HOW DO SOCIOECONOMIC FACTORS AND POLITICAL PARTY AFFILIATION IMPACT CONSUMER TRUST OF NEWS OUTLETS

Introduction & Rationale

My topic of interest would be exploring consumer trust in various forms of news outlets available in today's world. Things are becoming more digitalized than ever, and this trend is also reflected in the ways receive our news. With online social network sites, such as Facebook, quickly replacing traditional means of communications, such environments present an appealing platform to also share news. However, numerous studies demonstrate that not all news displayed on social media sites, such as Facebook or Twitter, can be classified as authentic content, and can even be fake, as they are not held with the highest standard in comparison to actual news agencies. With that in mind, I want to explore the different socioeconomic factors along with political party affiliation that contribute to consumer trust in various types of news outlet platforms.

Even though there has been a host of studies and articles that have already established an increasing level of distrust toward news on social media platforms following frequent media coverage of fake news, money-making schemes and hacking, people still continue to use these platforms to view their news. In fact, Facebook has been identified as the most popular platform in which "people in the US get their news" (Silverman, 2017). As such, I want to conduct a project that aim to look at the socioeconomic background and political views that make up people with differing level of trust toward various forms of news outlets. Considering the factors mentioned above, my research question would be, is there a significant correlation between socioeconomic factors and political party affiliations with the level of trust toward news outlets. In other words, do sociodemographic factors and political views impact a user's trust of news outlets?

Potential Implications & Benefit

If there is a significant correlation from a multi-variable analysis of socioeconomic factors to level of trust of news outlets, this offers important insights to not only social media and news outlets, but also marketers, government institutions, NPOs, and other organizations who use these platforms. Being able to predict which socioeconomic groups are more skeptical and less trusting of content online in general can offer invaluable actionable insights that can prove to be useful to various organizations.

To illustrate with the case of Facebook, the social media giant cannot collect detailed socioeconomic information from its users, such as income levels, household size, etc., from their private database alone. However, this can be collected through surveys from companies who have a vested interest, such as the poll conducted by IPSOS. With such information, Facebook can direct its privacy and fake content filtering efforts to more vulnerable groups to protect people from fake news, scam and/or deceptive practices often found online.

Other organizations who utilizes Facebook's platform can also make use of this data to drive their goals. Marketers can tailor their efforts toward groups that are more perceptive of the news content online. Similarly, content creators can cater toward an audience that is more active and avid users of Facebook from a certain socioeconomic background, in addition to their perceived topic of interest.

However, the most obvious downfall of such insights is that it can fall prey to the wrong hands. For example, new outlets who distribute fakes news to solely generate views and revenue, can also use analysis of data to find the most vulnerable groups that are willing to easily believe anything.

Hypothesis

1) The higher the income of the respondent, the more trusting of the content distributed by news outlets
2) The higher the education level of the respondent, the less trusting of the content distributed by news outlets
3) The greater the age of the respondent, the less trusting of the content distributed by news outlets
4) The more one affiliates with a right or left political party, the more trusting of the content distributed by news outlets

Data

I will be using two datasets provided by IPSOS Public Affairs. IPSOS is an independent market research company, which is based in France since 1975. It holds a reputable prestige as a worldwide research group in all areas, and has offices in 89 countries. It specializes in brand, advertising and media, customer loyalty, marketing, public affairs research and survey management. Considering the nature of the organization, they provide their survey-based research practice services to a number of prestigious American and international organizations, as well as other types of clients for businesses and professionals.

The dataset I will be using will be extracted from the IPSOS poll, which was intended for Buzzfeed's article, "Most Americans Who See Fake News Believe It", which was published on December 6, 2016. Its scope also includes adults from the US continent, including Alaska and Hawaii, and its purpose was to determine the extent to which an average American would trust the content of news in terms of its facts, which is directly related to my research interests of this project. It has a sample size of 4,135, and aims to represent the general US population using IPSOS calibration approach (i.e. raking-ratio adjustments).

Dependent Variables

Age: refers to the respondent's age

Income: refers to the respond's individual income, and has been encoded with the following labels in relation to their numeric tag. The numeric values are necessary to fit the variable into a model, whereas the labels are used for user-friendly description purposes.

Table 1: Income Coding

id <int>	income <fctr>
1	Less than $25,000
2	$25,000 to $34,999
3	$35,000 to $49,999
4	$50,000 to $74,999
5	$75,000 to $99,999
6	$100,000 to $149,999
7	$150,000 or more

Education: refers to the educational level of the respondent, and has been encoded with the following labels in relation to their numeric tag. The numeric values are necessary to fit the variable into a model, whereas the labels are used for user-friendly description purposes.

Table 2: Education Coding

id <int>	education <fctr>
1	Less than high school
2	High school graduate (includes equivalency)
3	Some college, no degree
4	Associate's degree
5	Bachelor's degree
6	Ph.D.
7	Graduate or professional degree

Party: refers to the political party affiliation in which the respondent identifies with, which range from Democrat, Republic, Independent and Others.

Independent Variables

Before getting into the details of the different models, it is important to establish the core concept being employed and measured, which is the level of trust. The definition of trust is built upon Lowry et al. study, which defines as "the degree to which a user's expectations, assumptions, or beliefs that the institution's actions will be beneficial, favorable, or not detrimental" (Lowry et al, 2015). In other words, trust is reflective of how the news outlet serves the intended interests of its users as a source of providing news.

Even though trust is a complex and subjective human concept, it can be measured quantitatively based on Lowry's definition in relation to the function of the news outlet. As such, I believe this dataset can provide a robust measure of consumer trust toward news, where trust is measured with rigor in relation to the actual content of the news. The variable, accuracy, measures how much a respondent believes in the authenticity of a specific news headline (11 headlines total) in terms of whether the content is fake or not. This is not a subjective question, but rather a direct straightforward answer that is measured with qualitative labels, which ranges from not at all accurate, not very accurate, somewhat accurate, and very accurate. These labels are then used to create a binary variable called 'trust', where if the respondent believe the news headline to be somewhat accurate or very accurate, then it warrants a numeric value of 1, which equates to trusting the contents of the news headline, whereas the other two values indicate sthe respondent does not trust it.

Table 3: Trust Encoding

id <int>	trust <fctr>
1	Don't Trust
2	Trust

Descriptive Statistics

Table 4 Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	4,135	44.674	16.142	18	32	58	102
income_numeric	4,135	3.242	1.331	1	2	4	5
education_numeric	4,135	5.307	1.483	1	4	7	7
trust	4,135	0.794	0.405	0	1	1	1

Age

There is a range between 18 to 102 years of age for all respondents, and the average is around 45 with standard deviation of 16 years. Moreover, 25% of the respondents are 32 years or younger, and 75% of the respondents are 58 years or younger. This seems like a good representative sample, without any noticeable distribution skewness for age.

Income

It seems like the respondents earn a wage of $35,000 to $49,999 on average, which is reasonable, as it is close to the reported US mean income per capita of $48,150 for year 2016 based on US consensus survey. Moreover, it seems like respondents are evenly split between the different income brackets, as 25% of the respondents fall within the second lowest income bracket and 75% of the respondents fall within the second highest income bracket.

Education

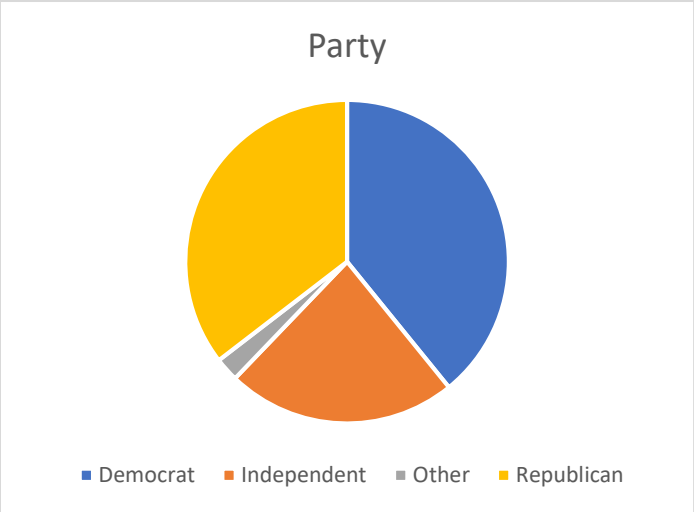
It seems like the average respondent has at least a bachelor's degree, which seems a bit higher than I expected. As such, I can expect a lot of the respondents to be well educated. However, this is also not surprising because people who are more educated tend to also pay attention to the news more. This is also demonstrated from the percentile range, where 25% of the respondents have at least an associate's degree and 75% of the respondents have a PhD degree.

Trust

It seems like people are generally more trusting of news headlines with a mean of 0.794.

Party

Chart 1: Political Party



Since it is a categorical variable, it cannot produce a descriptive statistic, instead I looked at the distribution of the different parties by count. There seems to be an even split between Democrats (1619 people) and Republicans (1465 people), with Democrats being a little bit more represented. There is a fair share of independents (952), with very minimal others (99 people).

Initial Models

Table #5: Linear Probability Model

```
lm(formula = trust ~ income_numeric + education_numeric + age +
    party, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8895  0.1389  0.1891  0.2244  0.3319

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.8295898  0.0314922  26.343  < 2e-16 ***
income_numeric     0.0099348  0.0050659   1.961 0.049930 *
education_numeric -0.0067684  0.0045459  -1.489 0.136589
age               -0.0009138  0.0003934  -2.323 0.020223 *
partyIndependent  -0.0390347  0.0166076  -2.350 0.018800 *
partyOther        -0.0750477  0.0417218  -1.799 0.072129 .
partyRepublican    0.0555720  0.0146738   3.787 0.000155 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4024 on 4128 degrees of freedom
Multiple R-squared:  0.01187,  Adjusted R-squared:  0.01043
F-statistic: 8.262 on 6 and 4128 DF,  p-value: 6.49e-09
```

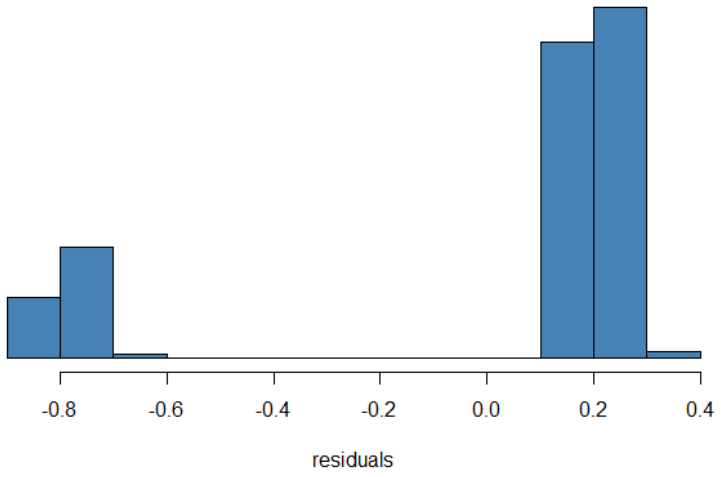
Since our target variable, trust, is binary in nature, I first used a linear probability model to identify any significant relationships between socioeconomic factors (income, education and age) and political party affiliation with trust. As a result, for every category increase in income brackets a person is 0.99 percentage points (statistically significant) more likely to trust the news headline on average, net of other factors. For every category increase in education a person is 0.67 percentage points (not statistically significant) less likely to trust the news headline on average, net of other factors. For every increase in age, the person is 0.09 percentage points (statistically significant) less likely to trust the news headline on average, net of other factors. Being in an independent political party (vs other political affiliation) makes a person 3.9 percentage points (statistically significant) less likely to trust the news headline on average, net of other factors. Being in an other political party (vs other political affiliation) makes a person

7.5 percentage points (not statistically significant) less likely to trust the news headline on average, net of other factors. Being in a Republican political party (vs other political affiliation) makes a person 5.6 percentage points (highly statistically significant) less likely to trust the news headline on average, net of other factors.

Model Improvement Stage 1

Linear probability models are not ideal due to their limitations. First, normality of errors assumptions is violated, as such p values might not be reliable, as seen from the graph below:

Chart #2: Residuals of Linear Probability Model



Secondly, expected values can be out of the prescribed, 0-1 range. In this case, it is not an issue.

Table #6: Range of Residuals of Linear Probability Model

```
[1] 0.6681227 0.8894856
```

Lastly, homoskedasticity might be violated. In this case, however, the test have a p value less than 0.05, and therefore cannot reject the null hypothesis that the variance of the residual is constant and that heteroscedasticity is present.

Table #7: BP Test of Linear Probability Model

```
studentized Breusch-Pagan test

data:  lm2
BP = 3.4714, df = 6, p-value = 0.7478
```

Although the linear probability model didn't perform poorly, I wanted to see if the model can be improved by fitting it with a logistic regression.

Table #8: Logistic Regression Model

```
Call:
glm(formula = trustv2 ~ income_numeric + education_numeric +
    age + party, family = binomial, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0487   0.5547   0.6432   0.7114   0.9185

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.571652   0.195037   8.058 7.74e-16 ***
income_numeric     0.060921   0.031202   1.952   0.0509 .
education_numeric -0.041370   0.028100  -1.472   0.1410
age               -0.005622   0.002416  -2.327   0.0199 *
partyIndependent  -0.214260   0.097077  -2.207   0.0273 *
partyOther        -0.396557   0.229750  -1.726   0.0843 .
partyRepublican    0.369399   0.094845   3.895 9.83e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4206.7  on 4134  degrees of freedom
Residual deviance: 4157.6  on 4128  degrees of freedom
AIC: 4171.6

Number of Fisher Scoring iterations: 4
```

Based on the results, for every category increase in income brackets, a person increases his/her logit by 0.061 (not statistically significant) of trusting the news headline on average, controlling for other factors. For every category increase in education a person decreases his/her logit by 0.041 (not statistically significant) of trusting the news headline on average, controlling for other factors.. For every increase in age, the person increases his/her logit by 0.006 (statistically significant) of trusting the news headline on average, controlling for other factors. Being in an independent political party (vs other political affiliation, the person decreases his/her logit by 0.214 (statistically significant) of trusting the news headline on average, controlling for other factors. Being in an other political party (vs other political affiliation), the person decreases his/her logit by 0.39 (not statistically significant) of trusting the news headline on average, controlling for other factors. Being in a Republican political party (vs other political affiliation), the person increase his/her logit by 0.37 (not statistically significant) of trusting the news headline on average, controlling for other factors.

Model Improvement Step 2

However, it is generally difficult to interpret the results of the logistic regression. As such, I calculations the odds-ratio for each of the variables.

Table #9: Odds Ratio of Logistic Regression Model

(Intercept)	income_numeric	education_numeric
4.8145939	1.0628154	0.9594741
age	partyIndependent	partyOther
0.9943933	0.8071381	0.6726318
partyRepublican		
1.4468645		

Based on these results, the odds of trusting a news headline goes up by 6% for every increase in the category of income bracket. The odds of trusting a news headline goes down by 4% for every increase in the category of education level. The odds of trusting a news headline goes down by 0.6% for every increase in age. The odds of trusting a news headline goes down by 19% when being in an independent. The odds of trusting a news headline goes down by 32% when being in an other political party. The odds of trusting a news headline goes up by 45% when being a republican. Net of political parties, for every category in the education level, income bracket, and age, their odds of a trusting a news headline goes up by 481%, which represents a proportionate increase.

Model Improvement Step 3 Variation A

The issue with odds-ration is that they are still hard to interpret, but probability model generates a more specific interpretation of the proportional effect of each of the predictor variables against the target variable's probability of occurrence, which is evaluated at the median for this model. With that in mind, it seems like the predictive power are strongest for political party affiliations in terms of the magnitude. I wanted to explore this future by using the prediction function demonstrated during Professor Greg's lecture.

Table #10: Predicted Probabilities using Median

income_numeric <dbl>	education_numeric <dbl>	age <int>	party <fctr>	PredictedProb <dbl>	0.5% <dbl>	99.5% <dbl>
3	6	40	Democrat	0.7826706	0.7532740	0.8094507
3	6	40	Republican	0.8389852	0.8100492	0.8642519
3	6	40	Independent	0.7440331	0.7032921	0.7809225
3	6	40	Other	0.7078036	0.5777180	0.8109307

Using the function, I derived the predictive probabilities displayed above. The predicted probability of trusting a news headline for someone who earns an median income of $35,000 to $49,999, has a median education of graduate or professional degree, a median age of 40 years old and is a Democrat is 78%. The predicted probability of trusting a news headline for someone who earns an median income of $35,000 to $49,999, has a median education of graduate or professional degree, a median age of 40 years old, and is a Republican is 84%. The predicted probability of trusting a news headline for someone who earns an median income of $35,000 to $49,999, has a median education of graduate or professional degree, a median age of 40 years old, and is an Independent is 74%. The predicted probability of trusting a news headline for someone who earns a median income of $35,000 to $49,999, has a median education of graduate or professional degree, a median age of 40 years old, and is not affiliated with any major political parties is 71%.

Model Improvement Step 3 Variation A

Table #11: Predicted Probabilities using Mean

	income_numeric <dbl>	education_numeric <dbl>	age <dbl>	party <fctr>	PredictedProb <dbl>	0.5% <dbl>	99.5% <dbl>
1	3.241596	5.307134	44.674	Democrat	0.7855657	0.7578215	0.8109244
2	3.241596	5.307134	44.674	Republican	0.8412822	0.8150075	0.8644457
3	3.241596	5.307134	44.674	Independent	0.7472766	0.7091679	0.7819274
4	3.241596	5.307134	44.674	Other	0.7113282	0.5819775	0.8134811

Using the sample, I wanted to evaluated the proportional effect at the mean, instead of the medium. From the results, it seems the results are still pretty consistent with the republican party having the highest probability for someone to trust the new headline, with other values set to ehir mean, while ‘other’ political party has the lowest probability.

Final Model Selection & Conclusion

When comparing the linear probability model with the different models of logistic regression, the logistic regression is a better choice. In an earlier section, I tested the linear regression model in various aspects, and it was surprisingly a good model. However, it does a mediocre job when predicting values at the extremes of the dependent variables, just due to the nature of the model, as mentioned by Professor Greg. Moreover, the linear probability model assumes the relationship between the dependent variables and the target variable to be linear.

Logistic model are more advantageous due to several factors. On one hand, it treats the target variable as a continues variable, which allows a cumulative probability structure with no floor or ceiling effects. On the other hand, it is better in predicting the values, and offers a more comphremsiove and holistic view when interepreting the relationship of the variables, as described above.

With that in mind, the political view of the person seems to have the highest predictive power in relation to whether the person will trust the content of the news headline or not. This was reflected in the results of both models. Moreover, all of my hypotehesis were proven true in terms of the direction of the impact of the predictor variables on the target variable, where people with more education and age are less likely to trust the content of the news headlines, while people with who identify as either Republican or Democrat and has higher income are more trusting. The direction of the variables were the same for both regression and logistic models as well.

In the future, I would definitely like to explore this dataset further in different approaches. First, I use the original coding of the target variable, where the four categories of

accuracy are retained. Then, I can conduct the same analysis for multi categorical variables. Secondly, since political parties is significant, I would like to explore this further by controlling for other variables, and control for the impact of socioeconomic factors. Lastly, I can use another variable, called guessed_correctly, to test another research question that is closely related to the one proposed here, where if the mistrust or trust of consumers are warranted or validated. Since I already have the true status of each of the headline, this aspect could be analyzed, and is a potential for another exciting project. Overall, this has been a very insightful independent project, which is very closely related to my master's thesis, which aims to identify what mechanisms are most effective to restore consumer trust of social media platforms, when their trust is impaired.

Citations:

Gregory, Eirich. "Lecture Class #6-10." Data Analysis for Social Sciences, Columbia University

Lowry , P.B., Clay, P., Bennett, R.B., Roberts, T.L. "Leveraging fairness and reactance theories to deter reactive computer abuse following enhanced organisational information security policies: An empirical study of the influence of counterfactual reasoning and organisational trust." Information Systems Journal, 25 (3) (2015), pp. 193-273.

Silverman, Craig. "People Read News On Facebook But They Don't Really Trust It, A Survey Found." BuzzFeed News, BuzzFeed, 19 Jan. 2017, https://www.buzzfeednews.com/article/craigsilverman/people-be-reading-but-not-trusting-news-on-facebook.