Ordinal Regression

Ryan James J. Martinez

Department of Mathematics and Statistics
Mindanao State University - Iligan Institute of Technology

April 2024

Outline

Introduction

In many research studies and real-world applications, outcomes of interest are not simply binary, but rather fall into multiple ordered categories. Understanding and modeling such ordinal outcomes is crucial for making decisions.

Background of the Model

Author

Peter McCullagh

- A Northern Irish-born American statistician
- Thesis - Analysis of Ordered Categorical Data (1977)

Motivation of the Study

Develop statistical methods that even after combining levels of responses, the validity of conclusion will not be affected by the new number of responses.

Ordinal Logistic Regression

Let Y denote the ordered response with k categories, with $k \geq 3$. Then $P(Y \leq j)$ is the cumulative probability of Y less than or equal to a specific category $j = 1, ..., k-1$.

The model can be defined as

$$log\left\{\frac{P(Y \leq j)}{P(Y > j)}\right\} = logit(P(Y \leq j)) = \beta_{j0} - \beta_1 X_1 - \cdots - \beta_p X_p$$

where $j = 1, ..., k-1$

Ordinal Logistic Regression

Assumptions

1. The dependent variable is ordered.
2. One or more of the independent variables are either continuous, categorical or ordinal.
3. Adequate cell count. No zero cell count.
4. No multicollinearity.
5. Parallel Regression Lines

Example 1

A study looks at factors that influence the decision of whether to apply to graduate school. College juniors are asked if they are unlikely, somewhat likely, or very likely to apply to graduate school. Data on parental educational, a binary variable indicating if at least one parent has attended graduate school and whether the undergraduate institution is public or private, and the current GPA of the student is also collected. The researchers have reason to believe that the "distances" between these three points are not equal. For example, the "distance" between "unlikely" and "somewhat likely" may be shorter than the distance between "somewhat likely" and "very likely".

Variables

apply (response variable)

A student feels they are "Unlikely"(1), "Somewhat likely"(2), or "Very likely"(3) to apply to graduate school.

pared

1 if at least one parent has a graduate degree; 0 otherwise

public

1 - public; 0 - private

gpa

gpa variable is the student's grade point average. (1-4)

```r
head(training_data)
```

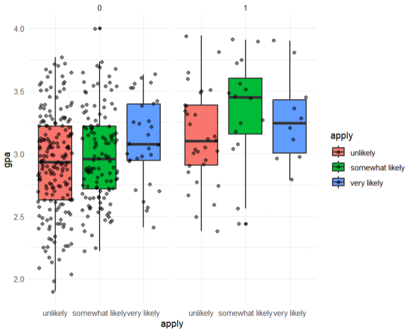
```
            apply pared public  gpa
1     very likely     0      0 3.26
2 somewhat likely     1      0 3.21
3        unlikely     1      1 3.94
4 somewhat likely     0      0 2.81
5 somewhat likely     0      0 2.53
6        unlikely     0      1 2.59
```

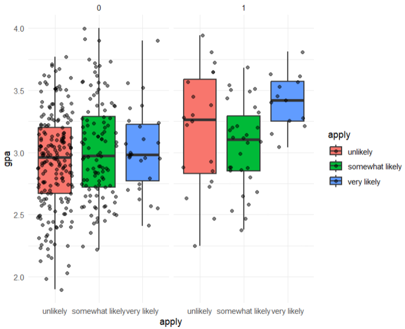
```r
# Checking for Missing values
sapply(training_data, function(x) sum(is.na(x)))
```

```
 apply  pared public    gpa
     0      0      0      0
```

```r
ftable(xtabs(~ public + apply + pared, data = training_data))
```

```
                      pared   0   1
public apply                       
0      unlikely             175  14
       somewhat likely       98  26
       very likely           20  10
1      unlikely              25   6
       somewhat likely       12   4
       very likely            7   3
```


0
1
4.0
3.5
3.0
2.5
2.0
gpa
unlikely
somewhat likely
very likely
apply
apply
unlikely
somewhat likely
very likely


0
1
4.0
3.5
3.0
2.5
2.0
gpa
unlikely
somewhat likely
very likely
apply
apply
unlikely
somewhat likely
very likely

```r
training_data$apply <- ifelse(training_data$apply == "unlikely", 1,
             ifelse(training_data$apply == "somewhat likely", 2, 3))
training_data$apply <- as.factor(training_data$apply)
head(training_data)
```

	apply	pared	public	gpa
1	3	0	0	3.26
2	2	1	0	3.21
3	1	1	1	3.94
4	2	0	0	2.81
5	2	0	0	2.53
6	1	0	1	2.59

```r
# Parallel Regression Assumption
par.reg <- polr(apply ~ pared + public + gpa, data = training_data)
# Brant's Test
brant(par.reg)
```

```
--------------------------------------------
Test for        X2      df      probability
--------------------------------------------
Omnibus         4.34    3       0.23
pared           0.13    1       0.72
public          3.44    1       0.06
gpa             0.18    1       0.67
--------------------------------------------

H0: Parallel Regression Assumption holds
```

```r
fit <- polr(apply ~ pared + public + gpa,
            data = training_data, Hess=TRUE)
summary(fit)
```

```
polr(formula = apply ~ pared + public + gpa, data = training_data,
    Hess = TRUE)

Coefficients:
          Value Std. Error t value
pared   1.04769     0.2658  3.9418
public -0.05879     0.2979 -0.1974
gpa     0.61594     0.2606  2.3632

Intercepts:
    Value   Std. Error t value
1|2  2.2039  0.7795     2.8272
2|3  4.2994  0.8043     5.3453

Residual Deviance: 717.0249
AIC: 727.0249
```

```r
coefs <- coef(summary(fit))
## Calculate CI
lb <- coefs[,1] - 1.96*coefs[,2]
ub <- coefs[,1] + 1.96*coefs[,2]
# Calculate P-value
p <- pnorm(abs(coefs[, "t value"]), lower.tail = FALSE) * 2
# Add Column
coefs <- cbind(coefs, "lower" = lb, "upper" = ub, "p value" = p)
coefs %>% round(., 4)
```

```
          Value Std. Error t value   lower  upper p value
pared1   1.0477     0.2658  3.9418  0.5267 1.5686  0.0001
public1 -0.0588     0.2979 -0.1974 -0.6426 0.5250  0.8435
gpa      0.6159     0.2606  2.3632  0.1051 1.1268  0.0181
1|2      2.2039     0.7795  2.8272  0.6760 3.7318  0.0047
2|3      4.2994     0.8043  5.3453  2.7229 5.8758  0.0000
```

The estimated model can be written as:

$$logit(\hat{P}(Y \leq 1) = 2.20 - 1.05 * \textbf{pared} - (-0.6) * \textbf{public} - 0.62 * \textbf{gpa}$$

$$logit(\hat{P}(Y \leq 2) = 4.30 - 1.05 * \textbf{pared} - (-0.6) * \textbf{public} - 0.62 * \textbf{gpa}$$

pared (Parental Education Status)

Holding all other variables constant, if a student's parent has attended graduate school (pared $= 1$) rather than not (pared $= 0$), the log-odds of the student being in a higher category (e.g., from "Unlikely" to "Somewhat likely", or from "Somewhat likely" to "Very likely") of applying to graduate school increase by approximately 1.05 units.

public (Institution Type)

Holding all other variables constant, there is no statistically significant effect of whether the undergraduate institution is public (public $= 1$) or private (public $= 0$) on the log-odds of a student being in a higher category of likelihood to apply to graduate school.

gpa (Grade Point Average)

Holding all other variables constant, for every one-unit increase in GPA, the log-odds of a student being in a higher category of likelihood to apply to graduate school increase by approximately 0.62 units.

Transition from "Unlikely" to "Somewhat likely"

Holding all other variables constant, the log-odds of a student transitioning from feeling "Unlikely" to "Somewhat likely" to apply to graduate school increase by approximately 2.20 units.

Transition from "Somewhat likely" to "Very likely"

Holding all other variables constant, the log-odds of a student transitioning from feeling "Somewhat likely" to "Very likely" to apply to graduate school increase by approximately 4.30 units.

```r
# convert coefficients into odds ratio, combine with CIs
round(exp(cbind(OR = coefs[,1], ci)), 4)
```

```
             OR   lower    upper
pared1   2.8511  1.6934   4.8001
public1  0.9429  0.5259   1.6905
gpa      1.8514  1.1108   3.0857
1|2      9.0604  1.9660  41.7552
2|3     73.6529 15.2241 356.3251
```

pared

For students whose parents did attend college, the odds of being more likely (i.e., very or somewhat likely versus unlikely) to apply is 2.85 times that of students whose parents did not go to college, holding constant all other variables.

public

There is no statistically significant difference in the odds of a student being in a higher category of likelihood to apply to graduate school between public and private undergraduate institutions. The odds ratio of 0.94 suggests that the odds are slightly lower for students from public institutions, but the 95% CI includes 1, indicating that the difference is not statistically significant.

gpa

For every one unit increase in student's GPA the odds of being more likely to apply (very or somewhat likely versus unlikely) is 1.85 times (85% increase), holding the other variables constant

Transition from unlikely to somewhat likely

The odds ratio of 9.0604 means that the odds of transitioning from "Unlikely" to "Somewhat likely" to apply to graduate school is approximately 9.0604 times higher.

Transition from somewhat likely to very likely

The odds of transitioning from somewhat likely to very likely to apply graduate school is 73.6529.

```r
## Assessing of Model Fit
# Nagelkerke's R-squared & The likelihood ratio test
nagelkerke(fit)
```

```
$Pseudo.R.squared.for.model.vs.null
                             Pseudo.R.squared
McFadden                            0.0326231
Cox and Snell (ML)                  0.0586601
Nagelkerke (Cragg and Uhler)        0.0695655

$Likelihood.ratio.test
 Df.diff LogLik.diff Chisq    p.value
      -3      -12.09 24.18 2.2905e-05
```

```r
lipsitz.test(fit)
```

```
        Lipsitz goodness of fit test for ordinal response models

data:  formula:  apply ~ pared + public + gpa
LR statistic = 8.5407, df = 9, p-value = 0.4807
```

```r
# Prediction
predicted_data <- predict(fit, testing_data, type = "class")
confusionMatrix(testing_data$apply, predicted_data$apply)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3
         1 54  1  0
         2 28  6  0
         3 10  1  0

Overall Statistics

               Accuracy : 0.6
                 95% CI : (0.4972, 0.6967)
```

Example 2

passengerClass (response variable)

1st, 2nd, or 3rd class.

survived

1 - yes; 0 - no

sex

1 - male; 0 - female

age

in years (and for some children, fractions of a year)

```r
# Check for Missing values
sapply(data2, function(x) sum(is.na(x)))
```

```
     survived           sex           age passengerClass
            0             0           263              0
```

```r
# Remove missing values
data2 <- na.omit(data2)
sapply(data2, function(x) sum(is.na(x)))
```

```
     survived           sex           age passengerClass
            0             0             0              0
```

```r
# Change to Factor
data2$passengerClass <- as.factor(data2$passengerClass)
data2$survived <- as.factor(data2$survived)
data2$sex <- as.factor(data2$sex)
```

```r
## Partitioning
# Set the seed for reproducibility
set.seed(123)
# Create an index for partitioning the data
# into training and testing sets
index <- createDataPartition(data2$passengerClass,
                p = 0.75, list = FALSE)
# Create training and testing sets using the index
training_data <- data2[index, ]
testing_data <- data2[-index, ]
```

```r
# Parallel Regression
par.reg <- polr(passengerClass ~ survived + sex + age,
                     data = training_data)
# Brant's Test
brant(par.reg)
```

```
--------------------------------------------
Test for        X2      df      probability
--------------------------------------------
Omnibus         6.07    3       0.11
survivedyes     3.09    1       0.08
sexmale         0.84    1       0.36
age             4.45    1       0.03
--------------------------------------------

H0: Parallel Regression Assumption holds
```

```r
## Modeling
fit <- polr(passengerClass ~ survived + sex + age,
                data = training_data, Hess=TRUE)
coefs <- coef(summary(fit))
# Calculate CI
lb <- coefs[,1] - 1.96*coefs[,2]
ub <- coefs[,1] + 1.96*coefs[,2]
# Calculate P-value
p <- pnorm(abs(coefs[, "t value"]), lower.tail = FALSE) * 2
# Add Column
coefs <- cbind(coefs, "p value" = p)
coefs %>% round(., 4)
```

```
              Value Std. Error  t value   lower   upper p value
survivedyes -1.6162     0.1840  -8.7858 -1.9768 -1.2557  0.0000
sexmale     -0.1451     0.1820  -0.7970 -0.5019  0.2117  0.4255
age         -0.0660     0.0057 -11.6273 -0.0772 -0.0549  0.0000
1st|2nd     -3.9800     0.2847 -13.9783 -4.5381 -3.4220  0.0000
2nd|3rd     -2.5978     0.2622  -9.9092 -3.1116 -2.0839  0.0000
```

The estimated model can be written as:

$$logit(\hat{P}(Y \leq 1) = -3.98 - (-1.6162) * \textbf{survivedyes} - (-0.1451) * \textbf{sexmale} - (-0.07) * \textbf{age}$$

$$logit(\hat{P}(Y \leq 2) = -2.60 - (-1.6162) * \textbf{survivedyes} - (-0.1451) * \textbf{sexmale} - (-0.07) * \textbf{age}$$

survivedyes

For passenger who survived, the log odds of being in a lower passenger class decrease by approximately 1.61622 units. This implies that among the survivors, there's a significant decrease in the log odds of being in a lower passenger class than that off being in a higher passenger class.

sexmale

For male passengers, the log odds of being in a lower passenger class decrease by approximately 0.14508 units compared to the log odds of being in a higher passenger class. This suggests that among male passengers, there's a decrease in the log odds of being in a lower passenger class relative to being in a higher passenger class.

age

For every one unit increase in age, the log odds of being in a lower passenger class decrease by approximately 0.06603 units. The older the passenger are the less likely they are in lower passenger class

1st—2nd

The intercept value for the transition from 1st to 2nd class is -3.9800. In other words, passengers are much less likely to be in the 1st class compared to the 2nd class.

2nd—3rd

The intercept value for the transition from 2nd to 3rd class is -2.5978. In other words, passengers are much less likely to be in the 2nd class compared to the 3rd class.

```r
# convert coefficients into odds ratio, combine with CIs
round(exp(cbind(OR = ceofs[,1, ci)), 4)
```

```
                OR   lower    upper
survivedyes 0.1986  1.6934   4.8001
sexmale     0.8650  0.5259   1.6905
age         0.9361  1.1108   3.0857
1st|2nd     0.0187  1.9660  41.7552
2nd|3rd     0.0744 15.2241 356.3251
```

survivedyes

For passengers who survived, the odds of being in a lower passenger class decrease by approximately 0.1986 times compared to the odds of being in a higher passenger class. This implies that among the survivors, there's a lower likelihood of being in a lower passenger class. The 95% confidence interval suggests that we are 95% confident that the true odds ratio lies between 0.1385 and 0.2849.

sexmale

For male passengers, the odds of being in a lower passenger class are approximately 0.8650 times the odds of being in a higher passenger class. The male passengers have lower odds of being in a lower passenger class relative to being in a higher passenger class. However, the confidence interval (0.6054 to 1.2358) includes 1, indicating that this effect may not be statistically significant at $\alpha = 0.05$ level.

age

For every one-unit increase in age, the odds of being in a lower passenger class decrease by approximately 0.9361 times. The older the passenger are the less likely they are in lower passenger class.

1st—2nd

The odds ratio of 0.0187 suggests that the odds of being in the 1st class are approximately 0.0187 times the odds of being in the 2nd class. In other words, passengers are much less likely to be in the 1st class compared to the 2nd class.

2nd—3rd

The odds ratio of 0.0744 suggests that the odds of being in the 2nd class are approximately 0.0744 times the odds of being in the 3rd class. In other words, passengers are much less likely to be in the 2nd class compared to the 3rd class.

```r
## Assessing of Model Fit
# Nagelkerke's R-squared & The likelihood ratio test
nagelkerke(fit)
```

```
$Pseudo.R.squared.for.model.vs.null
                             Pseudo.R.squared
McFadden                             0.144119
Cox and Snell (ML)                   0.261772
Nagelkerke (Cragg and Uhler)         0.298055

$Likelihood.ratio.test
 Df.diff LogLik.diff  Chisq    p.value
      -3     -119.12 238.25 2.2756e-51
```

```r
lipsitz.test(fit)
```

```
        Lipsitz goodness of fit test for ordinal response models

data:  formula:  passengerClass ~ survived + sex + age
LR statistic = 11.13, df = 9, p-value = 0.2669
```

```r
# Prediction
predicted_data <- predict(fit, testing_data, type = "class")
confusionMatrix(testing_data$passengerClass, predicted_data)
```

```
Confusion Matrix and Statistics

          Reference
Prediction 1st 2nd 3rd
       1st  43   0  28
       2nd  20   0  45
       3rd   8   0 117

Overall Statistics

               Accuracy : 0.613
                 95% CI : (0.551, 0.6724)
```