Individual Assignment 3

Eran Kan Kohen

13th October 2022

1 Panel Data Modelling

1.1 Setting up the Dataframe and Creating a Population Model

This section will focus on trying to find the factors which correlate with the time it takes to a country to export products. 3 separate variables were combined to create a hypothetical population model. The Carbon emissions for the transport sector, the percentage of STEM graduates, and the cost to export were all added into the model. It it expected that when more carbon emissions are emitted, the transport will be faster. More STEM graduates will indicate that the country is focusing more on efficiency and engineering, decreasing the time to export. Finally, more expensive exports has the potential to be the shortest. The following sections will investigate these relationships.

$$\begin{aligned}\text{timeExport}_{\text{it}} = \alpha_{it} &+ \beta_{1,it}\,\text{co2trans}_{\text{it}} \\ &+ \beta_{2,it}\,\text{percentageSTEMgrads}_{\text{it}} \\ &+ \beta_{3,it}\,\text{costExport}_{\text{it}} + \text{u}_{\text{it}}\end{aligned}$$

1.2 Data Processing and Presenting the Summary

The variables were imported from the World Bank database. Afterwards, in order to make sure the data is balanced and fits in with the desired time frame, several processing actions were taken. Data points outside the 2014-2019 time frame were removed. Later, rows without the dependent variable of time to export were deleted. Finally, the "iso" variables were removed since they were not going to be used.

A summary of the remaining variables is presented below:

Table 1:

Statistic	N	Mean	St. Dev.	Min	Max
Year	1,122	2,016.504	1.708	2,014	2,019
co2_trans	899	36.571	146.886	0.020	1,762.240
cost_export	1,122	396.525	354.247	0.000	2,222.692
time_export	1,122	58.118	59.413	0.000	515.038
perc_STEM_grad	411	22.094	7.213	0.000	46.270

As seen above, the data has complete cases with the variables of cost to export, time to export and the year. On the other hand, the variables concerning the carbon emission in the transportation sector and the percentage of STEM graduates are not are not complete. This has the potential to negatively affect the model. Other variables which would not have missing values were searched for the model however after trying more than 20 variables none were found, hence this model was kept.

Moreover, as seen through the means and the maximum values, most of the metrics are skewed to the right, with the mean value being much lower than the maximum value. This is a very interesting finding which indicates that a small proportion of countries have very extreme cases of being very slow in exporting.

1.3 Comparing Regression Models

Several different regression models were created to analyse the relationship. The models were added to a single table and compared below:

Table 2:

	Dependent variable:				
	time_export				
	OLS			*panel linear*	
	(1)	(2)	(3)	(4)	(5)
Constant	3.412	-3.869	0.854**		29.211***
	(6.842)	(12.033)	(0.402)		(6.709)
co2_trans	-0.019*	-0.020	-0.094	-0.130**	-0.034
	(0.012)	(0.022)	(0.079)	(0.053)	(0.022)
perc_STEM_grad	0.183	0.453	-0.476	-0.465***	-0.361*
	(0.268)	(0.466)	(0.293)	(0.178)	(0.192)
cost_export	0.115***	0.125***	0.037**	0.031***	0.091***
	(0.006)	(0.010)	(0.016)	(0.010)	(0.008)
Observations	368	116	368	368	368
R^2	0.509	0.608	0.030	0.104	0.356
Adjusted R^2	0.505	0.597	0.023	-0.300	0.351
Residual Std. Error	35.588 (df = 364)	38.375 (df = 112)	7.703 (df = 364)		
F Statistic	125.916*** (df = 3; 364)	57.823*** (df = 3; 112)	3.816** (df = 3; 364)	9.750*** (df = 3; 253)	152.190***
Note:					*p<0.1; **p<0.05; ***p<0.01

The columns above respectively correspond to pooled (1), between (2), within (3), fixed-effects (4), and random-effects (5) models. Unlike the expectations, the within model and the fixed-effect models are not the same, which is probably due to several missing values in the model.

Furthermore, investigating the table revealed that the cost to export was a significant variable in all of the models, highlighting the variable correlating with the time to export. It seems that as one increases, the other increases as well which is quite interesting. This indicates that the more money spent in the export, the longer it takes. It is possible that the initial hypothesis was wrong and that more money spent in export indicates a more cumbersome process, hence, the longer export periods.

Moreover, the R^2 values were quite high for the pooled and between models, and quite low for the within and fixed-effects models. This highlights that the variation is explained quite well with the initial two models, unlike the latter two models. The random effects model has a decent R^2 value. It

also included the percentage of STEM graduates as a significant variable.

Finally, it is seen that the fixed-effects model is the only model with all of the independent variables with significant relationships with the dependent variable.

1.4 Testing the Models

Following the comparison of the regression models, several tests were conducted to further investigate the models.

A pooltest between the pooled model and the fixed-effects model had a significant result. This indicates that the fixed-effects model should be preferred.

Afterwards, a Hausman test was used to compare the fixed and random effects models. Again, a highly significant test statistic and high chi square showed that the fixed-effects model should be preferred. Hence, even though it didn't have the highest R^2 value, the fixed-effects model is preferred among the other models.

2 Counts Data Modelling

2.1 Summary

The second part of this report will focus on data from the online website Mashable. It will aim to analyse the data to understand which factors might correlate with the number of times each article has been shared. 9 unique explanatory variables were selected and the other variables were removed from the data. The summary of the selected variables are presented below:

Table 3:

Statistic	N	Mean	St. Dev.	Min	Max
shares	39,644	3,395.380	11,626.950	1	843,300
is_weekend	39,644	0.131	0.337	0	1
rate_positive_words	39,644	0.682	0.190	0.000	1.000
rate_negative_words	39,644	0.288	0.156	0.000	1.000
num_imgs	39,644	4.544	8.309	0	128
num_videos	39,644	1.250	4.108	0	91
n_non_stop_words	39,644	0.996	5.231	0.000	1,042.000
n_tokens_title	39,644	10.399	2.114	2	23
num_hrefs	39,644	10.884	11.332	0	304
data_channel_is_entertainment	39,644	0.178	0.383	0	1

As seen above, a relatively large data set is available with information on almost 40,000 articles available. The variable shares is the dependent variable and the remaining 9 are independent variables. "is_weekend" variable is a binary variable indicating if the article was written on the weekend. As seen through the mean and standard deviation, most articles are not written on the weekends.

The variables "rate_positive_words" and "rate_negative_words" show the rate of positive and negative words among non-neutral tokens. This means that the values do not indicate the rate of positive or negative words among the whole article. The means values indicate that an article being positive is more likely than it being negative. Also, the maximum and minimum values indicate that articles exists on both ends of the spectrum.

Furthermore, the variables "numb_imgs", "numb_videos", and "num_hrefs" indicate the number of images, videos, and links respectively. The mean values indicate that on average articles use 4 images, 1 video, and 11 links. The standard deviation values show that these rates vary quite a bit.

The variable "n_non_stop_words" is a more interesting one. Natural Language Processing (NLP) models categorise words in articles. One of these categories is "stop-words" which are words like "a", "then", and "the" which do not add extra meaning into the context. The variable shows the rate of

words which are not classified as "stop-words". It is expected that higher rates show that the article is filled with more information. It is interesting that the minimum value is 0.000, showing that there is at least on article with no substance.

Moreover, the variable "n_tokens_title" shows how many tokens are in the title, basically indicating how long the title is". The maximum value shows that several articles have quite long titles. Also, the mean and standard deviation suggests that the values are evenly spread among the articles.

Finally, the final variable concerning the data channel is a binary variable indicating if the data channel is entertainment. The mean shows that most of the articles are not under the channel of entertainment.

Overall, it is expected that all of these variables will be a part of the relationship which will explain the variance in shares.

2.2 Relationship

The formal specification of the relationship between the dependent variable shares and independent variables is written below:

$$\begin{aligned}\text{shares} = \beta_0 &+ \beta_1 \text{is}_\text{w}\text{eekend} \\ &+ \beta_2 \text{rate}_\text{p}\text{ositive}_\text{w}\text{ords} \\ &+ \beta_3 \text{rate}_\text{n}\text{egative}_\text{w}\text{ords} \\ &+ \beta_4 \text{num}_\text{i}\text{mgs} \\ &+ \beta_5 \text{num}_\text{v}\text{ideos} \\ &+ \beta_6 \text{n}_\text{n}\text{on}_\text{s}\text{top}_\text{w}\text{ords} \\ &+ \beta_7 \text{n}_\text{t}\text{okens}_\text{t}\text{itle} \\ &+ \beta_8 \text{num}_\text{h}\text{refs} \\ &+ \beta_9 \text{data}_\text{c}\text{hannel}_\text{i}\text{s}_\text{e}\text{ntertainment}\end{aligned}$$

The variable concerning shares can be interpreted as a count variable, hence, several models were created. The summary of the models are present below:

Table 4:

	Dependent variable:	
	shares	
	Poisson	*negative binomial*
	(1)	(2)
Constant	8.198***	8.153***
	(0.001)	(0.041)
is_weekend	0.130***	0.154***
	(0.0002)	(0.016)
rate_positive_words	-0.427***	-0.441***
	(0.0005)	(0.033)
rate_negative_words	-0.449***	-0.519***
	(0.001)	(0.040)
num_imgs	0.010***	0.015***
	(0.00001)	(0.001)
num_videos	0.016***	0.027***
	(0.00002)	(0.001)
n_non_stop_words	-0.0001***	-0.0003
	(0.00001)	(0.001)
n_tokens_title	0.020***	0.021***
	(0.00004)	(0.002)
num_hrefs	0.008***	0.009***
	(0.00001)	(0.001)
data_channel_is_entertainment	-0.233***	-0.213***
	(0.0002)	(0.014)
Observations	39,644	39,644
Log Likelihood	-127,439,834.000	-360,835.700
θ		0.926***(0.006)
Akaike Inf. Crit.	254,879,687.000	721,691.400
Note:		*p<0.1; **p<0.05; ***p<0.01

The model above shows the likelihood change of shares as the independent variables change. All of the variables are significant, highlighting that the

variables chosen correlate with shares. The results show that articles with more pictures, videos, and links get more shares, probably hinting that articles with more substance and visual elements are shared more often. Also, articles published on weekends are shared at a higher rate. Furthermore, articles with longer titles happen to perform better as they see higher number of shares.

Among these variables, the article being published on the weekend seems to have the strongest correlation, and the number of links have the weakest one.

On the other hand, both the rate of negative and positive words seems to decrease the amount of shares. This indicates that having less words with a certain pull does not help the articles perform. Finally, the data channel of entertainment has a strong negative correlation with the amount of shares, indicating that articles under entertainment are not shared as often.

It should be highlighted that the rate of nonstop words, even though significant, has a very weak correlation. So it can be stated that articles with less substance and more stop words get more shares, even though it is a small difference.

Finally, as clearly seen above, the table compares two different models, Poisson and Negative Binomial. It seems like the models are very similar. However, the standard errors are much smaller in the Poisson model. Moreover, the AIC value for the Negative Binomial model is much lower, indicating a preference towards that model compared to the Poisson model.

Before moving into the next section, a likelihood ratio test was conducted, which revelaed that there is over dispersion in the data. With all of these pieces of information in mind, the negative binomial model was used for the analysis.

2.3 Comparing Models

2.3.1 Adding the OLS Model

After analysing the models, an OLS model was generated. Below, the table which contains the information on both models is present:

Table 5:

	Dependent variable:	
	shares	
	negative binomial	*OLS*
	(1)	(2)
Constant	8.153***	3,804.453***
	(0.041)	(454.230)
is_weekend	0.154***	450.596***
	(0.016)	(173.325)
rate_positive_words	-0.441***	-1,727.411***
	(0.033)	(367.314)
rate_negative_words	-0.519***	-1,831.004***
	(0.040)	(445.105)
num_imgs	0.015***	43.655***
	(0.001)	(7.597)
num_videos	0.027***	73.080***
	(0.001)	(14.615)
n_non_stop_words	-0.0003	-0.362
	(0.001)	(11.143)
n_tokens_title	0.021***	65.969**
	(0.002)	(27.911)
num_hrefs	0.009***	36.526***
	(0.001)	(5.649)
data_channel_is_entertainment	-0.213***	-760.573***
	(0.014)	(156.627)
Observations	39,644	39,644
R^2		0.005
Adjusted R^2		0.004
Log Likelihood	-360,835.700	
θ	0.926***(0.006)	
Akaike Inf. Crit.	721,691.400	
Residual Std. Error		11,600.810 (df = 39634)
F Statistic		20.873*** (df = 9; 39634)
Note:		*p<0.1; **p<0.05; ***p<0.01

As seen above, an OLS regression model was used to explain the relationship between the independent variables and shares. Even though this model reveals that every independent variable except "n_non_stop_words" has a significant correlation with the dependent variable, the R^2 value is quite low.

Also, in the Negative Binomial model, it was revealed that the number of non-stop words was a very weak indicator of shares. Combined with the findings from the OLS model, it can safely be stated that the variable is not a good predictor for the number of shares.

2.3.2 Partial Effects

Finally, the partial effects of the independent variables were investigated. The values of the effects per variable is present in the table below:

Table 6:

(Intercept)	is_weekend	rate_positive_words	rate_negative_words	num_imgs	num_videos	n_non_stop_words	n_tokens_title	num_hrefs	data_channel_is_entertainment
27,963.350	528.288	-1,512.379	-1,781.316	50.772	91.484	-1.005	71.295	32.208	-729.946

The table above adds on top of the information gathered by looking at the regression models. These values indicate the average changes in the mean counts from marginal changes in the independent variable.

For example, the variable regarding the data channel is a binary variable. The partial effect value shows that the average number of shares for articles in the entertainment channel is on average 790 times less than other articles. For non-binary variables such as the number of images, it shows that, on average, as the number of images increase by 1, the mean number of shares increase by 35.

This graph shares a more in-depth analysis to the findings gathered from the Poisson model. It highlights the negative correlation between the number of shares and the rate of negative or positive words. It should also be highlighted that these variables are scale variables, indicating that the partial effect values shows the marginal difference. Hence, the findings show that

the number of shares is expected to be drastically different when an article with a very high rate of negative words in compared to one with a low rate.

3 Binary Model

3.1 Summary

This part of the report will focus on Yelp reviews and aim to detect the reviews with 5 stars. In order to get a general understanding of the data, a summary was generated with the relevant variables.

Table 7:

Statistic	N	Mean	St. Dev.	Min	Max
review_stars	155,827	3.710	1.175	1	5
fans	155,827	32.497	83.984	0	722
years_elite	155,827	2.483	3.004	0	12
numb_friends	155,827	280.696	633.737	0	6,093
price_range	155,827	1.795	0.607	1	4
travel	155,827	0.104	0.305	0	1
dFiveStars	155,827	0.293	0.455	0	1

As seen above, the data has a great number of points with 155,827 cases. Moreover, it gives insight on a number of different variables. For example, the dFiveStars variable shows if a review gave 5 stars. As expected, the mean is lower than 0.5, indicating that most reviews do not give 5 stars.

Furthermore, the average Yelp user has around 32 fans, with a standard deviation of 83, showing that the number of fans range a great deal. Finally, on average, a Yelp user has 280 friends, which shows that the average Yelp user has around 280 friends more than I do.

3.2 Regression Model

3.2.1 ReviewStars

The model below includes the regression model which can be used to estimate the ReviewStars variable which indicates how many stars each review gave.

$$\begin{aligned}\text{ReviewStars} = \beta &+ \beta_1 \text{travel}_\text{i} \\ &+ \beta_2 \text{fans} \\ &+ \beta_3 \text{numbFriends} \\ &+ \beta_4 \text{priceRange} \\ &+ \beta_5 \text{yearsElite} \\ &+ \beta_6 \text{numbFriendstravel} \\ &+ \epsilon, \epsilon \sim n(0, \sigma)\end{aligned}$$

3.2.2 dFiveStars

Unlike the model above, this model focuses on the dFiveStars variable which is a binary variable which indicates if a review gave 5 stars.

$$\begin{aligned}\text{dFiveStars}_\text{i} = \beta i &+ \beta_1 \text{travel}_\text{i} \\ &+ \beta_2 \text{fans}_\text{i} \\ &+ \beta_3 \text{numbFriends}_\text{i} \\ &+ \beta_4 \text{priceRange}_\text{i} \\ &+ \beta_5 \text{yearsElite}_\text{i} \\ &+ \beta_6 \text{numbFriends}_\text{i} \text{travel}_\text{i} \\ &+ \epsilon_i, \epsilon \sim n(0, \sigma)\end{aligned}$$

3.3 Regression Models

3.3.1 ReviewStars

ReviewStars is an ordinal variable since the number of stars given, even though ordered by number, do not indicate linear distance between. Hence, an ordinal logistic regression model was created. The table below shows the regression model with the log likelihood and Akaike Information Criterion.

Table 8:

	Dependent variable:	
	review_stars	
	ordered logistic	*ordered probit*
	(1)	(2)
fans	-0.002***	-0.001***
	(0.0002)	(0.0001)
travel	0.166***	0.073***
	(0.017)	(0.010)
years_elite	-0.020***	-0.007***
	(0.002)	(0.001)
numb_friends	0.0002***	0.0001***
	(0.00003)	(0.00001)
price_range	0.079***	0.049***
	(0.008)	(0.005)
travel:numb_friends	-0.00000	0.00001
	(0.00004)	(0.00002)
mu.1	0.993	0.627
	0.016	0.009
mu.2	-0.502	-0.298
	0.015	0.009
mu.3	-1.526	-0.898
	0.016	0.009
mu.4	-2.553	-1.429
	0.018	0.01
lnL	-225457.043	-225561.207
AIC	450934.086	451142.415
Observations	155,827	155,827
Note:	*p<0.1; **p<0.05; ***p<0.01	

As seen through the table, the model consists of significant variables except the interaction variable. It seems like the two biggest actors in the

relationship are travel and price range. The person traveling or dining at a restaurant with an higher price range seems to yield more stars given in a review. On the other hand, users who have been elite for longer seems to give less stars in their reviews.

Moreover, the AIC values indicate that the logistic regression model is better compared to the probit model. The intercepts are also present for both of the models. Also, the intercepts are present for both of the models with the standard errors.

3.3.2 dFiveStars

After the equations were created, regression models were generated. The results of the models are presented below:

Table 9:

	Dependent variable:		
	dFiveStars		
	OLS	*logistic*	*probit*
	(1)	(2)	(3)
Constant	0.296***	-0.869***	-0.539***
	(0.004)	(0.018)	(0.011)
fans	-0.0005***	-0.003***	-0.002***
	(0.00003)	(0.0002)	(0.0001)
travel	0.064***	0.281***	0.172***
	(0.004)	(0.019)	(0.011)
years_elite	-0.018***	-0.093***	-0.055***
	(0.0004)	(0.002)	(0.001)
numb_friends	0.0001***	0.0003***	0.0002***
	(0.00000)	(0.00002)	(0.00001)
price_range	0.020***	0.096***	0.059***
	(0.002)	(0.009)	(0.006)
travel:numb_friends	-0.00001	-0.00004	-0.00002
	(0.00001)	(0.00004)	(0.00003)
Observations	155,827	155,827	155,827
R^2	0.021		
Adjusted R^2	0.021		
Log Likelihood		-92,482.340	-92,494.520
Akaike Inf. Crit.		184,978.700	185,003.000
Residual Std. Error	0.450 (df = 155820)		
F Statistic	568.260*** (df = 6; 155820)		
Note:			*p<0.1; **p<0.05; ***p<0.01

The fairly tall table above shows the different regression models. Interestingly, the likelihood of a 5 star review decreases as the number of fans and the years of being an elite member increases. The latter can be interpreted as members with more experience using a 5-star review more sparingly. Moreover, it seems that as the restaurant is more expensive, it seems to gen-

erate higher 5 star reviews, which is expected. Also, it seems that users who are traveling are more likely to give 5 star reviews. Finally, it seems that the interaction between traveling and number of friends did not yield interesting results since it is the only non-significant variable in the table.

Also, a table with the odds ratio is present below to better explain the logistic regression results.

Table 10:

(Intercept)	fans	travel	years_elite	numb_friends	price_range	travel:numb_friends
0.419	0.997	1.324	0.911	1.000	1.101	1.000

As seen above, the number of friends and the interaction variable barely has a correlation. The amount of years being an elite member, traveling, and the price range seems to be the factors which correlate strongest with the dependent variable.

The regression models presented above were tested with the suitable goodness-of-fit measures.

The McFadden R^2 value is 0.019, Efron R^2 value is 0.022, and Count R^2 value is 0.707. Overall, it doesn't seem like the model is a good measure to explain the variation in the dependent variable.

In order to further analyse the logistic model, the predicted scores by the model were compared to the actual scores. It seems like the model has a relatively high accuracy with 70%. However, the precision is low with 44%, indicating that the model has quite a bit false positives. Drastically, the model seems to predict that a review will not give 5 stars much more than it has to. There were almost 46,000 false negative values and only 61 false positives. It is clear that the logistic model is skewed towards the reviews not giving 5 stars.

Finally, the ROC curve and the area under the curve was investigated. The ROC graph is present below.

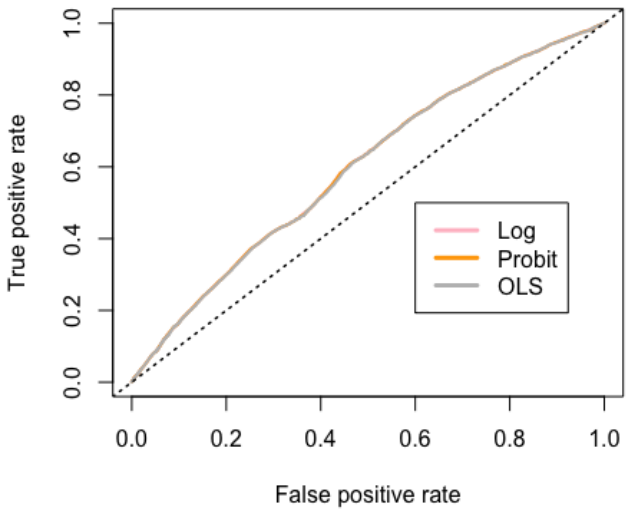

Figure 1: ROC Curve

As seen above, the curves are quite similar, almost the same, for the three regression models, highlighting the lack of difference between the models. The area under the curve is also the same with all models covering 60% of the area.