APPLICATION OF ORDINAL LOGISTIC REGRESSION TO MODEL CUSTOMER PRODUCT PURCHASE LIKELIHOOD

Adawia Ananda [1, a)], **Aurelio Naufal Effendy**[1, b)], **Favian Sulthan Wafi**[1, c)], **Rifqi Hafizuddin**[1, d)]

[1]*Departemen Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Indonesia, Jl. Prof. DR. Sudjono D. Pusponegoro, Pondok Cina, Kecamatan Beji, Kota Depok, 16424, Indonesia.*

[a)]*adawia.ananda@sci.ui.ac.id*
[b)]*aurelio.naufal@sci.ui.ac.id*
[c)]*favian.sulthan@sci.ui.ac.id*
[d)]*rifqi.hafizuddin@sci.ui.ac.id*

Abstract. *In the ever-evolving landscape of consumer behavior, understanding the intricate factors shaping customer product purchase likelihood is vital for businesses optimizing their strategies. This study employs the ordinal logistic model to analyze a dataset, focusing on the likelihood of a customer making a purchase measured on an ordinal scale. Model selection criteria, including Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), guide the assessment of model appropriateness. Among the predictor variables—possession of coupons, peer recommendations, and product quality—coupon usage and quality emerge as the most significant predictors, as indicated by the selected model's minimized AIC and BIC values. A one-unit increase in coupon possession raises the log odds of purchase likelihood by 0.739, while a similar increase in product quality elicits a 0.604 increase. In conclusion, coupon usage and quality of the product are the most significant predictor variables in affecting the customer product purchase likelihood in this dataset.*

Keywords: *Ordinal logistic regression, product purchase likelihood, coupon usage, product quality, peer recommendation.*

1. INTRODUCTION

In the dynamic landscape of consumer behavior, understanding the factors influencing customer product purchase likelihood is paramount for businesses seeking to optimize their strategies. Studies of customer behavior are gaining importance in marketing research and will increase greatly in the future[1]. The evergrowing size and complexities of mass markets cause mass marketing to be an expensive endeavor [2]. Hence, rather than extending identical incentive offers universally, a company can opt to target only those customers who fulfill specific profitability criteria aligned with their individual requirements and purchasing behavior[3]. In the long run, customer profiling aims to translate this comprehension into an automated interaction with their customers[3]. Hence, a comprehensive analysis is required to understand the complex contribution of variables affecting the decision-making process of customers.

Recognizing the crucial importance of comprehending customer product purchase likelihood, this article aims to analyze the underlying factors. Utilizing a simulated dataset from R-blogger, the response variable in focus is the likelihood of a customer purchasing a product, presented in an ordinal scale. Within the dataset, three predictor variables are going to be analyzed, comprising two categorical variables —possession of coupons and peer recommendations—and one numerical variable denoting the quality of the product on a scale of 1 to 5. The ordinal nature of the response variable necessitates the usage of an ordinal logistic regression as the chosen analytical tool. Ordinal logistic regression, applied in studies across various fields, proves its versatility. Das and Rahman (2011) used it for child malnutrition risk factors in Bangladesh[11], Liu and Koirala (2012) for educational data estimation[12], and Chang and Fan (2013) for identifying drivers of customer loyalty[13]. These examples affirm the method's credibility in handling data with an ordinal response variable.

Several analytical processes that are going to be discussed in this article include model building, parameter estimation, identification of significant predictors influencing the response variable, and model selection. Akaike Information Criterion (AIC) will be used as the criteria for best model selection. Through these methodological approaches, the aim is to create an optimal model that provides an understanding of the factors significantly influencing the likelihood of customer product purchase. In doing so, the writers aspire to contribute valuable insights that can inform strategic decision-making in the realm of customer relationship management and marketing.

2. THEORETICAL FRAMEWORK

2.1. Ordinal Logistic Regression Model

Ordinal logistic regression is a statistical analysis method that can be used to model the relationship between an ordinal response variable and one or more explanatory variables[4]. The model specification should take into consideration an obvious natural order among the response categories. Ordinal logistic regression is an extension of logistic regression, where instead of a binary response, the response has k variables with k-1 logits. Market research, opinion polls, and fields where soft measures are common such as psychiatry often presents ordinal responses[6].

In several cases, the response variable is conceptually a continuous variable z, and is therefore difficult to measure[6]. In this case, we utilize a crude method to identify "cut points" for the latent variables. This way, for example, response variables with small values are classified in the low category, those with larger values of z are classified in the middle category, and those with high values are classified in the high category. The cutpoints $C_1, \ldots, C_{J-1}$ define J ordinal categories with associated probabilities $\pi_1, \ldots, \pi_j$ (with $\sum_{j=1}^{J} \pi_j = 1$).

2.1.1. Cumulative Logit Model

Let y_i denote the response outcome category for subject i. That is, $y_i = j$ means that $y_{ij} = 1$ and $y_{ik} = 0$ for $k \neq j$, for the c multinomial indicators. To use category ordering, we express models in terms of the cumulative probabilities[5].

$$P(y_i \leq j) = \pi_{i1} + \ldots + \pi_{ij}, j = 1, \ldots, c \tag{1}$$

The cumulative logits are logits of these cumulative probabilities,

$$logit[P(y_i \leq j)] = log\frac{P(y_i \leq j)}{1-P(y_i \leq j)} \tag{2}$$

$$logit[P(y_i \leq j)] = log\frac{\pi_{i1}+\ldots+\pi_{ij}}{\pi_{i,j+1}+\ldots+\pi_{ic}} = x_j^T\beta_j \quad j = 1, \ldots, c-1 \tag{3}$$

Each cumulative logit uses all c response categories.

2.1.2. Proportional Odds Model

Assume that the linear predictor $x_j^T\beta_j$ in (3) has an intercept term β_{0j} which depends on the category j, then the model becomes the proportional odds model.

$$log\frac{\pi_1+\ldots+\pi_j}{\pi_{j+1}+\ldots+\pi_J} = \beta_{0j} + \beta_1 x_1 + \ldots + \beta_{p-1}x_{p-1} \tag{4}$$

The proportional odds model is based on the assumption that the effects of the covariates $x_1, \ldots, x_{p-1}$ are the same for all categories on the logarithmic scale. If some categories are amalgamated, this does not change the parameter estimates $\beta_1, \ldots, \beta_{p-1}$ in (4), although the terms β_{0j} will be affected due to the collapsibility property. This form of independence between the cutpoints C_j and the explanatory variables x_k requires a strong assumption that wherever the cutpoints are, the odds ratio for a one unit change in x is the same for all response categories. In addition, the proportional odds model is not affected if the labelling of the categories is reversed, only the signs of the parameters will be changed[6].

2.2. Parameter Estimation

There are several ways to parameterize a cumulative logit model.

Table 1. Parameterizations of the Ordinal Logistic Regression Model

Model	Parameterization
Model 1	$log(\frac{P(Y\leq j)}{1-P(Y\leq j)}) = \beta_0 - (\beta_1 x_1 + \ldots + \beta_{p-1}x_{p-1}), j = 1, \ldots, J-1$
Model 2	$log(\frac{P(Y\leq j)}{1-P(Y\leq j)}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1}x_{p-1}, j = 1, \ldots, J-1$
Model 3	$log(\frac{P(Y>j)}{1-P(Y>j)}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1}x_{p-1}, j = 2, \ldots, J-1$

Model 1 uses a negative sign to create a direct correspondence between the slope and the ranking. Therefore, a positive coefficient indicates that the likelihood of a higher ranking increases as the value of the explanatory variable increases. The most commonly used models are Model 1 and 2, where the outcome of interest is observing "Y less than or equal to *j*" where *j* is one of the ordered categories in the response variable[4].

A method that can be used to estimate regression parameters is the Maximum Likelihood Estimation (MLE). MLE identifies the values of model parameters that maximize the likelihood of observing the given data, essentially finding the parameter values that make the observed data most probable under a specified statistical model. The steps include transforming likelihood function into log-likelihood function and finding solution to the partial differential equation with the value 0.

Likelihood function is defined as such

$$L(\theta, \phi; y) = \prod_{i=1}^{n} f(y_i; \theta, \phi) \tag{5}$$

Equation (5) will the be transformed into log-likelihood function as such

$$ln(L(\theta, \phi; y)) = l(\theta, \phi; y) = \sum_{i=1}^{n} ln(f(y_i; \theta, \phi)) \tag{6}$$

Next, substitute exponential family into equation (6)

$$l(\theta, \phi; y) = \sum_{i=1}^{n} ln[exp(\frac{y_i\theta - b(\theta)}{a(\phi)}) + c(y_i, \phi)] = \sum_{i=1}^{n} \frac{y_i\theta - b(\theta)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi) \tag{6}$$

2.3. Tests of Statistical Significance

Conducting tests of statistical significance is essential to ensure that the estimated value of a parameter in a statistical model is significantly different from zero or follows a specific hypothesized value. This step helps determine the relevance and impact of the parameter in explaining variability in the data.

2.3.1. Likelihood Ratio Test

Likelihood Ratio Test (LRT) is used to compare two nested models. The statistic used for this test is formulized as such

$$LR = -2log(\frac{L_0}{L_1}) = -2(l_0 - l_1) \tag{7}$$

$$LR = -2(l_0 - l_1) \tag{8}$$

where l_0 and l_1 states the maximized log-likelihood function. L_0 represents the simpler model (the model that has fewer parameters) and L_1 represents the general model. Equation (8) shows that the LRT can be computed as a difference in the deviance for the two models[8].

The test statistic is distributed as a chi-square random variable asymptotically, with the degrees of freedom being equal to the difference in the number of parameters between the simpler and the general model. This test can be expanded for a case with several parameters. For example, for $\beta = (\beta_0, \beta_1)$, a null hypothesis is formulized as such H_0: $\beta_0 = 0$. Then, L_1 is the likelihood function that is calculated on the β, and L_0 is the likelihood function that is calculated when $\beta_0 = 0$ [5].

2.3.2. Wald Test

The Wald test is used to evaluate the significance of individual parameters in a regression model by comparing the estimated parameter value to its standard error. The test statistic is given as such

$$z = (\hat{\beta} - \beta_0)/SE \tag{9}$$

Equation (9) is called Wald statistic, used to test the null hypothesis H_0: $\beta = \beta_0$. It has an approximate standard normal distribution when $\beta = \beta_0$ and z^2 has an approximate chi-squared distribution with $df = 1$ [5].

For multiple parameters $\beta = (\beta_0, \beta_1)$, to test H_0: $\beta_0 = 0$, the Wald chi square statistic is

$$\hat{\beta}^T_0[\hat{var}(\hat{\beta}_0)]^{-1}\hat{\beta}_0, \tag{10}$$

where $\hat{\beta}_0$ is the unrestricted Maximum Likelihood estimate of β_0 and $\hat{var}(\hat{\beta}_0)$ is a block of the unrestricted estimated covariance matrix of $\hat{\beta}$ [5].

2.4. Model Selection Criterion

Model selection criteria are methods used to choose the most appropriate model among a set of candidate models. Three commonly used criteria are Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and pseudo R squared.

2.4.1. Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) judges a model by how close we can expect its sample fit to be to the true model fit [5]. Out of a set of reasonably fitting models, the optimal model minimizes

$$AIC = -2[L(\hat{\beta}_M) - number\ of\ parameters\ in\ M] \quad (10)$$

where M is an arbitrary model and $L(\hat{\beta}_M)$ is the maximum value of the model's likelihood function [5].

2.4.2. Bayesian Information Criterion (BIC)

Bayesian Information Criterion (BIC) penalizes more severely for the number of model parameters [5]. The BIC value is given as such

$$BIC = -log(n)[L(\hat{\beta}_M) - number\ of\ parameters\ in\ M] \quad (11)$$

where n is the number of observations on a random sample. The model with the lowest BIC value is chosen as the best model [9].

3. DATA CHARACTERISTICS

The dataset used in this research is a secondary cross-sectional dataset from R-Bloggers[7] containing 400 observations and 4 variables. The variables consist of a categorical response variable featuring 3 ordered categories denoting the likelihood of repeated customer purchases: "low probability", "medium probability", and "high probability". Additionally, it includes 1 numeric predictor variable and 2 categorical predictor variables with binary values. For a comprehensive understanding of these predictor variables, detailed explanations are provided in Table 2 and Table 3.

Table 2. Categorical Predictor Variable Description

Variable	Category	Description	Count
$X_1 = Coupon$	0	Didn't use coupon	337
	1	Used coupon	63
$X_2 = Peers$	0	Didn't receive recommendation by peers	343
	1	Received recommendation by peers	57

Data source: R-Bloggers[7]

Table 3. Numeric Predictor Variable Description

Variable	Description	Min	Mean	Max	Variance
$X_3 = Quality$	Quality of the product (scale of 1-5)	1.89	2.989	3.99	0.1583

Data source: R-Bloggers[7]

The majority of the sample in the dataset did not use any coupon during their initial product purchase (84.25/100%) and did not receive any recommendations from peers (85.75/100%). Additionally, their average rating for the product's quality was relatively low, with a mean of 2.989 and a variance of 0.1583.

To further analyze the relationship between response and predictor variables from the dataset, visualizations using plots will be conducted between the response and each of the predictor variables.

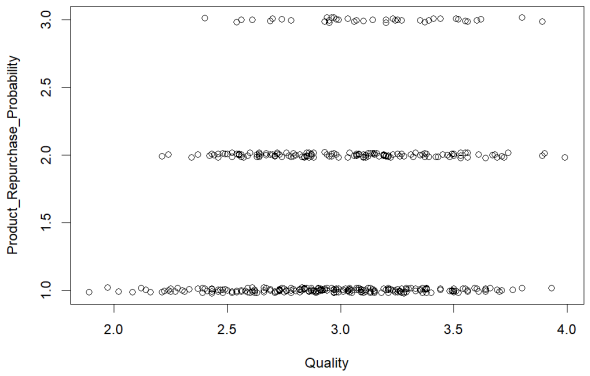

Figure 1. Scatterplot Illustrating the Relationship between Response and Quality Variable

Based on Figure 1, the concentration of data points for every "Y" response variable value appears notably higher around the “quality” scale of 3. Additionally, response variables with values 2 (medium probability) and 3 (high probability) demonstrate a smaller range with higher values, commencing from the “quality” scale around 2.5. In contrast, the response variable with a value of 1 (low probability) initiates at the “quality” scale around 2, indicating a distinct starting point compared to the other “Y” values. This indicates that there might be a positive relationship between the “Y” response variable and “quality” variable.

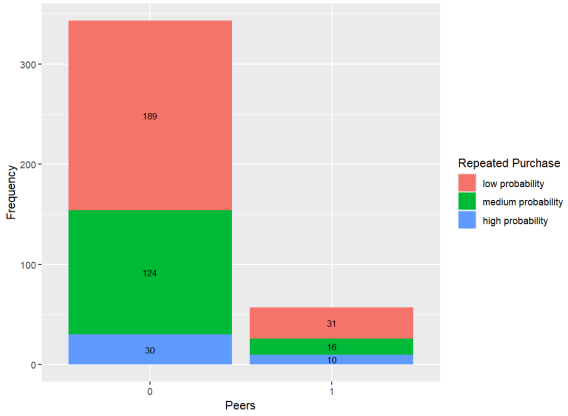

Figure 2. Stacked Barplot Illustrating the Relationship between Response and Peers Variable

In Figure 2, the “Y” response variable with value 1 (low probability) appears as the most prevalent across all categories within the “peers” variable. Subsequently, the response variable with value 2 (medium probability) and 3 (high probability) follow in frequency. This indicates that there might be no relationship between the “Y” response variable and “peers” variable. It can also be seen that a substantial number of samples did not receive any recommendations from peers.

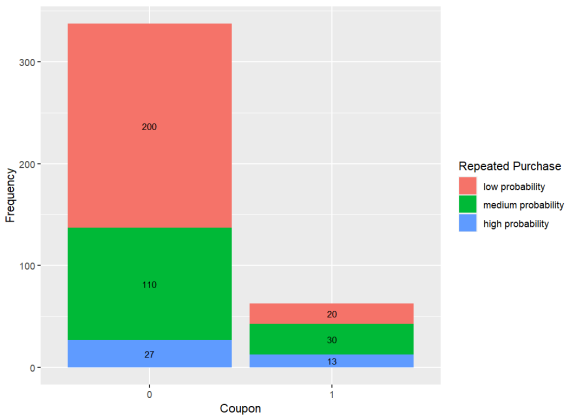

Figure 3. Stacked Barplot Illustrating the Relationship between Response and Coupon Variable

Figure 3 illustrates a significant proportion of samples indicating no use of coupons during their initial product purchase. Within the "coupon" variable category of value 0 (non-usage), the "Y" response variable with value 1 (low probability) appears as the most prevalent, followed by values 2 (medium probability) and 3 (high probability). Conversely, among samples that used coupons, value 2 (medium probability) of the "Y" response variable is more frequent, followed by values 1 (low probability) and 3 (high probability). This indicates that there might be a relationship between the "Y" response variable and "coupon" variable.

4. MODEL SELECTION AND INTERPRETATIONS

4.1. Data Preprocessing

In ordinal logistic regression the dependent variable is ordinal to consider the probability of an event and all the events that are below the focal event in the ordered hierarchy. Thus, we need to define the level order in the dependent variable "Y" and the relevant independent variables "X_1, X_2, X_3". We can do this by using the factor() function. Ignoring this step can result in analysis that is pointless.

4.2. Ordinal Logistic Regression Model

To find out the relationship between the ordinal response variable and the other explanatory variables, we will construct the ordinal logistic regression model using the R function, "polr". The summary is provided in table 4.

Table 4. Ordinal Logistic Regression Model With All Response Variables

Coefficients	Value	Std. Error	t value
coupon.L	0.74080	0.1879	3.942
peers.L	-0.04149	0.2106	-0.197
quality	0.61577	0.2606	2.363
Intercepts	**Value**	**Std. Error**	**t value**
low probability|medium probability	1.7028	0.8325	2.0454
medium probability|high probability	3.7982	0.8515	4.4605

Data source: (If data is taken from certain sources, it must be stated)

Table 4 displays the value of coefficients and intercepts, and corresponding standard errors and t values. The estimated ordinal logistic regression model can be written as:

$$logit(P(y_i \leq 1)) = 1.7028 - 0.74080 * coupon + 0.0415 * peers - 0.61577 * quality \quad (13)$$

$$logit(P(y_i \leq 2)) = 3.7982 - 0.74080 * coupon + 0.0415 * peers - 0.61577 * quality \quad (14)$$

Equation (13) and (14) can be interpreted as:
Log odds

1. A one unit increase in value of Coupon increases the expected value of rpurchase in log odds by 0.74, given all of the other variables in the model are held constant.
2. A one unit increase in value of Peers decreases the expected value of rpurchase in log odds by -0.041, given all of the other variables in the model are held constant.
3. A one unit increase in value of quality increases the expected value of rpurchase in log odds by 0.615, given all of the other variables in the model are held constant.

Odds ratio

1. For people who used coupon, the odds of having a higher probability (i.e., low probability to medium probability) to do repeated purchase is 2.0976 times that of people who didn't used coupon, holding constant all other variables.
2. For people who received recommendation by peers, the odds of having a higher probability (i.e., low probability to medium probability) to do repeated purchase is 0.9593 times that of people who didn't receive recommendation by peers, holding constant all other variables.
3. For every one unit increase in quality of product the odds of having a higher probability (i.e., low probability to medium probability) to do repeated purchase is multiplied 1.851 times (i.e., increases 85.1%), holding constant all other variables.

Next we will test the significance of each variable by finding the p value. We calculate the p value comparing the t-value against the standard normal distribution, like a z test.

Table 5. P value of All Response Variables

Coefficients	**p value**
coupon.L	8.0925e-05
peers.L	0.8438
quality	0.01814
Intercepts	**p value**
low probability|medium probability	0.0408
medium probability|high probability	8.1774e-06

With a significance level of $\alpha = 0.05$, based on the obtained p values, it can be concluded that the variables $X_1 = coupon$, $X_3 = quality$, and the intercepts are statistically significant, while $X_2 = peers$ is not statistically significant at the level $\alpha = 0.05$.

After identifying that all variables are not statistically significant in explaining the dependant variable $Y = rpurchase$, we will next construct another ordinal logistic regression model using only the statistically significant independent variables. The summary is provided in table 6.

Table 6. Ordinal Logistic Regression Model With Significant Response Variables

Coefficients	**Value**	**Std. Error**	**t value**
coupon.L	0.7394	0.1878	3.936
quality	0.6042	0.2539	2.379
Intercepts	**Value**	**Std. Error**	**t value**
low probability|medium probability	1.6474	0.7835	2.1025
medium probability|high probability	3.7427	0.8034	4.6584

Table 6 displays the value of coefficients and intercepts, and corresponding standard errors and t values. The estimated ordinal logistic regression model can be written as:

$$logit(P(y_i \leq 1)) = 1.6474 - 0.7394 * coupon - 0.6042 * quality \quad (15)$$

$$logit(P(y_i \leq 2)) = 3.7427 - 0.7394 * coupon - 0.6042 * quality \quad (16)$$

Equation (15) and (16) can be interpreted as:

Log odds

1. A one unit increase in value of Coupon increases the expected value of rpurchase in log odds by 0.739, given all of the other variables in the model are held constant.

2. A one unit increase in value of quality increases the expected value of rpurchase in log odds by 0.604, given all of the other variables in the model are held constant.

Odds ratio

1. For people who used coupon, the odds of having a higher probability (i.e., low probability to medium probability) to do repeated purchase is 2.0947 times that of people who didn't used coupon, holding constant all other variables.
2. For every one unit increase in quality of product the odds of having a higher probability (i.e., low probability to medium probability) to do repeated purchase is multiplied 1.8298 times (i.e., increases 82.9%), holding constant all other variables.

Next, we will test the significance of each variable by finding the p value. We calculate the p value comparing the t-value against the standard normal distribution.

Table 7. P value of All Response Variables

Coefficients	p value
coupon.L	8.2682e-05
quality	0.01734
Intercepts	**p value**
low probability|medium probability	0.0355
medium probability|high probability	3.186853e-06

With a significance level of $\alpha = 0.05$, based on the obtained p values, it can be concluded that all variables of the new model are statistically significant.

We are going to compare the full model that includes all covariates with the reduced model that excludes *peers*. First, we will check the AIC and the BIC from each model

Table 8. AIC and BIC for Full Model and Reduced Model

Model	df	AIC	BIC
Full Model (coupon + peers + quality)	5	727.0249	746.9822
Reduced Model (coupon + quality)	4	725.0638	741.0296

Based on table 8, the reduced model has lower AIC and BIC than the full model. This tells us that the reduced model with only coupon and quality as the covariates is better than the full model which included peers. To further convince us, we will perform a Likelihood Ratio Test to compare both full and reduced model, testing whether the peers variable is significant or not in the model.

Table 9. Likelihood Ratio Test for Full Model and Reduced Model

Model	df	LogLik	df	Chisq	Pr(>Chisq)
Full Model (coupon + peers + quality)	5	-358.51			
Reduced Model (coupon + quality)	4	-358.53	-1	0.0389	0.8436

We can see that the The Chi-Squared test-statistic is 0.0389 and the corresponding p-value is 0.8436, as shown in the output. We will not reject the null hypothesis at a significance level of 95% because the p-value is more than 0.05. This indicates that the full model and the reduced model fit the data equally well. As a result, we should employ the reduced model, because even though it has a less covariate than the full model, it can explain the data just as well as the full model, it complies with the principle of parsimony.

4.3. Proportional Odds Assumption

One of the assumptions underlying ordinal logistic regression is that the relationship between each pair of outcome groups is the same. In other words, ordinal logistic regression assumes that the coefficients that describe the relationship between the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category and all higher categories.

To test this assumption we will graph predicted logits from individual logistic regressions with a single predictor where the outcome groups are defined by either repeated purchase >= 2 and repeated purchase >= 3. If the difference between predicted logits for varying levels of a predictor are the same whether the outcome is defined by repeated purchase >= 2 or repeated purchase >=3, then we can be confident that the proportional odds assumption holds.

Table 10. Predicted Values without The Parallel Slopes Assumption

	Category	Count	Y ≥ 1	Y ≥ 2	Y ≥ 3
Coupon	0	337	Inf	-0.37833644	-2.440735
	1	63	Inf	0.76546784	-1.347074
Quality	[1.89.2.72)	102	Inf	-0.39730180	-2.772589
	[2.72,2.99)	99	Inf	-0.26415158	-2.302585
	[2.99,3.27)	100	Inf	-0.20067070	-2.090741
	[3.27,3.99)	99	Inf	0.06062462	-1.803594
Overall		400	Inf	-0.20067070	-2.197225

Table 10 above displays the (linear) predicted values we would get if we regressed our dependent variable on our predictor variables one at a time, without the parallel slopes assumption. We can evaluate the parallel slopes assumption by running a series of binary logistic regressions with varying cutpoints on the dependent variable and checking the equality of coefficients across cutpoints. We thus relax the parallel slopes assumption to check its tenability. We can use the values in this table to help us assess whether the proportional odds assumption is reasonable for our model.

- When coupon is equal to “no” the difference between the predicted value for repeated purchase greater than or equal to two and repeated purchase greater than or equal to three is -0.37833644 - (-2.440735) = 2.06239856, roughly 2.
- For coupon equal to “yes” the difference in predicted values for repeated purchase greater than or equal to two and repeated purchase greater than or equal to three is 0.76546784 - (-1.347074) = 2.11254184, which is also roughly 2.

This suggests that the parallel slopes assumption or the proportional odds assumption is reasonable for coupon covariate. Next, we will check this assumption for the quality variable,

- When the quality equal to [1.89.2.72) the difference between the predicted value for repeated purchase greater than or equal to two and repeated purchase greater than or equal to three is -0.39730180 - (-2.772589) = 2.3752872.
- For quality equal to [2.72,2.99) the difference between the predicted value for repeated purchase greater than or equal to two and repeated purchase greater than or equal to three is -0.26415158 - (-2.302585) = 2.03843342.
- For quality equal to [2.99,3.27) the difference between the predicted value for repeated purchase greater than or equal to two and repeated purchase greater than or equal to three is -0.20067070 - (-2.090741) = 1.8900703.
- For quality equal to [3.27,3.99) the difference between the predicted value for repeated purchase greater than or equal to two and repeated purchase greater than or equal to three is 0.06062462 - (-1.803594) = 1.86421862.

Notice that for all intervals, quality has nearly the same difference between the predicted value for repeated purchase greater than or equal to two and repeated purchase greater than or equal to three, specifically near 2. This means that to some extent, the parallel slopes or the proportional odds assumption does hold for the quality variable.

To help demonstrate this, we normalized all the first set of coefficients to be zero so there is a common reference point. If the proportional odds assumption holds, for each predictor variable, distance between the symbols for each set of categories of the dependent variable, should remain similar. Looking at the coefficients for the coupon variable in figure 4, we could see that the distance between the two sets of

coefficients is close. Similarly, the distances between the estimates for quality are also close, suggesting that the proportional odds assumption may hold.

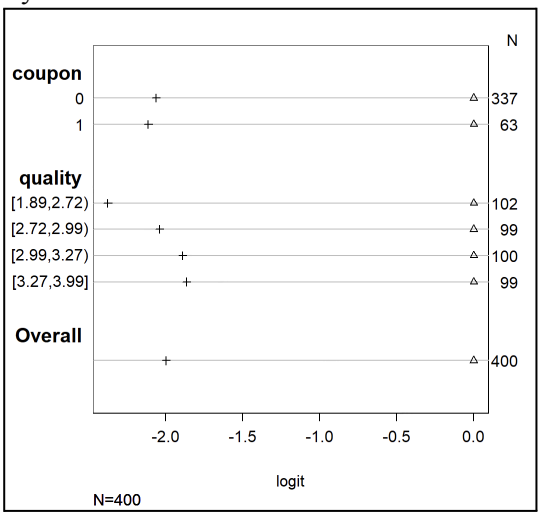

Figure 4. Distance Between Every Ordered Categories for Each Predictor

We could also do the Brant Test to assess whether the observed deviations from our Ordinal Logistic Regression model are larger than what could be attributed to chance alone or in other words test for proportional odds assumption with the null hypothesis being the proportional odds assumptions holds for the variable.

Table 11. Brant Test Result

Variable	chi-sq	df	pr(>chi)
Omnibus	0.729	2	0.69
coupon.L	0.091	1	0.76
quality	0.704	1	0.40

The Omnibus test in the context of the Brant test provides a global assessment of whether the proportional odds assumption is met. It considers the overall fit across all levels of the ordinal response variable. Failing to reject the null hypothesis in this case means that the proportional odds assumption holds. As we can see, all the variables failed to reject the null hypothesis, this means that the proportional odds assumption is fulfilled.

5. CONCLUSION AND RECOMMENDATIONS

The best model to explain the relationship between the ordinal response variable and the explanatory variables is

$$logit(P(y_i \leq 1)) = 1.6474 - 0.7394 * coupon - 0.6042 * quality \quad (17)$$

$$logit(P(y_i \leq 2)) = 3.7427 - 0.7394 * coupon - 0.6042 * quality \quad (18)$$

From the model, we can conclude that a one unit increase in *Coupon* increases the expected log odds of purchase by 0.739, while a similar increase in *Quality* yields 0.604 increase in the expected log odds of purchase. Conversely, whether someone receives a recommendation from their peers or not did not significantly affect the likelihood of repeated purchase.

However, recognizing the limitations, this study acknowledges the need for continued exploration. Future research could broaden the scope by incorporating additional variables like brand reputation, product cost, age, gender, and other potential influencers on customer behavior. This acknowledges the dynamic nature of customer behavior and decision-making, enhancing the comprehensiveness of predictive models in this field.

6. REFERENCE

[1] Applebaum, W. “Studying customer behavior in retail stores”. *Journal of marketing*, 16(2), 172-178, October 1951.

[2] Bounsaythip, C., Rinta-Runsala, E. “Overview of Data Mining for Customer Behavior Modelling”. *VTT Information Technology Research Report, Version, 1,* 1-53 , June 2001.

[3] PriceWaterHouseCoopers. “The CRM Handbook: from Group to multiindividual”. *PriceWaterHouseCoopers*, July 1999.

[4] Corcell Statistical Consulting Unit. “Ordinal Logistic Regression models and Statistical Software: What You Need to Know”. *Statnews*, 91, August 2020.

[5] A. Agresti, *Foundations of Linear and Generalized Linear Models*. New Jersey: John Wiley & Sons, Inc., 2015.

[6] A. J. Dobson, A. G. Barnett, *An Introduction to Generalized Linear Models*. Fourth Edition. Boca Raton: CRC Press, 2018.

[7] Perceptive Analytics, “Customer Repeated Purchase Likelihood Dataset”, *R-Bloggers*, 2019. [Online]. Available at: https://r-posts.com/wp-content/uploads/2018/12/data.txt. [Accessed: 9 December 2023]

[8] Colorado States University. “Likelihood Ratio Tests”. March 2017.

[9] D.C. Montgomery, E. A. Peck, and G.G. Vining, *Introduction to Linear Regression Analysis,* 5th Edition. New Jersey: John Wiley & Sons, Inc., 2012.

[10] J. H. Aldrich and F. D. Nelson, *Linear Probability, Logit, and Probit Models*. Beverly Hills: SAGE Publications, Inc., 1984

[11] Das, S., & Rahman, R. M. (2011). Application of ordinal logistic regression analysis in determining risk factors of child malnutrition in Bangladesh. Nutrition journal, 10(1), 1-11.

[12] Liu, X., & Koirala, H. (2012). Ordinal regression analysis: Using generalized ordinal logistic regression models to estimate educational data. Journal of modern Applied Statistical methods, 11(1), 21.

[13] Chang, E. C., & Fan, X. (2013). More promoters and less detractors: Using generalized ordinal logistic regression to identify drivers of customer loyalty. International Journal of Marketing Studies, 5(5), 12.

7. APPENDIX

7.1. Data

rpurchase	coupon	peers	quality
high probability	0	0	3,25
medium probability	1	0	3,2
low probability	1	1	3,93
medium probability	0	0	2,8
medium probability	0	0	2,52
low probability	0	1	2,58
medium probability	0	0	2,55
medium probability	0	0	2,72
low probability	0	0	2,99
medium probability	1	0	3,49
low probability	1	1	3,64
medium probability	0	0	2,83
high probability	0	1	3,89
medium probability	0	0	2,67
low probability	1	0	3,56
low probability	0	0	3,08
low probability	0	1	3,49
low probability	0	0	2,16
high probability	0	1	3,35
medium probability	0	0	3,39
high probability	0	0	2,74
medium probability	1	0	3,19
low probability	0	0	2,43
low probability	0	0	2,82
low probability	0	1	2,99
medium probability	0	1	3,26
medium probability	1	0	3,13
medium probability	0	0	3,36
high probability	0	1	2,78

low probability	0	0	2,89
medium probability	0	0	3,37
low probability	0	1	2,94
low probability	0	0	2,97
low probability	1	1	3,8
low probability	0	0	2,73
low probability	0	0	2,61
low probability	0	0	2,84
medium probability	0	0	2,49
medium probability	0	0	2,74
low probability	0	0	2,25
low probability	0	0	2,02
medium probability	1	0	2,84
low probability	0	0	2,71
medium probability	0	0	2,88
medium probability	1	0	2,46
low probability	0	0	3,03
low probability	0	0	3,09
low probability	0	0	2,56
low probability	0	0	2,08
medium probability	0	0	2,93
low probability	0	0	3,44
low probability	0	0	2,75
high probability	0	1	2,95
low probability	0	0	2,88
medium probability	0	0	2,96
medium probability	0	1	3,9
medium probability	0	0	2,76
low probability	0	0	2,5
low probability	0	0	3,23
low probability	0	0	2,43
medium probability	0	0	2,77

low probability	0	0	2,93
low probability	1	0	3,21
low probability	0	0	3,49
high probability	1	0	3,56
medium probability	0	0	3,16
low probability	0	1	3,22
low probability	0	0	2,9
low probability	0	0	3,27
low probability	0	1	3,31
low probability	0	0	3,61
low probability	0	0	2,54
low probability	0	0	2,96
high probability	1	0	3,62
low probability	0	1	3,01
low probability	0	1	3,59
low probability	0	0	2,94
high probability	1	1	3,8
medium probability	1	0	2,67
low probability	1	0	3,71
low probability	0	0	2,48
low probability	0	0	2,71
low probability	1	0	2,24
low probability	0	0	2,52
low probability	0	0	3,25
high probability	0	0	2,56
low probability	0	0	2,89
low probability	0	0	3,04
medium probability	0	0	2,84
medium probability	0	0	3,99
high probability	1	0	3,39
medium probability	0	1	3,23
low probability	0	1	3,5

high probability	0	0	2,54
low probability	0	0	3,05
high probability	0	0	2,99
medium probability	0	0	3,12
low probability	0	0	3,15
medium probability	0	0	2,46
medium probability	0	0	2,88
low probability	0	0	2,73
low probability	0	0	3,15
low probability	0	0	2,91
medium probability	0	0	3,56
medium probability	1	0	3,48
medium probability	0	1	2,55
medium probability	1	0	3,11
medium probability	1	0	2,64
medium probability	0	0	2,5
medium probability	0	0	3,07
medium probability	0	0	2,64
low probability	0	0	2,98
low probability	0	0	2,62
medium probability	0	0	2,79
medium probability	0	1	3,45
low probability	0	0	3,6
low probability	0	0	3,07
low probability	1	0	3,27
low probability	0	1	3,3
low probability	0	0	2,56
medium probability	0	0	3,38
low probability	0	0	2,25
low probability	0	0	2,41
low probability	0	0	2,59
medium probability	1	1	3,43

medium probability	0	0	2,97
medium probability	1	0	3,08
low probability	0	1	3,33
low probability	0	0	2,83
low probability	0	0	2,8
low probability	0	0	2,78
low probability	0	0	2,39
low probability	0	0	3,01
medium probability	0	0	2,64
medium probability	0	0	2,54
medium probability	1	0	3,11
low probability	0	0	2,97
high probability	1	0	3,6
low probability	0	0	2,97
low probability	0	0	3,18
low probability	0	0	3,5
low probability	0	0	3,09
low probability	0	0	3,54
low probability	0	0	2,97
high probability	0	0	2,98
low probability	0	0	3,04
low probability	0	0	2,98
low probability	0	0	2,31
low probability	0	0	2,49
low probability	0	1	2,89
high probability	1	1	3,27
medium probability	0	0	2,94
medium probability	0	0	3,53
medium probability	0	0	3,1
low probability	1	0	3,24
low probability	0	0	2,43
low probability	0	0	2,12

medium probability	0	0	3,21
medium probability	0	0	3,15
low probability	0	0	3,38
medium probability	0	0	2,69
medium probability	0	0	3,08
medium probability	0	0	3,15
low probability	0	0	2,27
low probability	0	0	2,9
low probability	0	0	3,64
low probability	0	0	2,85
low probability	0	1	3,38
low probability	0	0	3,7
low probability	0	0	3,24
low probability	0	0	3,13
low probability	0	0	2,4
medium probability	0	0	3,07
low probability	0	1	3,01
medium probability	0	0	3,14
high probability	0	0	2,94
low probability	0	0	2,21
medium probability	0	0	2,85
medium probability	1	0	2,87
high probability	0	0	2,61
low probability	0	0	3,36
medium probability	0	0	3,5
medium probability	0	0	3,64
high probability	1	0	3,41
high probability	0	0	2,4
high probability	0	1	3,2
low probability	0	0	3,21
low probability	0	0	2,52
medium probability	0	0	2,63

low probability	0	0	2,93
medium probability	0	0	2,55
low probability	0	1	3,11
medium probability	0	0	3,33
low probability	0	0	3,21
medium probability	0	0	3,04
medium probability	0	0	3,28
medium probability	0	0	2,7
low probability	0	0	2,86
low probability	0	0	3,28
low probability	0	0	3,35
low probability	1	1	2,84
low probability	0	0	2,78
medium probability	0	0	3,68
high probability	0	0	3,55
low probability	0	0	3,51
low probability	1	1	3,37
high probability	0	1	3,1
medium probability	1	0	3,19
low probability	0	0	2,82
low probability	0	0	3,07
high probability	0	1	2,96
low probability	0	0	2,63
low probability	0	0	2,46
high probability	0	0	3,06
medium probability	0	0	3,19
low probability	0	0	2,54
low probability	0	0	2,47
low probability	0	0	3,28
medium probability	0	0	2,63
medium probability	0	0	3,21
medium probability	1	0	2,79

medium probability	0	0	3,61
low probability	0	0	2,61
medium probability	0	1	3,03
low probability	0	1	2,48
low probability	0	1	3,09
high probability	1	0	3,14
medium probability	0	0	2,64
low probability	0	0	3,03
low probability	0	1	3,04
low probability	0	0	2,87
medium probability	0	0	2,85
medium probability	0	0	2,99
low probability	0	0	2,22
low probability	0	1	3,13
low probability	0	0	2,66
low probability	0	0	3,1
medium probability	0	0	3,72
medium probability	0	0	2,44
low probability	0	0	3,03
low probability	0	0	2,39
medium probability	0	0	3,42
medium probability	0	0	2,56
low probability	0	0	2,83
low probability	0	0	2,66
high probability	1	1	3,44
medium probability	1	0	2,87
medium probability	0	0	2,77
low probability	0	0	3,21
low probability	1	0	3,29
medium probability	1	0	2,85
medium probability	0	0	2,82
medium probability	0	0	3,69

medium probability	0	0	3,14
medium probability	1	1	3,07
low probability	0	0	2,91
low probability	0	0	2,88
high probability	1	0	3,52
low probability	0	0	3,09
medium probability	1	0	3,35
medium probability	0	0	2,73
medium probability	0	0	2,72
low probability	1	0	2,71
low probability	0	0	3,52
low probability	1	0	3,44
low probability	0	0	2,97
medium probability	0	0	2,87
high probability	1	0	3,2
low probability	0	0	3,05
low probability	0	0	2,39
medium probability	0	0	3,56
low probability	1	0	2,87
low probability	0	0	3,26
low probability	0	0	2,91
high probability	0	0	2,69
medium probability	1	0	2,52
medium probability	0	0	3,37
low probability	0	0	3,15
medium probability	0	0	2,59
low probability	0	0	2,75
low probability	0	0	3,55
low probability	0	0	2,86
medium probability	1	0	2,98
low probability	0	0	2,67
low probability	0	0	2,98

medium probability	0	0	2,95
low probability	1	1	2,76
medium probability	0	0	3,27
low probability	0	1	2,73
low probability	0	0	1,89
low probability	0	0	3,04
medium probability	0	1	2,43
medium probability	0	0	3,32
medium probability	0	1	3,55
low probability	0	0	2,14
low probability	0	0	2,82
medium probability	0	0	2,78
medium probability	0	0	2,71
high probability	1	0	3,03
low probability	0	1	3,02
low probability	0	0	2,87
medium probability	0	0	3,2
low probability	0	0	2,73
low probability	0	0	3,56
medium probability	0	0	2,42
low probability	0	0	3,1
low probability	0	0	2,75
medium probability	0	0	3,12
low probability	0	0	2,74
low probability	0	0	2,59
low probability	0	0	3,66
low probability	0	0	3,25
low probability	0	0	3,4
high probability	0	0	3,23
low probability	0	0	3,04
medium probability	1	0	2,37
medium probability	0	0	2,6

medium probability	1	1	3,17
low probability	0	0	3,33
medium probability	0	0	2,83
low probability	0	0	2,79
low probability	0	0	3,2
low probability	0	0	2,62
low probability	0	0	2,3
low probability	0	0	2,28
low probability	0	0	3,5
medium probability	0	0	2,86
low probability	0	0	3,08
medium probability	1	0	2,87
low probability	1	0	2,46
medium probability	0	1	3,71
low probability	0	0	3,19
medium probability	0	0	2,56
low probability	0	0	3,04
low probability	0	0	3,26
low probability	0	0	3,29
low probability	0	0	2,72
low probability	1	0	2,62
low probability	0	0	2,76
low probability	0	0	3,14
low probability	0	0	2,84
low probability	0	0	2,24
medium probability	0	0	3,55
low probability	0	0	2,69
high probability	0	0	2,93
medium probability	1	0	2,57
medium probability	0	0	2,96
high probability	0	0	3,37
high probability	0	0	2,95

low probability	0	0	2,43
medium probability	0	1	3,74
medium probability	0	0	3,06
medium probability	0	0	2,71
low probability	0	0	3,12
low probability	0	0	3,38
low probability	0	1	2,37
medium probability	0	0	2,86
low probability	0	1	2,9
high probability	1	0	3,24
medium probability	1	1	3,5
low probability	1	0	3,64
medium probability	0	0	3,1
high probability	0	0	2,7
low probability	0	0	3,04
low probability	0	0	2,92
medium probability	0	0	2,34
medium probability	0	0	2,48
low probability	0	0	3,25
low probability	0	0	3,76
low probability	0	1	3,48
low probability	0	0	3,38
low probability	0	0	3,36
low probability	0	0	1,97
low probability	1	0	2,92
low probability	0	0	3,35
low probability	0	0	3,11
medium probability	0	0	3,22
medium probability	0	1	3,89
low probability	0	0	2,69
low probability	0	0	2,33
low probability	0	0	3,29

medium probability	0	0	3,1
low probability	0	0	3,13
medium probability	1	0	3,67
medium probability	0	0	2,21
low probability	0	0	2,6
medium probability	0	1	3,47
high probability	0	0	3,07
low probability	0	0	2,78
low probability	0	0	3,11
low probability	0	1	2,66
medium probability	0	0	3,53
high probability	0	0	2,97
medium probability	1	0	3,31
medium probability	1	0	3,53
low probability	0	0	3,69
low probability	0	0	2,62
medium probability	0	0	2,24
medium probability	0	0	3,25
high probability	0	0	3,51

7.2. R Syntax

```
library(MASS)

View(data)

#Ordering the dependent variable
data$rpurchase = factor(data$rpurchase, levels = c("low probability", "medium probability",
                                                   "high probability"), ordered = TRUE)
data$peers = factor(data$peers, levels = c("0", "1"), ordered = TRUE)
data$coupon = factor(data$coupon, levels = c("0", "1"), ordered = TRUE)

#Summarizing the data
summary(data)
#Making frequency table
Summary_table1 <- table(data$rpurchase, data$coupon)
summary_table2 <- table(data$peers, data$rpurchase)
summary_df1 <- as.data.frame(summary_table1)
summary_df2 <- as.data.frame(summary_table2)

#visualize the  relationship between variables
Product_Repurchase_Probability <- jitter(as.numeric(data$rpurchase),0.1)
Quality <- data$quality
plot(Product_Repurchase_Probability~Quality)

ggplot(summary_df1, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_bar(stat = "identity") +
```

```
  geom_text(aes(label = Freq), position = position_stack(vjust = 0.5), size = 3, color = "black") +      labs(x = "Coupon", y = "Frequency", fill = "Repeated Purchase")

ggplot(summary_df2, aes(x = Var1, y = Freq, fill = Var2)) +
  geom_bar(stat = "identity") +
   geom_text(aes(label = Freq), position = position_stack(vjust = 0.5), size = 3, color = "black") +      labs(x = "Peers", y = "Frequency", fill = "Repeated Purchase")

#Build ordinal logistic regression model
model= polr(rpurchase ~ coupon + peers + quality , data = data, Hess = TRUE)
summary(model)
confint(model)
(ctable <- coef(summary(model)))
(p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2)
(ctable <- cbind(ctable, "p value" = p))
print(exp(coef(model)))

#model2
model1= polr(rpurchase ~ coupon  + quality , data = data, Hess = TRUE)
summary(model1)
(ctable1 <- coef(summary(model1)))
(p <- pnorm(abs(ctable1[, "t value"]), lower.tail = FALSE) * 2)
(ctable <- cbind(ctable1, "p value" = p))
print(exp(coef(model1)))

#Compare AIC & BIC
AIC(model,model2)
BIC(model,model2)

#LRtest
library(lmtest)
lrtest(model,model2)

library(Hmisc)
#Proporotional Odds Assumption
sf <- function(y) {
  c('Y>=1' = qlogis(mean(y >= 1)),
    'Y>=2' = qlogis(mean(y >= 2)),
    'Y>=3' = qlogis(mean(y >= 3)))
}

(s <- with(data, summary(as.numeric(rpurchase) ~ coupon + quality, fun=sf))) #print table

s[, 4] <- s[, 4] - s[, 3]
s[, 3] <- s[, 3] - s[, 3]

s #print table
plot(s, which=1:3, pch=1:3, xlab='logit', main=' ', xlim=range(s[,3:4]))

library(gofcat))
brant.test(model2)
```