Analysis of Ordinal Data

Yossi Levy

14.4.2019

Outline

- Data types and ordinal variables
- Univariate analysis
- Measures of association
- Ordinal response
- Ordinal explanatory variable

Four Levels of measurements

Qualitative / Categorical variables

- Nominal scale
- Ordinal scale

Quantitative variables

- Interval scale
- Ratio scale

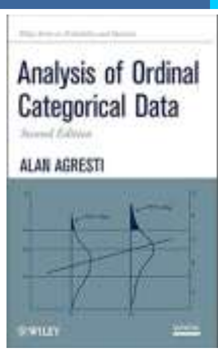
Ordinal variables

- An **ordinal variable** is a categorical variable in which the categories have natural order, but **the distances between the categories are not equal**.
- The ordinal scale is distinguished from the nominal scale by having a ranking.
- The ordinal scale differs from interval and ratio scales by not having category widths that represent equal increments of the underlying attribute

Examples of ordinal variables

- Cancer stage
- BMI class: Underweight/Normal/Overweight/Obese
- Socio-Economical Status
- Chili pepper heat level
- Class on the Titanic

A good start



Univariate Analysis

- Contingency table
- Location: median, quartiles, percentiles etc.
- Dispersion: IQR , range, etc.
- Goodness of fit: Chi-square test, KS test, etc.

Measures of association

- Spearman correlation coefficient
- Kendall's Tau
- Goodman and Kruskal's gamma
- M^2 coefficient
- Etc.

Response variable in regression model

Approaches for analysis

1. Ignore
2. Do something

Response variable - analysis by ranks

- Kruskal-Wallis test
- Friedman test
- Wilcoxon's signed ranks test
- Jonckheere test
- Mann-Whitney test
- Cochran–Armitage test for trend
- Etc.

All of these procedures are special cases of a regression model

Example – Titanic data

```
library(titanic)
library(plyr)
titanic=titanic_train
names(titanic)=tolower(names(titanic))
titanic=titanic[, c(1, 3, 7, 8)]
titanic$class_c=as.character(titanic$pclass)
titanic$class_c=revalue(titanic$class_c,
                        c("1"="First", "2"="Second", "3"="Third"))
titanic$class_o=ordered(titanic$class_c)
titanic$class_o=factor(titanic$class_o, levels(titanic$class_o)[3:1])
head(titanic)


  passengerid pclass sibsp parch class_c class_o
1           1      3     1     0   Third   Third
2           2      1     1     0   First   First
3           3      3     0     0   Third   Third
4           4      1     1     0   First   First
5           5      3     0     0   Third   Third
6           6      3     0     0   Third   Third
>
```


Model 1: ignore everything

```
> # first model - pclass as numeric
model1=glm(pclass ~ sibsp + parch, data=titanic)


             Estimate Std. Error t value  Pr(>|t|)
(Intercept)    2.280      0.032  71.094 <0.0001 *
sibsp          0.069      0.028   2.479  0.0133 *
parch         -0.020      0.038  -0.527  0.5986
>
```

Model 2: class as nominal

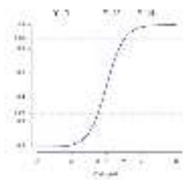
```
> # second model - multinomial logistic regression

> library(nnet)
> model2=multinom(class_c ~ sibsp + parch, data=titanic)


   class   covariate estimate    SE      t   p-value
1 Second (Intercept)    -0.164 0.115 -1.427  0.1537
2              sibsp    -0.036 0.124 -0.290  0.7715
3              parch     0.050 0.132  0.379  0.7049
4  Third (Intercept)     0.742 0.094  7.895 <0.0001 *
5              sibsp     0.192 0.093  2.059  0.0395 *
6              parch    -0.047 0.112 -0.418  0.6758
>
```


Ordinal response modelling

- Assume that the values of Y are determined by a latent random variable Y*
- Y=y if Y* is between some 2 values
- Y* is part of the model, not part of the data!
- Y* is a related to a function of some covariates



$P(Y \leq 1) = 0.27$
$P(Y \leq 2) = 0.88$

Formal definition

- Y is an ordinal variables that takes values 1,…C
- X_1,… X_p are explanatory variables
- β_1,… β_p are real valued parameters
- α_1,… α_{C-1} are real valued parameters such that $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_{C-1}$
- g is a link function
- Then an ordinal regression model is

$$P(Y \leq j\,|X) = g^{-1}(\alpha_j + \beta_1 X_1 + \cdots + \beta_p X_p)$$

- Note that positive β indicates negative effect!

Cumulative logit models

$$g(s) = \log(\frac{s}{1-s})$$

$$P(Y \leq j|x) = \frac{e^{\alpha_j + x\beta}}{1 + e^{\alpha_j + x\beta}}$$

$$\log OR = \log\frac{P(Y \leq j|x_1)/P(Y > j|x_1)}{P(Y \leq j|x_2)/P(Y > j|x_2)} = \beta_j(x_1 - x_2)$$

Model 3: class as ordinal

```
> # third model - ordinal regression
> library(MASS)
> model3=polr(class_o ~ sibsp + parch, data=titanic, Hess=TRUE,
method="logistic")


              estimate    SE      t   p-value
sibsp           -0.196 0.074 -2.654  0.0080 *
parch            0.057 0.089  0.640  0.5219
Third|Second     0.129 0.077  1.673  0.0943
Second|First     1.069 0.086 12.387 <0.0001 *
>
```


Explanatory ordinal variables

Example: Alcohol consumption during pregnancy and infant malformation

- alcohol_level: daily alcohol consumption during first three month of pregnancy, in 1-5 scale (1 indicates lowest level)
- malformation: presence or absence of congenital sex organ malformations

```
> head(gk87)

  alcohol_level malformation malformation01
1             1      Present              1
2             5       Absent              0
3             2       Absent              0
4             3      Present              1
5             1       Absent              0
6             1       Absent              0
>
```

Goal: predict malformation by daily alcohol consumption level

Source: B. I. Graubard and E. L. Korn, Biometrics, 43: 471–476, 1987.
Reprinted with permission in Agresti (2007), An Introduction to Categorical Data Analysis, page 42

Model 1: ignore

```
> # model1 1: treat alcohol level as a numeric variable
> model1=lm(malformation01~alcohol_level, data=gk87)

              Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.002      0.001   2.198 0.0279 *
alcohol_level   0.001      0.001   1.352 0.1764
>
```

19

19

Model 2: nominal variable

```
> head(gk87)
  alcohol_level malformation malformation01 alcohol_class
1             1      Present              1          None
2             5       Absent              0         Drunk
3             2       Absent              0           Low
4             3      Present              1    Reasonable
5             1       Absent              0          None
6             1       Absent              0          None

> # model 2: treat alcohol level as a categorical variable
> model2=glm(malformation01~alcohol_class, data=gk87,
+            family = binomial(link = 'logit'))

                        Estimate Std. Error z value  Pr(>|z|)
(Intercept)               -5.874      0.145 -40.640 <0.0001 *
alcohol_classLow          -0.068      0.217  -0.314  0.7538
alcohol_classReasonable    0.814      0.471   1.726  0.0843
alcohol_classHigh          1.037      1.014   1.023  0.3064
alcohol_classDrunk         2.263      1.024   2.210  0.0271 *
>
```

20

20

Model 3: ordinal variable

```
> gk87$alcohol_ordered=ordered(gk87$alcohol_class)
> head(gk87)
  alcohol_level malformation malformation01 alcohol_class alcohol_ordered
1             1      Present              1          None            None
2             5       Absent              0         Drunk           Drunk
3             2       Absent              0           Low             Low
4             3      Present              1    Reasonable      Reasonable
5             1       Absent              0          None            None
6             1       Absent              0          None            None

> head(gk87$alcohol_ordered)
[1] None        Drunk       Low         Reasonable None        None
Levels: None < Low < Reasonable < High < Drunk
```

21

21

Model 3: ordinal variable

```
> # model 3: treat alcohol level as an ordinal variable
> model3=glm(malformation01~alcohol_ordered, data=gk87)

                  Estimate Std. Error t value  Pr(>|t|)
(Intercept)          0.009      0.002   4.563 <0.0001 *
alcohol_ordered.L    0.017      0.006   2.909  0.0036 *
alcohol_ordered.Q    0.009      0.005   1.913  0.0557
alcohol_ordered.C    0.004      0.004   1.011  0.3121
alcohol_ordered^4    0.003      0.003   1.050  0.2938
>
```

22

22

What to do?

Strategy: replace categories by numerical scores that will portray the distances between the levels

Possible scoring systems

- Linear – practically ignore – model 1
- Orthogonal contrasts – model 3 – assuming linearity
- Mid ranks
- Mid range
- RC scores (Goodman, 1985)
- Canonical scores (Gilula and Haberman, 1986)
- And more…

23

23

Mid ranks applications

Consumption	N	Ranks	Mid rank	Score
None	17114	1 - 17114	(1+17114)/2	8557.5
Low	14502	11715 - 31616	(11715+31616)/2	24365.5
Reasonable	793	31616 - 32409	(31616+32409)/2	32013
High	127	32410 - 32536	(32410+32536)/2	32473
Drunk	38	32537 - 32574	(32537+32574)/2	32555.5
Total	32574			

24

24

Mid-range analysis

```
> head(gk87[, c(1,2,4,7,8)])
  alcohol_level malformation alcohol_class range midrange
1             1      Present          None     0      0.0
2             5       Absent         Drunk   >=6      7.0
3             2       Absent           Low    <1      0.5
4             3      Present    Reasonable   1-2      1.5
5             1       Absent          None     0      0.0
6             1       Absent          None     0      0.0

> # model 5: use midrange
> model5=glm(malformation01~midrange, data=gk87)

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)    0.002      0.000   6.837 <0.0001 *
midrange       0.002      0.001   2.563  0.0104 *
```

26

26

Mid-rank analysis

```
> # model 4: use midranks
> gk87$midranks=rank(gk87$alcohol_level, ties.method="average")
> head(gk87[, c(1,2,6)])
  alcohol_level malformation midranks
1             1      Present   8557.5
2             5       Absent  32555.5
3             2       Absent  24365.5
4             3      Present  32013.0
5             1       Absent   8557.5
6             1       Absent   8557.5

> model4=lm(malformation01~midranks, data=gk87)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.003      0.001   3.828 1e-040 *
midranks       0.000      0.000   0.593 0.5533
>
```

25

25

27

27