Ordered Logistic Regression

18 - WISE - Mengyuan Zhang

2021/5/6

1. Introduction to Ordered Logistic Regression

In statistics, the ordered logit model (also called ordered logistic regression or proportional odds model) is an ordinal regression model. Ordered logit models explain variation in an ordered categorical dependent variable (Y) as a function of one or more independent variables ($x_1, x_2, ..., x_M$). Categories must only be ordered (e.g., lowest to highest, weakest to strongest, strongly agree to strongly disagree)—the method does not require that the distance between the categories be equal.

Ordered logit models are typically used when the dependent variable has three to seven ordered categories. When the categories are more than that, we often use ordinary least squares regression instead. If the dependent variable only has two categories, the ordered logit model reduces to our simple logit regression.

1.1 Definitions

Let's establish some notations and introduce the concepts involved in ordered logistic regression. Let Y be an ordinal dependent variable with J categories, and x_i be the independent variable. Then $P(Y \leq j|x_i)$ is the cumulative probability of Y less than or equal to a specific category $j = 1, ..., J-1$ give x_i. The **odds** of being less than or equal a particular category can be defined as:

$$odds = \frac{P(Y \leq j|x_i)}{P(Y > j|x_i)}$$

for $j = 1, ..., J-1$ since $P(Y > J) = 0$ and dividing by zero is undefined. And $P(Y > j|x_i) = 1-P(Y \leq j|x_i)$.

The **log odds** can be written as:

$$log\frac{P(Y \leq j|x_i)}{P(Y > j|x_i)}$$

We define **log odds** as **logit**, so that:

$$logit(P(Y \leq j|x_i)) = log\frac{P(Y \leq j|x_i)}{P(Y > j|x_i)}$$

1.2 Ordered logistic regression model

After introducing the concepts, we can define our ordered logit model. The **ordinal logistic regression model** can be defined as:

$$logit(P(Y \leq j|x_i)) = \beta_{j0} + \beta_{j1}x_{1i} + ... + \beta_{jM}x_{Mi}$$

where β_{j0} are the intercepts for $j = 1, ..., J-1$, $\beta_{j1}, ..., \beta_{jM}$ are coefficients for the M independent variables.

Due to the **parallel lines assumption**, the intercepts are different for each category but the slopes are constant across categories, which simplifies the above equation to:

$$logit(P(Y \leq j|x_i)) = \beta_{j0} + \beta_1 x_{1i} + ... + \beta_M x_{Mi} \tag{1}$$

(**NOTE**: The key assumption in ordinal regression is that the effects of any explanatory variables are consistent or proportional across the different thresholds, that is, intercepts of different Js. Hence this is usually termed the assumption of **parallel lines assumption** or **proportional odds**.)

In R, the ordered logistic regression model (1) is parameterized as:

$$logit(P(Y \leq j|x_i)) = log\frac{P(Y \leq j|x_i)}{P(Y > j|x_i)} = \beta_{j0} - \eta_1 x_{1i} - ... - \eta_M x_{Mi} \tag{2}$$

where $\eta_i = -\beta_i$. The parameters β_{j0} , called thresholds or cutpoints, are in increasing order ($\beta_{1_0} < \beta_{2_0} < ... < \beta_{J-1_0}$).

To interpret the associated probability using the model given in R, We can define $P(Y \leq j|x_i) = p_{ji}$. From above equation, we can get:

$$logit(p_{ji}) = log\frac{p_{ji}}{1 - p_{ji}} = \beta_{j0} - \eta_1 x_{1i} - ... - \eta_M x_{Mi}$$

(**NOTE**: p_{ji} can be transformed as:

$$P(Y \leq j|x_i) = p_{ji} = \frac{exp(\beta_{j0} - \eta_1 x_{1i} - ... - \eta_M x_{Mi})}{1 + exp(\beta_{j0} - \eta_1 x_{1i} - ... - \eta_M x_{Mi})} = \frac{1}{exp(-\beta_{j0} + \eta_1 x_{1i} + ... + \eta_M x_{Mi})}$$

)

The ordered logit model is also known as the proportional odds model because the parallel regression assumption implies the proportionality of the odds of not exceeding the $j-th$ category $odds_{ji} = \frac{p_{ji}}{1-p_{ji}}$. The **ratio of the odds** for different observation of x, say i and k, is:

$$\frac{odds_{ji}}{odds_{jk}} = \frac{p_{ji}/1 - p_{ji}}{p_{jk}/1 - p_{jk}} = exp[-\eta_1(x_{1i} - x_{1k}) - ... - \eta_M(x_{Mi} - x_{Mk})] \tag{3}$$

To be more specific, we give an example to illustrate the above "odds" equation. Suppose we want to see whether a binary independent variable x_1 predicts an ordered outcome of our dependent variable Y.

Due to the parallel lines assumption, even though we have J categories, the coefficient of x_1 stays the same across the J categories. The the two equations for $x_1 = 1$ and $x_1 = 0$ are:

$$log\frac{P(Y \leq j|x_1 = 1)}{P(Y > j|x_1 = 1)} = logit(P(Y \leq j|x_1 = 1)) = \beta_{j0} - \eta_1 x_1 = \beta_{j0} - \eta_1$$

$$log\frac{P(Y \leq j|x_1 = 0)}{P(Y > j|x_1 = 0)} = logit(P(Y \leq j|x_1 = 0)) = \beta_{j0} - \eta_1 x_1 = \beta_{j0}$$

Then we can get:

$$logit\frac{P(Y \leq j|x_1 = 1)}{P(Y \leq j|x_1 = 0)} = logit(P(Y \leq j|x_1 = 1)) - logit(P(Y \leq j|x_1 = 0)) = -\eta_1$$

Since the exponent is the inverse function of the log, we can simply exponentiate both sides of this equation, and by using the property that $log(b) - log(a) = log(b/a)$, we can get:

$$\frac{P(Y \leq j|x_1 = 1)}{P(Y > j|x_1 = 1)} / \frac{P(Y \leq j|x_1 = 0)}{P(Y > j|x_1 = 0)} = exp(-\eta_1)$$

For simplicity of notation and by the proportional odds assumption, we can make:

$$odds_{j(x_1=1)} = \frac{P(Y \leq j|x_1 = 1)}{P(Y > j|x_1 = 1)} = p_1/(1 - p_1)$$

$$odds_{j(x_1=0)} \frac{P(Y \leq j|x_1 = 0)}{P(Y > j|x_1 = 0)} = p_0/(1 - p_0)$$

Then we can define the odds ratio as:

$$\frac{odds_{j(x_1=1)}}{odds_{j(x_1=0)}} = \frac{p_1/(1 - p_1)}{p_0/(1 - p_0)} = exp(-\eta_1) = \frac{1}{exp(\eta_1)}$$

which is in the same form as equation (3).

2. The Secret of Gay Supporters, White People, and Republican

To implement ordered logistic regression in R, we use data from the 2012 Cooperative Congressional Election Study (CCES 2012). The CCES 2012 includes a number of survey questions asking people their personal information, such as their gender, race, attitudes towards homosexuality, identification of party in United States and so on. Among all the questions, five independent variables and a dependent variable have been chosen (The statements in brackets are the corresponding questions). Our aim is to see whether these predictors can explain our dependent well:

Dependent variable:

View (CC422b): 1 = Strongly agree, 2 = Somewhat agree, 3 = Neither agree nor disagree, 4 = Somewhat disagree, 5 = Strongly disagree.

The dependent variable records whether respondents agreed or disagreed with the following statement:

Generations of slavery and discrimination have created conditions that make it difficult for Blacks to work their way out of the lower class.

Independent variables:

(1) Gender: 0 = Female, 1 = Male (Are you a male or female?)

(2) Race: 0 = Others, 1 = White (What race or ethnic group best describe you?)

(3) Child (CC303-6): 0 = No, 1 = Yes (Do you have a child?)

(4) Gay (CC326): 0 = Oppose, 1 = Favor (Do you favor or oppose allowing gays and lesbians to marry legally?)

(5) Party (pid7): 8 point scale where: 1= Strong Democrat, 2 = Not very strong Democrat, 3 = Lean Democrat, 4 = Independent, 5 = Lean Republican, 6 = Not very strong Republican, 7 = Strong Republican, 8 = Not sure (What party identification would you call yourself?)

2.1 Descriptive statistic

First step: define the ordering of the levels in the dependent variable and the relevant independent variables:

```r
dt <- read.csv("dataset.csv")
#Ordering the independent variables
dt$gender = factor(dt$gender, levels = c("0", "1"), ordered = TRUE)
dt$race = factor(dt$race, levels = c("0","1"), ordered = TRUE)
dt$child = factor(dt$child, levels = c("0", "1"), ordered = TRUE)
```

```
dt$gay = factor(dt$gay, levels = c("0", "1"), ordered = TRUE)
#Ordering the dependent variables
dt$view = factor(dt$view, levels = c("Strongly agree", "Somewhat agree",
  "Neither agree nor disagree","Somewhat disagree","Strongly disagree"), ordered = TRUE)
```

Next, we can create the summary statistics and frequency table:

```
#Summarizing the data
summary(dt)
```

```
##  gender    race      child     gay            party
##  0:22998   0: 8521   0:41665   0:21016   Min.   :1.000
##  1:20713   1:35190   1: 2046   1:22695   1st Qu.:2.000
##                                          Median :4.000
##                                          Mean   :3.925
##                                          3rd Qu.:6.000
##                                          Max.   :8.000
##                          view
##  Strongly agree            : 5220
##  Somewhat agree            :10202
##  Neither agree nor disagree: 6593
##  Somewhat disagree         : 8047
##  Strongly disagree         :13649
##
```

```
#Making frequency table
table(dt$gay,dt$view)
```

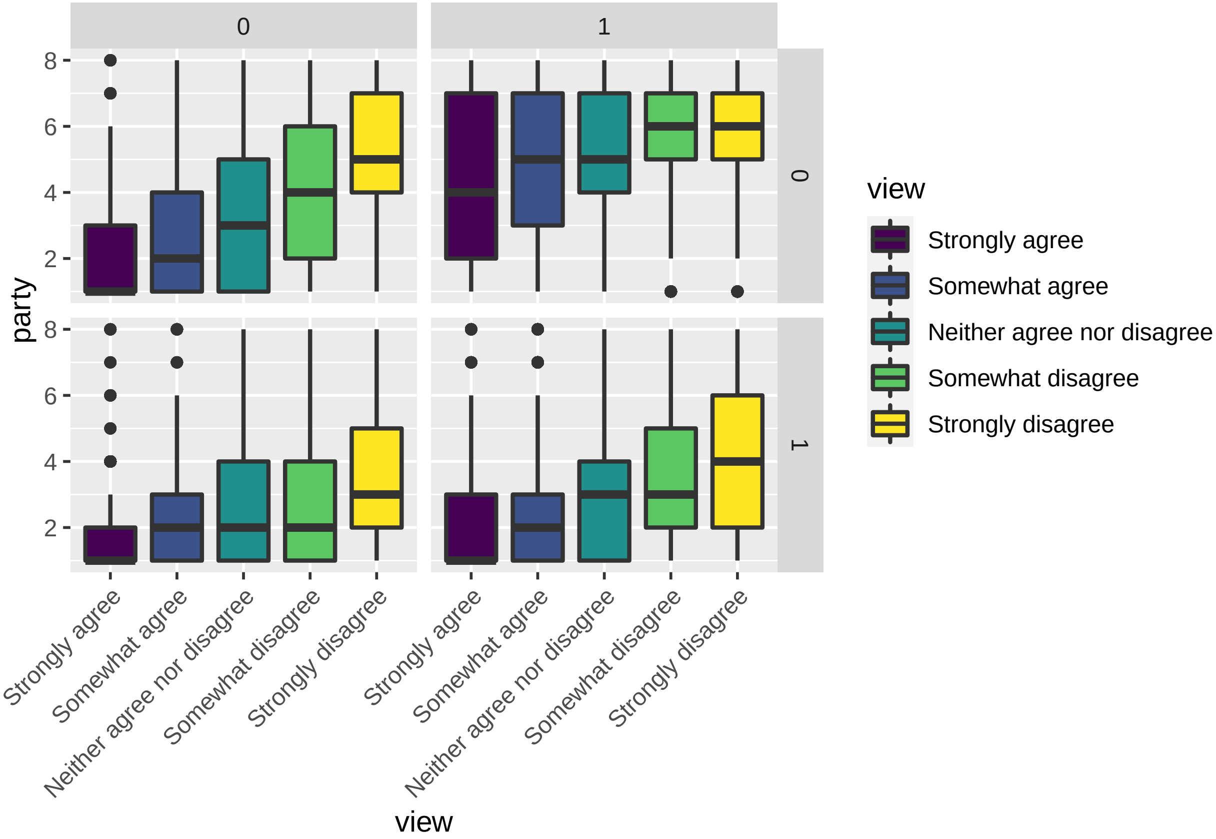
```
##
##      Strongly agree Somewhat agree Neither agree nor disagree Somewhat disagree
##   0           1086           2888                       2912              4446
##   1           4134           7314                       3681              3601
##
##      Strongly disagree
##   0              9684
##   1              3965
```

```
table(dt$party,dt$view)
```

```
##
##      Strongly agree Somewhat agree Neither agree nor disagree Somewhat disagree
##   1           3104           3950                       1506              1080
##   2            530           1520                        967               920
##   3            824           1656                        674               695
##   4            377            874                       1016               950
##   5             87            623                        580              1290
##   6             82            607                        604              1038
##   7            165            835                        860              1889
##   8             51            137                        386               185
##
##      Strongly disagree
##   1               994
##   2               881
##   3               555
##   4              1810
##   5              2881
##   6              1585
##   7              4673
##   8               270
```

To see the relationship among variables more clearly, we can plot the boxplot:

```
library(ggplot2)
ggplot(dt, aes(x = view, y = party, fill = view)) +
  geom_boxplot(size = .75) +
  facet_grid(gay ~ race, margins = FALSE) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))
```



The boxplot shows that for every level of view people have and party people are in, also for particular values of gay and race, what the distribution of our dataset looks like. For example, the box on the top left represents respondents whose race are not white and are against gay/lesbian. We can see that people who declare themselves to be Democrat tend to strongly agree with the views on historical racism's impact on blacks, and people who declare themselves to be Republican tend to strongly disagree with this statement.

2.2 Training and testing set

We partition the data into training and testing set. We will conduct our model using the training set and validate the model using the testing set:

```
#Dividing data into training and test set
#Random sampling
samplesize = 0.60*nrow(dt)
set.seed(100)
index = sample(seq_len(nrow(dt)), size = samplesize)
```

```r
#Creating training and test set
dttrain = dt[index,]
dttest = dt[-index,]
```

2.3 Conduct the ordered logistic regression model

Now, we conduct the model using the data in training set:

```r
#Build ordered logistic regression model
require(MASS)
model_fit <- polr(view~gender+race+child+gay+party, data = dttrain, Hess = TRUE)
summary(model_fit)
```

```
## Call:
## polr(formula = view ~ gender + race + child + gay + party, data = dttrain,
##     Hess = TRUE)
##
## Coefficients:
##             Value Std. Error t value
## gender.L -0.01778   0.016279  -1.092
## race.L    0.36771   0.020864  17.624
## child.L   0.04968   0.038476   1.291
## gay.L    -0.55382   0.018577 -29.812
## party     0.32257   0.006031  53.485
##
## Intercepts:
##                                               Value    Std. Error t value
## Strongly agree|Somewhat agree                  -1.0126   0.0376   -26.9303
## Somewhat agree|Neither agree nor disagree       0.6421   0.0363    17.6681
## Neither agree nor disagree|Somewhat disagree    1.4230   0.0373    38.1598
## Somewhat disagree|Strongly disagree             2.3747   0.0390    60.8513
##
## Residual Deviance: 73704.29
## AIC: 73722.29
```

The table above displays the value of coefficients and intercepts, and corresponding standard errors and t values. The interpretation for the coefficients is as follows: for example, holding everything else constant, an increase in value of gay (change from oppose to favor) by one unit decrease the value of people's view (change from disagree to agree) in log odds by -0.554.

We can also get the significance of coefficients and intercepts:

```r
summary_table <- coef(summary(model_fit))
pval <- pnorm(abs(summary_table[, "t value"]),lower.tail = FALSE)* 2
summary_table <- cbind(summary_table, "p value" = round(pval,3))
summary_table
```

```
##                                                     Value  Std. Error    t value
## gender.L                                      -0.01777645 0.016278873  -1.091995
## race.L                                         0.36770684 0.020864399  17.623649
## child.L                                        0.04967983 0.038476097   1.291187
## gay.L                                         -0.55381745 0.018576998 -29.811999
## party                                          0.32257450 0.006031077  53.485387
## Strongly agree|Somewhat agree                 -1.01261813 0.037601480 -26.930273
## Somewhat agree|Neither agree nor disagree      0.64214742 0.036344959  17.668129
## Neither agree nor disagree|Somewhat disagree   1.42303786 0.037291592  38.159751
## Somewhat disagree|Strongly disagree          2.37470290 0.039024659  60.851343
##                                             p value
## gender.L                                      0.275
## race.L                                        0.000
## child.L                                       0.197
## gay.L                                         0.000
## party                                         0.000
## Strongly agree|Somewhat agree                 0.000
## Somewhat agree|Neither agree nor disagree     0.000
## Neither agree nor disagree|Somewhat disagree  0.000
## Somewhat disagree|Strongly disagree           0.000
```

It can be seen that the coefficient of race, gay, party and all intercepts are significant at 1% level. (gender and child are not significant)

Formulation

As is introduced before, let J be the total number of categories of the dependent variable and M be the number of independent variables (In our example, $J = 5$ and $M = 5$). The formula of the ordered logistic model is given below:

$$logit(P(Y \leq j|x_i)) = log\frac{P(Y \leq j|x_i)}{P(Y > j|x_i)} = \beta_{j0} - \eta_1 x_{1i} - ... - \eta_M x_{Mi}$$

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

In our dataset:

j=1 refers to “Strongly agree”; j=2 refers to “Somewhat agree”; j=3 refers to “Neither agree nor disagree”; j=4 refers to “Somewhat disagree”; j=5 refers to “Strongly disagree”;

i=1 refers to “gender”; i=2 refers to “race”; i=3 refers to “child”; i=4 refers to “gay”; i=5 refers to “party”.

From the summary table by conducting the ordered logistic regression, we can get the following formula:

$$logit(P(Y \leq 1)) = -1.013 - (-0.018)gender - 0.368race - 0.050child - (-0.554)gay - 0.323party$$

$$logit(P(Y \leq 2)) = 0.642 - (-0.018)gender - 0.368race - 0.050child - (-0.554)gay - 0.323party$$

$$logit(P(Y \leq 3)) = 1.423 - (-0.018)gender - 0.368race - 0.050child - (-0.554)gay - 0.323party$$

$$logit(P(Y \leq 4)) = 2.375 - (-0.018)gender - 0.368race - 0.050child - (-0.554)gay - 0.323party$$

The results are interesting: maybe people who are gay or lesbian supporters have a more sympathetic view towards the impact of history on blacks. Also, people whose race are white and who are Republican instead of Democrat tend to against the views of the impact of history on blacks.

2.4 Compute the predicted probability

If we want to predict the probability corresponding to each perception for an individual, we can test the consumer with the following characteristics:

gender: 1, race: 1, child: 0, gay: 0, party: 7

NO MATTER - who you are, where you're from, your skin colour, your gender identity just: SPEAK YOURSELF, love yourself
LOVE IS LOVE
- Kim Namjoon (BTS)
Love is a terrible thing to Hate
THIS SINNER
WE WON

Figure 1: LGBT

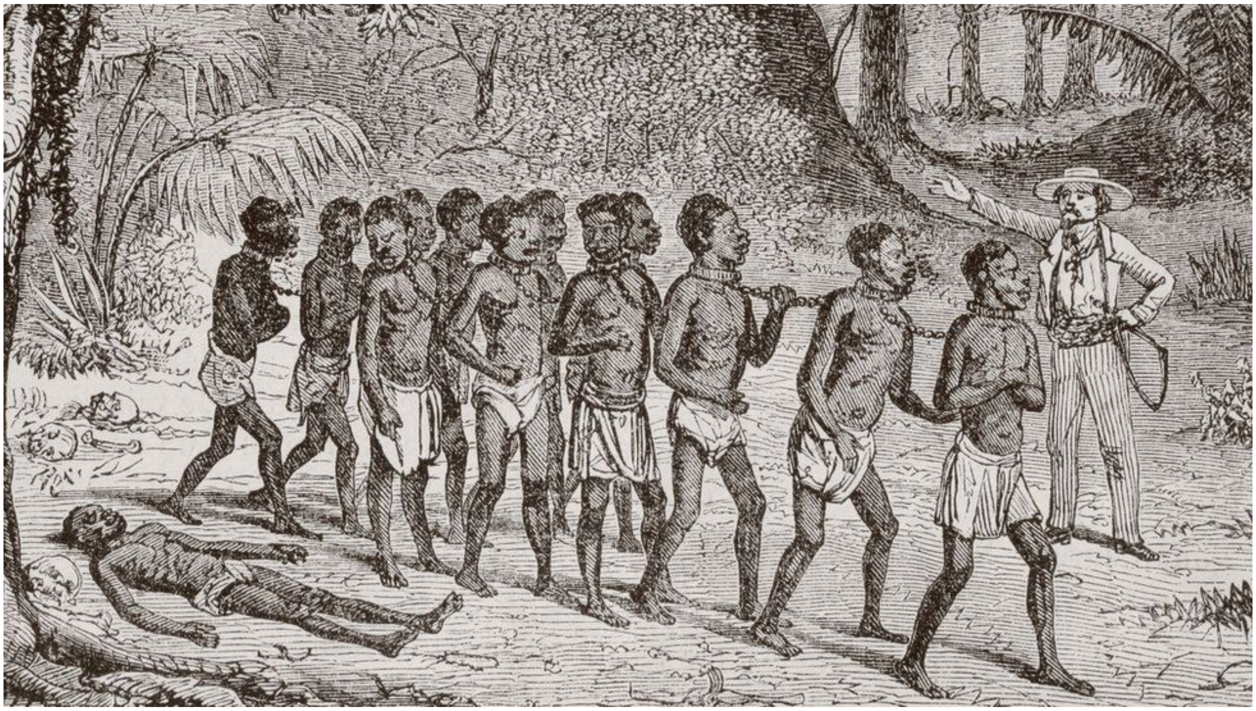


Figure 2: Black slavery in history

```
new_data <- data.frame("gender"= "1","race"="1","child"="0","gay"="0","party"=7)
round(predict(model_fit,new_data,type = "p"), 3)
```

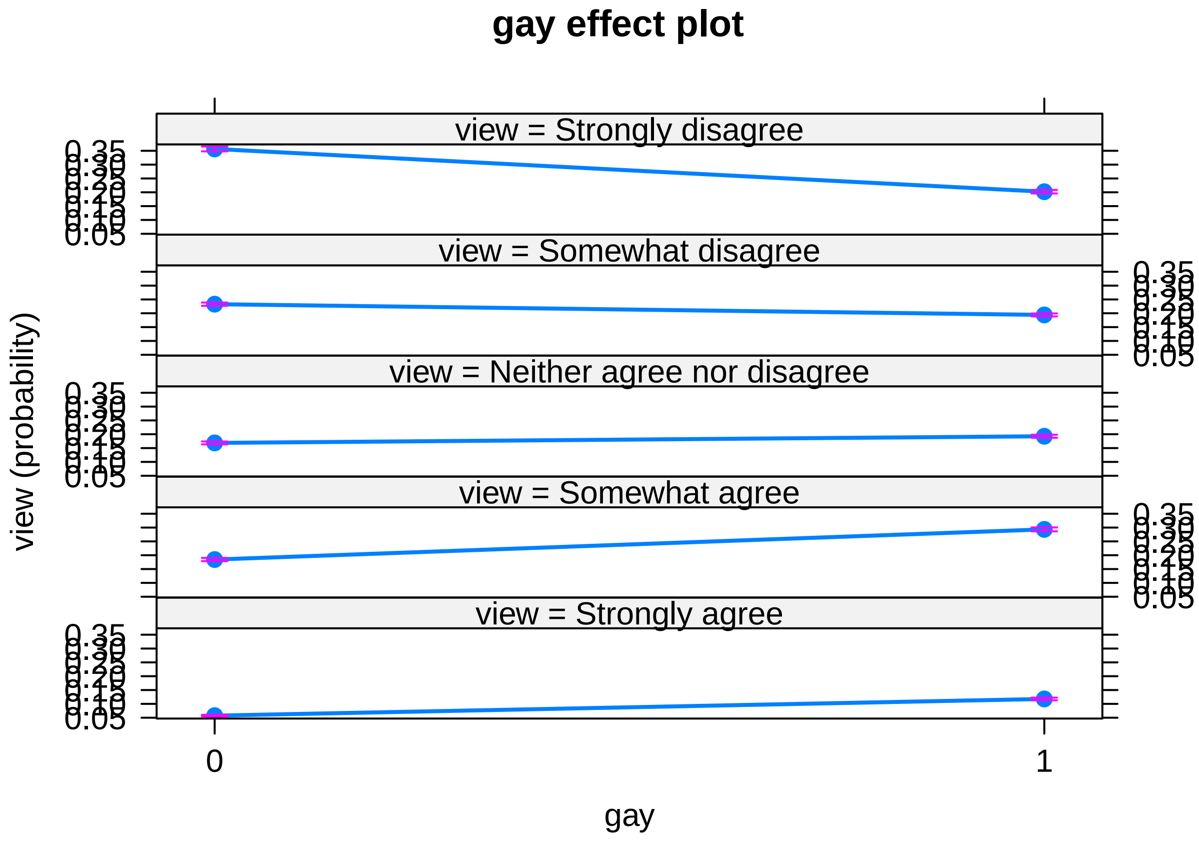
```
##            Strongly agree            Somewhat agree
##                     0.020                     0.078
## Neither agree nor disagree         Somewhat disagree
##                     0.094                     0.189
##          Strongly disagree
##                     0.619
```

The prediction says that: a white male respondent who is a strongly Republican with no child, opposed to gay/lesbian love, strongly disagree with the view of impact of history on blacks with the probability of 0.619!

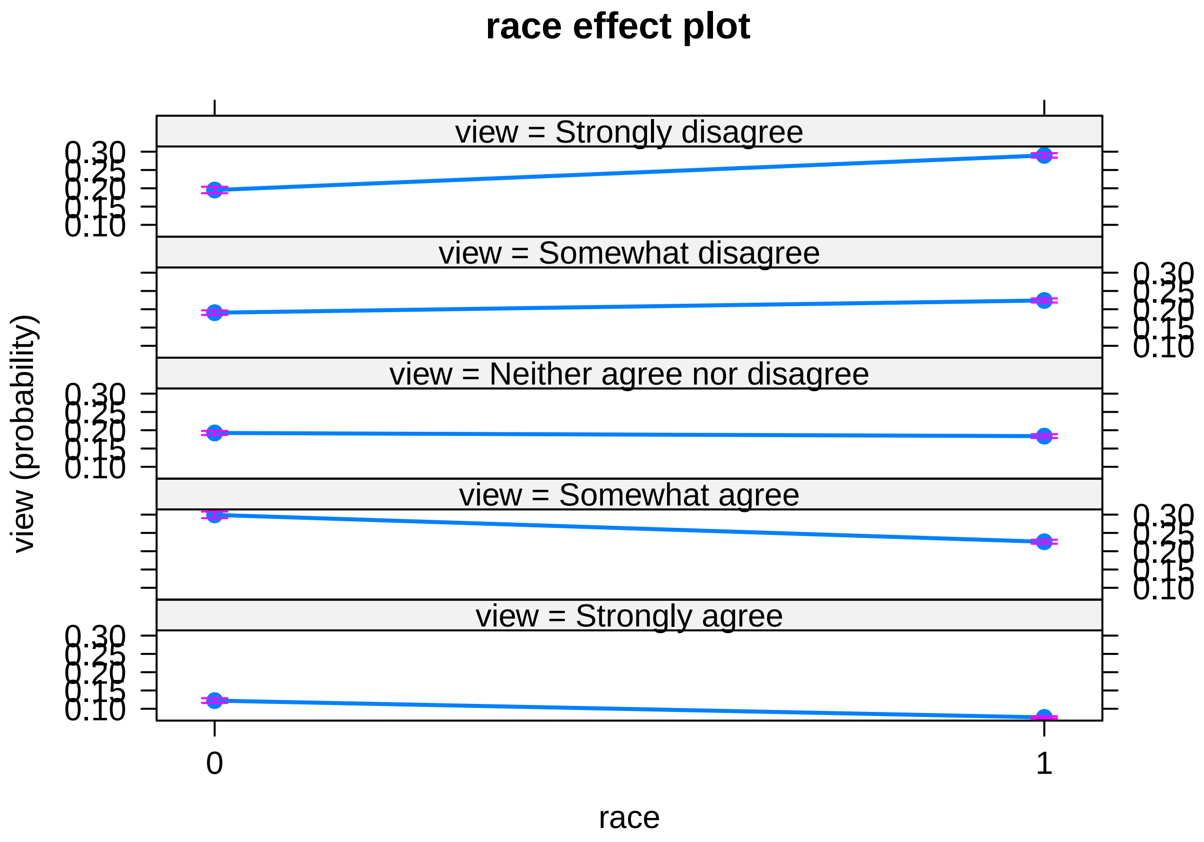
2.5 Interpret with graphs

To better see the effects of independent variables, we plot the effects in the figures below:

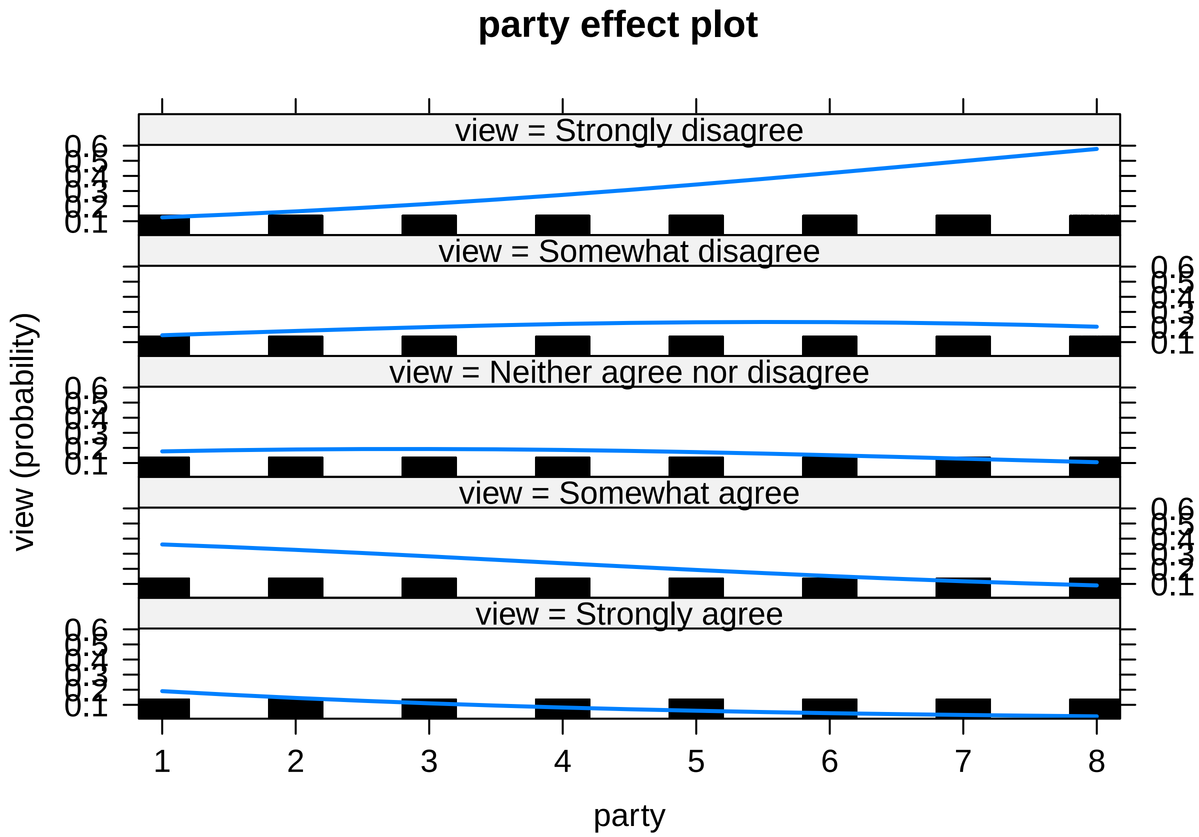
```
#Plotting the effects
library("effects")
plot(Effect(focal.predictors = "gay",model_fit))
```



```
plot(Effect(focal.predictors = "race",model_fit))
```



```
plot(Effect(focal.predictors = "party",model_fit))
```



Gay effect: In the first figure, which contains five small plots, it shows the effect of gay on the views. Gay supporters (denoted as 1) increase the likelihood of the view "strongly agree", while gay opponents increase the likelihood of the view "strongly disagree".

Race effect: The second figure shows the effect of race on views. White people (denoted as 1) increase the likelihood of the view "strongly disagree" and "Somewhat disagree", while people whose race is not white increase the likelihood of the view "Somewhat agree" and "Strongly agree".

Party effect: The third figure shows the effect of different identification of parties on views (1-7 denotes strongly Democrat to strongly Republican, 8 denotes Not sure). People who are Republican (5-lean, 6-not very strong, 7-strong) increase the likelihood of the view "Strongly disagree" and "Somewhat disagree", while people who are Democrat (3-lean, 2-not very strong, 1-strong) increase the likelihood of the view "Strongly agree" and "Somewhat agree".

2.6 Evaluate the ordered logistic model

The evaluation of the model is conducted on the test dataset:

```
#Compute confusion matrix and error
predictview = predict(model_fit,dt)
table(dt$view, predictview)
```

```
##                               predictview
##                                Strongly agree Somewhat agree
##   Strongly agree                            0           4707
##   Somewhat agree                            0           7643
##   Neither agree nor disagree                0           3648
##   Somewhat disagree                         0           3140
##   Strongly disagree                       0           3156
##                              predictview
##                               Neither agree nor disagree Somewhat disagree
##   Strongly agree                                      0                 0
##   Somewhat agree                                      0                 0
##   Neither agree nor disagree                          0                 0
##   Somewhat disagree                                   0                 0
##   Strongly disagree                                   0                 0
##                              predictview
##                               Strongly disagree
##   Strongly agree                           513
##   Somewhat agree                          2559
##   Neither agree nor disagree              2945
##   Somewhat disagree                       4907
##   Strongly disagree                      10493
```

```
mean(as.character(dt$view) != as.character(predictview))
```

```
## [1] 0.585093
```

The matrix shows the performance of the ordered logistic regression model. We observe that the model identifies “strongly agree”, “neither agree nor disagree” and “somewhat disagree” poorly. This happens maybe because of inadequate representation of these categories in the training data set. We also find that the misclassification error for our model is 58.51%. Too high! :(

3. Summary–Surprising Relationship: “Gay and lesbain supporters” with “Views towards black”

First, we discuss the theoretical foundation of ordered logistic regression. ordered logistic regression extends the simple logistic regression model to the situations where the dependent variable is ordinal, i.e. can be ordered.

Second, we do the empirical application to find out things that affect people’s views: the impact of discrimination on blacks in the United States. The dependent variable is views among respondents, which contains five degrees of agreements and disagreements. The independent variables (predictors) are gender, race, child, gay and party. The final results show that race, gay and party are significant predictors of the dependent variable. (So surprise for the significance of our predictor “gay”!)

Interesting empirical results: **maybe people who are gay or lesbian supporters have a more sympathetic view and more likely to admit the impact of history on blacks**. On the contrary, people whose race are white and who are Republican instead of Democrat tend to against the views of the impact of history on blacks.

While the former group of people, who are gay or lesbian supporters, are more likely to believes that history is to blame for the current condition of blacks in the United States, the latter group of people, who are mostly white and Republican, tend to think that it is the black themselves that make it difficult for them to work their way out of the lower class, not to blame on the history!

4. References

(1) How do I Interpret the Coefficient in an Ordinal Logistic Regression in R?

https://stats.idre.ucla.edu/r/faq/ologit-coefficients/

(2) How to Perform Ordinal Logistic Regression in R

https://www.r-bloggers.com/2019/06/how-to-perform-ordinal-logistic-regression-in-r/