2017 Predictive Analytics Symposium

Session 18, Ordinal Logistic Modeling: An Application

Moderator:
Benjamin David Kester, FSA

Presenter:
Marjorie A. Rosenberg, FSA

Ordinal Logistic Regression Models

Margie Rosenberg

University of Wisconsin – Madison

We acknowledge the Society of Actuaries CAE Research Grant for their partial support in this work.

Purpose

To motivate and explain logistic regression when the outcome variable is an ordered categorical variable.

Outline

1. Review of Logistic Regression
2. Ordinal Logistic Modeling
3. NAAJ paper
4. Conclusion

Review of Logistic Regression

Logistic Regression Model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}$$

Where:

$\pi_i = \Pr(Y_i = 1|\mathbf{x}_i)$

$\mathbf{x}_i$ = vector of covariates

$\boldsymbol{\beta}$ = vector of unknown parameters

$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathit{logit}(\pi_i)$

Some Review Questions

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}$$

Why is this called *logistic*?

Where is the error term?

What other link functions are possible in this case?

Logistic Model Link Function

$$g\left(E(Y_i)\right) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

(the canonical link)

$$E(Y_i) = \frac{1}{1+e^{-\mathbf{x}_i'\boldsymbol{\beta}}} \quad \text{(Logistic cdf)}$$

$$= \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}$$

$$= \Pr(Y_i = 1|\mathbf{x}_i) = \pi_i$$

Note: Not really answer question about *Why called logistic regression*

What Do We Know of Logistic Regression?

- Outcome variable has 2 levels: success/failure, disease/no disease
- Member of GLM family
- Write density in form of *exponential family*
- Logit link is *canonical link* that results from exponential family

Example: Predicting Health Status

You are the actuary and want to find a model using Age as a predictor to predict the probability that a person's perceived health status is Very Good or Excellent (VG/E) as contrasted to Poor/Fair/Good (P/F/G)

Dependent variable:

$$y_i = \begin{cases} 1 & i\text{th person is VG/E} \\ 0 & \text{otherwise} \end{cases}$$

Or could define dependent variables as :

$$y_i = \begin{cases} 1 & i\text{th person is P/F/G} \\ 0 & \text{otherwise} \end{cases}$$

Simple Example of Data (H156 MEPS 2011)

- One year of MEPS
- Ages 30 to 59
- Complete Cases
- 6,919 observations
- Results not adjusted for complex survey design

Observed Summary* of Data (H156 MEPS 2011)

P/F/G	VG/E
0.45	0.55

	Counts		% Row Total	
Age Cat	P/F/G	VG/E	P/F/G	VG/E
30s	957	1396	0.41	0.59
40s	1005	1264	0.44	0.56
50s	1137	1160	0.49	0.51

*Not adjusted for complex survey design

Example

Suppose want to predict *Perceived Health Status* of Very Good/Excellent vs. Poor/Fair/Good

$$\pi_i = 1 \quad \text{if person is VG/E}$$

With covariate whether person is in the 30s, 40s, or 50s (only these age groups)

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Age40s + \beta_2 Age50s$$

Three questions:

1. Where is Age30s covariate?
2. How interpret e_0^{β}?
3. How interpret e_1^{β}?

Hint: Recall $e^{log(x)} = x$

How interpret e^{β_0}?

Given that Age30s is the reference category, if person in their 30s, then:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0$$

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0}$$

Or, the odds of someone in their 30s reporting VG/E vs someone in the 30s reporting P/F/G

How interpret e^{β_1}?

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Age40s + \beta_2 Age50s$$

$$\begin{aligned} \log\left(\frac{\pi_{30}}{1-\pi_{30}}\right) &= \beta_0 \\ \log\left(\frac{\pi_{40}}{1-\pi_{40}}\right) &= \beta_0 + \beta_1 \end{aligned}$$

$$\log\left(\frac{\pi_{40}}{1-\pi_{40}}\right) - \log\left(\frac{\pi_{30}}{1-\pi_{30}}\right) = \beta_1$$

How interpret e^{β_1}? (Cont.)

$$\log\left(\frac{\pi_{40}}{1-\pi_{40}}\right) - \log\left(\frac{\pi_{30}}{1-\pi_{30}}\right) = \beta_1$$

$$\log\left(\frac{\frac{\pi_{40}}{1-\pi_{40}}}{\frac{\pi_{30}}{1-\pi_{30}}}\right) = \beta_1$$

$$\frac{\frac{\pi_{40}}{1-\pi_{40}}}{\frac{\pi_{30}}{1-\pi_{30}}} = e^{\beta_1}$$

Or, the odds *ratio* of someone in their 40s relative to someone in their 30s reporting VG/E vs someone reporting P/F/G

Logistic Regression Parameter Interpretation for Categorical Variable

Odds ratio = 1: Outcome of *success* equally likely to occur in both groups

Odds ratio $>$ 1: Outcome of *success* more likely for group referenced in numerator

Odds ratio $<$ 1: Outcome of *success* less likelyfor group referenced in numerator

Note: Relative risk = $\frac{\Pr(Y_i=1|x_{40}=1)}{\Pr(Y_i=1|x_{30}=0)}$

Two Examples of Impact of Changing Response Variable

1. Dependent variable of VG/E
2. Dependent variable of P/F/G

Logistic Results: Using VG/E

```
glm(formula = OHa ~ AgeCat, family =
   binomial(link = "logit"), data = dat1)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.37756    0.04197   8.997  < 2e-16
AgeCat40s   -0.14827    0.05956  -2.489   0.0128
AgeCat50s   -0.35754    0.05918  -6.041 1.53e-09

Null deviance: 9516.5  on 6918  degr of freedom
Residual deviance: 9479.5  on 6916  degr of freedom
AIC: 9485.5

Number of Fisher Scoring iterations: 4
```

Logistic Results: Using P/F/G

```
glm(formula = OHb ~ AgeCat, family =
  binomial(link = "logit"), data = dat1)

            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.37756    0.04197  -8.997  < 2e-16
AgeCat40s    0.14827    0.05956   2.489   0.0128
AgeCat50s    0.35754    0.05918   6.041 1.53e-09

Null deviance: 9516.5  on 6918  degr of freedom
Residual deviance: 9479.5  on 6916  degr of freedom
AIC: 9485.5

Number of Fisher Scoring iterations: 4
```

Latent Variable Representation

Define Y_i^* as unobserved continuous variable of Y_i

Where $Y_i^* = x_i'\beta + \epsilon_i$

Random error ϵ_i here assumed to have a standard logistic distribution (mean = 0)

$Y_i = 1$, if $Y_i^* > 0$

$Pr\left[Y_i = 1 \mid x_i\right] = Pr\left[Y_i^* > 0 \mid x_i\right]$

https://en.wikipedia.org/wiki/Logistic_regression

Ordinal Logistic Modeling

Introduction

- Instead of 2 outcome levels, there exist multiple outcome levels
- Include order of outcome
- Examples
  - Education
  - Perceived Health Status
  - Type of health care utilizer: Low, One-Time, Persistent
- Different link functions exist
- Different model forms exist

References

The ordinal logistic model was originally studied by Snell (1964) and Walker and Duncan (1967), extended by McCullagh (1980), and later by Anderson (1984).

Good references: Agresti (2010), Ananth and Kleinbaum (1997), Peterson and Harrell Jr (1990)

NHIS/MEPS Data

- Example from Kim & Rosenberg *The role of unhealthy behavior on perceived health status* accepted to NAAJ
- National Health Interview Survey (NHIS) linked to Medical Expenditure Panel Survey (MEPS)
- NHIS Sample Adult Questionnaire for adult health behavior data
- 3-year longitudinal data of adults aged 30 to 59 inclusive
- Total 12,160 adults representing 124,000,000 U.S. civilian non-institutionalized population from 2008 to 2012
- Results adjusted for complex survey design

Definitions

- Y_i represent perceived health status of individual i at end of first year of MEPS (dependent variable)
  - Five categories of $Y_i =$ Poor, Fair, Good, Very Good, and Excellent ($j = 1, 2, \ldots, 5$)
  - Poor $\leq$ Fair $\leq$ Good $\leq$ Very Good $\leq$ Excellent
- $X_i =$ vector of individual-level covariates from NHIS (unhealthy behaviors) and MEPS (other covariates)
- α_j be unknown intercept terms that separate the response categories
- β a vector of unknown regression parameters

Proportional Odds Model*

$\pi_i = Pr\,(Y_i \leq j | x_i, \alpha_j, \beta) =$ Cumulative probability of Y_i being equal to or less than category j, given the unknown parameters and the individual-level covariates

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_j - x_i'\beta \quad j = 1, \ldots, 4$$

Note:

1. $\alpha_j =$ Cutpoints $(-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_j = \infty)$
2. β constant
3. Relationship to latent framework

*Note: Know your software to verify which representation

Why Called Proportional Odds Model?

Suppose two different people i and k had same values of Y, but different x

$$\begin{aligned} log\left(\frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_k}{1-\pi_k}}\right) &= log\left(\frac{\pi_i}{1-\pi_i}\right) - log\left(\frac{\pi_k}{1-\pi_k}\right) \\ &= \alpha_j - x_i'\beta - (\alpha_j - x_k'\beta) \end{aligned}$$

Odds ratio not depend on j:

$$\frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_k}{1-\pi_k}} = e^{-\left(x_i' - x_k'\right)\beta}$$

Odds Ratio

Suppose two different people *i* and *k* had same values of Y, but one is in their 30s and other in their 40s respectively

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_j - \beta_1 Age40s - \beta_2 Age50s$$

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_j$$

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_j - \beta_1 Age40s$$

$$\frac{\frac{\pi_{40}}{1-\pi_{40}}}{\frac{\pi_{30}}{1-\pi_{30}}} = e^{-\beta_1}$$

- As with logistic regression, interpret regression parameters β using an odds ratio
- But with defined structure, e^{β} reflects ratio of survival probability to cumulative probability of one category relative to the reference category (See next slide)

Odds Ratio (Cont.)

Look at:

$$\begin{aligned} log\left(\frac{\pi_i}{1-\pi_i}\right) &= log(\pi_i) - log(1-\pi_i) \\ &= -\left(log(1-\pi_i) - log(\pi_i)\right) \end{aligned}$$

$$log(1-\pi_i) - log(\pi_i) = -\alpha_j + \beta_1 Age40s + \beta_2 Age50s$$

$$\frac{\frac{1-\pi_{40}}{\pi_{40}}}{\frac{1-\pi_{30}}{\pi_{30}}} = e^{\beta_1}$$

- Here e^{β}_{40} calculates odds ratio of being in a higher category for a person in the forties relative to a person in their thirties.
- In our model, interpretation of odds ratio for $\beta > 0$ is that people report that they are in better perceived health as compared to those in the reference category and in worse perceived health when $\beta < 0$

Calculate Individual Probabilities

$$
\begin{aligned}
Pr\left(Y_i = 1\right) &= \exp\left(-\left(\alpha_1 - X_i'\beta\right)\right)^{-1} \\
&\quad \text{for } j = 1 \\
Pr\left(Y_i = j\right) &= \exp\left(-\left(\alpha_j - X_i'\beta\right)\right)^{-1} - \exp\left(-\left(\alpha_{j-1} - X_i'\beta\right)\right)^{-1} \\
&\quad \text{for } j = 2, 3, 4 \\
Pr\left(Y_i = 5\right) &= 1 - Pr\left(Y_i \leq 4\right) \quad \text{for } j = 5
\end{aligned}
$$

Latent Variable Framework

Define Y_i^* as unobserved continuous variable of Y_i

Where $Y_i^* = X_i' \beta + \epsilon_i$

Random error ϵ_i here assumed to have a logistic distribution

$Y_i = j$, if $\alpha_{j-1} < Y_i^* \leq \alpha_j$

Thus Y_i is assigned level j, when Y_i^* is within this interval

$Pr[Y_i \leq j \mid X_i] = Pr\left[Y_i^* \leq \alpha_j \mid X_i\right]$

Agresti (2010)

Interpretation of Output

1. Order of dependent variable (E to P or P to E)
2. Function Used (e.g. in R)
   - polr (in MASS) uses $\alpha_j - X_i'\beta$
   - clm (in ordinal) uses $\alpha_j - X_i'\beta$
   - vglm (in VGAM) uses $\alpha_j + X_i'\beta$

Output Differences Depending on Order of Outcome Variable

H156 Output using polr function (P/F/G/VG/E)

```
polr(formula = OH1 ~ AgeCat, data = dat1, Hess = TRUE)

Coefficients:
            Value Std. Error t value
AgeCat40s -0.2028    0.05260  -3.855
AgeCat50s -0.4562    0.05302  -8.605

Intercepts:
                    Value    Std. Error t value
Poor|Fair            -3.3430   0.0676   -49.4488
Fair|Good            -1.7940   0.0447   -40.1773
Good|Very Good       -0.4254   0.0387   -10.9944
Very Good|Excellent   0.9196   0.0403    22.8207

Residual Deviance: 20151.38
AIC: 20163.38
```

H156 Output using polr function (E/VG/G/F/P)

```
polr(formula = OH2 ~ AgeCat, data = dat1, Hess = TRUE)

Coefficients:
           Value Std. Error t value
AgeCat40s 0.2028    0.05260   3.855
AgeCat50s 0.4563    0.05302   8.606

Intercepts:
                        Value    Std. Error t value
Excellent|Very Good  -0.9196   0.0403    -22.8212
Very Good|Good        0.4254   0.0387     10.9943
Good|Fair             1.7940   0.0447     40.1777
Fair|Poor             3.3431   0.0676     49.4496

Residual Deviance: 20151.38
AIC: 20163.38
```

Output Using Different R functions

H156 Output using clm function (P/F/G/VG/E)

```
formula: OH1 ~ AgeCat
data:    dat1

 link  threshold nobs logLik    AIC      niter max.grad cond.H
 logit flexible  6919 -10075.69 20163.38 5(0)  4.59e-09 3.6e+01

          Estimate Std. Error z value Pr(>|z|)
AgeCat40s -0.20276    0.05260  -3.855 0.000116
AgeCat50s -0.45625    0.05302  -8.606  < 2e-16

Threshold coefficients:
                    Estimate Std. Error z value
Poor|Fair           -3.34309    0.06761  -49.45
Fair|Good           -1.79399    0.04465  -40.18
Good|Very Good      -0.42536    0.03869  -10.99
Very Good|Excellent  0.91965    0.04030   22.82
```

H156 Output using vglm function (P/F/G/VG/E)

```
vglm(formula = OH1 ~ AgeCat, family = propodds, data = dat1)

               Estimate Std. Error z value Pr(>|z|)
(Intercept):1  3.34309    0.06767  49.400  < 2e-16
(Intercept):2  1.79399    0.04495  39.908  < 2e-16
(Intercept):3  0.42536    0.03894  10.925  < 2e-16
(Intercept):4 -0.91965    0.04035 -22.789  < 2e-16
AgeCat40s     -0.20276    0.05290  -3.833 0.000127
AgeCat50s     -0.45625    0.05286  -8.632  < 2e-16

Residual deviance: 20151.38 on 27670 degrees of freedom

Log-likelihood: -10075.69 on 27670 degrees of freedom

Number of iterations: 3

Exponentiated coefficients:
AgeCat40s AgeCat50s
0.8164735 0.6336551
```

Outcome Variable and Covariates of NAAJ Paper

- **Purpose:** Explore the role of unhealthy behaviors in influencing the perceived health status of an individual
- **Perceived health status:** *In general, compared to other people of your age, would you say your health is Excellent/ Very good/ Good/ Fair/ Poor ?*
- **Unhealthy Behaviors:** Inadequate sleeping, inadequate physical activity, smoking, current heavy drinker

Additional Covariates

- **Predisposing:** Age, gender, race-ethnicity, marital status, education, employment
- **Enabling:** Income level, insurance coverage, region, MSA, usual source of care, transportation
- **Needs:** Diagnosed medical conditions, functional limitations

Summary of Unhealthy Behaviors

# Unhealthy Behaviors	%Pop	Perceived Health Status (%)				
		P	F	G	VG	E
0	28.3	1.0	6.2	24.9	36.6	31.3
1	41.4	2.5	9.2	29.8	33.4	25.1
2	23.5	4.5	14.2	32.2	31.4	17.7
3	6.4	13.0	16.7	34.1	23.3	12.9
4	0.4	4.5	18.3	30.6	28.6	18.0

Odds Ratio

Relative to Reference category: 0

# Unhealthy Behaviors	Odds Ratio	Std. Error	p-value
1	0.83	0.045	0.001
2	0.67	0.044	< 0.001
3	0.47	0.040	< 0.001
4	0.62	0.264	0.263

Prediction of Perceived Health Status

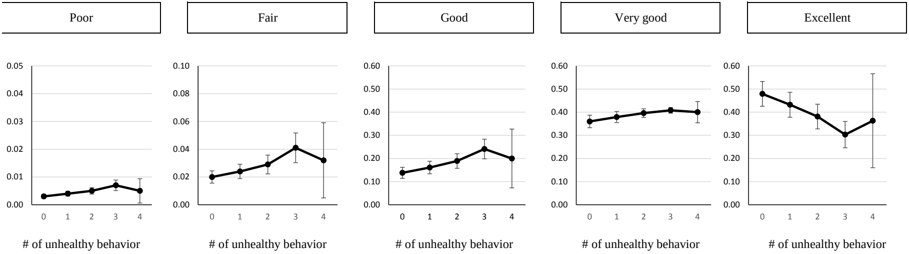
- Two profiles with differing degree of health
- All calculations are based on survey weights and standard errors are based on Taylor-linearized methods
- 95% confidence intervals for the probability estimates
- y-axes differ to account for smaller probabilities of outcomes

Profile A

Note: Categories chosen based on modal valued category except for income quantile (middle quantile)

- White female in 40's
- Employed with total income at the middle quantile of the population
- Living in South Metropolitan Statistical Area
- Some college education
- Private insurance
- Usual source of care within 15 minutes reach
- **No** hospital expenditure nor medical/perceived needs
- MEPS panel 16

Profile A

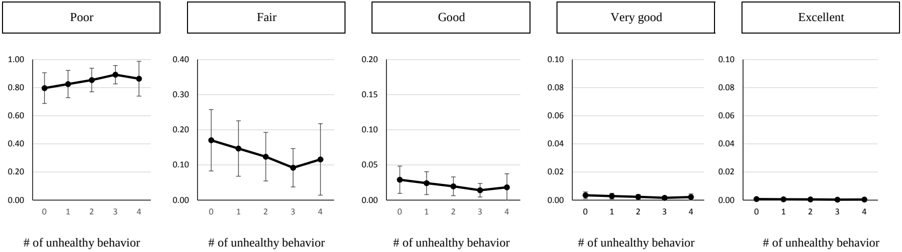

Poor
Fair
Good
Very good
Excellent
0.05
0.04
0.03
0.02
0.01
0.00
0.10
0.08
0.06
0.04
0.02
0.00
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0 1 2 3 4
# of unhealthy behavior

Profile B

Note: Categories chosen based on modal valued category except for income quantile (middle quantile)

- White female in 40’s
- Employed with total income at the middle quantile of the population
- Living in South Metropolitan Statistical Area
- Some college education
- Private insurance
- Usual source of care within 15 minutes reach
- **Has** hospital expenditure and medical/perceived needs (for years spent with diagnosis, weighted sample mean values)
- MEPS panel 16

Profile B


Poor
Fair
Good
Very good
Excellent
1.00
0.80
0.60
0.40
0.20
0.00
0.40
0.30
0.20
0.10
0.00
0.20
0.15
0.10
0.05
0.00
0.10
0.08
0.06
0.04
0.02
0.00
0 1 2 3 4
# of unhealthy behavior

Conclusion

- Reviewed logistic regression as preview for ordered logistic regression
- Covered only proportional odds model with logistic link
- Care taken with interpretation given definition of outcome variable and software function used
- Could explore other forms of ordered logistic regression
  - Other models like continuation ratio and adjacent categories
  - Other link functions like probit and complementary log-log
  - Non-constant regression parameters across levels

Helpful Resources for Ordinal Modeling in R

- http://www.stat.ufl.edu/~aa/ordinal/R_examples.pdf
- https://cran.r-project.org/web/packages/ordinal/ordinal.pdf
- https://www.researchgate.net/profile/Thomas_Yee3/publication/46515756_The_VGAM_Package_for_Categorical_Data_Analysis/links/55bea8e808ae9289a099d9ec/The-VGAM-Package-for-Categorical-Data-Analysis.pdf
- http://dwoll.de/rexrepos/posts/regressionOrdinal.html

Bibliography I

Agresti, A. (2010). *Analysis of ordinal categorical data*, Volume 656. John Wiley & Sons.

Ananth, C. V. and D. G. Kleinbaum (1997). Regression models for ordinal responses: a review of methods and applications. *International journal of epidemiology 26*(6), 1323–1333.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–30.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological) 42*(2), 109–142.

Peterson, B. and F. E. Harrell Jr (1990). Partial proportional odds models for ordinal response variables. *Applied statistics*, 205–217.

Bibliography II

Snell, E. (1964). A scaling procedure for ordered categorical data. *Biometrics*, 592–607.

Walker, S. H. and D. B. Duncan (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika 54*(1-2), 167–179.