RESEARCH

Introduction to the profile areas of data sciences: project 7

Eva Aßmann* and Paul Vogler

*Correspondence:
eva.assmann@fu-berlin.de
Institute of Mathematics and Informatics, Berlin, Takustr. 9, Berlin, Germany
Full list of author information is available at the end of the article

Abstract

Goal of the project: Train an compare multiple regression models with different assumptions and terms in order to estimate panel data. Decide which model is more appropriate for the structure of the used data set.

Main results of the project:

Personal key learnings: We learned about the different regression methods to fit panel data, how to evaluate them and decide which method works best for the data of interest.

Estimated working hours: 9

Project evaluation: 4

Number of words: 1549 Words

Suggestions for improvements of the project:

Keywords: sample; article; author

Background and goal of the project

In social sciences, the purpose of a survey research is to collect information from a sample of individuals and measure certain attributes or behaviours. In order to understand a population as a whole, conclusions are made based on a representative sample. Panel studies combine cross-section and time series data by collecting information for the same group(s) of observed individuals at various time points. Because panel data is pooled over time and space, it can take account of individual-specific heterogeneity, gives more data variation, less collinearity and more degrees of freedom. Panel data thus enables the study of more complex behavioural models that measure effects and capture dynamic change better than based on cross-sectional or time series-data. This project's goal was to analyse panel data from a social science study using different regression models. The fitted models were evaluated in order to determine which one estimated the panel data best.

1 Description of the data

The panel data from Dahlberg's and Johansson's *"Examination of the dynamic behaviour of local governments using GMM bootstrapping"* was used in this project. They investigated the dynamic relationship between local government revenues and expenditures on a panel of 265 Swedish municipalities over the period 1979 to 1987 [1]. The research focus is an interesting aspect of political and economic science and thus lies within the scope of social sciences. The data was gathered by Statistics Sweden and obtained from Financial Accounts for the Municipalities. The panel data

set contained information on total expenditures, own-source revenues and received intergovernmental grants for the municipalities (see Tab. 1). The total expenditures contained both capital and current expenditures. All values were expressed in million SEK. The series was in per capita form and was deflated using a municipality-specific price index obtained by dividing total local consumption expenditures at current prices by total local consumption expenditures at fixed (1985) prices.

Attribute	**Description**	range
ID	ID number for municipality	114,..,2584 (not consistently increasing)
YEAR	Year (additionally included as categorical column 'year')	1979-1987
EXPEND	Total expenditures	0.012226-0.033883
REVENUE	Total own-source revenues	0.006228-0.029142
GRANTS	Intergovernmental grants received by a municipality	0.001571-0.012589

Table 1: Panel data set on swedish municipalities.

1.1 Preprocessing the panel data

The data was then reformatted into a panel format holding ID and YEAR as Multi-Index and REVENUE, GRANTS and EXPEND as columns. The YEAR values were added as a categorical column in order to facilitate the creation of dummy values in following analysis steps. Pairwise correlation was analysed for all feature columns. In case of highly correlated feature variables, feature selection normally is applied, since high feature correlation can decrease a model's accuracy. The GRANTS variable showed weak correlation with the EXPEND variable, while REVENUE showed to be strongly correlated with EXPEND. Since EXPEND was the target variable (dependent variable) in the following analysis steps, these correlation results did not require any feature selection. For the independent variables REVENUE and GRANTS no significant correlation was observed with each other or the YEAR values.

	EXPEND	GRANTS	REVENUE
EXPEND	1.000000	0.348488	0.875631
GRANTS	0.348488	1.000000	-0.054213
REVENUE	0.875631	-0.054213	1.000000

Figure 1: Pairwise correlation of panel data variables.

2 Results

Basic pooled OLS regression

The expenditures were modeled by the independent variables REVENUE and GRANTS and the time dummy variable 'year' using the linearmodels library. In the first analysis step, the data was estimated using a basic pooled OLS regression (PooledOLS). The fitted PooledOLS model was summarised with an P-value for the F-statistic (FP-value) of 0.0 and an overall R-squared value of 0.9262. PooledOLS does not take account of the variation across municipalities, thus further regression models based on different assumption were explored.

Estimating models with uncorrelated effects

Error component models add additional error terms to the basic pooled OLS regression. A random effects model (RE) assumes that the additional terms across the panel data are uncorrelated with the regressors. A RE model was trained with the same input as the PooledOLS model and returned the same values for FP-value and overall R-squared value. The variance decomposition showed that no variance within the expenditures was explained by the added random effects.

The between estimator (BE) is an alternative that first computes the time averages of the dependent and independent variables and then runs a simple regression using these averages. The year dummies were dropped since the averaging removes differences due to the year. A BE model was trained on all independent variables. The trained model showed an FP-value of 0.0 and an overall R-squared value of 0.9212.

When comparing all models with uncorrelated effects, both models fitted the data significantly better than an intercept-only model, judged by the FP-values. The BE model explained slightly more variance between panels, while the RE model explained more variance within panels and overall, judged by the overall R-squared value (Fig. 2).

Model Comparison

	BE	RE
Dep. Variable	EXPEND	EXPEND
Estimator	BetweenOLS	RandomEffects
No. Observations	265	2385
Cov. Est.	Unadjusted	Unadjusted
R-squared	0.9925	0.9262
R-Squared (Within)	0.6870	0.7139
R-Squared (Between)	0.9925	0.9907
R-Squared (Overall)	0.9212	0.9262
F-statistic	1.731e+04	2977.4
P-value (F-stat)	0.0000	0.0000

Figure 2: Caption

Estimating models with correlated effects

When effects are correlated with the regressors, the RE and BE estimators are not consistent. There are two ways to fit a fixed effects model (FE).

First, a model was fitted by including a constant dummy for each municipality (entity). Since there were no time invariant independent variables in the data that, no data had to be removed and the model was trained on REVENUE, GRANTS and 'years'. Time-invariant variables could be included when using entity effects since, once demeaned, these would all be 0. The trained model yielded an F-test for poolability value of 1.8456, an FP-value of 0.0 and an overall R-squared value of 0.8845.

In the second approach, in addition to the fixed entity effects, time effect was added into the model. Therefore, the previously created time dummy was removed and the model was trained only on REVENUE and GRANTS. The estimation summary showed an F-test for poolability value of 1.8456, the FP-value 0.0 and an overall R-squared value of 0.8791. When comparing the models with time dummies and added time effects, it was observed that the only change occured in the test statistic for

poolability since now the "effects" include both municipality and time, whereas before only municipality were included.

As another alternative to using fixed effects, first differencing (FD) was applied. There were no time-invariant variables to remove. While in all previous training approaches, a constant was input in addition to the training variables, this was refrained for FD. This variable would soak up all time-trends in the data, and so interpretations of these variable would be challenging. The FD model was trained on REVENUE and GRANTS and yielded the FP-value 0.0 and an overall R-squared value of 0.9146.

When comparing all models with correlated effects, all FP-values showed that all of them significantly fitted the data better than an intercept-only model. While the model with time dummies and added entity effects explained the most data variance within and between the panels, the first differences model showed the highest overall R-squared score and thus explained the most variance on average (Fig. 3).

Model Comparison

	FirstDiff	TimeDummies	TimeEffects
Dep. Variable	EXPEND	EXPEND	EXPEND
Estimator	FirstDifferenceOLS	PanelOLS	PanelOLS
No. Observations	2120	2385	2385
Cov. Est.	Unadjusted	Unadjusted	Unadjusted
R-squared	0.5934	0.7427	0.5630
R-Squared (Within)	0.7162	0.7427	0.7196
R-Squared (Between)	0.9159	0.9276	0.9276
R-Squared (Overall)	0.9146	0.8845	0.8791
F-statistic	1545.6	609.14	1359.0
P-value (F-stat)	0.0000	0.0000	0.0000

Figure 3: Caption

Comparing fixed and random effects model

The FD, RE and the model with fixed entity effects and time dummies all significantly fitted the data better than an intercept-only model. The RE model showed the best overall and between-panel variance explanation score, while the time dummy model explained the most variance within panels (Fig. 4). In general, RE models are more appropriate to use for large data sets with little time points and random samples as cross-sectional groups. FE models are easier to compute and more appropriate for small data sets with a lot of time points.

Model Comparison

	FirstDiff	Pooled	RE	TimeDummies
Dep. Variable	EXPEND	EXPEND	EXPEND	EXPEND
Estimator	FirstDifferenceOLS	PooledOLS	RandomEffects	PanelOLS
No. Observations	2120	2385	2385	2385
Cov. Est.	Unadjusted	Unadjusted	Unadjusted	Unadjusted
R-squared	0.5934	0.9262	0.9262	0.7427
R-Squared (Within)	0.7162	0.7139	0.7139	0.7427
R-Squared (Between)	0.9159	0.9907	0.9907	0.9276
R-Squared (Overall)	0.9146	0.9262	0.9262	0.8845
F-statistic	1545.6	2977.4	2977.4	609.14
P-value (F-stat)	0.0000	0.0000	0.0000	0.0000

Figure 4: Caption

Perform Hausman test

The Hausman test is a test on the correlation of the explaining variables and the error term. The prerequisite for the random effects model to be efficient is that

there is (almost) no correlation. The test uses the parameter values for the fixed and random effects, as well as their covariance matrices. It assumes a underlying Chi^2 statistic and rejects the application of the random effects model at a P-value of over 0.05. For this dataset the P-value was 8.55, so the random effects model was rejected and the fixed effects model is preferable.

Covariance Options

In a last step, the options for different covariance settings was explored, using the robust method when not using fixed effects and a clustered covariance using either entity or time (or both) as clusters. Clustering on entity reduced the T-statistics across the board. This suggests there is important correlation in the residuals per entity. Clustering by both also decreases the t-stats which suggests that there is cross-sectional dependence in the data.
Other clusters can be given as direct input to the function. The data was clustered by occupation, which is probably not a reliable variable to cluster on since there are only 9 groups and the usual theory for clustered standard errors requires that the number of clusters is large.

3 Discussion

In contrast to previous projects, this week's project was very statistical and was quite tedious, in parts. On the other hand, it was easier to understand the content and context of the data, since even without any profile background, social sciences mostly are more intuitive than biological data.

References

[1] Matz Dahlberg & Eva Johansson, 2000. "An examination of the dynamic behaviour of local governments using GMM bootstrapping methods," Journal of Applied Econometrics, John Wiley & Sons, Ltd., vol. 15(4), pages 401-416.