Application of fixed effect regression method in some examples

م . علي لطيف عارف
جامعة ذي قار - كلية العلوم

الخلاصة :

عدة مجاميع من البيانات تكتب بشكل مشاهدات متكررة في عينة من المواضيع (موضوع واحد لكل صف) ولتسهيل عملية تحليل البيانات يصار إلى إعادة تركيب هذه البيانات وبعدة طرق منها (طريقة الخلط) . استخدمت طريقة (انحدار التأثيرات الثابت) ولحالتين يكون فيها تكرار المشاهدات مرة ثنائي وأخرى ثلاثي ولهذه الطريقة القدرة في السيطرة على استقرار مميزات الأفراد المقاسة أو غير المقاسة والتي غالبا ماتكون ضمن أصناف مثل العمر أو الجنس أو الزمن -- الخ . كما إن هناك طريقة بديلة تعطي نفس النتائج تدعى طريقة الانحراف المتوسط تم تطبيقها على نفس المثال . كما تناول البحث استخدام الأنموذج اللوجستي لميزة المتغير المعتمد الثنائي وكان التطبيق لبيانات موسمية لسنتين متتاليتين ليصبح المتغير المعتمد يمثل مشاهدتين لكل مفردة .تم تحليل البيانات باستخدام البرنامج الإحصائي الجاهز (SAS 9.1)

Abstract

Several groups of data are written in a form of repeated observations in a sample of subjects (one subject for each range) .To make easy data analysis process , it is to restructure these data in several methods (one of them is a mixing one) . Method use the fixed effects regression method for two cases of which repeated observations are binary for one time and trinary for another time. This method has the capability to control the stability of the individuals measured and unmeasured characteristics – which are written in the types of age , sex , time , etc . Also there is another substitute method that will give the same results; it is known as (mean deviation method) applied to the same example . This research , also , deals with logistic model for concerning the sign of the dependent binary variable . The data of our application are seasonal of two successive years so that the dependent variable can represent two observations for each individual The analysis of all these data is complemented by using the available statistical programmed (SAS 9.1) .

Introduction

In any empirical researcher knows that randomized experiments have major advantages over observational studies in making causal inferences. Randomized of subjects to different treatment conditions ensures that the treatment groups are identical with respect to all possible characteristics of the subjects , regardless of whether those characteristics can be measured or not , the treatment groups produced by randomization will be approximately equal with respect to such easily measured variables as sex , ethnicity, region of birth and age , and also approximately equal for more problematic variables like intelligence , aggressiveness ,creativity , parents child- rearing practices , and genetic makeup .But in non experimental research , the classic way to control the potentially confounding variables is to measure them and put them in some kinds of regression model such as logistic and Poisson to count data ,or propensity scores. While statistical control can certainly be a useful tactic , it has two things First no matter how many variables you control for someone can always criticize by suggesting that left out some crucial variable . the omission of covariate can be lead to severe bias in estimating the effects of the variables that are included , Second to statistically control for a variable , you have to measure it and explicitly include it in some kind of model ,the problem is that some variables are difficult to measure ,and if the measurement imperfect ,this can also lead to biased estimates .[1]

Repeated measure

Repeated measures are responsing outcomes measured on the same experimental unit. Usually, these measurements are made over a period of time, such as blood pressure measured once a week for a month. However, repeated measures can also refer to multiple measurements on an experimental unit, such as left and right eye visual acuity. The experimental units are most often called subjects. Time is called a within-subject effect because there are different measurements at different times on the same subject. Explanatory variables such as treatment or sex are called between-subject effects because their values change only from subject to subject; there is not a different value for them at different times for the same subject. Questions you may want to answer in a repeated measures analysis may be whether treatment influences the response, whether time, the repeated factor, influences the response, and whether there is a treatment by time interaction. The repeated measures for the same subject are correlated, and this correlation must be taken into account analysis, therefore, You need to specify a covariance structure of an individual subject ,the same covariance structure is used for all subjects. another possible covariance structure is called unstructured, which means that no particular structure is placed on the covariance and variances. this task uses the mixed models approach of analyzing repeated measures.[4]

Structure of data

The data for this task must be structured in a particular way. Each measurement must be contained as a separate row of the data table. In addition, you must be able to identify each subject with a single variable or an effect constructed of multiple variables. If you are familiar with the repeated measurements analysis of variance in the GLM procedure, you may be used to having your multiple repeated measurements in the same row, for example, time measurements named Time1, Time2, and Time 3. But for this task, if your data are not structured with a single repeated measurement per row, you must create a new data table with this structure. suppose we have the data in table 1 .[3]

Table 1: Multiple Measures Per Row

Subject	Sex	Year1	Year2	Year3
1	M	22.1	22.3	21.2
2	M	23.1	22.4	22.5
3	M	20.1	19.2	25.4
4	F	19.2	22.3	22.3
5	F	22.4	24.2	21.6
6	M	19.2	17.4	23.4

We can choose year1 , year2 and year3 as the stacking variables like ' Time' for the stacked column name, and let the default name ' Source' be the source column name
The new data showing in table 2

Table 2: One Measure Per Row

Subject	Sex	Time	Source	Subject	Sex	Time	Source
1	M	22.1	YEAR1	4	F	19.2	YEAR1
1	M	22.3	YEAR2	4	F	22.3	YEAR2
1	M	22.2	YEAR3	4	F	22.3	YEAR3
2	M	23.1	YEAR1	5	F	22.4	YEAR1
2	M	22.4	YEAR2	5	F	24.2	YEAR2
2	M	22.5	YEAR3	5	F	21.6	YEAR3
3	M	20.1	YEAR1	6	M	19.2	YEAR1
3	M	19.2	YEAR2	6	M	17.4	YEAR2
3	M	25.4	YEAR3	6	M	23.4	YEAR3

Fixed Effects Method

The term “fixed effects model” treats with un-observed differences between individuals data as a set of fixed parameters that can either be directly estimated, or partialed out of the estimating equations . Fixed effects regressions are very important because data often fall into categories such as industries, states and families, , you will normally want to control for characteristics of those categories that might affect the left hand side variable. Unfortunately, you can never be certain that you have all the relevant control variables, so if you estimate OLS model, you will have to worry about unobservable factors that are correlated with the variables that you included in the regression. Omitted variable bias would result. If you believe that these unobservable factors are time-invariant, then fixed effects regression will eliminate omitted variable bias .. And within that project the most difficult problem is how to statistically control the variables that cannot be observed . fixed effects models are nice precaution even if you think you might not have a problem with omitted variable bias. In the basic fixed effects model, the effect of each predictor variable (i.e., the slope) is assumed to be identical across all the groups, and the regression merely reports the average within-group effect. What if you believe the slopes differ across all groups? finally, you have to estimate a different regression for each group. This will generate a different slope for each predictor variable in each market, which can quickly get out of hand. A more parsimonious solution is to estimate a single fixed effects regression but include slope dummy .

A dummy variable is a binary variable that is coded to either 1 or zero. It is commonly used to examine group and time effects in regression analysis [8] [4]

The form

$$y_{it} = (\alpha + \mu_i) + \beta_{it} x_{it} + \varepsilon_{it} \qquad \text{-------------------(1)}$$

Where y_{it}: be the dependent variable .

x_{it}: the vector of predictor variables.

α, μ_i : the baseline or considered part of the intercept.

β_{it}: the vectors of coefficients .

ε_{it}: the random variation at each point in time .

Intercept varying across groups (time) , the slope and error variance are constant

There are two basic of data requirements for using fixed effects methods First the dependent variables must be measured for each individual on at least two occasions. Those measurements must be directly comparable , that’s they must have the same meaning and metric. Second the predictor variables of interest must change in value across those two occasions for some substantial portion of the sample .[2]

These methods are useless in case of data for estimating the effects of variable that don't change over time. And data in which the dependent variable is measured on an interval scale and is linearly independent on a set of predictor variables .The name fixed effects is a source of considerable confusion consider we have a set of individuals (i = 1,2,……,n) each of whom is measured at two or more points in time (t = 1,2,…..,t) . let y_{it} be the dependent variable and a set of predictor variables that vary over time represented by the vector x_{it} and another set of predictor variables z_i that don't vary over time consider the linear model

$$y_{it} = \mu_i + \beta_i x_{it} + \gamma_i z_{it} + \alpha_i + \varepsilon_{it} \qquad \text{………………………(2)}$$

Where i subscript refers to different persons and t refers to different measurements within persons .we can use these model as single variables .because the two "errors" term α_i and ε_{it} behave somewhat differently . There is a different ε_{it} for each point in time , but α_i only varies across individuals not over time . i.e ε_{it} represents purely random variation at each point in time

The two period – case

Equation (2) is practically easy when the variables are observed at only two points in time (T = 2). Then the two equation are :-

$$y_{i1} = \mu_1 + \beta_i x_{i1} + \gamma_i z_{i1} + \alpha_i + \varepsilon_{i1}$$

$$y_{i2} = \mu_2 + \beta_i x_{i2} + \gamma_i z_{i2} + \alpha_i + \varepsilon_{i2} \qquad \text{…………………(3)}$$

If we subtract the first equation from the second , we get the " first difference " equation :

$$y_{i2} - y_{i1} = (\mu_2 - \mu_1) + \beta(x_{i2} - x_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1}) \qquad \text{…………….. (4)}$$

Which can be rewritten as :

$$y_i^* = \mu^* + \beta x_i^* + \varepsilon_i^* \qquad \text{……………(5)}$$

Where the asterisks indicate difference scores.

Table 3 : data for four cities contain price , quantity and year

Location	***Year***	***Price***	***Quantity***
Chicago	2003	$75	2.0
Chicago	2004	$85	1.8
Peoria	2003	$50	1.0
Peoria	2004	$48	1.1
Milwaukee	2003	$60	1.5
Milwaukee	2004	$65	1.4
Madison	2003	$55	0.8
Madison	2004	$60	0.7

Table 4: output of fixed effects regression in SAS

		Parameter Estimates			
		Parameter	Standard		
Variable	DF	Estimate	Error	t Value	Pr > |t|
Intercept	1	3.86884	0.15134	25.56	0.0015
price	1	-0.02534	0.00205	-12.33	0.0065
year	1	0.03904	0.01275	3.06	0.0921
Chicago	0	0	.	.	.
Peoria	1	-1.63562	0.06490	-25.20	0.0016
Milwaukee	1	-0.89349	0.03804	-23.49	0.0018
Madison	1	-1.72021	0.04787	-35.93	0.0008

Where
Intercept : of the demand curve in the omitted city
Price : Each 1$ increase in price fall by 0.025
Year : in 2004 is higher than 2003 by 0.039

This is known as " fixed effects " regression because it hold (fixes) the average effects of each city By including fixed effects (group dummies), you are controlling for the average differences across cities in any observable or unobservable predictors The fixed effect coefficients soak up all the across-group action. What is left over is the within-group action, which is what you want. You have greatly reduced the threat of omitted variable bias.[6]

The three – period case

When there are three or more observations per individual , a different method is needed . The dummy variables method requires a data set with a rather difference structure , one record for each observation for each individual. [5]

Table 5 :data of three observations per individual.

Id	gender	Time	Age	distance
1	1	1	10	20.0
1	1	2	12	21.5
1	1	3	14	23.0
2	1	1	10	21.5
2	1	2	12	24.0
2	1	3	14	25.5
3	1	1	10	24.0
3	1	2	12	24.5
3	1	3	14	26.0
4	1	1	10	24.5
4	1	2	12	25.0
4	1	3	14	26.5
5	1	1	10	23.0
5	1	2	12	22.5
5	1	3	14	23.5

6	2	1	10	25.0
6	2	2	12	29.0
6	2	3	14	31.0
7	2	1	10	22.5
7	2	2	12	23.0
7	2	3	14	26.5
8	2	1	10	22.5
8	2	2	12	24.0
8	2	3	14	27.5
9	2	1	10	27.5
9	2	2	12	26.5
9	2	3	14	27.0
10	2	1	10	23.5
10	2	2	12	22.5
10	2	3	14	26.0

Table 6: output of fixed effects

	fixed effects			**Conventional OLS**		
	Coefficient	Std .Err	P	Coefficient	Std .Err	P
Gender	1.933	2.896	0.0001	2.183	10.700	0.0001
Age	0.712	1.448	0.1438	0.712	0.214	0.0026
Time	0	0				
Id	-0.05	0.25	0.0031			

Both "gender" and "age" are significant at the .05 level. This means that "gender" and "age" are potentially important predictors of the dependent variable

For comparison, the right-hand panel of Table 6 gives the OLS estimates of the coefficients As we saw in the two-period case .

The mean deviation method

There are alternative algorithm that produces exactly the results . It is working as for each person and for each time-varying variable (both response and predictor variables) , we compute the means over time for that person:

$$\bar{y}_i = \tfrac{1}{n_i}\sum_t y_{it} \qquad \bar{x}_i = \tfrac{1}{n_i}\sum_t x_{it} \qquad \text{-----------------------(6)}$$

Where n_i is the number of measurements for person i .then subtract the person –specific means from the observed values of each variable:

$$y_{it}^* = y_{it} - \bar{y}_i \qquad x_{it}^* = x_{it} - \bar{x}_i \qquad \text{-----------------------(7)}$$

Finally , we regress y^* on x^* , plus variables to represent the effect of time. This is sometimes called a "conditional" method because it conditions out the coefficients for the fixed effects dummy variables

Examining the data in table (3) will be concerned with the following

Variations in prices around the mean price for each city ·

Variations in quantities around the mean quantity for each city.

Table (7) reports the mean prices and quantities of massages in each of the four cities, where P* and Q* denote these means.[7]

Table 7: the variation around the means

Location	Year	Price	P*	**P - P***	Quantity	Q*	**Q - Q***
Chicago	2003	75	80	**-5.0**	2.0	1.9	**0.1**
Chicago	2004	85	80	**5.0**	1.8	1.9	**-0.1**
Peoria	2003	50	49	**1.0**	1.0	1.05	**-.05**
Peoria	2004	48	49	**-1.0**	1.1	1.05	**.05**
Milwauke e	2003	60	62.5	**-2.5**	1.5	1.45	**.05**
Milwauke e	2004	65	62.5	**2.5**	1.4	1.45	**-.05**
Madison	2003	55	57.5	**-2.5**	0.8	0.75	**.05**
Madison	2004	60	57.5	**2.5**	0.7	0.75	**-.05**

The final step is to regress **Q - Q*** on **P - P*** Note that by subtracting the means, we have restricted all of the action in the regression to within-city action. Thus, we have eliminated the key source of omitted variable bias, namely, unobservable across-city differences in quality, sophistication, or whatever!

Table 8:output of mean deviation method

		Parameter Estimates			
		Parameter	Standard		
Variable	DF	Estimate	Error	t Value	Pr > |t|
Intercept	1	0	0.00604	0.00	1.0000
price	1	-0.02078	0.00195	-10.67	<.0001

Logistic model for data with two observation.

The logistic regression model is frequently used in epidemiologic studies , yielding odds ratio (dichotomous outcomes) . the response variable is a dichotomy and there are exactly two observations for each individual .

Let y_{it} be the value of the response variable for individual i on occasion t , but now y is to have values of either 0 or 1 and sometimes 1 or 2 .

Let p_{it} be the probability that $y_{it} = 1$. It is convenient to assume that the dependence of p_{it} on possible predictor variables is described by a logistic regression [1]

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = \mu_t + \beta x_{it} + \gamma z_i + \alpha_i \qquad \text{---------------------(8)}$$

Where z_i is a column vector of variables that describe the individuals but do not vary over time , and x_{it} is a column vector of variables that vary both over individuals and over time for each individual . in this equation μ_t is an intercept that is allowed to vary with time , β and γ are row vectors of coefficients. α_i represents all differences between persons that are stable over time and not otherwise accounted for by z_i .

Additionally ,assume that for a given individual i y_{1i} and y_{2i} are independent

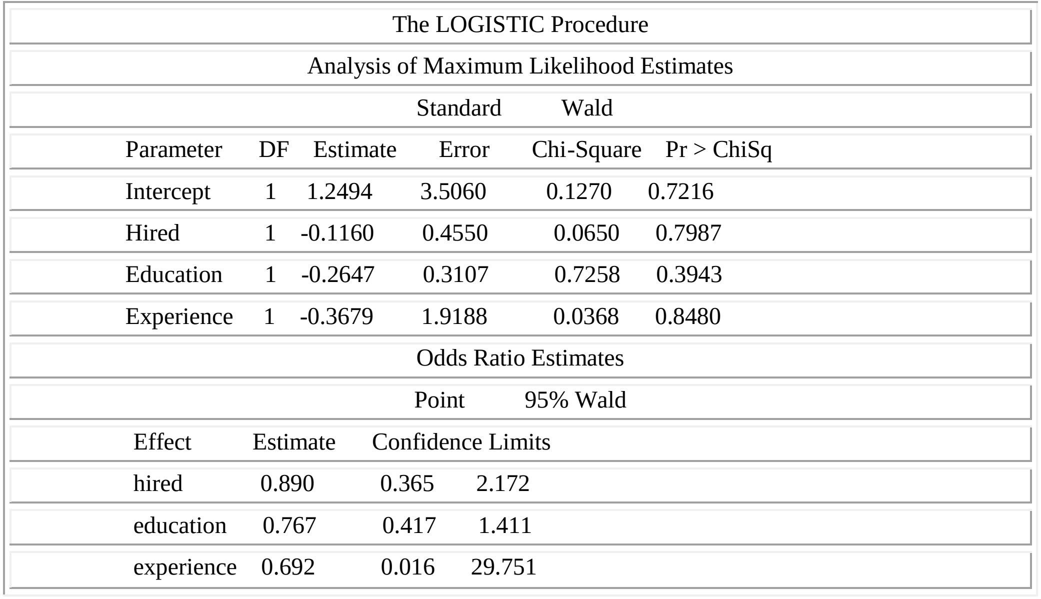
$$\log\left(\frac{pr(y_{i1}=0,y_{i2}=1)}{pr(y_{i1}=1,y_{i2}=0)}\right)=(\mu_2-\mu_1)+\beta(x_{i2}-x_{i1}) \qquad \text{------------------(9)}$$

Thus, as we found for the linear model, both z_i and α_i have been "differenced " out of the equation..

Table 9 : data of 28 persons hiring practice of particular firm. At two years.

2003				2004			
hired	education	Experience	sex	hired	education	experience	sex
1	8	1	0	0	6	2	0
0	6	1	1	0	4	0	1
0	8	4	0	1	6	6	1
1	4	2	1	1	6	3	1
1	6	3	1	0	4	1	0
1	8	5	0	1	8	3	0
0	4	7	1	0	4	2	1
0	6	0	1	0	4	4	0
1	4	3	0	0	6	1	0
1	8	5	0	1	8	10	0
0	4	0	1	0	4	2	1
0	6	9	0	0	8	5	0
0	4	0	9	0	6	7	0
1	8	10	1	1	4	5	1
0	4	0	0	0	6	4	0
1	8	1	1	0	8	0	1
1	6	12	0	1	6	1	1
0	8	5	0	0	4	7	0
1	8	6	1	0	4	1	1
0	4	7	0	0	4	5	0
0	4	9	1	0	6	0	1
1	6	2	1	1	8	5	1
0	6	10	0	0	4	9	0
1	4	8	0	0	8	1	0
0	4	7	1	0	6	1	1
1	6	5	1	1	4	10	1
1	8	9	0	1	6	12	0
1	6	2	1	0	8	5	0

Table 10: logistic output for regression on difference score



```
                              The LOGISTIC Procedure

                       Analysis of Maximum Likelihood Estimates

                                    Standard        Wald
          Parameter     DF   Estimate      Error     Chi-Square   Pr > ChiSq

          Intercept      1    1.2494      3.5060       0.1270      0.7216

          Hired          1   -0.1160      0.4550        0.0650      0.7987

          Education      1   -0.2647      0.3107        0.7258      0.3943

          Experience     1   -0.3679      1.9188        0.0368      0.8480

                                    Odds Ratio Estimates

                                    Point        95% Wald

           Effect        Estimate     Confidence Limits

           hired          0.890         0.365      2.172

           education      0.767         0.417       1.411

           experience     0.692         0.016      29.751
```

Table 10 gives the results. Although the time-varying predictors are expressed as difference scores, their coefficients should be interpreted as they appear in equation (8), that is, as the effect of the value of the variable in a given year on the probability of poverty in that same year.

The coefficients (and odds ratios) for education and experience must be interpreted somewhat differently. According to equation (9), as time-invariant predictors these variables shouldn't even be in the model. In fact, they represent interactions between time-invariant predictor variables and time itself, so that the rate of change in the odds of hired depends on the value of these variables

Conclusion

1- it is possible to control for all possible characteristics of the individuals in the study—even without measuring them .

2- If the dependent variable is quantitative, then fixed effects methods can be easily implemented using ordinary least squares linear regression .

3- fixed effects regression methods provide a relatively easy and effective way to control for stable variables that cannot be measured .

4- Fixed effects regressions are very important because data often fall into categories such as industries, states, families, etc .

5- We could just include dummy variables for all but one of the units. If we have longitudinal data, this sacrifices a lot of degrees of freedom. And with so many units and very few time periods .

References

1- Allison, P. D. (1999), *Logistic Regression Using the SAS System*, Cary, NC: SAS Institute Inc.

2- Baltagi, Badi H. 2001. *Econometric Analysis of Panel Data*. Wiley, John & Sons.

3- Fuller, Wayne A. and George E. Battese. 1973. "Transformations for Estimation of Linear Models with Nested-Error Structure." *Journal of the American Statistical Association*, 68(343) (September): 626-632 .

4- LaMotte, L. R. (1983), "Fixed-, Random-, and Mixed-Effects Models," in *Encyclopedia of Statistical Sciences,* eds. S .

5- Suits, Daniel B. 1984. "Dummy Variables: Mechanics V. Interpretation." *Review of Economics & Statistics* 66 (1):177-180.

6- Singer J.D. (1998). "Using SAS PROC MIXED to fix multilevel models, hierarchical models and individual growth models." Journal of Educational and Behavioral Statistics, 24:323-355.

7- Sobel, M.E. (2000), "Causal Inference in the Social Sciences," Journal of the American Statistical Association, 95, 647–651 .

8- Uyar, Bulent, and Orhan Erdem. 1990. "Regression Procedures in SAS: Problems?" *American Statistician* 44(4): 296-301..