Faculty of Economics, University of Niš
18October 2018

49th International Scientific Conference
QUANTITATIVE AND QUALITATIVE ANALYSIS IN ECONOMICS

PRACTICAL USE OF PANEL DATA, PROBIT AND LOGIT MODELS IN ECONOMY

Nastasja Stašević, MSc*

Abstract: *The main purpose of this paper is to show how panel, probit, and logit models can be used in the economic analysis. The first part of the paper focuses on panel data model, which possesses a cross-sectional and time dimension. This type of data provides information over time for individuals. Panel data example has been completed with two software, Gretl, and RStudio. The second part focuses on probit and logit model, and the differences between them. These models are mostly used in research such as buying product – yes or no. Probit and Logit analysis has been completed with RStudio software.*

Keywords*: panel data analysis, probit and logit model, time series, entrepreneurs.*

1. Introduction

The main problem of this paper is the use of longitudinal data in economics. Panel model is suitable because it collects data in at least two dimensions. There has been a rising interest for Panel model, which collects data by individuals and by time. There are a lot of advantages of using Panel data models; according to Baldev and Badi (1992) the main two advantages are improved efficiency of estimators and the depth of the analysis. Panel data can be very useful for economics analysis. Nowadays panel data models increased its usage especially in the developing countries. In this paper, panel data analysis has been completed with 5 periods of time and 12 individuals. The data has been collected by survey which focuses on entrepreneurs in Serbia and their taxation. The entrepreneurs are especially important for developing countries, such as Serbia, because they produce innovation and they reduce unemployment.

Probit and Logit models are helpful in the analysis; they overcome the limitation of linear regression. These models have dichotomy variable as a dependent variable and their probability is limited between 0 and 1. The main difference between Logit and Probit models is in distribution. Logit uses logistic distribution, and Probit uses the standard normal one. Probit and Logit models are just a base for multinomial and ordered models.

* PhD student at University of Novi Sad, Faculty of Economics in Subotica, Serbia;
✉ nastasja.stasevic@gmail.com
UDC 519.8:330

The mentioned models have a great usage in sociology, psychology and political science.[1] Usually Logit is preferred over Probit, and the main result is that Logit can be calculated by using a calculator.

2. Panel data

Panel data models have a great usage in the economy. For example, for labor economics, it shows how education, experience, marriage status over time affects the income of an individual; how the number of children, education, and age affects savings over time. The difference between panel data and time-series is that panel data collects data for more than one individual, and the time series collect changes over time for just one individual. The best example of a time series is inflation over the years for one country.

Panel data model can be balanced. This happens when there is data for all individuals over time. Also, it happens that we could not collect data for all individuals over the time periods, and that is what we call unbalanced data. Perhaps the biggest problem in panel models is unbalanced data. Panel data have a lot of advantages, one of them is heterogeneity. When we talk about Panel data, we assume clustering. That means that we have one individual and n- number of time periods. In this example, we have two individuals over two-time periods.

Table 1: Panel data example

Individuals	Time	Days worked	Income
1	1	355	38 520$
1	2	352	66 280$
2	1	320	68 520$
2	2	290	36 820$
3	1	311	47 736$
3	2	248	75 460$

Source: Example given by author

There are three types of panel data:

1. Short panel data – few periods of time
2. Long panel data – few individuals
3. Both – in this type of data there are a lot of periods of time and a lot of individuals. (Katchova (2013), Panel data models)

Also, Katchova (2013) defines three types of regressors:

1. Varying regressor – changes over time for an individual (savings for the individual over the years)
2. Time-invariant - these regressors are the same over time (gender)
3. Individual – invariant regressors are changing over time, for example, inflation. "Panel models describe the individual behavior both across time and individual." (Katchova A, Panel data models, (2013), Retrieved from: https://sites.google.com/site/econometricsacademy/econometrics-models/panel-data-models, Accessed on August 1st, 2018.)

[1]Aldrich J., Nelson F., AdlerS., (1984), Linear Probability, Logit, and Probit Models, Sage university paper.

Katchova (2013) in Panel data models described three types of panel models:

1. Pooled - which shows total variations,
2. Fixed – which shows variations between individuals,
3. Random effects – this model shows time variations.

The first test, L-M test shows if it is better to choose a pooled model or random effect model. The second, Hausman test is used to show which model, fixed or random effect is better for the data.

3. Probit and Logit model

Probit and Logit models are usually used in product analysis, for example, to determine whether the consumer bought the product or not, or whether the employer employed the candidate or not. The main difference between linear regression and Probit or Logit model is that the dependent variable is a dichotomy, which means the answer is either yes or no. These answers are usually coded with numbers 1 and 0. "Logistic analyses for binary outcomes attempt to model the odds of an event's occurrence and to estimate the effects of independent variables on these odds.". (O'Connell A.,(2006), Logistic regression models for original response variables)

In the Probit and Logit analysis, it is possible to interpret coefficients and marginal effects. Coefficients and marginal effects need to have the same sign, but the interpretation is slightly different. Marginal effects show more likelihood or less likelihood of the happening event. Coefficients are interpreted just as a magnitude.

Probit and Logit models are base for bivariate and multinomial Probit and Logit. In the multinomial models, it is possible that the dependent variable has more than two alternatives, meaning that it is a categorical variable and an individual can only choose one of them.

The main difference between Probit and Logit models are in distribution, Probit has standard normal distribution and Logit has logistic distribution. These models are usually used in the applied economy, sociology, political science. Probit or Logit analysis can be called qualitative analysis, because "dependent variable takes a discrete number of mutually exclusive values" (Borooah V.K.,(2002), Logit and Probit: Ordered and Multinomial Models). Borooah (2012) claims that social scientists were always interested in the choices between mutually exclusive options and that is what makes Probit and Logit so interesting.

4. Panel data model example

I have collected this data from the survey. The data was collected from April 13th to July 6th, 2018. It focuses on entrepreneurs and the taxation system in Serbia. Nowadays, entrepreneurs play a very important role in transitional economies because they are the key for innovations and prosperity. In this data, there are 5 periods of time starting from 2013 to 2017. In the sample were examined entrepreneurs which were active all these years. The survey is divided into five parts, one for every year. Entrepreneurs needed to provide information about income, expenditures, employees, and profit on a yearly basis. It was assumed that the number of employees affected the amount of the tax, which was proven

wrong. The dependent variable is income tax, which is calculated as 10% of the profit, and it is expressed in dinars. The independent variables are profit, also expressed in dinars and in the number of employers.

First, it is necessary to define the variable in R using function c bind. Also there should be installed "plm" package which provides tools to do panel analysis. After that, it should be defined time and individual dimension. In R it is possible to calculate Fixed, Random, First difference, Between, Pooled estimator model. Coefficients are usually similar.

Table 2. Descriptive statistics for dependent variables

Profit	Employers
Min. : 54489	Min. : 0.000
1st Qu.: 179230	1st Qu.: 1.000
Median : 247205	Median : 1.000
Mean : 354379	Mean : 2.517
3rd Qu.: 282295	3rd Qu.: 2.000
Max. :2447036	Max. :17.000

Source: Research results

The mean profit is 354 379 dinars, and the mean of employers is 2.51, which is 3 employees. First Quartile shows that 25% of the individuals have less than, or equal to 179 230 dinars of profit. The third Quartile shows that 75% of the individuals have less than or equal to 282 295 dinars of profit.

First Quartile shows that 25% of the individuals have less than or equal to employeers. The third Quartile shows that 75% of the individuals have less than or equal to 2 employees.

Table 3. Standard deviation, Coefficient of Variance, Skewness and Ex. Kurtosis

Variables	Std. Dev.	C.V.	Skewness	Ex. Kurtosis
Employment	3.8861	1.5441	2.4598	5.1335
Income	3.22E+07	2.0759	3.2948	10.477
Profit	4.36E+05	1.2317	3.3283	11.053
Tax Total	48317	1.3497	3.8879	16.292

Source: Research results

Standard deviation shows how data concentrated around the mean. If the data is more concentrated the results are better, which is this case with this example. Skewness is showing the lack of symmetry and Ex. Kurtosis is showing if the distribution is heavy-tailed or light-tailed. In this data, there is a necessity for transformation which has been done by logging the variables. After transformation of the variables, the test of normality showed negative results, so the analysis has been completed with the original data.

Picture 1. Test of normality

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
inco	.321	85	.000	.452	85	.000
prof	.353	85	.000	.488	85	.000
employ	.385	85	.000	.547	85	.000
taxtotal	.375	85	.000	.440	85	.000

a. Lilliefors Significance Correction

Source: Research results

Table 4. Coefficients of the Panel data Model

Tax	POOLED OLS	BETWEEN	FIXED	FIRST DIFF	RANDOM
Profit	1.0145e-01 ***	8.8261e-02 ***	1.0651e-01 ***	1.1207e-01 ***	1.0171e-01 ***
Employees	1.0344e+03 *	2.2056e+03 *	1.0107e+03	2.5657e+02	1.0131e+03 *
R2	0.97085	0.99318	0.9266	0.78862	0.9698
Adj R2	0.96982	0.99166	0.90586	0.78402	0.96874

Source: Research results

In the data, there are 12 individuals - entrepreneurs in five periods of time. This panel data is balanced, which means there are data for each individual and for every period of time. The Adjusted R-squared is 0.96 which is great, and means that 96% of variations are explained with the model.

Comment for pooled model: Across entrepreneurs over time, an additional unit of profit leads to a higher tax for 10%. Also, an additional employee leads to a higher tax, but variable x2- employees is not significant.

Between estimators shows similar results. The average tax is 8% higher for entrepreneurs with one unit of profit more than average. In this case too, the number of employees is not significant. For additional unit of profit above average for entrepreneurs leads to 10, 15% higher tax.

First difference estimator tells that for each additional unit of profit ; from one year to the next leads to 11% higher fax.

Hausman Test

Data: Y ~ X

chisq = 4.1129, df = 2, p-value = 0.1279

Alternative hypothesis: one model is inconsistent

The result of Hausman test shows that it should be used a fixed effect model. The same result is possible to get form Gretl software, which can be easier to use than R-software. It is necessary to import data and to check box panel data. In the upper right corner, we choose Model, Panel, and estimator. With this software it is possible to get the results for fixed effect, OLS and between estimators.

5. Probit and Logit model example

The next example was downloaded from the Kaggle data base. It consists of data related to buying the product, gender of consumers, their estimated salary.

Table 5. Descriptive statistics for dependent variable

Y- Did the consumer bought the product?
Min. :0.0000
1st Qu.:0.0000
Median :0.0000
Mean :0.3575
3rd Qu.:1.0000
Max. :1.0000

Source: Research results

Table 6. Descriptive statistics for independent variables

V1- Gender	V2- Age	V3-Salary
Min. :0.00	Min. :18.00	Min. : 15000
1st Qu.:0.00	1st Qu.:29.75	1st Qu.: 43000
Median :0.00	Median :37.00	Median : 70000
Mean :0.49	Mean :37.66	Mean : 69743
3rd Qu.:1.00	3rd Qu.:46.00	3rd Qu.: 88000
Max. :1.00	Max. :60.00	Max. :150000

Source: Research results

First Quartile shows that 25% of the individuals have less than or equal to 43000 dollars salary. The third Quartile shows that 75% of the individuals have less than or equal to 88000 dollars salary.

First Quartile shows that 25% of the individuals are less than or equal to 29.75 years old. The third Quartile shows that 75% of the individuals are less than or equal to 46 years old.

Table 7. Percentage of consumers

Y	
0	1
0.6425	0.3575

Source: Research results

64% of consumers did not bought the product. Age and salary as independent variables are significant, but gender is not. The interpretation of the coefficients is just in magnitude. Individuals with the higher salary and older individuals are more likely to buy the product, in comparison to the individuals who have a lower salary and are younger. The results of coefficients are similar, one unit in Logit coefficients equals to 1.6 units of Probit.

Table 8. Coefficients and marginal effects

Consumption of the product	Logit coefficient	Logit marginal effect	Probit coefficients	Probit marginal effect
Gender	3.338e-01	3.684644e-02	1.975e-01	3.887923e-02
Age	3.338e-01 ***	2.615442e-02	1.274e-01 ***	2.508013e-02
Salary	3.644e-05 ***	4.022030e-06	1.946e-05 ***	3.829556e-06

Source: Research results

If we concentrate on the marginal effects, the results are pretty similar. For each additional unit in salary, individuals are 0.003% more likely to buy the product. For each additional year, individuals are 2.51% more likely to buy the product. Interpretation for Logit marginal effect is the same.

The model predicts 85% correct which is great, and the rest are misclassified. Usually for good model fit, the limit is 60% correct predictions. McFadden's Pseudo R-square is 0.4665987 which means that independent variables have explanatory power, which is not great, but it is acceptable.

6. Conclusion

Panel data or longitudinal data have spatial and time dimensions which makes it so useful for economic analysis, especially labor analysis.

In the United States, panel data have been collected since 1966. Now panel data are available in developing countries, especially data about income, aging, health.

The advantages of using panel data are many, among which its proven that it increases the degrees of freedom and improves the estimates. However, panel data is not perfect data. One of the problems is that explaining individual behavior can be very hard because every individual can be subjected to the different influences.

Probit and Logit functions are very similar - their predictions are mostly the same. Tom Futing Liao gives the advantage to Logit model because predicted probabilities can be reached by hand with the calculator. However, there should be noted that Probit and Logit are not just concentrated on the binary dependent variable. For example, if you want to find out which type of credit an individual uses, which type of yogurt, Multinomial Logit, and Probit model should be used.

Nowadays these models are essential in economics analysis.

References

Aldrich J., Forrest N., Adler S., (1984*), Linear Probability, Logit, and Probit Models*, Sage Publications.

Baldev R., Badi H. B.,(1992*), Panel data analysis*, Physical – Verlag Heidelberg.

Charbounneau, K., (2017), *Multiple fixed effects in binary response panel data models*, The Econometrics journal Vol 20 Issue 3.

Croissant Y., Millo G., (2007), *Introduction to plm*, Retrieved from: ftp://ftp.uni-bayreuth.de/pub/math/statlib/R/CRAN/doc/vignettes/plm/plmEN.pdf , Accessed on 1 October 2018.

Futing Liao T., (1994), *Interpreting Probability Models: Logit, Probit, and Other,* Sage University paper

O'Connell A., (2006), *Logistic regression models for original response variables*, Sage publications.

Hsiao C. (2014), *Analysis of Panel Data,* Cambridge.

Katchova A, (2013), *Panel data Models*, Retrieved from: https://sites.google.com/site/econometricsacademy/home, Accessed on 1 August 2018.

Katchova A, (2013), *Probit and Logit Models*, Retrieved from: https://sites.google.com/site/econometricsacademy/home, Accessed on 1 August 2018.

PRAKTIČNA UPOTREBA PANEL, PROBIT I LOGIT MODELA U EKONOMIJI

***Rezime**: Osnovna svrha rada je da pokaže kako panel, probit i logit modeli mogu biti korišćeni u ekonomskim analizama. Prvi deo rada odnosi se na panel model koji priazuje dve dimenizij - prva se odnosi na individue koje se posmatraju, a druga na vremensku dimenziju. Panel podaci u radu obrađeni su u dva softwera, u RStudiu i Gretl-u. Drugi deo rada, fokusira se na Logit i Probit modele, kao i na razlike među njima. Ovi modeli najčešće se koriste u analizama kao što je: kupovina proizvoda – da ili ne. Takođe, Probit i Logit analiza izvršena je u RStudiu.*

***Keywords:** Panel podaci, logit i probit analiza, vremenske serije.*