Advanced Statistics and Programming - Individual Assignment 3

Thanh Dat Nguyen - 532618tn

14 October 2022

Panel data modeling: time to export

Time to export is an important metric for both countries and exporting firms, as it increases costs and the risks of deteriorated goods. In the following analysis, a panel data model is utilized to analyze the determinants of time to export between countries and over time.

Population model

For the analysis, four independent variables were chosen to be studied as possible determinants. Firstly, there is the country GDP per capita (given in constant 2015 US dollars), which we assume decreases (negatively affects) time to export (as a country's economy grows, we assume it opens itself up more to export products). Then there is the tax on exports (expressed as a percentage of tax revenue), which is assumed to positively affect the dependent variable (higher tax might indicate a governmental policy that discourages exports, hence longer time to export). Thirdly, there is the foreign direct investment, which is assumed to have a negative relationship with time to export, as this indicator is generally positively correlated with export performance (we assume that if there are more exports from a country, it should also take less time to export from the given country). This is expressed by the equation below:

$$TimeToExport_{it} = \alpha_{it} + \beta_{1,it} GDP_{it} + \beta_{2,it} ExportTax_{it} + \beta_{3,it} FDI_{it} + u_{it} \tag{1}$$

Where i indicates the country, t indicates the year and u_{it} stands for the disturbance term.

Summary Statistics

In this section, a balanced panel data set is created, with complete observations across the years 2014 to 2019. The summary table is presented below. FDI is transformed from dollars to millions and GDP per capita into GDP per thousand capita for table results to be more legible.

Table 1: Summary Table of the Panel Data

Statistic	N	Mean	St. Dev.	Min	Max
date	624	2,016.500	1.709	2,014	2,019
ExportTax%	624	1.143	4.733	0.000	43.730
TimeToExport	624	44.412	48.919	0.000	312.000
FDImillions	624	−3,338.464	31,556.580	−344,331.000	166,208.100
GDPkcap	624	16.640	20.914	0.380	108.570

When we limit ourselves to the chosen variables and only complete observations, there are 624 observations remaining within the data set. Dividing the number of observations by six (the number of years between 2014 and 2019) gives us a total of 104 countries to compare in the dataset. This also confirms that this is a balanced panel data set, as there is the same number of observations per individual (country).

An interesting observation is that FDI can go into extremely negative numbers (minus 344 billion dollars) - this is most likely due to developed economies investing more into other countries than the investments they receive from other countries. The independent variable export tax ranges anywhere from no tax to almost 44% tax in the data set, although the average still hovers around 1% - this shows a large diversity when it comes to export policy between the countries in the set. From the average, however, it seems that most countries in the data prefer a less restrictive approach towards exports.

The time to export dependent variable also follows a similar distribution - it is interesting to see that there are countries, for which the time to export is 0 hours. This is most likely due to countries that have trading agreements with the countries they export to (such as the EU), which allows for free travel of goods and therefore no time spent at customs. GDP per capita also varies greatly across time and countries as the

minimum is 380 dollars while the maximum is 108570 dollars - basing on the average, however, it seems a large portion of the countries are more on the developed side (this average is not the most reliable, however, as the standard deviation is almost three times the mean).

Model Results

Now that our balanced panel data set is ready, we can perform the pooled, between, fixed-effects and random-effects models and compare the results. These are presented in Table 2:

Table 2: Results of the different panel data models

	Dependent variable:			
	TimeToExport			
	(Pooled)	(Between)	(FE)	(RE)
Constant	55.992***	55.979***		55.356***
	(2.344)	(5.754)		(5.257)
ExportTax	2.177***	2.294**	0.081	0.354
	(0.376)	(0.942)	(0.348)	(0.328)
FDImillions	0.00002	0.00005	−0.00001	−0.00001
	(0.0001)	(0.0002)	(0.00002)	(0.00002)
GDPkcap	−0.842***	−0.843***	−0.120	−0.684***
	(0.085)	(0.208)	(0.344)	(0.178)
Observations	624	104	624	624
R^2	0.191	0.200	0.001	0.026
Adjusted R^2	0.187	0.176	-0.204	0.021
F Statistic	48.715*** (df = 3; 620)	8.345*** (df = 3; 100)	0.127 (df = 3; 517)	16.533***
Note:				*p<0.1; **p<0.05; ***p<0.01

As can be seen, there are vast differences in the estimated coefficients across the different models. Only the pooled model and the between model have have similar coefficients, although this is to be expected since the pooled model is made up largely of the between variation (and in a smaller part out of the within variation). We can also see that the Between model only has 104 observations, which is equivalent to the number of countries in the data set, since the model runs the regression on the averages of each of the variables per country.

Studying these two models in more detail, there are two variables that have significant coefficients - the ExportTax (expressed as a percentage of the tax revenue) and the GDPkcap variable (GDP per thousand capita). A percentage increase in the export tax is associated with a 2 hour increase in time to export ($p < 0.01$), while a unit increase in the GDP per thousand capita is associated with a 0.84 hour decrease in time to export ($p < 0.01$). This is in line with the assumptions in the population model section. Both models are jointly significant ($p < 0.01$) and have a relatively low R-squared (around 0.2).

Looking at the fixed-effects model, none of the coefficients are statistically significant - the standard error for each of the coefficients is high and in all cases at least twice as large as the coefficients. The values of the coefficients are low, with an export tax increase only being associated with a 0.08 hour increase in time to export and GDP per (thousand) capita increase being associated with a 0.1 decrease. The variables are also not jointly significant. Overall the model suggests no relationship between the chosen variables and time to export. The random effects model only has the variable GDPkcap as significant, with a unit increase being associated with a 0.68 hour decrease in time to export ($p < 0.01$). The model does remain jointly significant though ($F = 16.533, p < 0.01$).

Model Comparisons

We run a number of tests to compare the generated models and to decide which is the most suitable one. Firstly, the partial F-test is performed to compare the Between and Pooled model - since the result of the test is not significant, the pooled model is not rejected in favor of the Between model. A similar test is also run to compare the fixed-effects model with the pooled model - the results this time, however, are significant (p-value is close to zero), therefore the pooled model is rejected in favor of the fixed country effects model. Finally, we run Hausman's specification test to compare the fixed-effects and random-effects models. As the result is significant ($p = 0.03424$), it suggests a correlation between the disturbance of the model and explanatory variables, therefore only the fixed-effects model is consistent and is chosen as the final model (even though it suggests no relationship between the chosen variables).

Counts data modeling: the Mashable case

In the following section, an analysis will be performed to determine the variables, which determine the number of shares of Mashable articles, an online entertainment site.

Variable Selection and Summary Table for the Mashable study

There is a total of 58 possible variables that can be used to explain the variation in the number of shares a Mashable article receives in the data set. Based on a previous, similar study (Karnowski, Leiner, Sophie Kümpel, & Leonhard, 2021), which suggested that the length of an article, the news section, visualizations and the humour/emotional reactions to an article may determine whether an article is shared, we choose 9 corresponding variables for the analysis - the number of words in an article, the number of references, whether an article is in the entertainment section, number of images, number of videos, the polarity of the title, number of positive and negative words, and finally the total subjectivity of the article. The summary statistics for these is presented below:

Table 3: Summary Table of the Chosen Variables for the Mashable case

Statistic	N	Mean	St. Dev.	Min	Max
shares	39,644	3,395.380	11,626.950	1	843,300
words	39,644	546.515	471.108	0	8,474
references	39,644	10.884	11.332	0	304
entertainment	39,644	0.178	0.383	0	1
images	39,644	4.544	8.309	0	128
videos	39,644	1.250	4.108	0	91
title_polarity	39,644	0.071	0.265	−1.000	1.000
positive_words	39,644	0.040	0.017	0.000	0.155
negative_words	39,644	0.017	0.011	0.000	0.185
subjectivity	39,644	0.443	0.117	0.000	1.000

There are 39644 total articles in the data set. Looking at the number of shares, there is a large range in the number of shares an article gets in the data set, with the maximum reaching a staggering 843000 shares, while the minimum is 1 (possibly because there is always at least 1 share by the author of the article). On average, however, an article gets an impressive 3400 shares (although the standard deviation around this average is very high, more than triple the value, suggesting there are observations that skew the shares). It is surprising to see that the database has articles, which have no words (possibly an article only made up of videos or images) or have an absurdly large amount of words (more than 8000) (this could possibly be a pseudo-research paper). The first hypothesis seems to be confirmed when studying the maximum number of videos and images an article has, as there are articles, which have around 100 images/videos. The second seems to be confirmed when looking at the maximum number of references an article has in the database,

which is 304, which may be indicative of a more comprehensive study.

The entertainment statistics tell us that 17.8% of the articles in the data set belong in the entertainment section. When looking at the four variables, which were chosen to evaluate the emotional response an article may invoke (polarity, positive and negative words, subjectivity), we can see that most articles are written in relatively neutral language, with the average title polarity and rate of positive and negative words hovering around 0 (although more articles are written in a positive language). There is a large deviation around the title polarity mean, however, so it seems there are plenty of polarizing titles present as well. The articles are relatively subjective, scoring on average a 0.44 out of 1, which is understandable for an entertainment outlet.

Counts Data Model Specification and Results

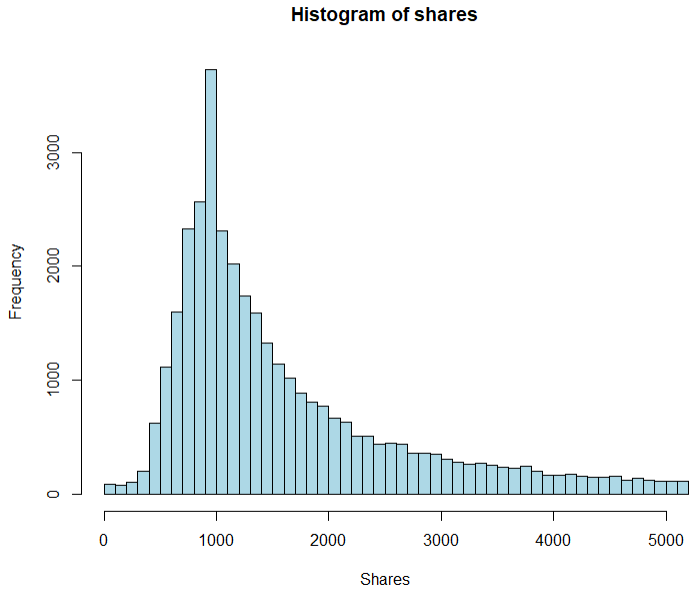

Figure 1: Histogram of the Shares Dependent Variable

Since our dependent variable depicts a frequency of an event (shares), only has positive values and is right skewed (see Figure 1), a counts data regression model using the Poisson distribution is deemed the most suitable. We can formally specify this relation with the below equation using log mean shares of an article:

$$\begin{aligned} ln\mu_i = \beta_0 + \beta_1 words + \beta_2 references + \beta_3 entertainment + \beta_4 images + \\ \beta_5 videos + \beta_6 title\_polarity + \beta_7 positive\_words + \beta_8 negative\_words + \\ \beta_9 subjectivity \end{aligned} \tag{2}$$

Where the μ stands for the mean of shares, i represents an individual article and β are the unobserved parameters estimated using the maximum likelihood method. Given that the Poisson assumptions (of the mean and variance of the dependent variable being equal) are quite strict and seem to be violated from the summary table (the standard deviation for shares is more than triple the mean), we also run the model with robust standard errors, a Quasi-Poisson model and a negative binomial model to account for the overdispersion. Now that we have the specification, we can run the actual model, which is presented in Table 4.

As can be seen, a large portion of the variables is highly significant across the models, although the exact number varies. The regular Poisson model estimates all the dependent variables as highly significant

Table 4: Counts Data Regression Model Results

	Dependent variable:			
	shares			
	Poisson		*glm: quasipoisson link = log*	*negative binomial*
	(Poisson)	(Robust)	(Quasi)	(NB)
Constant	7.976***	7.976***	7.976***	8.023***
	(0.0005)	(0.092)	(0.090)	(0.026)
words	−0.0002***	−0.0002**	−0.0002***	−0.0001***
	(0.00000)	(0.0001)	(0.00004)	(0.00001)
references	0.008***	0.008***	0.008***	0.009***
	(0.00001)	(0.001)	(0.001)	(0.001)
entertainment	−0.212***	−0.212***	−0.212***	−0.200***
	(0.0002)	(0.039)	(0.047)	(0.014)
images	0.013***	0.013***	0.013***	0.015***
	(0.00001)	(0.002)	(0.002)	(0.001)
videos	0.016***	0.016***	0.016***	0.026***
	(0.00002)	(0.003)	(0.003)	(0.001)
title_polarity	0.127***	0.127*	0.127**	0.129***
	(0.0003)	(0.071)	(0.063)	(0.020)
positive_words	−2.152***	−2.152*	−2.152*	−2.444***
	(0.006)	(1.112)	(1.117)	(0.347)
negative_words	0.615***	0.615	0.615	−0.817
	(0.008)	(1.289)	(1.606)	(0.516)
subjectivity	0.862***	0.862***	0.862***	0.666***
	(0.001)	(0.179)	(0.172)	(0.053)
Observations	39,644	39,644	39,644	39,644
Log Likelihood	-127,110,805.000	-127,110,805.000		-360,865.300
θ				0.925***(0.006)
Akaike Inf. Crit.	254,221,631.000	254,221,631.000		721,750.700
Note:				*p<0.1; **p<0.05; ***p<0.01

($p < 0.01$), while the model with the robust standard errors sees the rate of negative words as insignificant ($p = 0.615$), and the rate of positive words ($\beta = -2.152, p = 0.054$) and title polarity ($\beta = 0.127, p = 0.073$) as only moderately significant. These results are mostly mirrored by the Quasi-Poisson model, although in this model the title polarity is slightly more significant ($p = 0.044$) The negative binomial model, on the other hand, views all the variables as significant, except for the rate of negative words ($p = 0.114$).

What all the models agree on, however is the direction of these effects (e.g. number of words, rate of positive words, and entertainment section negatively affects the number of shares; references, images videos etc. positively affect the number of shares). As all these parameters are generated by the maximum likelihood, however, interpreting them in their current state is not straightforward and the partial effects need to be generated from them instead, which will be done in the following section. Before proceeding, however, we will choose the negative binomial distribution model, as the goodness of fit measures for maximum likelihood models, the Log Likelihood and AIC, are highest and lowest respectively for this model. I also performed the likelihood ratio test to compare the Poisson and Negative Binomial model, which turned a significant result, therefore the NB model is preferred.

OLS results and Average Partial Effects of Counts Model

Below, a table comparing the results of the OLS regression and the average partial effects of the negative binomial distribution model is constructed. The significance levels cannot be computed in R for the APE's so the significance symbols are borrowed from Table 4 results for easier comparison.

Table 5: OLS and Count Model Results Comparison

	Dependent variable:	
	shares	
	(OLS)	(APE - count)
Constant	$2,896.240^{***}$	$27,383.250^{***}$
	(292.100)	
words	-0.731^{***}	-0.450^{***}
	(0.143)	
references	37.007^{***}	29.968^{***}
	(5.985)	
entertainment	-710.845^{***}	-781.020^{***}
	(155.804)	
images	53.158^{***}	50.240^{***}
	(7.848)	
videos	76.266^{***}	89.104^{***}
	(14.821)	
title_polarity	451.444^{**}	439.961^{***}
	(224.864)	
positive_words	$-8,195.923^{**}$	$-8,342.625^{***}$
	(3,872.031)	
negative_words	1,409.028	$-2,788.534$
	(5,760.047)	
subjectivity	$2,854.390^{***}$	$2,272.947^{***}$
	(593.632)	
Observations	39,644	
R^2	0.005	
Adjusted R^2	0.005	
Residual Std. Error (df = 39634)	11,598.330	
F Statistic (df = 9; 39634)	22.765^{***}	
Note:		*p<0.1; **p<0.05; ***p<0.01

Comparing the results, we can see that the direction of the effects and the significance levels of the variables are the same across the two models ($p < 0.01$ for all variables except negative words). What differs, however, is the size of the effects (although certain variables, such as the number of images and title polarity have very similar effects across models). Overall, the R^2 of the OLS model is very low (0.005), suggesting that the linear regression model is not the best fit for the given data set.

First let us examine the negative effects. A unit increase in the number of words leads to 0.7 less shares under the OLS model, but only a 0.4 decrease under the count model. If an article is in the entertainment section, it gets on average 700-800 less shares (compared to the other article types such as social media, lifestyle etc.) across both models. Finally, a marginal increase in the rate of positive words actually leads to a staggering decrease of more than 8000 shares in both models (holding everything else constant). This suggests it may be more beneficial for articles to be in a neutral rather than positive language.

For the positive effects, these can be interpreted similarly, although some of the variables have large

differences in effects across models. In the counts model, the effect of an additional reference is 30 shares compared to OLS' 37, and a marginal increase in subjectivity is associated with 2272 extra shares compared to 2854 shares given by the OLS model. On the other hand, the count model suggests that videos have a higher effect on the number of shares (89 shares) than the OLS model (only 76 extra shares), holding all else constant. On the whole, it could be said that the count model is slightly more conservative with its estimates, but most of the coefficients are relatively close to each other.

Ordinal logistic data modeling: the Yelp case

In the following section, the influence of a number of variables on the online ratings of restaurants will be studied, using data from the site Yelp.

Summary Statistics of the Yelp data

Besides the dependent variable, which are the review stars in the data set, an additional dummy variable Five Stars is created for analysis, which takes the value 1 if the review receives 5 stars and 0 if the review receives less - this is to account for the positively skewed nature of most rating scales. Additional to the mandated travel variable, four other independent variables are chosen for analysis:

- length of the review - we assume that the length is negatively correlated with the ratings (i.e. negative reviews are more likely to be longer since the person is experiencing a stronger emotional response)
- number of fans - we expect the number of fans to be negatively correlated with the ratings as the user may start feeling more responsible for his fans and is therefore stricter in their reviews
- years the user has been part of the Elite program - this indicates the level of experience the user has had with eating at restaurants and we assume that with more experience, their reviews become more negative/stricter
- price range of the restaurant - we expect restaurants to have better food and provide a better experience the more expensive they are, which should result in more positive reviews

For the travel variable, it is assumed to be positively correlated with the reviews, as travelers may be more lenient with their reviews, and usually tend to seek the best restaurants at their travel destination. As a caveat, although the price range of restaurants is technically an ordinal variable, in the analysis it is treated as a quantitative variable for the sake of simplicity. The summary statistics for our chosen variables are presented in Table 6:

Table 6: Summary Table of the Chosen Variables for the Yelp case

Statistic	N	Mean	St. Dev.	Min	Max
review_stars	156,521	3.710	1.176	1	5
dFiveStars	156,521	0.293	0.455	0	1
travel	156,521	0.104	0.305	0	1
length	156,521	752.267	652.334	1	5,000
fans	156,521	32.526	84.051	0	722
years_elite	156,521	2.485	3.005	0	12
price_range	155,827	1.795	0.607	1	4

There are a total of 156521 reviews in the data set. Looking at the average of review stars dependent variable, we see that the initial assumption that reviews are generally skewed towards the higher end holds true (as the number is close to 3.7 compared to the expected 3.0). As expected, the reviews can also take on ratings between 1 and 5. The Five Stars dummy variable also confirms this observation as almost 30% of all reviews have a 5-star rating compared to the expected 20%. It also seems that a majority of the reviews

takes place in the reviewers' home city, judging based on the mean of the travel variable (0.104). From the length variable, we can see that the mandatory length of a review is at least one word, and the average review is around 750 words long. It is interesting to see that there even exist 5000 word restaurant reviews.

The number of fans greatly varies from user to user, but it is surprising to observe that the largest amount of fans a user has is 722 since with the rise of social media and Internet usage, I would have expected the number to be higher. Studying the elite years variable, we see the highest number of years a user has in the program is 12 in the data set. From the average, it also seems most users have some experience with the program (2.485 years), but based on the standard deviation, this may be because of highly influential extremes. The price range statistics suggest that most of the reviewed restaurants are on the less expensive side, with the mean being 1.8 on a 4-point scale.

Formal Specification of the Models

Since both our dependent variables are not continuous and rather ordinal and binary variables respectively, we cannot apply the OLS model and rather use the ordinal response and binary choice regression models. In the below specifications, we also include an interaction term between the travel and elite years variable to study how reviewing a restaurant outside the home city affects the relationship between a user's expertise and their ratings. Both these models are based on maximum likelihood estimators, which have a corresponding latent regression model and an index function, presented below.

The ordinal response model:

$$\begin{aligned} review\_stars* = \beta_0 + \beta_{11} years\_elite + \beta_{12} years\_elite : travel + \beta_2 travel + \beta_3 length + \\ \beta_4 fans + \beta_5 price\_range + \epsilon \end{aligned} \tag{3}$$

$$review\_stars = \begin{cases} 1, \text{if } review\_stars* \leq \mu_1 \\ 2, \text{if } \mu_1 < review\_stars* \leq \mu_2 \\ 3, \text{if } \mu_2 < review\_stars* \leq \mu_3 \\ 4, \text{if } \mu_3 < review\_stars* \leq \mu_4 \\ 5, \text{if } review\_stars* > \mu_4 \end{cases} \tag{4}$$

This ordinal response model has four intercepts due to there being five possible ratings.

The binary choice model:

$$\begin{aligned} dFiveStars* = \beta_0 + \beta_{11} years\_elite + \beta_{12} years\_elite : travel + \beta_2 travel + \beta_3 length + \\ \beta_4 fans + \beta_5 price\_range + \epsilon \end{aligned} \tag{5}$$

$$dFiveStars = \begin{cases} 1, \text{if } dFiveStars* > c \\ 0, otherwise \end{cases} \tag{6}$$

Comparison of the Models and Interpretation of Results

Now that we have our formal specifications with the variables to be included in the regression, we can run the two models both as logit (using logistic distribution for the maximum likelihood estimator) and probit models (using normal distribution for the maximum likelihood estimator). For the ordinal model, we also add in the four different intercepts, which separate the five rating options (between 1 and 2, 2 and 3 etc.) as well as the goodness of fit estimators of log likelihood and AIC. Results are presented below in Table 7:

Table 7: Logit and Probit Models of the Ordinal Response and Binary Choice Models

	Dependent variable:			
	review_stars		dFiveStars	
	ordered logistic	*ordered probit*	*logistic*	*probit*
	(1)	(2)	(3)	(4)
Constant			-0.748^{***}	-0.475^{***}
			(0.019)	(0.011)
years_elite	-0.002	0.004^{***}	-0.073^{***}	-0.044^{***}
	(0.002)	(0.001)	(0.002)	(0.001)
travel1	0.144^{***}	0.053^{***}	0.270^{***}	0.169^{***}
	(0.019)	(0.011)	(0.020)	(0.012)
length	-0.0003^{***}	-0.0002^{***}	-0.0003^{***}	-0.0002^{***}
	(0.00001)	(0.00000)	(0.00001)	(0.00001)
fans	0.0003^{***}	0.0002^{***}	-0.0004^{***}	-0.0002^{***}
	(0.0001)	(0.00004)	(0.0001)	(0.0001)
price_range	0.130^{***}	0.079^{***}	0.139^{***}	0.084^{***}
	(0.008)	(0.005)	(0.009)	(0.006)
years_elite:travel1	-0.010^{*}	-0.0004	-0.021^{***}	-0.014^{***}
	(0.006)	(0.004)	(0.008)	(0.005)
mu.1	-2.693	-1.512		
	0.018	0.01		
mu.2	-1.66	-0.977		
	0.016	0.009		
mu.3	-0.627	-0.371		
	0.016	0.009		
mu.4	0.881	0.56		
	0.016	0.009		
Observations	155,827	155,827	155,827	155,827
Log Likelihood	-224518.855	-224614.906	-92,050.510	-92,093.810
Akaike Inf. Crit.	449057.71	449249.811	184,115.000	184,201.600

Note: *p<0.1; **p<0.05; ***p<0.01

Let us first examine the ordered model results. The estimated logit and probit models have both different values and even different significance levels for certain variables. For example, the effect of the travel and price_range on the rating is around twice as large numerically under the logit model than under the probit model. Both models return the travel, length, fans and price range variables as highly significant ($p < 0.01$), while disagreeing on the significance levels of the years_elite variable (highly significant under probit but insignificant under logit) and the interaction term (moderately significant under the logit model, insignificant under the probit model). Both of these models have comparable AIC and Log Likelihood scores, although the AIC is lower for the logit model, making it the slightly preferred model.

Examining the results of the binary choice model, here the results are more comparable and both models

agree on the direction and significance levels of all variable coefficients as they are all highly significant ($p < 0.01$) - similarly to before the AIC is slightly lower for logit, making it the preferred model. While it seems some effects are stronger under one model than the other, not much else can be said, as it is difficult to interpret the exact effects on our dependent variable without also calculating the marginal effects - these are presented in Table 8 below:

Table 8: APE's for the binary choice model

	APElogitbin.1	APEprobitbin.1
(Intercept)	-0.151	-0.160
years_elite	-0.015	-0.015
travel1	0.047	0.048
length	-0.0001	-0.0001
fans	-0.0001	-0.0001
price_range	0.028	0.028
years_elite:travel1	-0.004	-0.005

The above table gives us the average effect of a marginal increase in an explanatory variable on the probability of the dependent binary variable dFiveStars - for example, we see that an additional year on the Elite program decreases the probability of a review getting five stars by 1.5% (for restaurant reviews in the user's home city). Interestingly, for restaurant reviews outside the home city, this effect is slightly stronger, with an additional elite year decreasing the probability by 1.9%. Other variables can be interpreted similarly. An interesting observation is also that, despite the logit and probit model results differing quite a bit in Table 7, their APE's are almost the same. All the APE's follow the direction set in the initial assumptions.

Table 9: APE's for the Ordinal Logit Model

	years_elite	travel	length	fans	price_range	years_elite:travel
P(y=1)	0.0001	−0.009	0.00002	−0.00002	−0.008	0.001
P(y=2)	0.0001	−0.011	0.00003	−0.00002	−0.010	0.001
P(y=3)	0.0002	−0.013	0.00003	−0.00002	−0.012	0.001
P(y=4)	−0.00004	0.003	−0.00001	0	0.003	−0.0002
P(y=5)	−0.0004	0.029	−0.0001	0.00005	0.027	−0.002

Table 10: APE's for the Ordinal Probit Model

	years_elite	travel	length	fans	price_range	years_elite:travel
P(y=1)	−0.001	−0.007	0.00003	−0.00002	−0.010	0.00005
P(y=2)	−0.0005	−0.006	0.00002	−0.00002	−0.009	0.00005
P(y=3)	−0.0005	−0.007	0.00002	−0.00002	−0.010	0.00005
P(y=4)	0.0001	0.001	−0.00001	0	0.002	−0.00001
P(y=5)	0.001	0.018	−0.0001	0.00005	0.027	−0.0001

When inspecting the two tables, similarly to the binary model APE table, the shared significant variables travel, length, fans and price_range have similar effects between the probit and logit APE's. For example, in both the probit and logit models, an increase in fans leads to a 0.002% decrease in the probability of a review having 1 star. Interestingly, though, it leads to a 0.005% increase in the probability of a review having 5 stars in both tables, which goes against our initial assumptions. This positive effect for the probability of 5 stars can be also observed across the variables travel and price range, which is in line with our assumptions. Length also decreases this probability by 0.01%, which follows initial assumptions.

References

Karnowski, V., Leiner, D. J., Sophie Kümpel, A., & Leonhard, L. (2021). Worth to share? how content characteristics and article competitiveness influence news sharing on social network sites. *Journalism & Mass Communication Quarterly*, *98*(1), 59–82.

Appendix - R code

```
####Session set-up####
library(tidyverse)
library(stargazer)
library(AER)
library(plyr)
library(plm)
library(curl)
library(wbstats)
library(MASS)
library(sandwich)
library(extraDistr)
library(pscl)
library(lmtest)



####Defining paths and directories####
dir <- "C:/Users/tnguyen10/Private/School/Block 1/ASAP/Assignment 3/"
dirProg <- paste0(dir, "Programs/")
dirData <- paste0(dir, "Data/")
dirResults <- paste0(dir, "Results/")

#-------------------------------------------------------------------------------
# Part 1
#-------------------------------------------------------------------------------

####Part 1 - Task 1 - Data preparation####

# Download selected data from the portal and store
# the data in dataframe dfTime2Export

dfTime2Export <-
  wb_data(indicator = c("IC.EXP.TMBC",
                        "NY.GDP.PCAP.KD",
                        "GC.TAX.EXPT.ZS","BN.KLT.DINV.CD"),
          country = "countries_only",
          start_date = 1960,
          end_date = 2021)
sum(is.na(dfTime2Export$IC.EXP.TMBC))
sum(is.na(dfTime2Export$IS.SHP.GOOD.TU))
sum(is.na(dfTime2Export$TRD.ACRS.BRDR.EXPT.COST.CD.DB0615.DFRN))
sum(is.na(dfTime2Export$GC.TAX.EXPT.ZS))

dfIndicators <- wb_indicators()

save <- dfTime2Export
##Indicators to work with: TRD.ACRS.BRDR.EXPT.COST.CD.DB0615.DFRN (Cost to export
#per container), 6.0.GDP_usd (GDP in 2005 US £), LP.LPI.CUST.XQ (Efficiency of customs clearance)

##Renaming the columns##
colnames(dfTime2Export)[colnames(dfTime2Export) == "IC.EXP.TMBC"] <- "TimeToExport"
colnames(dfTime2Export)[colnames(dfTime2Export) == "NY.GDP.PCAP.KD"] <- "GDPcap"
```

```r
colnames(dfTime2Export)[colnames(dfTime2Export) == "GC.TAX.EXPT.ZS"] <- "ExportTax%"
colnames(dfTime2Export)[colnames(dfTime2Export) == "BN.KLT.DINV.CD"] <- "FDI"

##removing unneeded columns
dfTime2Export$iso2c <- NULL
dfTime2Export$iso3c <- NULL


####Part 1 - Task 2####
##Subsetting on non-missing values of TimeToExport
dfTime2Export.sub <- filter(dfTime2Export, !is.na(TimeToExport))
##Taking only complete observations for this subset
dfTime2Export.sub <- dfTime2Export.sub[complete.cases(dfTime2Export.sub),]
##Limiting the complete observations to the time between 2014 and 2019
dfTime2Export.sub <- dfTime2Export.sub[(dfTime2Export.sub$date >= 2014 & dfTime2Export.sub$date <= 2019)
##Creating a new column, which counts the number of observations per country
dfTime2Export.sub <- dfTime2Export.sub %>%
  group_by(country) %>%
  dplyr :: mutate(count = n())
##Creating a balanced dataset - only complete cases, where each country has 6 observations (2014-2019)
dfTime2Export.sub <- dfTime2Export.sub[dfTime2Export.sub$count == 6,]

##Changing the units of some variables to make regressions more legible
dfTime2Export.sub$FDImillions <- dfTime2Export.sub$FDI/1000000
dfTime2Export.sub$GDPkcap <- dfTime2Export.sub$GDPcap/1000

##Saving the subsetted and making changes
dfexporttime <- subset(dfTime2Export.sub, select = -c(FDI,GDPcap))
dfexporttime$iso2c <- NULL
dfexporttime$iso3c <- NULL
dfexporttime <- dfexporttime[,c(1,2,4,3,6,7,5)]
colnames(dfexporttime)[colnames(dfexporttime) == "ExportTax%"] <- "ExportTax"



##Creating a stargazer table
stargazer(as.data.frame(dfexporttime))


####Part 1 - Task 3####
##Formulating the model##
mdl <- TimeToExport ~ ExportTax + FDImillions + GDPkcap

##All the regression models##
rsltPool <- plm(mdl, data = dfexporttime, model = "pooling")
rsltBetween <- plm(mdl, data = dfexporttime, model = "between")
rsltFE <- plm(mdl, data = dfexporttime, model = "within")
rsltRE <- plm(mdl, data = dfexporttime, model = "random")
stargazer(rsltPool, rsltBetween, rsltFE, rsltRE, align = TRUE, no.space = TRUE,
          intercept.bottom =  FALSE, type = "text")
#only GDP stays consistent over the between, pooled, random,

####Part 1 - Task 4####
pFtest(rsltBetween, rsltPool)
```

```r
#results of the test not significant - pooled model is not rejected in favor of the between model
pFtest(rsltFE, rsltPool)
#results of the test significant - pooled model rejected in favor of the fixed country effects model
phtest(rsltFE, rsltRE)
#the assumption of the independence of the unobserved individual effects and explanatory variables
#(the countries) is incorrect and leads to RE model being inconsistent; FE model is chosen


#------------------------------------------------------------------------------
# Part 2
#------------------------------------------------------------------------------

####Part 2 - Task 1####
##Loading the data##
dfmash <- read.csv(file = paste0(dirData, "OnlineNewsPopularity.csv"), header = TRUE)
dfmash.sub <- subset(dfmash, select = c(shares, n_tokens_content, num_hrefs,
                                        data_channel_is_entertainment,
                                       num_imgs, num_videos, title_sentiment_polarity,
                                       global_rate_positive_words, global_rate_negative_words,
                                       global_subjectivity))
colnames(dfmash.sub) <- c("shares","words","references","entertainment","images",
                          "videos","title_polarity","positive_words","negative_words",
                          "subjectivity")
stargazer(dfmash.sub, type = "text")
str(dfmash.sub)

####Part 2 - Task 2####
hist(dfmash.sub$shares, breaks = 12000, col = "lightblue", main = "Histogram of shares",
     xlim = c(0,5000), xlab = "Shares", ylab = "Frequency")
?hist
##as can be seen from the histogram, not normal distribution - distribution is
#skewed to the right
#the dependent variable is also a discrete and positive number, which denotes a
#frequency of an event (sharing the article)
#therefore a counts data regression model is chosen


##define the model based on the chosen variable
mdlA <- shares ~ words + references + entertainment + images + videos +
  title_polarity + positive_words + negative_words + subjectivity

##Poisson Model
rsltPoisson <- glm(mdlA, data = dfmash.sub, family = c("poisson"))

#May not be the best choice to use the strict Poisson model when we look at the
#standard deviation of the dependent variable shares
#robust standard models, quasipoisson and negative binomial distribution models
rsltQuasi <- glm(mdlA, data = dfmash.sub, family = c("quasipoisson"))
rsltNegBin <- glm.nb(mdlA, data = dfmash.sub)
seWhite <- sqrt(diag(vcovHC(rsltPoisson, type = "HC0")))

stargazer(rsltPoisson, rsltPoisson, rsltQuasi, rsltNegBin, align = TRUE,
          no.space = TRUE, intercept.bottom = FALSE,
          se = list (NULL, seWhite, NULL, NULL), report=('vc*p'), type = "text")
```

```r
lrtest(rsltPoisson, rsltNegBin)

#from the histogram, you can also see that up to 1000, the shares are given in
#units, after that point in hundreds - this causes a large jump
#at 1000 shares, and a change in the distribution from that point forward - we
#could possibly do hurdle model



####Part 2 - Task 3####
#We need to calculate partial effects to be able to interpret the results (using
#the negative binomial distribution)
estBeta <- coef(rsltNegBin)
estBeta

#Calculating the average partial effects
APE <- mean(exp(predict.glm(rsltNegBin, type = "link")))*estBeta
round(APE, 3)
#average changes in the mean counts due to a small change in the explanatory variable

#We have one dummy variable (entertainment), so we have to calculate the partial
#effect for this one separately
tmp <- dfmash.sub

tmp$entertainment <- 1
tmpAPE.1 <- mean(exp(predict.glm(rsltNegBin, newdata = tmp, type = "link")))

tmp$entertainment <- 0
tmpAPE.0 <- mean(exp(predict.glm(rsltNegBin, newdata = tmp, type = "link")))

APE.private <- tmpAPE.1 - tmpAPE.0
APE.private

#replacing the coefficient for entertainment with the true APE
APE["entertainment"] <- APE.private

##creating an OLS model for comparison
rsltOLS <- lm(mdlA, data = dfmash.sub)

##creating a complete table comparing the two
stargazer(rsltOLS, rsltOLS, APE, align = TRUE, no.space = TRUE,
          intercept.bottom = FALSE, type = "text")
save <- APE
APE <- save
names(APE)[names(APE) == "(Intercept)"] <- "Constant"

?stargazer

#-------------------------------------------------------------------------------
# Part 3
#-------------------------------------------------------------------------------
####Part 3 - Task 1####
##Loading in the data and creating a summary table##
```

```
dfyelp <- read.csv(file = paste0(dirData, "online_ratings_travel.csv"), header = TRUE, sep = ";")
dfyelp$dFiveStars <- case_when(dfyelp$review_stars == 5 ~ 1, dfyelp$review_stars < 5 ~ 0)
dfyelp.sub <- subset(dfyelp, select = c(review_stars, dFiveStars, travel, length, fans, years_elite, pri
str(dfyelp.sub)

stargazer(dfyelp.sub, type = "text")
dfyelp.sub$review_stars <- as.factor(dfyelp.sub$review_stars)
```

```
####Part 3 - Task 2####
##Formalizing the model - can study the interaction of travel on years_elite -
#maybe it weakens the relationship?##
mdlOrd <- review_stars ~ years_elite*travel + length + fans + price_range
mdlBin <- dFiveStars ~ years_elite*travel + length + fans + price_range
```

```
####Part 3 - Task 3####
##Ordinal response model##
rsltOrd.Logit <- polr(mdlOrd, data = dfyelp.sub, method = "logistic")
rsltOrd.Probit <- polr(mdlOrd, data = dfyelp.sub, method = "probit")

summary(rsltOrd.Logit)
stargazer(rsltOrd.Logit, type = "text")
##Four intercepts - four logistic regressions
#interaction term is significant and negative - elite years seems to have a
#small negative non-significant effect on the review stars when at home/not travelling
#but when travelling this effect is more negative - so elite years in that case
#actually decreases the ratings more it seems like

add.lnL <- c("lnL", round(logLik(rsltOrd.Logit),3),
                    round(logLik(rsltOrd.Probit),3))
add.Aic <- c("AIC", round(AIC(rsltOrd.Logit),3),
                    round(AIC(rsltOrd.Probit),3))

#Creating intercepts
est.Logit <- summary(rsltOrd.Logit)$coefficients
est.Probit <- summary(rsltOrd.Probit)$coefficients

add.mu1.est <- c("mu.1",
                 round(est.Logit[nrow(est.Logit)-3,  "Value"], 3),
                 round(est.Probit[nrow(est.Logit)-3, "Value"], 3)
)
add.mu1.std <- c("",
                 round(est.Logit[nrow(est.Logit)-3,  "Std. Error"], 3),
                 round(est.Probit[nrow(est.Logit)-3, "Std. Error"], 3)
```

```
)
add.mu2.est <- c("mu.2",
                 round(est.Logit[nrow(est.Logit)-2,  "Value"], 3),
                 round(est.Probit[nrow(est.Logit)-2, "Value"], 3)
)
add.mu2.std <- c("",
                 round(est.Logit[nrow(est.Logit)-2,  "Std. Error"], 3),
                 round(est.Probit[nrow(est.Logit)-2, "Std. Error"], 3)
)
add.mu3.est <- c("mu.3",
                 round(est.Logit[nrow(est.Logit)-1,  "Value"], 3),
                 round(est.Probit[nrow(est.Logit)-1, "Value"], 3)
)
add.mu3.std <- c("",
                 round(est.Logit[nrow(est.Logit)-1,  "Std. Error"], 3),
                 round(est.Probit[nrow(est.Logit)-1, "Std. Error"], 3)
)
add.mu4.est <- c("mu.4",
                 round(est.Logit[nrow(est.Logit),  "Value"], 3),
                 round(est.Probit[nrow(est.Logit), "Value"], 3)
)
add.mu4.std <- c("",
                 round(est.Logit[nrow(est.Logit),  "Std. Error"], 3),
                 round(est.Probit[nrow(est.Logit), "Std. Error"], 3)
)

stargazer(rsltOrd.Logit, rsltOrd.Probit,
          align=TRUE, no.space = TRUE, intercept.bottom = TRUE, type="text",
          add.lines = list(add.mu1.est, add.mu1.std,
                           add.mu2.est, add.mu2.std,
                           add.mu3.est, add.mu3.std,
                           add.mu4.est, add.mu4.std,
                           add.lnL, add.Aic))




##Binary model##
rsltBin.Logit <- glm(mdlBin, data = dfyelp.sub,
                     family = binomial(link = "logit"))
rsltBin.Probit <- glm(mdlBin, data = dfyelp.sub,
                      family = binomial(link = "probit"))


##Putting all the results into stargazer##
stargazer(rsltOrd.Logit, rsltOrd.Probit, rsltBin.Logit, rsltBin.Probit,
          align = TRUE, no.space = TRUE, intercept.bottom = FALSE, type = "text",
          add.lines = list(add.mu1.est, add.mu1.std,
                           add.mu2.est, add.mu2.std,
                           add.mu3.est, add.mu3.std,
                           add.mu4.est, add.mu4.std,
                           add.lnL, add.Aic))
```

```r
##Interpretation table - how much does the probability increase by change of the variable?
##Ordinal APE - code frome the lecture
prb.Logit  <- as.data.frame(predict(rsltOrd.Logit,  type="probs"))
prb.Probit <- as.data.frame(predict(rsltOrd.Probit, type="probs"))

# Calculate the cumulative probabilities (as an intermediate
# step, for code transparency)
cdf.Logit.1 <- prb.Logit[, 1]
cdf.Logit.2 <- prb.Logit[, 1] + prb.Logit[, 2]
cdf.Logit.3 <- prb.Logit[, 1] + prb.Logit[, 2] + prb.Logit[, 3]
cdf.Logit.4 <- prb.Logit[, 1] + prb.Logit[, 2] + prb.Logit[, 3] + prb.Logit[,4]
cdf.Logit.5 <- prb.Logit[, 1] + prb.Logit[, 2] + prb.Logit[, 3] + prb.Logit[,4] +
               prb.Logit[, 5]

cdf.Probit.1 <- prb.Probit[, 1]
cdf.Probit.2 <- prb.Probit[, 1] + prb.Probit[, 2]
cdf.Probit.3 <- prb.Probit[, 1] + prb.Probit[, 2] + prb.Probit[, 3]
cdf.Probit.4 <- prb.Probit[, 1] + prb.Probit[, 2] + prb.Probit[, 3] + prb.Probit[,4]
cdf.Probit.5 <- prb.Probit[, 1] + prb.Probit[, 2] + prb.Probit[, 3] + prb.Probit[,4] +
                prb.Probit[, 5]

# ... cdf.Logit.5 and cdf.Probit.5 should be equal to 1

# Calculate density parts of the effects (Greene, p.910)
prb.Logit$pdf.1 <-
  -dlogis(qlogis(cdf.Logit.1))
prb.Logit$pdf.2 <-
  dlogis(qlogis(cdf.Logit.1)) - dlogis(qlogis(cdf.Logit.2))
prb.Logit$pdf.3 <-
  dlogis(qlogis(cdf.Logit.2)) - dlogis(qlogis(cdf.Logit.3))
prb.Logit$pdf.4 <-
  dlogis(qlogis(cdf.Logit.3)) - dlogis(qlogis(cdf.Logit.4))
prb.Logit$pdf.5 <-
  dlogis(qlogis(cdf.Logit.4))


prb.Probit$pdf.1 <-
  -dnorm(qnorm(cdf.Probit.1))
prb.Probit$pdf.2 <-
  dnorm(qnorm(cdf.Probit.1)) - dnorm(qnorm(cdf.Probit.2))
prb.Probit$pdf.3 <-
  dnorm(qnorm(cdf.Probit.2)) - dnorm(qnorm(cdf.Probit.3))
prb.Probit$pdf.4 <-
  dnorm(qnorm(cdf.Probit.3)) - dnorm(qnorm(cdf.Probit.4))
prb.Probit$pdf.5 <-
  dnorm(qnorm(cdf.Probit.4))

# Determine the average effects (apart from the estimated
# effects)
avgAPE.Logit  <- colMeans(prb.Logit[c("pdf.1", "pdf.2", "pdf.3", "pdf.4", "pdf.5")])
```

```r
avgAPE.Probit <- colMeans(prb.Probit[c("pdf.1", "pdf.2", "pdf.3", "pdf.4", "pdf.5")])

# Extract the estimated effects from the logit
# and probit objects
coef.Logit  <- coef(rsltOrd.Logit)
coef.Probit <- coef(rsltOrd.Probit)

# Determine the APE for ordinal
dfAPE.Logit  <- as.data.frame(round(avgAPE.Logit %*% t(coef.Logit), 5))
dfAPE.Probit <- as.data.frame(round(avgAPE.Probit %*% t(coef.Probit), 5))

rownames(dfAPE.Logit) <- rownames(dfAPE.Probit) <-
  c("P(y=1)", "P(y=2)", "P(y=3)", "P(y=4)", "P(y=5)")

stargazer(dfAPE.Logit, summary = FALSE,
          align = TRUE, no.space = TRUE, type="text")
stargazer(dfAPE.Probit, summary = FALSE,
          align = TRUE, no.space = TRUE, type="text")
```

```r
##APE table for binary
betaBinLogit <- coefficients(rsltBin.Logit)
betaBinProbit <- coefficients(rsltBin.Probit)
APElogitbin.1 <- mean(dlogis(predict(rsltBin.Logit, type = "link")))*betaBinLogit
APEprobitbin.1 <- mean(dnorm(predict(rsltBin.Probit, type = "link")))*betaBinProbit

#creating it separately for travel
tmp <- dfyelp.sub
str(tmp)

tmp$travel <- 0

tmpAPEbinlogit.0 <- mean(predict(rsltBin.Logit, newdata = tmp, type = "response"), na.rm = TRUE)
tmpAPEbinprobit.0 <- mean(predict(rsltBin.Probit, newdata = tmp, type = "response"), na.rm = TRUE)

tmp$travel <- 1

tmpAPEbinlogit.1 <- mean(predict(rsltBin.Logit, newdata = tmp, type = "response"), na.rm = TRUE)
tmpAPEbinprobit.1 <- mean(predict(rsltBin.Probit, newdata = tmp, type = "response"), na.rm = TRUE)

APElogitbin.1["travel1"] <- tmpAPEbinlogit.1 - tmpAPEbinlogit.0
APEprobitbin.1["travel1"] <- tmpAPEbinprobit.1 - tmpAPEbinprobit.0

stargazer(round(cbind(APElogitbin.1, APEprobitbin.1),4), type = "text")
```