BM01BAM - Advanced Statistics & Programming

Individual Assignment 3:

Time to Export, News Article Sharing, and Online Reviews

The third, and last, individual assignment is about applying your new skills to the analysis of panel data, count data, and ordinal data. You will put these skills to use to examine three questions: what factors influence the time to export (*panel data*); what drives online information sharing (*count data*); and hat factors influence online reviews (*ordinal data*)? All parts of the assignment can be made with the material taught in the lectures and the tutorials of weeks 5 to 7. Much of the tutorial code can be re-used. Always motivate your answers, be precise, no hand-waving. This assignment is due on **Friday, October 14, 2022, 11:59am**. [Expected length of your submission is 6.25–8.75 pages, all inclusive]

Before starting the assignment, make a designated folder with sub-folders: *Data*, for all data sets; *Programs*, for code; and *Results*, for results, e.g., figures and tables. This will enhance transparency of the work process, add to developing hygienic coding skill, and help us to help you in the case of questions.

Submit your individual answers as a pdf-file on Canvas, with R-script in the appendix (in a not too large font). Alternatively, you may opt to write your assignment using R Markdown (or R Sweave), which integrates the code and your answers. The use of Latex for this submission, as explained in the course manual, is encouraged (see *this* or *this* link for helpful resources). A decent layout is part of the assessment, as well as a proper motivation of your claims and findings.

1 Panel data modeling: time to export

The time to export, a measure of border compliance, is an important indicator of the quality of trade conditions of countries. The longer the time to export the larger the burden and costs on behalf of exporting firms, the larger the risks of

obsolescence and deteriorated goods. This purpose of the concluding task is to estimate the effects of potential determinants of variation in time to export between countries and over time. All analyses are based on data provided by the World Bank.

1. Download data about dependent variable 'time to export' (*IC.EXP.TMBC*) and three or *at most* four other variables of your liking, which hypothetically influence variation in the time to export, from the World Bank's open data platform. The following code snippet installs and loads package *wbstats* and subsequently downloads the dependent and some other variable:

```
# Install package wbstats to facilitate access to the
# World Bank's data portal
install.packages("wbstats", dependencies = TRUE)
library(wbstats)

# Download selected data from the portal and store
# the data in dataframe dfTime2Export
dfTime2Export <-
   wb_data(indicator = c("IC.EXP.TMBC",
                         "IS.SHP.GOOD.TU"),
           country = "countries_only",
           start_date = 1960,
           end_date = 2021)
```

   Function `wb_indicators` can be used to obtain a table with available variables. Download the data, assign meaningful column names to the dependent and independent variables. Present the population model that reflects the assumed effects. [About 0.5 page]

2. Construct a balanced data set of the downloaded data by implementing the following steps: (*i*) remove records with missing values of the dependent variable 'time to export'; (*ii*) select all records that have been completely observed during the period 2014-2019. Use function `stargazer` to make a table with summary statistics of the dependent and independent variables of the model, and concisely discuss the main insights obtained from this table (be precise, no hand waving). [About 0.5-1 page]

3. Estimate the previously formulated model based on pooled regression, between regression, fixed-effect regression, and random-effect regression. Use function `stargazer` to make a table of the combined results. Concisely discuss the main insights obtained from the table (be precise, no hand waving). [About 0.75-1.25 page]

4. Use appropriate statistical tests to decide which of the estimated models is the preferred specification for the analysis of 'time to export'. Discuss your findings. [About 0.25 page]

2 Counts data modeling: the Mashable case

In today's high-speed low-attention society, it is key to platform managers to understand which variables drive the number of news articles shares. This is of particular interest to Mashable, a popular online site for entertainment, culture and social media; see `https://mashable.com/`. Close to finishing your first block of courses in the business analytics master, you are in an excellent position to analyze the news sharing behavior using real-world data. To this end, the company has made a data set available, which contains a wide range of variables on online news articles. Arguably the most interesting variable of this data set is the target variable *shares*.

Download the Mashable data set from *this link*, and read the accompanying documentation. Use these data to elaborate on the following questions.

1. Based on your understanding of the topic, decide about the nine most important variables that explain the sharing of online news articles. Use `stargazer` to make a table with summary statistics of ten variables (including the dependent variable *shares*) and provide a concise, but proper description of the data. [About 0.5 - 1 page]

2. Present the formal specification of the relation between the number shares (*shares*) and the nine explanatory variables selected in the preceding question. Estimate the unobserved parameters of this relationship with the most appropriate statistical method given the nature of the dependent variable.[1] Present the estimation results in a table. Describe the key results of the model in a concise, but meaningful way. [About 1 - 1.5 pages]

3. Compute the partial effects of the independent variables in the preferred model.[2] Present them in a suitable table together with the effects of OLS regression applied to the same model specification. What differences can you see in the results between the linear model and the count data model? Discuss. [About 0.75-1 page]

3 Ordinal logistic data modeling: the Yelp case

Yelp is an online platform where individuals can post or find reviews of restaurants, mostly in urban environments; see `https://www.yelp.com`. Variation in reviews is as inevitable as desirable, since it helps individuals to select restaurants of their liking. At the same time it is desirable to identify sources that systematically contribute to variation in these reviews. In order to explore the relations, Yelp has published a data set that contains online ratings from their restaurant rating platform. A special feature of this data set is that a distinction

[1] Note that various such statistical methods have been discussed during the lecture and the tutorials, and that, in principle, they all bear relevance to this question.

[2] In the preceding question you are likely to have estimated several models. Here, you can confine your efforts to the preferred model.

can be made between contributors who review restaurants located outside their home city, i.e., while traveling, or within their home city. This is captured by the indicator variable *travel*. To optimize the recommendations of restaurants to website visitors, it is essential to know which variables play a role in the online rating behavior of people. It is your task to analyze the influence of certain variables (such as *travel*) on the online rating. A description of the data set is in Table 1.

Table 1: Description of variables in *online_rating_travel.csv*

Variable	Variable Description
business	ID of the business being rated;
user	ID of the user giving the rating;
review_stars	The rating of the user for that restaurant (from 1 = lowest to 5 = highest) ;
travel	Travel indicator (1 = reviewer was traveling, 0 = was in home city);
date	The date the rating has been given;
length	The length of the review in characters;
*votes_**	Review received upvotes by other users in the domain 'funny', 'cool', or 'useful';
**_sent*	User sent upvotes to other reviews in the domain 'funny', 'cool', or 'useful';
reviews_in_city_so_far	Number of reviews a user as given up to this point in time in this city so far;
review_id	ID of the review;
yelping_since	The data the user joined Yelp;
fans	Number of fans a user has on Yelp;
years_elite	Number of years a user has been member of the Elite program of Yelp;
numb_friends	Number of friends a user has on Yelp;
category	Category of the restaurant;
price_range	The price range of the restaurant (from 1 = inexpensive to 4 = expensive);
*good_for_**	Indicator target use (1 = good for a certain group of customers, as in variable name);
years_elite	Number of years a user has been member of the Elite program of Yelp.

Download the Yelp dataset from Canvas. Use these data to elaborate the following questions.

1. In addition to dependent variable *review_stars* construct a binary dependent variable *dFiveStars* that has value 1 if *review_stars* is equal to 5, and value 0 otherwise. The background of this extra dependent variable is that

review ratings are generally skewed towards the higher end, and that the dichotomous *dFiveStars* may help to capture the really positive reviews.

Use `stargazer` to make a table with summary statistics of seven variables (including *review_stars*, *dFiveStars*, *travel* and four other explanatory variables of choice). Provide a concise description of the results in this table. [About 0.5 - 1 page]

2. Present the formal specification of the relation between the review ratings (once for *review_stars* and once for *dFiveStars*) and the five explanatory variables selected in the preceding question; also, add at least one interaction term that involves *travel* and a quantitative variable of choice. [About 0.25 - 0.5 page]

3. Estimate the appropriate binary-choice and ordered-response regression models associated with the specification(S) in the previous question, both as probit and logit models. Present the results in a table, including appropriate goodness-of-fit measures. Concisely describe the key insights from this table. [About 0.75-1 page]