Estimating Dynamic Panel Models When There are Gaps in the Dependent Variable

Kåre Vernby†
Karl-Oskar Lindgren‡

December 3, 2009

Abstract

This short note explains and improves on Iversen and Soskice’s (2006) method for estimating dynamic panel data models when there are gaps in the dependent variable but complete or nearly complete data on the independent variables. It also describes a Stata program, dvgreg, developed by the authors for conveniently estimating such models.

*We wish to thank Torben Iversen, Noam Lupu and Åsa Vernby for comments on the note and the accompanying code and the Jan Wallander and Tom Hedelius Foundation for financial support. Torben Iversen and David Soskice graciously provided the data that are distributed together with this note.

†Researcher, Uppsala University. Email: kare.vernby@statsvet.uu.se
‡Researcher, Uppsala University. Email: karl-oskar.lindgren@statsvet.uu.se
Dynamic panel models are one of the most common methods of conducting quantitative comparative analysis in political science. However, while researchers may feel that a phenomenon of interest should be modeled dynamically, the data collected often fall short of meeting the necessary requirements. Arguably, numerous and unevenly spaced gaps on the dependent variable constitutes one of the most serious obstacles to conducting dynamic panel analysis. This short note deals with the problem of estimating dynamic panel models in this context.

In comparative politics, the canonical example of a data set that fails to meet the requirements of dynamic panel analysis is the Luxembourg Income Study (LIS). LIS publishes high-quality data on the wage and income distributions of the advanced industrial countries. In Table 1, the structure of the LIS-data used in Bradley et al. (2003) and Iversen and Soskice (2006) are described. As can be seen data for a country is typically collected each fourth to sixth year. However, there is a great deal of variation in the length of the gaps, both between and within countries.

The type of data structure described in Table 1 poses a serious challenge to researchers who wish to explain movements in highly persistent series. Confronted with this type of scenario, most sidestep the issue of dynamics. For instance, in their important paper on the determinants of distribution and redistribution in advanced industrial democracies, Bradley et al. (2003), do not use a dynamic model. By contrast, Iversen and Soskice (2006) devise a method for dealing with the problem.

In this note we describe Iversen and Soskice’s (2006) approach to estimating dynamic panel data models when there are gaps in the dependent variable but complete or nearly complete data on the independent variables. We also improve on their approach, introducing a solution to the problem of heteroskedasticity that is induced by their particular method. Finally, the note also describes a Stata program, dvgreg, written by the authors, which can be used for conveniently estimating such models in Stata. To illustrate the use of the Stata program, we reanalyze the data used by Iversen and Soskice in their 2006 article, using the method described in this note. For the purpose of allowing the reader to familiarize herself with how the program works, the file apsr06long.dta, which contains Iversen and Soskice’s original data in long format, is also distributed with this note.
Table 1: LIS-waves by country as used in Bradley et al. (2003) and Iversen and Soskice (2006)

<table>
<thead>
<tr>
<th>Country</th>
<th>Years of measurement</th>
<th>Length of Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Min</td>
</tr>
<tr>
<td>Belgium</td>
<td>1985, 1988, 1992</td>
<td>3</td>
</tr>
<tr>
<td>Denmark</td>
<td>1987, 1992</td>
<td>5</td>
</tr>
<tr>
<td>Switzerland</td>
<td>1982, 1992</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Background

Assume that we are interested in the outcome $y$, whose long run equilibrium level is

$$y_{i,t}^* = \alpha + \beta' X_{i,t}$$

where $i$ indexes the panels (or cross-sectional units) and $t$ indexes time. That is, the equilibrium level of $y$ is determined by an intercept, $\alpha$, and a number of other factors, $X_{i,t}$, that vary both within panels and across time.

However, the level of $y$ in a particular country at a particular time, $y_{it}$ is a function of $y_{it-1}$, plus an adjustment towards the equilibrium level $y_{i,t}^*$ such that

$$y_{i,t} = y_{i,t-1} + \rho(y_{i,t}^* - y_{i,t-1}) + \varepsilon_{i,t}$$

where $\rho$ lies in the (0, 1) interval and measures the rate at which actual level of $y$ adjusts towards the equilibrium level, $y_{i,t}^*$ and $\varepsilon_{i,t}$ is the i.i.d. error term.

When there are few or no gaps in the dependent variable one would simply substitute (1) into (2) and rearrange to obtain

$$y_{i,t} = (1 - \rho)y_{i,t-1} + \rho(\alpha + \beta' X_{i,t}) + \varepsilon_{i,t}$$

which can be estimated directly. However, if the the dependent variable contains gaps, such as when it is not collected on an annual basis, this is not possible. Repeated substitution of (3) and rearranging yields

$$y_{i,t} - (1 - \rho)^T y_{i,t-T} = \rho \sum_{s=0}^{T-1} (1 - \rho)^s \alpha
+ \rho \sum_{s=0}^{T-1} (1 - \rho)^s \beta' X_{i,t-s} + \varepsilon_{i,t,T}$$

where $T$ stands for the number of time periods (e.g. years) one has to go back to find the next observation of the dependent variable. This repeated substitution results in an equation containing a composite error term

$$\varepsilon_{i,t,T} = \sum_{s=0}^{T-1} (1 - \rho)^s \varepsilon_{i,t-s}$$

Equation (4) is identical to the estimating equation in Iversen and Soskice (2006) and can be used to obtain estimates of the parameters of interest: $\alpha$ and $\beta'$. However, note that the ‘independent variables’ in this equation are really weighted sums of the independent variables. Therefore, one will need complete or nearly complete time-series data on the independent variables.
To exemplify, if one has observations on \( y_{i,t} \) and \( y_{i,t-3} \), one needs data on \( X_{i,t} \), \( X_{i,t-1} \) and \( X_{i,t-2} \). Second, although there exist methods for directly estimating models that are non-linear in the parameters, such as the one in (4), they are difficult to apply in most contexts since \( T \) will typically vary across observations.\(^1\) This is indeed the case with the LIS-data used in Iversen and Soskice’s (2006) study. Instead, they proceed by iteratively estimating (4) using different values of \( \rho \) and choosing the specification with the best model fit. This will be the approach taken here as well.

The final problem associated with estimating (4) is that the error term is heteroskedastic. Specifically, the variance of the composite error term depends on the number years between the current and the previous observation on the dependent variable in the following fashion

\[
\sigma^2_{\bar{\varepsilon}_t} = \left[ \sum_{s=0}^{T-1} (1 - \rho)^s \sigma^2 \right]^2 \quad (6)
\]

where \( \sigma \) is the standard deviation of the original i.i.d. error terms, \( \varepsilon_{i,t} \).

Iversen and Soskice (2006) attempt to deal with this by correcting the standard errors for panel heteroskedasticity of an unknown form. However, since \( T \) not only varies within, but also between, panels this does not solve the problem. In fact, as (6) shows, the variance is constant within, but varies between, each group of observations defined by \( T \). Another alternative would be to use something like robust standard errors clustered on \( T \). Although it is preferable to the approach used by Iversen and Soskice (2006), it is unnecessary to apply a standard error fix such as this in the present context. These fixes yield statistically inefficient estimates, and when the form of heteroskedasticity is, in fact, known one can obtain efficient estimates by the appropriate use of weighted least squares (see, e.g., Cameron and Trivedi 2005).

In the present context, this is accomplished by multiplying (4) through by the weight \( \tilde{\rho}_T = 1/\sum_{s=0}^{T-1} (1 - \rho)^s \), which yields the following estimating equation:

\[
\tilde{\rho}_T [y_{i,t} - (1 - \rho)^T y_{i,t-T}] = \rho \alpha + \tilde{\rho}_T \rho \left[ \sum_{s=0}^{T-1} (1 - \rho)^s \beta' X_{i,t-s} \right] + \tilde{\rho}_T \bar{\varepsilon}_{i,t,T} \quad (7)
\]

\(^1\)If \( T \) is the same across all observations, non-linear least squares could plausibly be used. Another alternative, if \( T \) varies across observations, would be to use multi-level modeling, which is becoming increasingly popular in political science (see, e.g., Gelman and Hill 2007). Note, however, that this would not be practically feasible in most applications, since the researcher would have to allow all coefficients vary across the groups defined by \( T \), which would place great strains on the data.
where the new composite error term, $\tilde{\sigma}_T \tilde{z}_{i,t,T}$, has variance

$$\sigma^2 \tilde{\sigma}_T \tilde{z}_T = \sigma^2$$

and is therefore no longer heteroskedastic.

**Estimation in Stata**

Creating the data set required to estimate the model described in (7) is very cumbersome. And our solution to the problem of heteroskedasticity adds yet another step to this process. To this is added the fact that once the researcher manages to set up the data, she still has to go through the iterative process of estimating a new model for each value of $\rho$ she wishes to try out. Finally, model selection involves eyeballing the results for each value of $\rho$ that the researcher has tried to find the model that best fits the data. All these problems are worsened as soon as the researcher wants $\rho$ to be estimated with some degree of precision. For instance, faced with the practical problems associated with ‘manually’ setting up the data required for estimating (4), Iversen and Soskice (2006) only considered values of $\rho$ between .1 and .9 in increments of .1.

In order to eliminate hurdles to estimating (7) and obtaining more precise parameter estimates, we have created a Stata program that automates this procedure. In the following we give a short description of its use.

**The Data**

Note that the data set has to be set up in long format, with each line representing a cross-sectional unit at a certain point in time, for the following code to be applicable. The data used in the following examples are stored in the file *apsr06long.dta*, which accompanies this note. As was described in Table 1, these contain numerous gaps on the dependent variable. However, as necessitated by the method, the vector of independent variables, $X_{i,t}$, contains no missing values.²

²In Iversen and Soskice’s (2006) data, the voter turnout variable was missing for two country-years and the earnings inequality variable—which measures the ratio of earnings at the 90th decile to those at the 50th—was missing for more country-years. They imputed these data. The rule was that if the country-year for which the variable was missing was prior to the first observation on the same, the value for the first observation was used. If, on the other hand, the country-year for which the variable was missing was subsequent to the last observation on the same, the value for the last observation was used. For missing data internal to the time-series of a country, linear interpolation was used.
The Program

The command `dvgreg` along with its help file can be installed in Stata by downloading the files `dvgreg.ado` and `dvgreg.hlp` and locating them in Stata’s (PLUS) ado directory. To see where this directory is located on your computer, launch Stata and type `sysdir`. When you have saved the two files to the correct location and (re)started Stata you can access the help file by typing `help dvgreg` in the command window.

To illustrate the use of `dvgreg` we are now going to replicate parts of the analysis in Iversen and Soskice (2006). We begin by replicating the analysis in column 1 of Table 5 of Iversen and Soskice (2006), but this time using the estimating equation in (7) rather than that the original one in (4). To do this enter

```
dvgreg ginich p9050 fempar rgdppc unempl, id(ccode) time(year)
```

where the `id()` and `time()` are required arguments and should contain the the panel and time variables in numeric format.

After this is done, Stata reports the following output:

```plaintext
Model bestfit (rho=.8)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 47</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1285.11687</td>
<td>4</td>
<td>321.279216</td>
<td>F( 4, 42) = 15.75</td>
</tr>
<tr>
<td>Residual</td>
<td>856.575067</td>
<td>42</td>
<td>20.3946444</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>2141.69193</td>
<td>46</td>
<td>46.5585203</td>
<td>R-squared = 0.6000</td>
</tr>
</tbody>
</table>

ginich | Coef. Std. Err. t P>|t| [95% Conf. Interval]
--------|-------------------|---------|-------|------------------|
p9050 | -15.58112 5.820448 -2.68 0.011 -27.32725 -3.834977 |
fempar | .6664333 .1158587 5.75 0.000 .432621 .9002455 |
rgdppc | -.0015052 .0005625 -2.68 0.011 -.0026403 -.00037 |
unempl | .9599858 .2860611 3.36 0.002 .382691 1.537281 |
_cons | 25.65244 10.8219 2.37 0.022 3.812972 47.49191 |
```

Although the size of the coefficients and their standard errors differ from those found in Iversen and Soskice (2006), the substantive conclusions that can be drawn are roughly the same. The fact that our standard errors differ is not surprising given the difference in our approaches to correcting heteroskedasticity. The discrepancies between our coefficients and theirs can not, however, be explained by this, since the weighting procedure only affects the standard errors. As it turns out, Iversen and Soskice do not really estimate equation (4). Looking at the equation, it can be seen that the
constant term is multiplied by $\sum_{t=0}^{T-1}(1 - \rho)^s$. In other words, this is a variable that should be included in the regression. Iversen and Soskice do not, with model misspecification and bias as a result. Therefore, our coefficient estimates differ. Most notably, our estimated $\rho$ is .8 rather than .4.

To illustrate how to obtain more precise estimates of $\rho$, we replicate column 2 of Table 5 in Iversen and Soskice (2006). The default setting for `dvgreg` is to consider values of $\rho$ between .1 and .9 in increments of .1. The following replicates the analysis allowing greater precision:

```
.dvgreg ginich p9050 fempar rgdppc unempl cogcom union turnout pr veto, time(year) id(ccode)
> rhomin(.001) rhostep(.001) rhomax(.999)
```

```
Model bestfit (rho=.982)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 47</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2365.86402</td>
<td>9</td>
<td>262.873781</td>
<td>F( 9, 37) = 11.36</td>
</tr>
<tr>
<td>Residual</td>
<td>855.911896</td>
<td>37</td>
<td>23.132754</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>3221.77592</td>
<td>46</td>
<td>70.038607</td>
<td>R-squared = 0.7343</td>
</tr>
</tbody>
</table>

ginich | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval]
---|-------|-----------|---|-----|------------------|
| p9050 | 13.66768 | 10.25918 | 1.33 | 0.191 | -7.119385 | 34.45476 |
| fempar | .4104825 | .2264687 | 1.81 | 0.078 | -.0483866 | .8693516 |
| rgdppc | -.0007603 | .0007129 | -1.07 | 0.293 | -.0022047 | .0006841 |
| unempl | 1.09132 | .3121301 | 3.50 | 0.001 | .458884 | 1.723755 |
| cogcom | -.988968 | 3.460057 | -2.86 | 0.007 | -.16.90031 | -2.878827 |
| union | .1458925 | .0967366 | 1.51 | 0.140 | -.0501145 | .3418995 |
| turnout | .0512305 | .1187805 | 0.43 | 0.669 | -.1894417 | .2919028 |
| pr | 5.1822 | 2.435586 | 2.13 | 0.040 | .2472344 | 10.11717 |
| veto | -1.211942 | .6884557 | -1.76 | 0.087 | -2.606886 | .1830017 |
| _cons | -27.83307 | 22.39269 | -1.24 | 0.222 | -73.20498 | 17.53883 |
```

Again the results differ from those of Iversen and Soskice (2006), although the substantive conclusions remain similar. The main difference is that unionization, which is significant in their paper, is non-significant here.\(^3\) Note that if the user wants even more precise estimates of $\rho$ she will have to proceed iteratively since `dvgreg`, in its current incarnation, cannot handle searching through more than approximately 1500 steps of $\rho$. In the present example this would imply setting `rhomin(.981)`, `rhomax(.983)`, and then, e.g., `rhostep(.0001)`.

\(^3\)The size of the coefficient for cabinet center of gravity is much higher here, but this is due to the fact that Iversen and Soskice have standardized the variable.
Discussion

This short note has explained and improved on Iversen and Soskice’s (2006) method for estimating dynamic panel data models when there are gaps in the dependent variable but complete or nearly complete data on the independent variables. In addition, it introduces the program \texttt{dvgreg}, which was written for Stata by the authors, for conveniently implementing the suggested model. Although the method and code described in this note was developed in the context of using LIS data, its applicability ranges across many of political scientists’ favorite topics.

For instance, the researcher interested in institutional change frequently encounters data with numerous and unevenly spaced gaps on her dependent variable. One example is Rueda (2005), who wishes to explain variations in employment protection legislation using a dynamic model. However, because there are long and unevenly spaced gaps on what he considers to be the conceptually preferable measure, he resorts to using a suboptimal indicator.

Another potential area of applicability is the comparative study of mass attitudes. One such example is Inglehart and Welzel’s (2005) already much cited study of the impact of economic modernization, via culture, on democratisation. As part of this endeavour, they attempt to dynamically model movements in their index of self-expression values, which is aggregated by country, but fail to take into account the fact that it, being based on the World Values Surveys, is measured on an irregular basis.

A number of questions do, however, remain. As for the method’s properties, preliminary Monte Carlo studies conducted by the authors suggest that the estimates of the coefficients and standard errors are accurate as long as $\rho$ and $R^2$ are reasonably high, and the gaps are not too long. However, an extensive and systematic evaluation of the method’s properties has yet to be conducted. In the context of such a study, it would also be interesting to compare it to other methods. In particular, since we, when doing actual empirical research, never know the true data generating process, it would be of great practical use to know how this method compares to that which assumes no dynamics, both in situations when the data generating process exhibits dynamics, and when there are little or no dynamics. A similar comparison of the model presented here to the techniques of multiple imputation popularized in political science by King et al. (2001) would also be interesting.
References


