Take-Home Assignment in Applied Econometrics

Joaquin Barrutia Alvarez*

VT 2023

1 Exercise 1

1.1 Question (ii)

The first specification presented in Table 1 is a Linear Probability Model (LPM) estimating the probability of failure in the MSc of Finance at the University of Reading. The only variables which are significant at least at the 5% level are *Agrade*, *WorkExperience*, and *Year2006*. In short, what we can say about the results presented is that having an A grade at the undergraduate level decreases the probability of failing by 8.2 percentage points holding all the other variables fixed. Having work experience decreases the probability of failing by 6.5 percentage points holding all the other variables fixed. Being en the 2006 cohort increases the probability of failing by 14.2 percentage points holding all the other variables fixed. The main advantage of using a LPM is that the results are straight-forward to interpret, coefficient estimates in this model are interpretable as marginal effects. LPM is just a special case of OLS, thatÂ´s why the interpretation and implementation of this model is easy. The disadvantages would be that it assumes a linear relationship between the independent variables and the dependent variable, which is not the case in practice, meaning that the functional form will be wrongly specified which can lead to biased estimates. LPM is also sensitive to extreme values which can affect the estimated coefficients.

1.2 Question (iii)

The *Agrade* coefficient is not statistically significant but the *WorkExperience* coefficient is statistically different than zero at the 5 % level.

1.3 Questions (iv) & (v)

We can interpret the results in Table 2 as a unit increase in *BelowBGrade* (i.e., going from having a B grade or above to having a grade below a B) is associated with a 5.62 or a 6.17 percentage point increase in the probability of failure, holding all other variables constant. Similarly, a unit increase in *Year2006* (i.e., going from being in a cohort different than 2006 to being in the 2006 cohort) is associated with a 13.5 or a 13.6 percentage point increase in the probability of failure, holding all other variables constant. The results differ a little depending on which model we are using, Logit or Probit. It's important to notice that for both models only the *Year2006* effects are

*Email: joaquin.barrutia-alvarez.1895@student.uu.se

statistically different than zero. These estimates suggest that being in the year 2006 is associated with a higher probability of failure. There are small differences in the marginal effects depending on which model we are using this is because the two models have different underlying assumptions about the distribution of the error term. The Logit model assumes that the error term follows a logistic distribution and the Probit model assumes that the error term follows a normal distribution leading to differences in the estimated marginal effects. The advantages of using these models are that Probit and Logit are both robust models that can handle non-linear relationships between the independent variables and the dependent variable, Both models are bounded between 0 and 1 like probabilities. On the other side, these models are more computationally intensive than a LPM and they are also harder to interpret.

1.4 Question (vi)

The results reported in Table 3 show that only the *Year2006* and the intercept coefficients are statistically significant. the intercept coefficient of 0.098 implies that the predicted probability of failure when all the independent variables are zero is 9.8% (i.e., when the student is not from the 2006 cohort, and does not have postgraduate studies). The coefficient for *Year2006* of 0.175 means that, holding all other variables constant, the probability of failure for a student who started the program in 2006 is 17.5 percentage points higher than a student who started the program in other years. The coefficient for *postgrad* of 0.033 means that, holding all other variables constant, the probability of failure for a student who has a postgraduate degree is 3.3 percentage points higher than a student who does not have a postgraduate degree. The coefficient for *post_grad06* of -0.163 means that, holding all other variables constant, the probability of failure for a student who has a postgraduate degree and joined the MSc program in 2006 is 16.3 percentage points lower than a student who does not have a postgraduate degree and did not join the MSc program in 2006. It's important to take into account that these results were obtained from a reduced sample beacause we filtered country 8 out of our dataset.

2 Exercise 2

2.1 Question (i)

For the following questions we wil be using a panel dataset, in whcih each individual is a country and each time period is a year. Our dataset has 18 different countries and 19 years (i.e., from 1960 to 1978), we have 342 observations meaning that our panel data is balanced. Table 4 shows some descriptive statistics from all the variables in our panel data. The *overall* statistics reported provide information about the overall distribution of each variable in the dataset, meaning that these statistics are calculated across all time periods and cross-sectional units. The *within* statistics reported provide information about the variation of each variable within each cross-sectional unit. For example, the mean and standard deviation can be used to describe the average level and variability of each variable within each country. Similarly, *between* statistics reported provide information about the variation of each variable between cross-sectional units. For example, the mean and standard deviation can be used to describe the average level and variability of each variable between countries.

2.2 Question (vi)

The results for this question are presented in Table 5. We can summarize the main results for each estimation:

- Pooled OLS: For a 1% increase in per capita income, holding all other variables constant, we would expect a 0.890% increase in gasoline consumption per car. For a 1% increase in relative gasoline price, holding all other variables constant, we would expect a 0.892% decrease in gasoline consumption per car. For a 1% increase in cars per capita, holding all other variables constant, we would expect a 0.763% decrease in gasoline consumption per car.

- Random Effects: For a 1% increase in per capita income, holding all other variables constant, we would expect a 0.555% increase in gasoline consumption per car, on average across all countries. For a 1% increase in relative gasoline price, holding all other variables constant, we would expect a 0.420% decrease in gasoline consumption per car, on average across all countries. For a 1% increase in cars per capita, holding all other variables constant, we would expect a 0.607% decrease in gasoline consumption per car, on average across all countries.

- Fixed Effects: For a 1% increase in per capita income, holding all other variables constant, we would expect a 0.662% increase in gasoline consumption per car within a particular country. For a 1% increase in relative gasoline price, holding all other variables constant, we would expect a 0.322% decrease in gasoline consumption per car within a particular country. For a 1% increase in cars per capita, holding all other variables constant, we would expect a 0.640% decrease in gasoline consumption per car within each country.

- First-Differences: A 1% increase in per capita income is associated with a 0.242% increase in gasoline consumption per car, on average, within each country over time. A 1% increase in relative gasoline price is associated with a 0.252% decrease in gasoline consumption per car, on average, within each country over time. Finally, a 1% increase in cars per capita is associated with a 0.656% decrease in gasoline consumption per car, on average, within each country over time.

We can see that the magnitudes in the coefficients are different depending on which specification we are using, but the signs in all the specifications are the same, we should not make any causal statements based on these magnitudes but given of that there is robustness on the sign of the coefficients across all specifications we can at least have more certainty on the direction of the effects.

2.3 Question (vii)

The results using robust standard errors are presented in Table 6. It can be observed that standard errors presented in Table 6 are larger than the ones presented in Table 5, decreasing the level of significance for some coefficients. I would report robust standard errors because they take into account the possibility of heteroscedasticity and autocorrelation in our data, this is why they tend to be larger. Standard OLS standard errors assume that the variance of the errors is constant across all observations and that there is no correlation between the errors for different observations.

2.4 Question (viii)

The identification assumption of the Pooled OLS is the following:

$$\mathbb{E}[u_{it}|X_{it}] = 0$$

Which implies that the error term is uncorrelated with the explanatory variables, requiring that there are no omitted variables that are correlated with both the explanatory variables and the error term. It is easy to think of examples of omitted variable bias in this case which will lead to biased estimates.

2.5 Question (viii)

The identification assumption of Fixed Effects estimator is known as strict exogenity:

$$\mathbb{E}[u_{it}|X_{i1}, ..., X_{iT}] = 0$$

This means that the explanatory variables in each time period must be uncorrelated with the error term in each time period.

The identification assumption of First-Differences estimator is weaker than the strict exogeneity assumption:

$$\mathbb{E}[\Delta u_{it}|\Delta X_{it}] = 0$$

Changes in the error term must be uncorrelated with changes in the explanatory variables.

Table 1: Determinants in the Probability of Failure

	(LPM) Fail	(Logit) Fail	(Probit) Fail
Age	-0.003	-0.025	-0.012
	(0.01)	(0.07)	(0.03)
Agrade	-0.082**	-1.182*	-0.629*
	(0.03)	(0.50)	(0.25)
BelowBGrade	0.084	0.509	0.309
	(0.06)	(0.39)	(0.22)
2.CountryCode	0.045	0.314	0.209
	(0.09)	(0.90)	(0.48)
3.CountryCode	-0.101	-	-
	(0.09)	(.)	(.)
4.CountryCode	-0.023	-0.360	-0.177
	(0.08)	(0.90)	(0.46)
5.CountryCode	0.022	0.220	0.134
	(0.15)	(1.50)	(0.77)
6.CountryCode	0.182	1.405	0.817
	(0.14)	(1.06)	(0.57)
7.CountryCode	0.186	1.290	0.776
	(0.12)	(1.00)	(0.53)
8.CountryCode	0.047	0.379	0.227
	(0.07)	(0.68)	(0.35)
9.CountryCode	0.057	0.642	0.365
	(0.11)	(1.05)	(0.54)
10.CountryCode	0.033	0.186	0.133
	(0.08)	(0.86)	(0.45)
WorkExperience	-0.065*	-0.625*	-0.346*
	(0.03)	(0.30)	(0.16)
Year2004	0.067	0.682	0.380
	(0.04)	(0.50)	(0.25)
Year2005	0.001	-0.151	-0.108
	(0.03)	(0.53)	(0.26)
Year2006	0.142**	1.224**	0.680**
	(0.05)	(0.47)	(0.24)
Year2007	0.085	0.822	0.433
	(0.05)	(0.49)	(0.26)
post_grad	0.050	0.425	0.237
	(0.05)	(0.45)	(0.23)
Female	-0.038	-0.405	-0.235
	(0.03)	(0.38)	(0.20)
English	-0.021	-0.195	-0.091
	(0.07)	(0.64)	(0.34)
constant	0.165	-1.596	-1.015
	(0.21)	(1.97)	(0.91)
Robust standard errors	Yes	Yes	Yes
R-sqr	0.095		
Observations	500	468	468

Note: Robust standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Marginal Effects

	(Logit)	(Probit)
	dy/dx	dy/dx
BelowBGrade	0.0562	0.0617
	(1.30)	(1.39)
Year2006	0.135**	0.136**
	(2.60)	(2.83)
N	468	468

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Determinants in the Probability of Failure Filtering Country = 8

	(LPM)
	Fail
post_grad	0.033
	(0.05)
Year2006	0.175**
	(0.06)
post_grad06	-0.163
	(0.12)
constant	0.098***
	(0.02)
Robust standard errors	Yes
R-sqr	0.035
Observations	340

Note: Robust standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Summary Statistics

		Mean	Std. Dev.	Min	Max	N/n/T-bar
Country_Code	overall	9.5	5.195729	1	18	342
	between	.	5.338539	1	18	18
	within	.	0	9.5	9.5	19
Year	overall	1969	5.485251	1960	1978	342
	between	.	0	1969	1969	18
	within	.	5.485251	1960	1978	19
log_gaspcar	overall	4.296242	.5489071	3.380209	6.156644	342
	between	.	.5150439	3.729646	5.766355	18
	within	.	.2236768	3.545347	5.591887	19
log_incomepcap	overall	-6.139425	.6345925	-8.072524	-5.221232	342
	between	.	.6094852	-7.816214	-5.44856	18
	within	.	.2254859	-6.876993	-5.599641	19
log_relgasprice	overall	-.5231032	.6782225	-2.896497	1.125311	342
	between	.	.684445	-2.709171	.7393789	18
	within	.	.1274722	-1.057113	-.1371714	19
log_carpcap	overall	-9.041805	1.218896	-13.47518	-7.536176	342
	between	.	1.1142	-12.45886	-7.78109	18
	within	.	.5565938	-11.3318	-7.69075	19

Table 5: Panel Data Specifications: Determinants on Gas Consumption

	(Pooled OLS) log_gaspcar	(Random Effects) log_gaspcar	(Fixed Effects) log_gaspcar	(First-Differences) d_log_gaspcar
log_incomepcap	0.890***	0.555***	0.662***	
	(0.04)	(0.06)	(0.07)	
log_relgasprice	-0.892***	-0.420***	-0.322***	
	(0.03)	(0.04)	(0.04)	
log_carpcap	-0.763***	-0.607***	-0.640***	
	(0.02)	(0.03)	(0.03)	
d_log_incomepcap				0.242**
				(0.08)
d_log_relgasprice				-0.252***
				(0.03)
d_log_carpcap				-0.656***
				(0.04)
constant	2.391***	1.997***	2.403***	0.020***
	(0.12)	(0.18)	(0.23)	(0.00)
Indivudal fixed effects	No	No	Yes	No
Year fixed effect	No	No	No	No
Robust standard errors	No	No	No	No
R-sqr	0.855		0.840	0.469
Observations	342	342	342	324
vce	ols	conventional	conventional	ols

Note: Standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Panel Data Specifications: Determinants on Gas Consumption

	(Pooled OLS) log_gaspcar	(Random Effects) log_gaspcar	(Fixt Effects) log_gaspcar	(First-Differences) d_log_gaspcar
log_incomepcap	0.890***	0.555***	0.662***	
	(0.04)	(0.12)	(0.16)	
log_relgasprice	-0.892***	-0.420***	-0.322*	
	(0.04)	(0.12)	(0.13)	
log_carpcap	-0.763***	-0.607***	-0.640***	
	(0.02)	(0.09)	(0.10)	
d_log_incomepcap				0.242*
				(0.11)
d_log_relgasprice				-0.252***
				(0.03)
d_log_carpcap				-0.656***
				(0.06)
constant	2.391***	1.997***	2.403***	0.020***
	(0.12)	(0.53)	(0.60)	(0.00)
Indivudal fixed effects	No	No	Yes	No
Year fixed effect	No	No	No	No
Robust standard errors	Yes	Yes	Yes	Yes
R-sqr	0.855		0.840	0.469
Observations	342	342	342	324
vce	robust	robust	cluster	robust

Note: Robust standard errors in parentheses.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$