UNIVERSITY OF THE PHILIPPINES DILIMAN

Professional Master's in Data Science (Analytics)

Ryan Emmanuel S. Dela Paz

Modelling Reliance on Social Media for Political Information: An Ordinal Logistic Regression Approach Based on Social Class and Income Level

Submitted to:

Czarinne Antoinette A. Antonio

STAT 218 – Statistical Machine Learning

Date of Submission:

May 2025

I. Introduction and Related Literature

Social Media is currently a dominant platform on how many individuals consume news and public information. Even now, digital age is still in the process of emerging. With the increasing invasion of internet access and digital platforms, especially in developing countries like the Philippines, the use of social media is no longer just a tool for entertainment or connection between people, it also serves as a primary source of information for some, shaping opinions, beliefs, and specially voting behavior which is timely today. This shift in information consumption raises questions about who relies on these platforms and why.

A lot of these concerns formed a body of research having focus on socioeconomic status as a significant role in shaping access to and engagement with information sources. Individuals from lower income brackets or social classes may lack access to traditional media such as cable news, newspapers and other premium articles, so instead, opt to a more accessible, low-cost platforms like Facebook, YouTube, or even TikTok. However, the use of these platforms as information sources is clearly not without consequences. Numerous concerns have been raised about the quality, bias, and reliability of news disseminated on social media. Understanding the demographic and SES associated with social media dependence is of importance for information dissemination.

This study seeks to examine the relationship between income level and perceived social class to the frequency of using social media as a source of information. The researcher will approach this study be creating a model to estimate probability of an individual's frequency of using social media daily, weekly, or never, depending on their SES. This is important for identifying patterns of media reliance across different strata of society, and can serve as a guide for strategies on disseminating news, campaigning and other forms of media regulation.

In this study, the researcher will contribute to this literature by modeling Ordinal Logistic Regression to chosen data from the World Values Survey, incorporating income level (Q287) and subjective social class (Q288R) as the predictors to predict the chosen variable, frequency of social media use for information (Q207).

Beyond this, by identifying which segments of the population are most reliant on social media, stakeholders, including policymakers, advocacy groups, and media organizations can effectively leverage social media platforms to target outreach, awareness campaigns,

and public engagement strategies tailored to the informational behaviors of specific social groups.

Related Literature

Prior studies indicates a clear stratification in digital consumption behaviors. Ucar et al. (2021) revealed that low-income areas are more likely to engage in social media, and less likely to consume news or engage in information searching, or streaming. This implies that social media may have different roles across different SES, which reinforces existing inequalities in information access.

A study from Kalogeropoulos (2018) noted that lower social grade individuals use fewer sources of online news, and are more reliant on data through social media. This shows a pattern that contributes to a gap in media literacy and diversity between different SES. Kalogeropoulos (2018) also discussed that these disparities will grow as digital media becomes even more prevalent today.

A literature also highlights the difference in trust of individuals towards online information. Hussain et al. (2023) found that university students with rural backgrounds are more likely to trust information on Facebook than their urban counterparts. This also signals a potential divide not just in usage but also in critical engagement with content.

Looking at the study from a methodological perspective, it will leverage the use of ordinal logistic regression to model the perceived income class which is usually an ordinal outcome variable. This is reinforced in line with the findings of Lelisho et al. (2022), ordinal regression is particularly suitable for this type of analysis. It is also relevant pointing out that they recommend the Partial Proportional Odds Model (PPOM) when the proportional odds assumption is violated for any covariates, an extension that this study considers when assessing model fit.

These related studies constructs a theoretical framework to justify the variables selected for the model. Social media usage frequency, social class, income levels and trust in social media emerge as key predictors based on the literature. This research contributes to the growing field of digital inequality by contextualizing global findings within the Philippines only.

II. Methodology

Description of Variables

This study focuses on identifying the relationship between socioeconomic status (SES) and the frequency of individuals that use social media as their source of information. The analysis is based on data drawn from the World Values Survey. The primary objective is to regress and have estimate the likelihood that individuals of different income levels and social class utilize social media to obtain news and information. The variables selected for the analysis are as follows:

Response Variable

Q207: Frequency of Using Social Media as an Information Source

Respondents were asked: "*People learn what is going on in this country and the world from various sources. For each of the following sources, please indicate whether you use it to obtain information daily, weekly, monthly, less than monthly or never: Social media (Facebook, Twitter, etc.).*" The original responses were categorized into five ordinal levels:

1 - Daily

2 - Weekly

3 - Monthly

4 - Less than monthly

5 - Never

The values "Don't know," "No answer," "Not asked," and "Missing" are treated as missing observations and excluded from analysis. This variable captures the frequency of social media use for gathering information, which is clearly an ordinal outcome for regression modeling. The gradation in frequency reflects underlying behavioral differences in digital engagement, which may be influenced by SES-related constraints.

Predictor Variables

Q287: Subjective Social Class

This variable measures the respondents' social class which is self-perceived. Respondents were asked to identify their social class from the following options:

1 - Upper class

2 - Upper middle class

3 - Lower middle class

4 - Working class

5 - Lower class

Subjective social class is a grounded measure often used to capture an individual's perception of their social position, beyond just income. This is chosen because perceptions of class may influence both media preferences and their behaviors when seeking for information. For example, those that are identified in lower classes may be go towards free accessible platforms such as Facebook or TikTok, (Kalogeropoulos, 2018).

Q288R: Income Level (Recoded)

Income was collected in a scale with 10 options in the original WVS Q288, but was recoded for this study into three levels for easier interpretation:

1 - Low income (Steps 1–3)

2 - Medium income (Steps 4–7)

3 - High income (Steps 8–10)

Income level represents objective economic status and complements the more subjective social class variable. As stated in a previous study by Ucar et al. (2021), there is a link with lower income levels with increased dependence on social media and decreased usage of more traditional or premium news sources, making this a theoretically appropriate predictor.

Data Analysis Approach

To investigate the influence of socioeconomic variables on individuals' reliance on social media as a source of political information, a structured and rigorous analytical pipeline combining exploratory and inferential statistical techniques are employed. All analyses were conducted using **R.**

The World Values Survey (WVS) dataset was extracted. As per the WVS codebook, responses coded as -1 (Don't know), -2 (No answer), -4 (Not asked), and -5 (Missing; Not available) were treated as missing and converted to **NA**. Only complete cases for the variables of interest were retained for modeling, resulting in a final sample of 1,199 observations after excluding 328 incomplete cases.

The primary response variable, Q207 (Use of social media as an information source), was treated as an ordinal variable with five levels ranging from "Never" to "Daily". Two predictors were included: Q287 (Perceived social class) and Q288R (Self-reported income level), both also coded as ordered factors.

```
wvs <- na.omit(wvs[c("Q207", "Q287", "Q288R")])

# Q207 - Social media information source: Ordinal variable
wvs$Q207 <- factor(wvs$Q207,
                   levels = c(5, 4, 3, 2, 1),
                   ordered = TRUE)


# Q287 - Social class: Factor with ordered levels
wvs$Q287 <- factor(wvs$Q287,
                   levels = c(5, 4, 3, 2, 1),
                   ordered = TRUE)

# Q288R - Income level: Factor with ordered levels
wvs$Q288R <- factor(wvs$Q288R,
                   levels = c(1, 2, 3),
                    ordered = TRUE)

str(wvs)
```

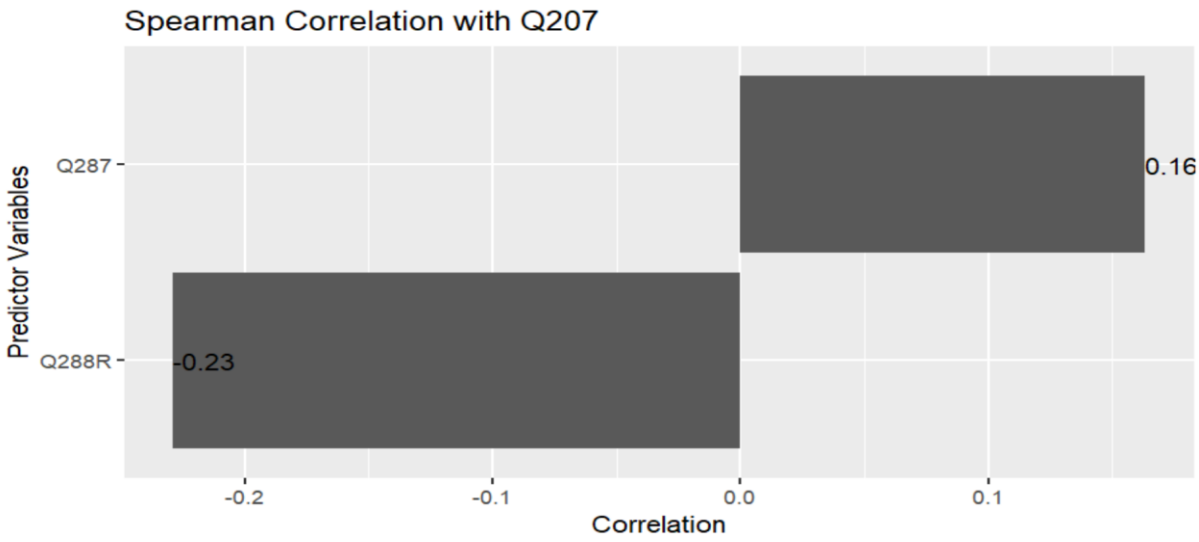
```
## tibble [1,199 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Q207 : Ord.factor w/ 5 levels "5"<"4"<"3"<"2"<..: 1 4 5 5 5 4 2 5 4 5 ...
##  $ Q287 : Ord.factor w/ 5 levels "5"<"4"<"3"<"2"<..: 3 1 3 1 1 3 3 2 3 3 ...
##  $ Q288R: Ord.factor w/ 3 levels "1"<"2"<"3": 2 2 2 1 1 2 2 2 2 2 ...
##  - attr(*, "na.action")= 'omit' Named int 328
##   ..- attr(*, "names")= chr "328"
```

As part of the exploratory data analysis, **Spearman's rank correlation coefficients** is compued to assess the monotonic association between the ordinal response variable (Q207) and each predictor (Q287 and Q288R). This nonparametric measure was appropriate given the ordinal nature of the variables. The results provided initial insight into the direction and strength of the relationships, helping to justify the inclusion of both predictors in the subsequent regression analysis.

```
cor_results <- sapply(vars[-1], function(var) {
  cor(wvs_vars$Q207, wvs_vars[[var]], use = "pairwise.complete.obs", method = "spearman")
})

cor_sorted <- sort(cor_results, decreasing = TRUE)

print(round(cor_sorted, 3))
```

```
##   Q287  Q288R
##  0.163 -0.229
```

```
cor_df <- data.frame(
  Predictors = names(cor_sorted),
  Correlation = as.numeric(cor_sorted)
)

ggplot(cor_df, aes(x = reorder(Predictors, Correlation), y = Correlation)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  geom_text(aes(label = round(Correlation, 2)), hjust = -0.1) +
  labs(title = "Spearman Correlation with Q207",
       x = "Predictor Variables")
```



An **Ordinal Logistic Regression** is performed using the polr() function from the MASS package. The model specified Q207 as the dependent variable and included Q287 and Q288R as independent variables:

$$\log\left(\frac{P(Y \geq j)}{P(Y < j)}\right) = \theta_j - (\beta_1 Q287 + \beta_2 Q288R), \qquad j = 2, \ldots .5$$

```
model <- polr(Q207 ~ Q287 + Q288R, data = wvs, method = "logistic")
summary(model)
```

A crucial assumption of the ordinal logistic regression model is the **proportional odds assumption**. This is evaluated using the Brant test via the brant package. To gain further insight into the ordinal structure of the response of proportional odds, a stratified summary statistics are also computed. A custom function was implemented to calculate cumulative log-odds for each threshold of the response variable across levels of the predictors. Although no formal inferences were drawn from this step, the resulting table of cumulative logits will serve as a diagnostic check to evaluate if the linear predictors differ across the different thresholds.

As reinforcement to the verification of the proportional odds assumption, a set of binary logistic regressions, each corresponding to a different dichotomization of the ordinal response (e.g., Q207 ≥ 2, Q207 ≥ 3, etc.) are conducted. For each threshold, a GLM with a binomial distribution and logit link was fitted using the same predictor variables. By comparing the coefficient estimates across these models, the study could assess whether the effects of the predictors remain constant, which is the implication of the proportional odds assumption.

Finally, the predicted probabilities are then computed from the valid ordinal logistic model for every possible combination of the predictor variables. These predicted probabilities were plotted to illustrate how the likelihood of each response level varies across categories of perceived social class and household income.

III. Results and Discussion

To examine how social class and household income influence the individuals' reliance on social media for political information, an ordinal logistic regression is modeled using the the polr() function with a logit link. The response variable Q207, was modeled as a five-level

ordered outcome, and two predictors: Q287 - subjective social class, and Q288R - income level were included as ordered factors.

The **ordinal logistic regression model** was chosen due to the ordered categorical nature of the response variable and the appropriateness of the proportional odds assumption for modeling cumulative probabilities across ordinal thresholds. The model offers an interpretable framework for understanding how shifts in social class and income influence the reliance of social media for political information.

Ordinal Logistic Regression Results

```
model <- polr(Q207 ~ Q287 + Q288R, data = wvs, method = "logistic")
summary(model)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = Q207 ~ Q287 + Q288R, data = wvs, method = "logistic")
##
## Coefficients:
##            Value Std. Error t value
## Q287.L   0.52919     0.2287  2.3140
## Q287.Q  -0.09452     0.1992 -0.4745
## Q287.C   0.14404     0.1622  0.8881
## Q287^4  -0.09569     0.1152 -0.8305
## Q288R.L  0.74703     0.1651  4.5236
## Q288R.Q -0.15610     0.1110 -1.4060
##
## Intercepts:
##     Value   Std. Error t value
## 5|4 -0.6041  0.0977    -6.1850
## 4|3 -0.0113  0.0960    -0.1181
## 3|2  0.1415  0.0961     1.4721
## 2|1  0.4912  0.0972     5.0550

##
## Residual Deviance: 3108.34
## AIC: 3128.34
```

The model revealed a statistically significant positive linear effect of both predictors. Specifically, the first-order social class Q287.L had a coefficient of 0.53, indicating that individuals that perceive themselves as belonging to a higher social class, they are more likely to use social media as a source of political information. On the other hand, the linear effect of income level of Q288R.L was 0.75, suggesting that higher income is also associated with more frequent social media use for political information. These positive coefficients are interpreted on the log-odds scale and support the hypothesis that both socioeconomic indicators are positively associated with the outcome.

The model's residual deviance was 3108.34, and the AIC was 3128.34, indicating acceptable model fit relative to the sample size.

Testing the Proportional Odds Assumption

To assess the validity of the proportional odds assumption inherent in ordinal logistic regression, Brant Test is conducted. This diagnostic check evaluates the effect of each predictor if it is consistent across all thresholds of the ordinal response variable.

```
brant_test <- brant(model)
```

```
## ------------------------------------------
## Test for X2  df  probability
## ------------------------------------------
## Omnibus      8.83    18  0.96
## Q287.L       2   3   0.57
## Q287.Q       0.71    3   0.87
## Q287.C       0.76    3   0.86
## Q287^4       3.21    3   0.36
## Q288R.L      3.9 3   0.27
## Q288R.Q      2.35    3   0.5
## ------------------------------------------
##
## H0: Parallel Regression Assumption holds

## Warning in brant(model): 7 combinations in table(dv,ivs) do not occur. Because
## of that, the test results might be invalid.
```

The test resulted to (χ^2 = 8.83, p > 0.05), indicating that, the model satisfies the proportional odds assumption and that all individual predictors, including the linear contrasts for subjective social class (Q287.L) and income level (Q288R.L) had non-significant test statistics (p > 0.05), further supporting that the PO assumption holds.

These findings validate the use of ordinal logistic regression model and remains theoretically and statistically justified for addressing the research question.

To further investigate the proportional odds structure, the stratified summary statistics of the cumulative log-odds of the response variable across levels of the predictors using a custom cumulative logit function is computed. This function transformed the proportions of responses into log-odds.

```r
sf <- function(y) {
  eps <- 1e-6  # small constant to avoid 0 or 1
  p <- function(k) mean(y >= k)
  qlogis_safe <- function(p) qlogis(pmax(eps, pmin(1 - eps, p)))

  c('Y>=1' = qlogis_safe(p(1)),
    'Y>=2' = qlogis_safe(p(2)),
    'Y>=3' = qlogis_safe(p(3)),
    'Y>=4' = qlogis_safe(p(4)),
    'Y>=5' = qlogis_safe(p(5)))
}
```

```r
s <- with(wvs, summary(as.numeric(Q207) ~ Q287 + Q288R, fun=sf))
s
```

```
## as.numeric(Q207)      N= 1199
##
## +-------+-+----+--------+----------+----------+----------+----------+
## |       | |   N|    Y>=1|      Y>=2|       Y>=3|       Y>=4|       Y>=5|
## +-------+-+----+--------+----------+----------+----------+----------+
## |   Q287|5| 259|13.81551|-0.1469822|-0.69894430|-0.87875281|-1.24286190|
## |       |4| 170|13.81551| 0.4300364|-0.07061757|-0.23638878|-0.58047402|
## |       |3| 512|13.81551| 0.5780779|-0.06252036|-0.18805223|-0.53590531|
## |       |2| 225|13.81551| 0.7334211| 0.25921961| 0.09785579|-0.22314355|
## |       |1|  33|13.81551| 0.9808293| 0.18232156| 0.06062462|-0.18232156|
## +-------+-+----+--------+----------+----------+----------+----------+
## |  Q288R|1| 393|13.81551|-0.1580892|-0.75096675|-0.88268899|-1.22839133|
## |       |2| 725|13.81551| 0.6972880| 0.10215772|-0.03586591|-0.37680985|
```

The table above shows how the cumulative log-odds vary across categories of perceived social class Q287 and income level Q288R. Looking at the results, individuals reporting the lowest social class for example (287 = 1), had consistently higher log-odds of frequent social media reliance compared to those in higher social classes, which is a positive association between subjective class and reliance on social media for political information. A similar trend was observed for the income, with higher income levels (Q288R = 3) corresponding to greater log-odds of frequent use. This ordering provides informal yet useful evidence that the parallel assumption is not violated, complementing the formal Brant test results.

To further assess the appropriateness of the proportional odds model,separate binary logistic regressions are fitted for adjacent cumulative splits of the response variable:

```
# Logistic regression for Q207 >= 2
model1 <- glm(I(as.numeric(Q207) >= 2) ~ Q287 + Q288R,
              family = "binomial", data = wvs)

# Logistic regression for Q207 >= 3
model2 <- glm(I(as.numeric(Q207) >= 3) ~ Q287 + Q288R,
              family = "binomial", data = wvs)

# View model summaries
summary(model1)
```

```
##
## Call:
## glm(formula = I(as.numeric(Q207) >= 2) ~ Q287 + Q288R, family = "binomial",
##     data = wvs)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.641177   0.117367   5.463 4.68e-08 ***
## Q287.L       0.579134   0.277312   2.088   0.0368 *
## Q287.Q      -0.008563   0.241686  -0.035   0.9717
## Q287.C       0.206170   0.190703   1.081   0.2796
## Q287^4      -0.007175   0.132690  -0.054   0.9569
## Q288R.L      0.817125   0.204324   3.999 6.36e-05 ***
## Q288R.Q     -0.119845   0.134925  -0.888   0.3744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1607.5  on 1198  degrees of freedom
## Residual deviance: 1542.5  on 1192  degrees of freedom
## AIC: 1556.5
##
## Number of Fisher Scoring iterations: 4

##
## Call:
## glm(formula = I(as.numeric(Q207) >= 3) ~ Q287 + Q288R, family = "binomial",
##     data = wvs)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.05102    0.10636   0.480    0.631
## Q287.L       0.39945    0.25680   1.555    0.120
## Q287.Q      -0.15264    0.22380  -0.682    0.495
## Q287.C       0.09533    0.17975   0.530    0.596
## Q287^4      -0.16664    0.12694  -1.313    0.189
## Q288R.L      0.89719    0.18903   4.746 2.07e-06 ***
## Q288R.Q     -0.07761    0.12572  -0.617    0.537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1657.2  on 1198  degrees of freedom
## Residual deviance: 1590.1  on 1192  degrees of freedom
## AIC: 1604.1
##
## Number of Fisher Scoring iterations: 4
```

The estimated coefficients for key predictors Q287 and Q288R were compared across these models. It is evident that the log-odds for Q288R.L remained consistently positive and statistically significant across thresholds. The coefficients for Q287.L also showed comparable values across splits, though it does have a slight varying significance levels which may be due to sample variation. All of these results align well with the proportional odds assumption, reinforcing the decision to use the ordinal logistic regression model for interpretation.

It is also worth noting that although the separate binary logistic models under thresholds yielded lower residual deviances and AIC values, does not mean they serve as a better model. The two models only capture partial information from the ordinal outcome and are not directly comparable to the cumulative model. The ordinal logistic regression model, despite a higher total deviance and AIC, offers a framework that models the entire ordinal response efficiently.

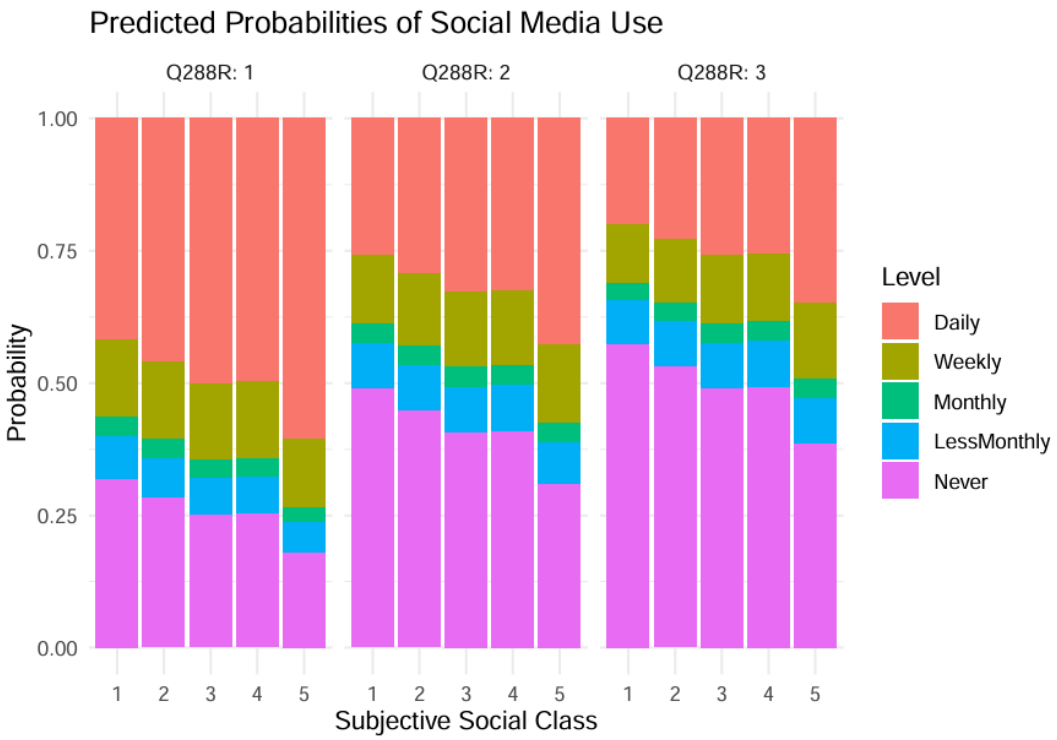
Interpretation of Predicted Probabilities and Final Discussion

To complement the coefficient-based interpretation of the Ordinal Logistic Regression OLR model, predicted probabilities were computed for each category of social media use Q207 across combinations of social class Q287 and income class Q288R. The results, visualized as stacked bar plots, illustrate how the likelihood of different usage frequencies changes with socioeconomic indicators.

```
# Create all combinations of Q287 and Q288R
newdat <- expand.grid(
  Q287 = factor(1:5, levels = 1:5, ordered = TRUE),
  Q288R = factor(1:3, levels = 1:3, ordered = TRUE)
)

# Add predicted probabilities
newdat <- cbind(newdat, predict(model, newdat, type = "probs"))

# View first few rows
View(newdat)
```

```
# Rename the levels to be interpretable
colnames(newdat)[3:7] <- c("Daily", "Weekly", "Monthly", "LessMonthly", "Never")

# Convert wide to long format
longdat <- melt(newdat, id.vars = c("Q287", "Q288R"),
                variable.name = "Level", value.name = "Probability")

# View
View(longdat)
```

```r
ggplot(longdat, aes(x = Q287, y = Probability, fill = Level)) +
  geom_bar(stat = "identity", position = "stack") +
  facet_wrap(~ Q288R, labeller = label_both) +
  labs(title = "Predicted Probabilities of Social Media Use",
       x = "Subjective Social Class",
       y = "Probability") +
  theme_minimal()
```



The estimated probabilities of social media use reveal important patterns across combinations of subjective social class and income class. It is worth pointing out that the greatest variation appears in the "Daily" and "Never" categories, suggesting that engagement polarities are most sensitive to class dynamics.

When viewed, the whole probability table (longdat) is interpreted as follows. Daily social media use increases with higher social class at every income level. For the middle-income individuals (Q288R = 2), the probability goes up from 0.259 (Q287 = 1) to 0.428 (Q287 = 5), which indicates a strong class gradient. A similar trend can also be observed in the low-income group (Q288R = 1). In the high-income group (Q288R = 3), the trend is flatter, which implies that while economic access is available, social class still shapes engagement intensity.

In contrast, inverse pattern is observed for the "Never" use category. Among high-income individuals with low social class, the probability of never using social media peaks at 0.572, which is the highest across all combinations. This finding highlights a sort of "disconnect" between material wealth and digital engagement. The "Weekly," "Monthly," and

"Less than Monthly" use categories shows uniform probabilities which typically ranges between 0.03 and 0.15, with minimal variation across classes. This indicates that the clearest differences lie at the extremes.

Therefore, the results support the claim that both subjective social class and income jointly shape patterns of media consumption, offering empirical evidence for the nuanced role of social stratification in digital behaviors. In a practical sense, the study can prove importance in highlighting potential targeted digital engagement strategies. Since the results tell us that social media use is more pronounced among certain income and class combinations, stakeholders such as public health communicators, advocacy groups, or policymakers, can more effectively tailor outreach by leveraging the platforms most used by these key social segments.

In comparison to the related literature of Kalogeropoulos et al. (2021), their similar predictive modeling to assess news consumption stratification in Europe, is adapted to the Philippines' context for this study. Both shows that social class based inequalities manifest strongly in consumption frequency, even after controlling for basic access. The findings also align with Uçar et al. (2021), who stresses how social class, not just income, predicts digital behavior.

IV. References

- Auxier, B., & Anderson, M. (2021). Social media use in 2021. Pew Research Center. https://www.sciencedirect.com/science/article/abs/pii/S0099133323000848
- Kalogeropoulos, A., & Nielsen, R. K. (2021). Factsheet: Social inequalities in news consumption. Reuters Institute for the Study of Journalism. https://www.digitalnewsreport.org/publications/2018/factsheet-social-inequalities-news-consumption/
- Ucar, I., Gramaglia, M., Fiore, M., Smoreda, Z., & Moro, E. (2021). News or social media? Socio-economic divide of mobile service consumption. https://royalsocietypublishing.org/doi/full/10.1098/rsif.2021.0350
- UCLA Institute for Digital Research and Education. (n.d.). Ordinal logistic regression. https://stats.oarc.ucla.edu/r/dae/ordinal-logistic-regression/
- R-statistics.co. (n.d.). *Ordinal logistic regression with R*. https://r-statistics.co/Ordinal-Logistic-Regression-With-R.html

- Restore Project. (n.d.). *Module 5 – Ordinal regression*. https://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/modules/mod5/module_5_-_ordinal_regression.pdf
- McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological), 42*(2), 109–142. https://sci-hub.se/https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2347760

V. Appendices

All other relevant tables, plots, and code snippets are presented within the Methodology and Results sections to enhance clarity and immediacy of interpretation.