Student alcohol consumption

Vadym Dudarenko 444820

Vladimir Shargin 437981

04.06.2022

Abstract

The goal was to research and understand how young generation consume alcohol and what factors influence on the high or low using. For this assessment, we use ordered choice model, that helps to estimate the probability of it. The data is valid for Portuguese students in secondary school.

The economic consequences of alcohol consumption can be severe, particularly for the poor. Apart from money spent on drinks, heavy drinkers may suffer other economic problems such as lost employment opportunities, increased medical and legal expenses.

The study showed that alcohol is a part of students' lives. These findings show a few factors that has a direct impact on the student. It will help to analyze current problem and avoid it in the future.

Introduction

Alcohol abuse among students is an unhealthy lifestyle. In different countries, the large number of underage students who drink alcohol has created many problems and consequences for learning. Causes of alcohol abuse are usually peer pressure, poor life and stress. Students who abuse alcohol may suffer from health problems, poor performance or legal consequences.

Using the data of students math and Portuguese language courses in secondary school, we built the ordered choice model, that help to estimate factors of high or low consumption alcohol.

Literature review

1. Drinking at European universities?

https://www.sciencedirect.com/science/article/pii/S0306460310001802?casa_token=OfwQA8iWH7EAAAAA:2vnpAX0grrhwJ8aXii5fzkOVRbq2K0tq_1bKkVDPAKn4L6NNOIyaFh3gPWubj1svYQXG8xihv_c

These studies generally indicate that alcohol is consumed by a very large proportion of the student population, hence the stress which the authors place on the importance of prevention and intervention efforts

Looking at the characteristics of alcohol-consuming university students in Europe, four patterns of association have been consistently reported: a) male students consume alcohol both more frequently and in higher quantities, including RSOD; although in the UK and Nordic countries some studies did not find gender differences; b) students consume alcohol mostly for social and enhancement motives during social gatherings; c) students living in a "prototypical", less controlled situation and without family obligations are more likely to consume alcohol more frequently, in higher quantities or to engage in RSOD; d) students tend to overestimate the extent of their fellow students' alcohol consumption, a bias that is more pronounced among those with higher alcohol consumption

2. Alcohol consumption among university students in Ireland and the United Kingdom from 2002 to 2014

https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-016-2843-1

Almost two thirds of students reported a hazardous alcohol consumption score on the AUDIT scale. Over 20 % reported alcohol problems over their lifetime using CAGE while over 20 % exceed sensible limits each week. Noteworthy is the narrowing of the gender gap throughout the past decade.

3. The effect of physician advice on alcohol consumption: count regression with an endogenous treatment effect https://onlinelibrary.wiley.com/doi/full/10.1002/jae.596?casa_token=jXWSqWvxUMMAAAAA%3AJUkTNGwTeKpzMl5crjsmr8sChDw5FqdQg8TVigvCMRg_wTnSUh7SgrzvMRHdyitY_-cytBIVgHvhVXM

In this paper they use a microeconometric model and non-experimental data to estimate the effectiveness of physician advice on alcohol consumption.

The results support contention that the receipt of advice should be considered a potentially endogenous explanatory variable. When we correct for endogeneity, we find evidence that physician advice can lead to a reduction in alcohol consumption.

This evidence suggests that the efficacy of physician advice as demonstrated in clinical trials may translate into effectiveness in everyday practice.

IIIllustrative calculations suggest that policies to encourage physician advice about drinking to patients with hypertension are likely to yield substantially more benefits than costs.

Data

The data were obtained in a survey of student's math and Portuguese language courses in secondary school. It contains a lot of interesting social, gender and study information about students and their alcohol consumption and health condition. This dataset shows how different factors influence on the students.

We take the data from Kaggle source: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption

Attributes for student-mat.csv datasets:

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)

5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)

These grades are related with the course subject, math or Portuguese:

1. G1 - first period grade (numeric: from 0 to 20)
2. G2 - second period grade (numeric: from 0 to 20)
3. G3 - final grade (numeric: from 0 to 20, output target)

We are going to estimate workday alcohol consumption ("Dalc" according to the dataset). There is an order (level of consumption) 1 – very low, 2 – low, 3 – medium, 4 – high, 5 – very high. Our independent variables are sex, studytime, G3, famsize, goout, famrel, freetime, health, Mjob, traveltime, absences.

Diagrams show real dependences between variables, and we can analyze it.

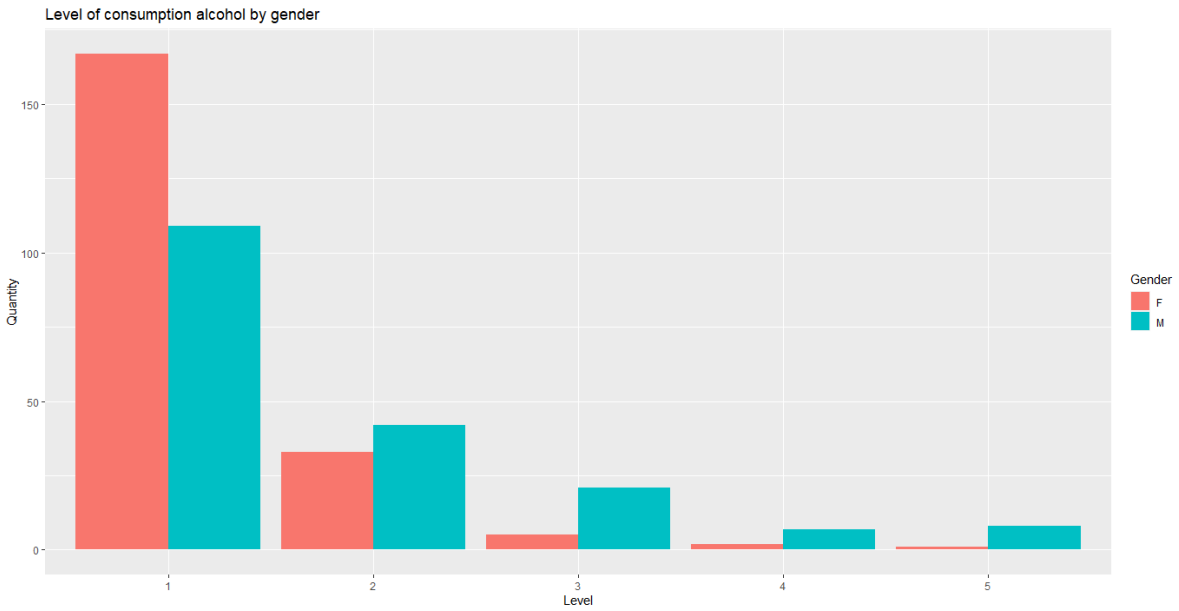

Figure 1 – Level of alcohol consumption by gender

From first figure it is obvious that male have higher alcohol consumption than female. Diagram also shows that female have low level of consuming.

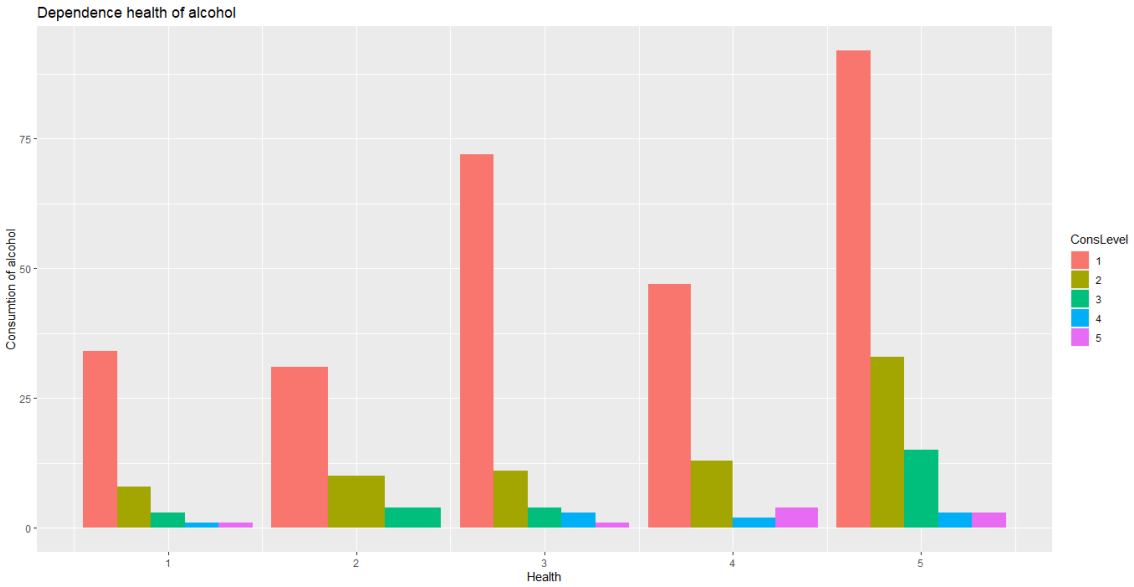

Figure 2 – Dependence the health condition of alcohol consumption

From second figure, we can see that there is a dependence between alcohol consumption and health condition. It shows that people who consume less alcohol have better health and oppositive who have a high level of consumption

Method/Model. General – to – Specific model:

Step 1

Estimate general model. Define all variables that are significant

```
reg1 = lm(as.numeric(Dalc)~ sex +studytime + G3+ famsize + goout + famrel
        + freetime + health + as.factor(Mjob)+
         traveltime + absences, data=student.mat)
summary(reg1)
```

```
Call:
lm(formula = as.numeric(Dalc) ~ sex + studytime + G3 + famsize +
    goout + famrel + freetime + health + as.factor(Mjob) + traveltime +
    absences, data = student.mat)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4799 -0.5044 -0.1756  0.2311  3.4200

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             0.716135   0.318189   2.251 0.024978 *
sexM                    0.353770   0.091664   3.859 0.000134 ***
studytime              -0.078750   0.052433  -1.502 0.133945
G3                     -0.003998   0.009453  -0.423 0.672574
famsizeLE3              0.120963   0.091466   1.322 0.186802
goout                   0.172964   0.039080   4.426 1.26e-05 ***
famrel                 -0.117544   0.046688  -2.518 0.012225 *
freetime                0.097529   0.044634   2.185 0.029492 *
health                  0.034011   0.030228   1.125 0.261236
as.factor(Mjob)health  -0.220529   0.181051  -1.218 0.223962
as.factor(Mjob)other    0.019512   0.128020   0.152 0.878939
as.factor(Mjob)services 0.037858   0.136760   0.277 0.782067
as.factor(Mjob)teacher  0.002907   0.156444   0.019 0.985184
traveltime              0.128458   0.060458   2.125 0.034253 *
absences                0.012189   0.005191   2.348 0.019387 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8107 on 380 degrees of freedom
Multiple R-squared:  0.2012,    Adjusted R-squared:  0.1717
F-statistic: 6.835 on 14 and 380 DF,  p-value: 1.611e-12
```

Test whether all insignificant variables all jointly insignificant

```
reg1a = lm(as.numeric(Dalc)~ sex +goout + famrel + freetime +
           traveltime + absences, data=student.mat)

anova(reg1, reg1a)
```

```
Analysis of Variance Table

Model 1: as.numeric(Dalc) ~ sex + studytime + G3 + famsize + goout + famrel +
    freetime + health + as.factor(Mjob) + traveltime + absences
Model 2: as.numeric(Dalc) ~ sex + goout + famrel + freetime + traveltime +
    absences
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    380 249.72
2    388 255.32 -8   -5.6002 1.0652 0.3868
```

We cannot reject the null hypothesis H_0. It means that all insignificant variables are jointly insignificant, and we can remove all insignificant variable in one step.

Results

Estimate ordered probit and ordered logit, selection of the covariates.

```
Ordered Logit Regression
Log-Likelihood: -318.3936
No. Iterations: 5
McFadden's R2: 0.1213263
AIC: 672.7872
                          Estimate Std. error t value  Pr(>|t|)
sexM                      0.909744   0.264746  3.4363 0.0005897 ***
studytime                -0.293391   0.158498 -1.8511 0.0641589 .
G3                       -0.026226   0.026785 -0.9791 0.3275070
famsizeLE3                0.471010   0.249098  1.8909 0.0586428 .
goout                     0.463446   0.113629  4.0786 4.531e-05 ***
famrel                   -0.388684   0.132877 -2.9251 0.0034430 **
freetime                  0.284879   0.129596  2.1982 0.0279345 *
health                    0.115653   0.088142  1.3121 0.1894794
as.factor(Mjob)health    -0.882685   0.576726 -1.5305 0.1258904
as.factor(Mjob)other     -0.081476   0.368278 -0.2212 0.8249090
as.factor(Mjob)services  -0.032767   0.392989 -0.0834 0.9335497
as.factor(Mjob)teacher    0.073765   0.430280  0.1714 0.8638816
traveltime                0.198947   0.168452  1.1810 0.2375899
absences                  0.030942   0.013561  2.2817 0.0225050 *
----- Threshold Parameters -----
                  Estimate Std. error t value  Pr(>|t|)
Threshold (1->2)  2.32970    0.89899  2.5915  0.009557 **
Threshold (2->3)  3.81246    0.91569  4.1635 3.135e-05 ***
Threshold (3->4)  4.87522    0.94228  5.1739 2.293e-07 ***
Threshold (4->5)  5.63777    0.97707  5.7701 7.924e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Firstly, we can interpret only value that has only significant level, it means where p-value > 5%. According to the ordered logit model we should estimate value for lowest and highest point. It means for daily consumption Dalc =1 and Dalc = 5.

For the lowest point (Dalc = 1):

Male will have less probability to have low alcohol consuming than female

People who are going out will have less probability of low alcohol consumption

Variable(farmel) quality of relationship will be better for low consuming

Free time will have negative effect of low consumption.

Number of absences during the education process will have less effect on the low consumption.

For the highest point (Dalc = 5):

Male will have bigger probability to have high alcohol consuming than female

People who are going out will have higher probability of alcohol consumption

Variable(farmel) quality of relationship will be worst for high consuming

Free time will have positive effect of high consumption.

Number of absences during the education process will have effect on the high consumption.

Calculation and interpretation of marginal effects for the final model (from the general-to specific approach)

```
Marginal Effects on Pr(Outcome==1)
             Marg. Eff Std. error t value  Pr(>|t|)
sexM        -0.2203748  0.0470558 -4.6833 2.823e-06 ***
goout       -0.0892430  0.0214200 -4.1663 3.095e-05 ***
famrel       0.0740040  0.0248185  2.9818  0.002866 **
freetime    -0.0578671  0.0249046 -2.3235  0.020150 *
traveltime  -0.0550909  0.0319172 -1.7261  0.084337 .
absences    -0.0065559  0.0026196 -2.5027  0.012326 *
-------------------------------------
Marginal Effects on Pr(Outcome==2)
             Marg. Eff Std. error t value  Pr(>|t|)
sexM         0.1325704  0.0303924  4.3620 1.289e-05 ***
goout        0.0553774  0.0145880  3.7961  0.000147 ***
famrel      -0.0459212  0.0161551 -2.8425  0.004476 **
freetime     0.0359079  0.0158742  2.2620  0.023695 *
traveltime   0.0341852  0.0201140  1.6996  0.089211 .
absences     0.0040681  0.0016777  2.4248  0.015318 *
-------------------------------------
Marginal Effects on Pr(Outcome==3)
              Marg. Eff  Std. error t value  Pr(>|t|)
sexM         0.05313272  0.01476618  3.5983 0.0003203 ***
goout        0.02072766  0.00599551  3.4572 0.0005458 ***
famrel      -0.01718823  0.00638378 -2.6925 0.0070922 **
freetime     0.01344027  0.00620312  2.1667 0.0302581 *
traveltime   0.01279546  0.00774468  1.6522 0.0985015 .
absences     0.00152268  0.00065945  2.3090 0.0209437 *
-------------------------------------

Marginal Effects on Pr(Outcome==4)
              Marg. Eff  Std. error t value Pr(>|t|)
sexM         0.01785116  0.00692289  2.5786 0.009921 **
goout        0.00680116  0.00266598  2.5511 0.010739 *
famrel      -0.00563980  0.00258907 -2.1783 0.029383 *
freetime     0.00441002  0.00236049  1.8683 0.061725 .
traveltime   0.00419845  0.00275686  1.5229 0.127782
absences     0.00049962  0.00025434  1.9644 0.049486 *
-------------------------------------
Marginal Effects on Pr(Outcome==5)
              Marg. Eff  Std. error t value Pr(>|t|)
sexM         0.01682043  0.00656337  2.5628  0.01038 *
goout        0.00633680  0.00247908  2.5561  0.01058 *
famrel      -0.00525474  0.00241808 -2.1731  0.02977 *
freetime     0.00410892  0.00218646  1.8793  0.06021 .
traveltime   0.00391179  0.00255723  1.5297  0.12609
absences     0.00046551  0.00023612  1.9715  0.04867 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We will estimate marginal effect for high alcohol consumption (Outcome =5). We should omit variable with p-value > 5 %

Male will have bigger probability to consume high amount of alcohol than female by 1.6 percentage point

People who go out more often will have high level consumption by 0.6 percentage point

Students that have good relationship in the family will have less probability to have high level consumption by 0.5 percentage point

Number of absences during the educational process will increase probability of high consumption by 0.04 percentage point.

Estimation of marginal effect for medium alcohol consumption (Outcome =3). We should omit variable with p-value > 5 %.

Male will have bigger probability to consume medium amount of alcohol than female by 5.3 percentage point

People who go out more often will have medium level consumption by 2.03 percentage point

Students that have good relationship in the family will have less probability to have medium level consumption by 1.7 percentage point

Number of absences during the educational process will increase probability of medium consumption by 0.15 percentage point.

Estimation of marginal effect for low alcohol consumption (Outcome =1). We should omit variable with p-value > 5 %.

Male will have less probability to consume low amount of alcohol than female by 22.03 percentage point

People who do not go out will have low level consumption by 8.9 percentage point

Students that have good relationship in the family will have bigger probability to have low level consumption by 7.4 percentage point.

Free time will have negative effect on the low consumption by 5.7 percentage point

Number of absences during the educational process will decrease probability of low consumption by 0.65 percentage point.

Perform the linktest and interpret the result

```
Likelihood ratio test

Model 1: as.factor(Dalc) ~ sex + goout + famrel + freetime + traveltime +
    absences
Model 2: as.factor(Dalc) ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  10 -325.71
2   4 -362.36 -6 73.284  8.652e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We use this test to estimate significant level, if p-value < 5%, it means we should reject null hypothesis, they are jointly significant.

Perform the Hosmer-Lemeshow test, the Lipsitz, and the Pulkstenis-Robinson tests

```
        Hosmer and Lemeshow test (ordinal model)

data:  student.mat$Dalc, fitted(ologit.unrestricted)
X-squared = 30.707, df = 35, p-value = 0.6755
```

In this case we cannot reject the null hypothesis, because our p-value = 67 %. It means that our model is fine and has an appropriate form.

```
        Pulkstenis-Robinson chi-squared test

data:  formula:  as.factor(Dalc) ~ sex + goout + famrel + freetime + traveltime +
 formula:      absences
X-squared = 9.6372, df = 10, p-value = 0.4729


        Lipsitz goodness of fit test for ordinal response models

data:  formula:  Dalc ~ sex + goout + famrel + freetime + traveltime + absences
LR statistic = 7.2727, df = 9, p-value = 0.6088
```

It works for all test, that we applied. It shows that our model is fine and does not have any problem.

Check the proportional odds assumption

```
Call:
polr(formula = as.factor(Dalc) ~ sex + goout + famrel + freetime +
    traveltime + absences, data = student.mat)

Coefficients:
              Value Std. Error t value
sexM        1.10909    0.24300   4.564
goout       0.45027    0.10902   4.130
famrel     -0.37337    0.12615  -2.960
freetime    0.29196    0.12612   2.315
traveltime  0.27795    0.16058   1.731
absences    0.03308    0.01324   2.498

Intercepts:
    Value   Std. Error t value
1|2  2.9706  0.7004     4.2416
2|3  4.4047  0.7248     6.0768
3|4  5.4614  0.7576     7.2089
4|5  6.2229  0.7992     7.7862

Residual Deviance: 651.4297
AIC: 671.4297
```

It gives an information about standard error and t value, AIC and residual Deviance.

Findings

We also can solve this problem using models with Binary Dependent Variables. In this case we need to change our order and replaces the numbers than we have only two choices for alcohol consuming (Yes (1) or No (0)). Process will be the same, but of course using with appropriate function. In this solution, we can estimate only if students consume alcohol or not. In order model it gives a better picture of level of consumption, but this is another way to solve the problem. Of course, this is valid only to this dataset.

Literature

1. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008
2. William H. Greene, David A. Hensher, Modeling Ordered Choices, Department of Economics, Stern School of Business, New York University, New York, 2009
3. Vani K. Borooah, Logit and Probit: Ordered and Multinomial Models, November 2001
4. Fullerton, A.S., & Xu, J., Ordered Regression Models Parallel, Partial, and Non-Parallel Alternatives. New York: Chapman and Hall/CRC., 2010
5. Greene, W. H., & Hensher, D. A., Modeling ordered choices: A primer. Cambridge: Cambridge University Press, 2010