text{LAG}i\left[Z_j\right]$ for $j = 1, ...., k$ and $i = q_j, ..., p_j$. Include an intercept if $\alpha_0 \neq 0$.