POLM086 Quantitative Data Analysis (30 credit module)

Question1:

A specification of the research topic, theoretical background/issues and your research questions. A list of some relevant data (give URL links to the data if possible), a statement of how the data could be used to address your research questions, a discussion of the opportunities for research and/or limitations of current data sources. Comment on whether and how variables in the dataset measure the main concepts of interest to you in your research (for example are they valid and reliable measures. Consider the dataset(s) potential for considering relationships of interest as dependent and independent variables for statistical analysis in a multivariate regression model. Comment on any limitations of the data for using regression to address your research questions.

For my MA Dissertation I would like to focus on the rise of anti-vaccine movement in the context of the current Covid19 Pandemic. I believe that this project is important as identifying trends of vaccine hesitancy within population is an important part of any vaccination program. Large sums of money, time and effort is being put into the development of the vaccination program. Unfortunately, if a big portion of the populace is sceptical towards vaccines and refuses to take the jab, these efforts and sacrifices become futile. In order to address misinformation and vaccine scepticism it may be useful for us to understand the context behind medical scepticism. My Dissertation will aim to locate the relationship between populist movements and how they feed into the distrust sentiments towards the government of the "anti-vax" movement. This Dissertation draws on the existing literature such as "Medical populism and the COVID-19 pandemic" by Lasco G and "The International Health Regulations (2005), the threat of populism and the COVID-19 pandemic" by Wilson, Kumanan Halabi, Sam Gostin, Lawrence O.[1,2] These authors suggest that populism and scepticism towards medicine go hand in hand.

[1] Gideon Lasco. "Medical Populism and the COVID-19 Pandemic." Taylor &amp; Francis, www.tandfonline.com/doi/full/10.1080/17441692.2020.1807581.

[2] Wilson, K., Halabi, S. & Gostin, L.O. The International Health Regulations (2005), the threat of populism and the COVID-19 pandemic. Global Health 16, 70 (2020). https://doi.org/10.1186/s12992-020-00600-4

For this particular assignment I chose to focus on “COVID-19 Vaccine Hesitancy Surveys” presented by Wagner Abraham on the “openICPSR” website.[3] These surveys were conducted across United States, China, Taiwan, Indonesia, India, and Malaysia These surveys are aiming to trace Vaccine Hesitancy Trends before and during the vaccine rollout program, therefore, it is being updated on a rolling basis. The Dataset I chose to work with for the purposes of this assignment contains the most recent results from the USA that were published on the 07.01.2020.[4] For this dataset a sample of 693 individuals has been selected. As I mentioned earlier, this project is meant to be updated on a rolling basis. At this moment the vaccination program is still at its early phases and it goes without saying that it may be too early to make any strong claims regarding vaccine hesitancy. Across many nations a lot of people who desire to get vaccinated don’t have the opportunity to get the jab. Perhaps, with more people getting vaccinated public perception about Covid19 Vaccines will shift accordingly. Because of this we should review our results in light of the limitations of available data.

To explore the relationship between Vaccine Hesitancy and Populism I would like to use this Dataset from the US and focus on 3 variables when answering further questions of this assignment. These variables are:

1. Vaccine Hesitancy Scale Score **(lo_sum)**
2. Monthly income level categories of participants **(d_inc_4cat)**
3. Age categories of participants **(d_agecat)**

The assumption I am trying to test is that the demographic that finds populism appealing tends to consist of younger individuals with lower earnings. This is because the populist message addresses the grievances of the “ordinary people” that are usually ignored and particularly resonates in the online sphere. The assumption that populism mostly appeals to the lower income individuals can possibly be deemed as problematic, however it is a commonly shared assumption within the literature. Many consider populism to be defined as “an appeal to `the people' against both the established structure of power and the dominant ideas and values of the society”.[5] Others, go as far as suggesting that financial grievances of the ordinary people is what sustains political populism in the first place as “the middle and working classes of once-

[3] Wagner, Abram. COVID-19 Vaccine Hesitancy Surveys. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2021-01-08. https://doi.org/10.3886/E130422V1

[4] Wagner, Abram. COVID-19 Vaccine Hesitancy Surveys: us202003_icpsr.sas7bdat. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2021-01-08. https://doi.org/10.3886/E130422V1-74280

[5] Margaret Canovan, “Trust the People! Populism and the Two Faces of Democracy,” *Political Studies* 47, no. 1 (March 1999): 2–16, https://doi.org/10.1111/1467-9248.00184. p3

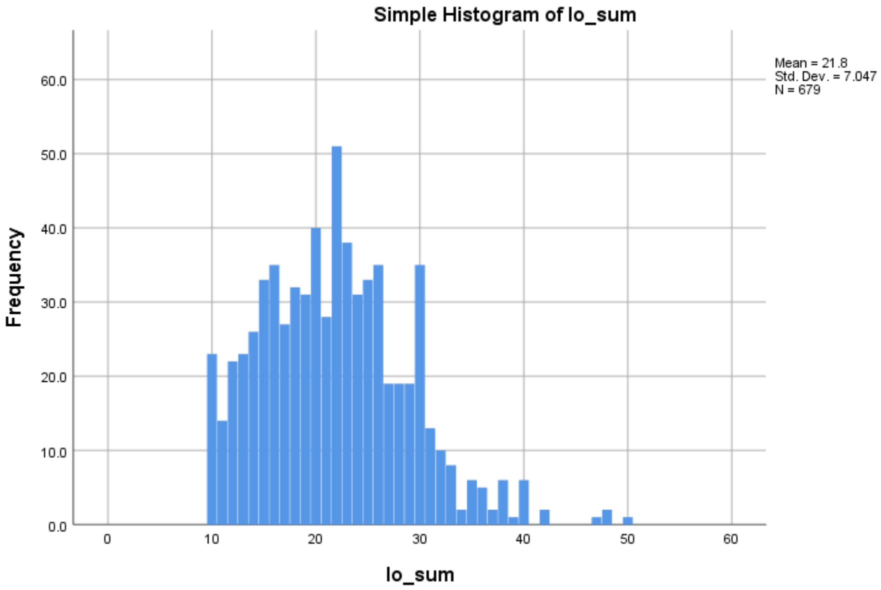
prosperous countries have seen living standards stagnate and economic security disappear.[6] There is no denying that populist supporters can come from all facets of society, but for simplicity's sake we can accept the common definition and assumptions about populism as being the voice of " the people". And thus, to test the assumptions behind my thesis, I will attempt to answer question 2 and 3 using the Independent variables of **(d_inc_4cat)**; **(d_agecat)** and locate their relationship with the dependent variable **(lo_sum)** in the USA during the earliest stages of vaccine rollout.

[6] Craig Calhoun, "'Brexit Is a Mutiny against the Cosmopolitan Elite', New Perspectives Quarterly, 33 (3), 50-58," 2016, https://onlinelibrary.wiley.com/doi/10.1111/npqu.12048. P54

Question 2

For one interval/ratio variable in a dataset that measures a concept of interest to you in addressing your research question(s), plot a histogram of the variable's distribution. Report the mean, median, mode values and a measure of skewness. Comment on the overall shape of the distribution you see visually in the histogram (for example does it resemble the normal distribution, can you 'see' any skew?). Comment on whether there is missing data for this variable and whether there are outliers. Discuss whether and how the results of this analysis help you address your research question and the usefulness of this variable in addressing the question.

I would like to answer this question focusing on the "Vaccine Hesitancy" **(lo_sum)** variable. This Variable is going to be an independent variable in my further analysis as we are trying to understand how it is being affected by other variables. The "Vaccine Hesitancy" responses are measured on a scale from 10 – 50, with 10 being "least hesitant" and 50 being "most hesitant" as it is mentioned in the Master Codebook which comes with the dataset.[7] The variable **(lo_sum**) is an interval variable and thus we can plot a histogram for this dataset and find the mean, median and mode. This will allow us to understand how prevalent the "anti-vax" beliefs within this sample of the population are. Below is the histogram of the **(lo_sum)** variable:



[7] Wagner, Abram. COVID-19 Vaccine Hesitancy Surveys: Master Codebook.docx. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2021-01-08. https://doi.org/10.3886/E130422V1-73442

Statistics

lo_sum

N	Valid	679
	Missing	14
Mean		21.80
Median		22.00
Mode		22
Skewness		.588
Std. Error of Skewness		.094

For the Vaccine Hesitancy **(lo_sum)** variable the mean is 21.80, the median is 22.00 and the mode is 22. For these results we also must account the missing values of 14 individuals who did not give their responses. This explains the discrepancy between the sample size of 693 and the N value of 679 on the plotted histogram.

The skewness of the histogram is equal to 0.588 which is closer to 1 as opposed to 0. From this data and from the overall shape of the histogram we can see that the distribution is not symmetrical, however it has a general resemblance of a bell-shaped normal distribution. In fact, this histogram is having a negative skew, with most of the responses clustering around lower Vaccine Hesitancy scores.

This means that an average response of the individuals from this sample was that they were "somewhat hesitant" about the vaccines. As I mentioned earlier, this is not surprising given that the USA is currently in the early stages of the vaccination program. Perhaps, the speed at which this vaccine was developed arises reasonable doubts within the participants. Nonetheless, that lack of confidence in the vaccine, however small it may be, is ought to be addressed for the success of the vaccination program. Therefore, even though "anti-vax" beliefs are not particularly prevalent amongst the sample of this population, the average responses indicate presence of a degree of suspicion towards the Covid19 Vaccine which for our interests is higher than ideal.

Question 3

Use multivariate linear regression to analyse a relationship between a Y dependent variable and more than one X, independent, variables. The variables you choose for inclusion as Y and X variables in this analysis should include the variable you described in section 2 above. Report a two tailed t test of the coefficients of the X variables with the null that the coefficient is equal to zero. Comment on the results and the implications, if any, for your research questions. Discuss how the inclusion of multiple variables in your regression (compared to a bivariate model with only one of the X variables you included) affects your results. Discuss the strengths and limitations of the regression model you have applied for assessing causal relationships between the X variables and the Y variable. In your discussion include the definition, and consider the potential relevance to your analysis, of omitted variable bias.

As previously mentioned, for the purposes of this assignment I have chosen the variables of Income Category **(d_inc_4cat)**; Age Category **(d_agecat)** and Vaccine Hesitancy **(lo_sum)**. In question 2 we saw that the Vaccine Hesitancy scores, despite being overall low, are still higher than desirable. Perhaps, to better our understanding we can see what variables affect Vaccine Hesitancy scores and can serve as a good enough predictor.

Before proceeding with this exercise, I must elaborate on the role of this Multivariate Linear Regression model in relation to my thesis. The variables I am choosing for my Multivariate Linear Regression model are continuous. Nonetheless, because of the nature of this dataset, the values of these variables are limited to the number of categories individuals could choose. Secondly, this exercise is not going to directly address the premise behind my thesis as the connection between age, income and populism is not primary. However, these variables are of the greater interest to me compared to others as I am curious about the possible impact that income and age may have on Vaccine Hesitancy Scores Overall. Therefore, this model may not support my thesis directly, but hopefully it will better my understanding of Vaccine Hesitancy sentiments amongst the population.

Below is the Multivariate Regression Model for this dataset:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.358[a]	.128	.126	6.609

a. Predictors: (Constant), d_inc_4cat, d_agecat

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4312.433	2	2156.216	49.358	.000[b]
	Residual	29269.026	670	43.685		
	Total	33581.459	672			

a. Dependent Variable: lo_sum

b. Predictors: (Constant), d_inc_4cat, d_agecat

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	28.895	.784		36.876	.000	27.356	30.433
	d_agecat	-.119	.016	-.264	-7.253	.000	-.151	-.087
	d_inc_4cat	-.426	.075	-.207	-5.693	.000	-.573	-.279

a. Dependent Variable: lo_sum

Given this we can now perform a two tailed t test of the coefficients of the X variables with the null that the coefficient is equal to zero.

Below is the multivariate regression equation with **(d_agecat)** variable being x_1 and **(d_inc_4cat)** variable being $x_{2:}$

$$f(x_1, x_2) = 28.895 – 0.119x_1 – 0.426x_2$$

The Coefficients of this multivariate regression equation are b_0=28.895, b_1=-0.119, b_2=-0.426. The standard error in each one of them is 0.784, 0.016 and 0.075 respectively.

For this analysis we will focus on the gradients of variables x_1 and x_2 as they are relevant for our purposes.

Let's suppose the two possible values of b_1 and b_2 for a level of significance of 0.05. b = 0 meaning f and x are not correlated (Null Hypothesis) and b not equal to 0 meaning that they are (Alternate Hypothesis)

The t value of b_1 is -7.253 and the p value is lower than 0.000, the t value of b_2 -5.693 with the same p 0.000. This means that we can reject the null hypothesis and assume that there is a correlation between the two variables chosen and the vaccine hesitancy score **(lo_sum).**

And thus, we have conducted a multivariate regression analysis on the variables present in this dataset. Though age and income categories have an impact on vaccine hesitancy scores this could also be attributed to the omitted-variable bias. Omitted variable bias refers to the possibility of missing out on important variables that could provide better prediction for the values of the dependent variables. In this way, the effects of the omitted variable could be attributed to the variables that have been selected in the chosen model, giving an illusion of false causal relation. As mentioned earlier, this dataset does not include sufficient information that could explain vaccine hesitancy scores. I chose the most seemingly fitting variables, but nonetheless, the correlation we observe could be attributed to some variable that is not included in my analysis. In addition to this, there will be more data available in the future that would give us a better insight into the factors that affect Vaccine Hesitancy in the given population.

In conclusion, this multivariate regression model pointed out the drawbacks in the dataset, showing it to be insufficient for my thesis. At the same time, we have observed that there is a presence of some vaccine hesitancy in the US population considering the variables chosen.

Bibliography:

1. Craig Calhoun, "'Brexit Is a Mutiny against the Cosmopolitan Elite', New Perspectives Quarterly, 33 (3), 50-58," 2016, https://onlinelibrary.wiley.com/doi/10.1111/npqu.12048.

2. Gideon Lasco. "Medical Populism and the COVID-19 Pandemic." Taylor &amp; Francis, www.tandfonline.com/doi/full/10.1080/17441692.2020.1807581.

3. Margaret Canovan, "Trust the People! Populism and the Two Faces of Democracy," *Political Studies* 47, no. 1 (March 1999): 2–16, https://doi.org/10.1111/1467-9248.00184.

4. Wagner, Abram. COVID-19 Vaccine Hesitancy Surveys: us202003_icpsr.sas7bdat. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2021-01-08. https://doi.org/10.3886/E130422V1-74280

5. Wilson, K., Halabi, S. & Gostin, L.O. The International Health Regulations (2005), the threat of populism and the COVID-19 pandemic. Global Health 16, 70 (2020). https://doi.org/10.1186/s12992-020-00600-4