Breakfast and Energy Level Study:

Application of Ordinal Logistic Regression Model

Jinju Park

University of Minnesota

Abstract

Proportional-odds cumulative logit model is one of the most common types of ordinal regression models used in medical research to find out the association of explanatory variables with an ordered categorical response. The main purpose of this study is to investigate the effects of having breakfast on the level of energy using proportional-odds model. The results of this study show that the odds of having higher energy level when eating breakfast are 1.9 times greater than the odds when not eating breakfast.

Keywords: Ordinal logistic regression, logistic regression modeling, proportional odds model, odd ratio

Breakfast and Energy Level Study:

Application of Ordinal Logistic Regression Model

Ordinal logistic regression is a useful statistical method to predict an ordered categorical response variable. It is commonly used in medical research where the results are typically measured in ordered scales; the author's strong interest in biostatistics was the motivation to study this method. The main purpose of the study is to examine the effect of breakfast on the level of energy using proportional odds (PO) models. The paper is organized as follows. In Literature Review section, three statistical papers about ordinal regression are briefly discussed. In Methodology section, PO model and assumption as well as data collection process are described. Data analysis section provides detailed interpretation about the regression results, followed by Conclusions section which provides a brief summary of this study, limitations and discussions about future research.

Literature Review

Abreu, Siqueira, Cardoso, and Caiaffa (2008) presented a review of ordinal logistic regression models including PO model and the application of the models for quality of life study. Abreu et al. (2008) also provided a sample size formula for ordinal data, which will be used later in this paper. Brant (1990) addressed an approach to assessing the goodness of fit of PO model (proportional assumption) by comparing separate fits to the binary logistic models underlying the overall model. Bouwmeester, Beurden-Tan, Bennison, and Heeg (2014) compared the PO model and multinomial logistic model for Network meta-analysis in ordered categorical data and found out that PO model produced better results as long as data satisfied the PO assumption.

Methodology

Given the findings in Bouwmeester et al. (2014), this study adopted the PO model to analyze the association between having breakfast (binary explanatory variable) and the level of energy (ordered categorical response). In terms of data collection for this study, all the data were collected from an open-ended virtual environment called the *Island*. The top three largest villages from each island were selected to collect samples that represent the entire population of Islanders. Population group for this study is individuals whose age is between 20 and 50. Interview was conducted on each individual to gather information about gender, village where they live, level of energy, minutes of exercise, age, hours of sleep, and whether they had breakfast this morning. Interview questions can be found in Appendix B.

$$OR_j = \frac{\frac{P(Y \le Y_j|x_A)}{P(Y>Y_j|x_A)}}{\frac{P(Y \le Y_j|x_B)}{P(Y>Y_j|x_B)}} = \frac{odds(A)}{odds(B)} \qquad n = \frac{6*\left[\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2/(logOR)^2\right]}{(1-\sum_{i=1}^{k}\overline{\pi_j}^3)} \qquad (1)$$

In order to calculate the sample size for the study, formulas in (1) above adopted from Abreu et al. (2008) were used. OR in (1) refers to the odd ratio between the two odds of the two groups (A and B) that we would like to compare, which are usually control and treatment group respectively. K is the number of orders of the response categories, α is significance level, β is the probability of type 2 error, and $\overline{\pi_j}$ is the mean proportion of subjects in category j of the two groups, A and B (Abreu et al, 2008). Although this study does not contain a control group, it was assumed that control group (not having breakfast) existed to calculate the sample size. Table 1 below shows rough estimates calculated under the general assumption that not having breakfast is associated with relatively lower energy level. Since the odd ratio is same for all response groups, cumulative odds of medium energy level group for control and treatment group were used to

calculate the odd ratio.

	Odds(control)	Odds(treatment)	OR	$\overline{\pi_{Low}}$	$\overline{\pi_{Med}}$	$\overline{\pi_{High}}$	α	β	n
Estimates	0.7/0.3	0.6/0.4	1.555	0.3	0.5	0.2	0.05	80%	75

Table 1. Estimates for sample size calculation

As a result, it turned out that around 75 samples were needed from each response group, meaning that 225 samples were needed in total. The estimated sample size was within reason so was selected for this study, although it was not calculated as ideally. For the sake of simplicity, 75 samples from each of the three islands, 25 samples from each of the top three villages, were collected.

The energy level which was originally measured in a scale of 1 to 10 was categorized into "Low" if it falls within level 1 to 4, "Medium" for level 5 or 6, and "High" for level 7 to 10 in order to reduce the complexity of interpretation. Independent variables, other than breakfast, include gender (Male/Female), hours of exercise yesterday (continuous), and hours of sleep last night (continuous). Regression equations (2a) and (2b) below were tested for the analysis. j in the cumulative probability refers to the order of the response.

$$\text{logit}(\text{P}(\text{Y} \leq \text{j} \,|X)) = \alpha_j - \beta_1 breakfast - \beta_2 exercise - \beta_3 sleep - \beta_4 gender \quad (2a)$$

$$\text{logit}(\text{P}(\text{Y} \leq \text{j} \,|X)) = \alpha_j - \beta_1 breakfast \quad (2b)$$

The left side of the equations (2a) and (2b) above is called a cumulative logit, which is log-odds of two cumulative probabilities. This cumulative logit "measures how likely the response to be in category j or below versus in a category higher than j" (8.4 - The Proportional-Odds Cumulative Logit Model, n.d.). Since there are three ordered categories in this study, two different logit equations are obtained from the analysis; when j is equal to 1 and when j is equal to 2. The equation (3a) measures the log-odds of response being "Low" versus "Medium" or "High" combined, while

the equation (3b) measures the log-odds of response being “Low” or “Medium” combined versus “High”.

$$L_1 = \text{logit}(\text{P}(\text{Y} \leq 1 \,|X)) = \log\left(\frac{P(Y \leq 1 \mid X)}{P(Y > 1 \mid X)}\right) = \log\left(\frac{\pi(1)}{\pi(2)+\pi(3)}\right) \quad (3a)$$

$$L_2 = \text{logit}(\text{P}(\text{Y} \leq 2 \,|X)) = \log\left(\frac{P(Y \leq 2 \mid X)}{P(Y > 2 \mid X)}\right) = \log\left(\frac{\pi(1)+\pi(2)}{\pi(3)}\right) \quad (3b)$$

Notice that the intercept terms of the equations (2a) and (2b), α_j, change as the category of the response changes, whereas the beta coefficients of the explanatory variables stay the same. This is called **proportional odds assumptions**, which indicates that the coefficients of the lowest group of the response versus all higher groups are the same as the coefficients of the next lowest group versus all higher categories.

Data Analysis

The regression results of equations (2a) and (2b) are given in Table 2 and Table 3 below. According to the results of likelihood ratio test of the two regression models, model (2b) was selected as a final model.

Coefficients	**j = 1**	**j = 2**	**p-value**
Intercept	0.5751 (2.46)	4.2364 (2.48)	0.816 (j =1), 0.088 (j =2)
Breakfast(yes)	0.65956 (0.334)		**0.0484**
Exercise	0.01875 (0.063)		0.766
Sleep	0.38189 (0.303)		0.232
Gender(female)	0.21124 (0.282)		0.453
Residual deviance	363.1676		
AIC	375.1676		

Table 2. Regression results for model (2a). Standard errors in the parentheses.

Coefficients	**j = 1**	**j = 2**	**p-value**
Intercept	-2.423 (0.369)	1.218 (0.3006)	0 (j = 1), 0 (j =2)
Breakfast(yes)	**0.6463** (0.3311)		**0.0509**
Residual deviance	364.8723		
AIC	370.8723		

Table 3. Regression results for model (2b). Standard errors in the parentheses.

Using regression results from Table 3, The coefficient of breakfast, *0.6463*, indicates that when the breakfast increases by 1 unit, which means going from 0 (not having breakfast) to 1 (having breakfast), energy level increases by 0.6463 in the log-odds scale. Using these results and the formulas (4) – (6) below, we can compute the cumulative probability for each category j (j = 1 or 2), odds and odd ratio between having breakfast and not having breakfast.

Cumulative probability: $P(Y \leq j \mid X) = \frac{\exp(\alpha_j - \beta_1 breakfast)}{1+\exp(\alpha_j - \beta_1 breakfast)}$ (4)

Odds: $\frac{P(Y \leq j \mid X)}{P(Y > j \mid X)}) = \exp(\alpha_j - \beta_1 breakfast)$ (5)

Odd ratio: $\frac{\frac{P(Y \leq Y_j \mid no\ breakfast)}{P(Y > Y_j \mid no\ breakfast)}}{\frac{P(Y \leq Y_j \mid breakfast)}{P(Y > Y_j \mid breakfast)}} = \frac{odds(no\ breakfast)}{odds(breakfast)}$ (6)

	j=1 (Low)		j=2 (Low + Medium)	
	No breakfast	breakfast	No breakfast	Breakfast
Cumulative probability	0.081	0.044	0.772	0.639
Odds	0.089	0.046	3.379	1.770
Odd ratio	**1.9085 = exp(0.6463)**			

Table 4. Cumulative probability, odds and odd ratio for each category and breakfast group

Table 4 above shows the cumulative probability, odds, and odd ratio for each response category. It is noticeable that cumulative probability of those who do not have breakfast falling into the "Low" energy level group (0.081) is higher than that of those who have breakfast falling into the "Low" energy level group (0.044). Using the cumulative probabilities in Table 4, we could also calculate predicted probabilities of particular energy level depending on whether individuals have breakfast or not as shown in Table 5 below.

probability	Low	Medium	High
Breakfast = 1 (yes)	0.044	0.595	0.361
Breakfast = 0 (no)	0.081	0.690	0.228

Table 5. Predicted probability of particular energy level

Notice that the value of odd ratio in Table 4, 1.9085, is an exponent of the beta coefficient of "breakfast" variable, 0.6463 in Table 3. The odd ratio indicates that the estimated odds that those who have breakfast are in the higher level of energy group are 1.9 times the estimated odds of those who do not have breakfast. In other words, when breakfast increases by 1 unit (from not having breakfast to having breakfast), the odds of moving from "Low" energy level to "Medium" or "High" energy level combined (or moving from "Low" or "Medium" energy combined to "High" energy level) are 1.9 times greater.

In order to check if the proportional odd assumptions hold, likelihood ratio test was performed to compare the PO model with a multinomial logit model, since it is a nested model of a multinomial logit model. The likelihood ratio test result for this data is 0.39967, which means that "the proportional odds model fits as well as the more complex multinomial logit model" (University of Virginia Library Research Data Services Sciences, n.d.).

Conclusions

This study aimed to identify the association between having breakfast and energy level using ordinal logistic regression models. The regression results indicate that the odds of individuals who eat breakfast having higher energy level are 1.9 times greater than the odds of individuals who do not eat breakfast. One limitation of this study is the lack of information about exercise hours. If it had been possible to collect data about how many hours on average individuals spend exercising per week, instead of hours spent on exercise yesterday (which was the only available data), study results might have differed. In addition, other habits of individuals such as smoking, drinking, or stress level could be further studied.

References

Abreu, M. N., Siqueira, A. L., Cardoso, C. S., & Caiaffa, W. T. (2008). Ordinal logistic regression models: Application in quality of life studies. *Cadernos De Saúde Pública, 24*(Suppl 4). doi:10.1590/s0102-311x2008001600010

Bouwmeester, W., Beurden-Tan, C. V., Bennison, C., & Heeg, B. (2014). The Proportional Odds Model Is More Efficient Than The Multinomial Logistic Model For Network Meta-Analyses Of Ordered Outcomes. Value in Health, 17(7). doi:10.1016/j.jval.2014.08.1883

Brant, R. (1990). Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. *Biometrics,* 46(4), 1171. doi:10.2307/2532457

8.4 - The Proportional-Odds Cumulative Logit Model. (n.d.). Retrieved March 22, 2018, from https://onlinecourses.science.psu.edu/stat504/node/176

University of Virginia Library Research Data Services Sciences. (n.d.). Retrieved March 22, 2018, from http://data.library.virginia.edu/fitting-and-interpreting-a-proportional-odds-model/

Appendix A

```
dat = read.csv("data.csv", header = TRUE)
library(MASS)
m = polr(energylevel ~ breakfast + sex + exercise.hr + sleep.hr, data = dat2, Hess = TRUE)
summary(m)

## Call:
## polr(formula = energylevel ~ breakfast + sex + exercise.hr +
##     sleep.hr, data = dat2, Hess = TRUE)
##
## Coefficients:
##                    Value Std. Error t value
## breakfast1     0.65956     0.33415    1.9738
## sex1           0.21124     0.28198    0.7491
## exercise.hr 0.01875     0.06309    0.2972
## sleep.hr       0.36189     0.30259    1.1960
##
## Intercepts:
##                    Value    Std. Error t value
## low|medium    0.5751 2.4688      0.2330
## medium|high 4.2364 2.4839      1.7056
##
## Residual Deviance: 363.1676
## AIC: 375.1676

m1 = polr(energylevel ~ breakfast,data = dat2, Hess = TRUE)
summary(m1)

## Call:
## polr(formula = energylevel ~ breakfast, data = dat2, Hess = TRUE)
##
## Coefficients:
##                    Value Std. Error t value
## breakfast1 0.6463      0.3311     1.952
##
## Intercepts:
##                    Value     Std. Error t value
## low|medium    -2.4234    0.3695     -6.5580
## medium|high    1.2175    0.3006      4.0495
##
## Residual Deviance: 364.8723
## AIC: 370.8723
```

```
anova(m1, m, test = "Chisq") ##not significant

## Likelihood ratio tests of ordinal regression models
##
## Response: energylevel
##                                                 Model Resid. df Resid. Dev      Test
## 1                                             breakfast       222      364.8723
## 2 breakfast + sex + exercise.hr + sleep.hr          219      363.1676 1 vs 2
##           Df LR stat.      Pr(Chi)
## 1
## 2         3 1.704728 0.6358832
ctable <- coef(summary(m))

p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE)*2
(ctable <-cbind(ctable, "p value" =p))

##                              Value Std. Error      t value       p value
## breakfast1    0.65955690 0.33415209 1.9738225 0.04840193
## sex1               0.21124343 0.28198292 0.7491356 0.45377550
## exercise.hr 0.01874998 0.06309011 0.2971936 0.76631873
## sleep.hr       0.36189157 0.30259016 1.1959793 0.23170464
## low|medium    0.57514591 2.46881921 0.2329640 0.81578939
## medium|high 4.23644327 2.48386531 1.7055849 0.08808538

ctable1 <- coef(summary(m1))

p1 <- pnorm(abs(ctable1[, "t value"]), lower.tail = FALSE)*2
(ctable <-cbind(ctable1, "p value" =p1))

##                              Value Std. Error      t value          p value
## breakfast1      0.6463004    0.3311210    1.951856 5.095533e-02
## low|medium    -2.4234037    0.3695357 -6.557968 5.454573e-11
## medium|high    1.2174844    0.3006475    4.049541 5.131825e-05
nobrf= exp(m1$zeta)/(1+exp(m1$zeta)) ##no breakfast cumulative probability

##    low|medium medium|high
##             0.081             0.772

brf= exp(m1$zeta - m1$coefficients)/(1 + exp(m1$zeta - m1$coefficients)) ##breakfast
cumulative probability
##    low|medium medium|high
##             0.044             0.639
nobrf.odd = nobrf/(1-nobrf) ##no breakfast odds
##    low|medium medium|high
```

```
##             0.089             3.379
brf.odd= brf/(1-brf) ##breakfast odds

##    low|medium medium|high
##             0.046             1.770
or = nobrf.odd/brf.odd ##odd ratio = 1.908

##assumption check
library(nnet)
mlm <- multinom(energylevel ~ breakfast, data=dat2)

## # weights:    9 (4 variable)
## initial    value 247.187765
## iter    10 value 180.962013
## iter    10 value 180.962013
## final    value 180.962013
## converged

M1 <- logLik(m1)
M2 <- logLik(mlm)
G <- -2*(M1[1] - M2[1])

pchisq(G,3,lower.tail = FALSE)

## [1] 0.39967
```

Appendix B

Questionnaire

1. Are you male or female?
2. Where do you live?
3. On a scale from 1 to 10 how energetic do you feel right now?
4. How many minutes did you spend doing any moderate physical activity in the last week?
5. Do you eat breakfast?
6. How many years old are you?
7. How many hours did you sleep last night?