Class 8: Chapter 9

EPH 705

Min Lu

Division of Biostatistics
University of Miami

Spring 2017

Overview

1. Object:
   - Logit Models for Nominal Responses
   - Plot and Understand Model Specification
2. R Exercise
   - Reshape Data and Plot ROC Curve
   - Logit Model for Nominal Responses
   - In class exercise
   - Take home exercise

Logit Models for Nominal Responses

Multinomial logistic regression uses a linear predictor function $f(k, i)$ to predict the probability that observation i has outcome k, of the following form:

$$f(k, i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + \cdots + \beta_{M,k}x_{M,i},$$

where $\beta_{m,k}$ is a regression coefficient associated with the mth explanatory variable and the kth outcome. The regression coefficients and explanatory variables are normally grouped into vectors of size M+1, so that the predictor function can be written more compactly:

$$f(k, i) = \boldsymbol{\beta}_k \cdot \mathbf{x}_i,$$

where $\boldsymbol{\beta}_k$ is the set of regression coefficients associated with outcome k, and $\mathbf{x}_i$ (a row vector) is the set of explanatory variables associated with observation i.

Logit Models for Nominal Responses

Using the fact that all K of the probabilities must sum to one, we find:

$$\Pr(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}}$$

We can use this to find the other probabilities:
The fact that we run multiple regressions reveals why the model relies on the assumption of independence of irrelevant alternatives described above.

Estimating response probabilities

$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol{\beta}_1 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = 2) = \frac{e^{\boldsymbol{\beta}_2 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}}$$

$$\cdots\cdots$$

$$\Pr(Y_i = K-1) = \frac{e^{\boldsymbol{\beta}_{K-1} \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\beta}_k \cdot \mathbf{X}_i}}$$

As a set of independent binary regressions

One fairly simple way to arrive at the multinomial logit model is to imagine, for K possible outcomes, running $K-1$ independent binary logistic regression models, in which one outcome is chosen as a "pivot" and then the other $K-1$ outcomes are separately regressed against the pivot outcome. This would proceed as follows, if outcome K (the last outcome) is chosen as the pivot:

Baseline-category logits

$$\ln \frac{\Pr(Y_i = 1)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_1 \cdot \mathbf{X}_i$$

$$\ln \frac{\Pr(Y_i = 2)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_2 \cdot \mathbf{X}_i$$

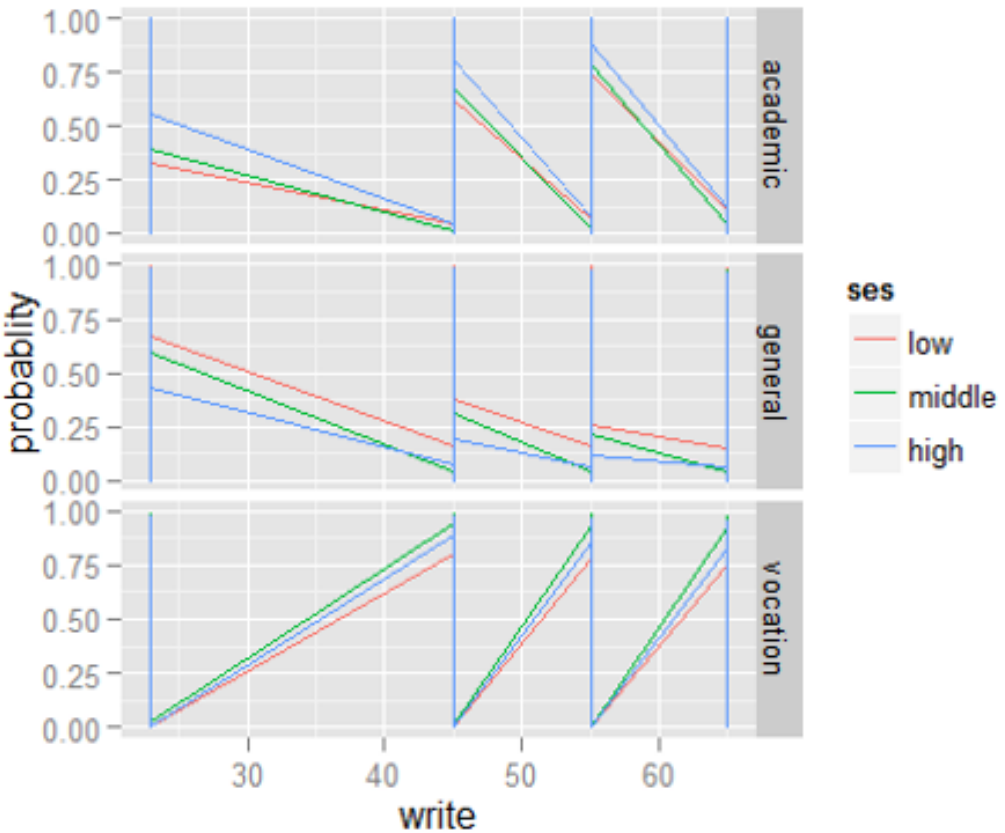
$$\cdots\cdots$$

$$\ln \frac{\Pr(Y_i = K-1)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_{K-1} \cdot \mathbf{X}_i$$

Note that we have introduced separate sets of regression coefficients, one for each possible outcome.

Model Specification Example

The modeling of program choices made by high school students can be done using Multinomial logit. The program choices are general program, vocational program and academic program. Their choice can be modeled using their writing score and their social economic status.



Reshape Data and ROC Curve

reshape a data into binary outcome

```
dat <- matrix(c(0, 10, 25, 0, 20, 35, 0, 21, 45, 7, 15, 55, 6, 11, 65, 17, 25, 75), 6, 3,
    byrow = T)
colnames(dat) <- c("disease", "n", "age")
dat <- as.data.frame(dat)
# mod=glm(formula = disease/n~age,weights=dat$n, family = binomial(link = logit), data
# = dat)

Melt <- function(data, yes, n) {
    x <- do.call(rbind, lapply(1:nrow(data), function(i) {
        cbind(data[i, -c(yes, n)], c(rep(1, data[i, yes]), rep(0, (data[i, n] - data[i, yes]))))
    }))
    colnames(x) = c(colnames(data)[-c(yes, n)], colnames(data)[yes])
    x
}
### yes=number of colume you want to covert to 1, no=.......0
head(Melt(data = dat, yes = 1, n = 2))
```

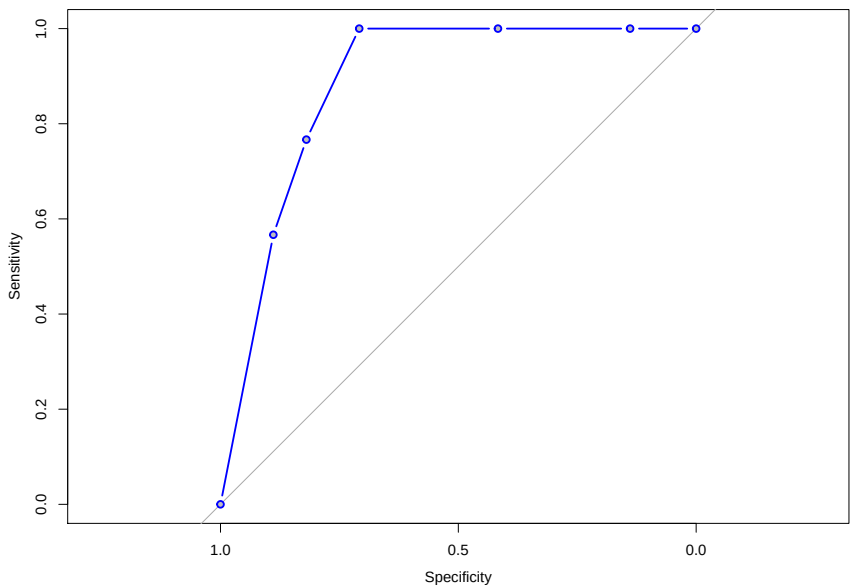
```
##      age disease
## [1,]  25       0
## [2,]  25       0
## [3,]  25       0
## [4,]  25       0
## [5,]  25       0
## [6,]  25       0
```

Reshape Data and Plot ROC Curve

ROC curve (only works for binary outcome)

```
dat1 <- as.data.frame(Melt(data = dat, yes = 1, n = 2))
mod <- glm(disease ~ age, family = binomial, data = dat1)

library(pROC)
rocobj <- plot.roc(dat1$disease, fitted(mod), type = "n")
lines(rocobj, type = "b", pch = 21, col = "blue", bg = "grey")
```



Western Collaborative Group Study data

Data:The Western Collaborative Group Study (WCGS), a prospective cohort studye, recruited middle-aged men (ages 39 to 59) who were employees of 10 California companies and collected data on 3154 individuals during the years 1960-1961. These subjects were primarily selected to study the relationship between behavior pattern and the risk of coronary hearth disease (CHD). A number of other risk factors were also measured.

variable name	discreption
id	Subject ID:
age0	Age: age in years
height0	Height: height in inches
weight0	Weight: weight in pounds
sbp0	Systolic blood pressure: mm Hg
dbp0	Diastolic blood pressure: mm Hg
chol0	Cholesterol: mg/100 ml
behpat0	Behavior pattern:
ncigs0	Smoking: Cigarettes/day
dibpat0	Dichotomous behavior pattern: 0 = Type B; 1 = Type A
chd69	Coronary heart disease event: 0 = none; 1 = yes
typechd	to be done
time169	Observation (follow up) time: Days
arcus0	Corneal arcus: 0 = none; 1 = yes

Logit Model for Nominal Responses

Logistic regression for nominal outcome

```
wcgs <- read.csv("wcgs.csv")[,-1]
library("nnet")
wcgs$behpat0 <- as.factor(wcgs$behpat0)
wcgs$chd69 <- as.factor(wcgs$chd69)
wcgs$behpat1 <- relevel(wcgs$behpat0, ref = "1")
model <- multinom(behpat1 ~ age0 + chd69, data = wcgs)
```

```
## # weights:  16 (9 variable)
## initial  value 4372.372415
## iter  10 value 3707.459619
## final  value 3698.301245
## converged
```

```
summary(model)
```

```
## Call:
## multinom(formula = behpat1 ~ age0 + chd69, data = wcgs)
##
## Coefficients:
##   (Intercept)        age0      chd691
## 2    3.017701 -0.02995064  0.04913322
## 3    3.984092 -0.05135421 -0.77081969
## 4    3.046124 -0.05818738 -0.72683151
##
## Std. Errors:
##   (Intercept)       age0    chd691
## 2   0.5637161 0.01186314 0.2146806
## 3   0.5719645 0.01207527 0.2365233
## 4   0.6939913 0.01481671 0.3125612
##
## Residual Deviance: 7396.602
## AIC: 7414.602
```

Logit Model for Nominal Responses

Logistic regression for nominal outcome

```
beta <- summary(model)$coefficients
beta.se <- summary(model)$standard.errors
colnames(beta.se) <- paste(colnames(beta), ".SE", sep = "")
round(cbind(beta, beta.se), 2)
```

```
##   (Intercept)  age0 chd691 (Intercept).SE age0.SE chd691.SE
## 2        3.02 -0.03   0.05           0.56    0.01      0.21
## 3        3.98 -0.05  -0.77           0.57    0.01      0.24
## 4        3.05 -0.06  -0.73           0.69    0.01      0.31
```

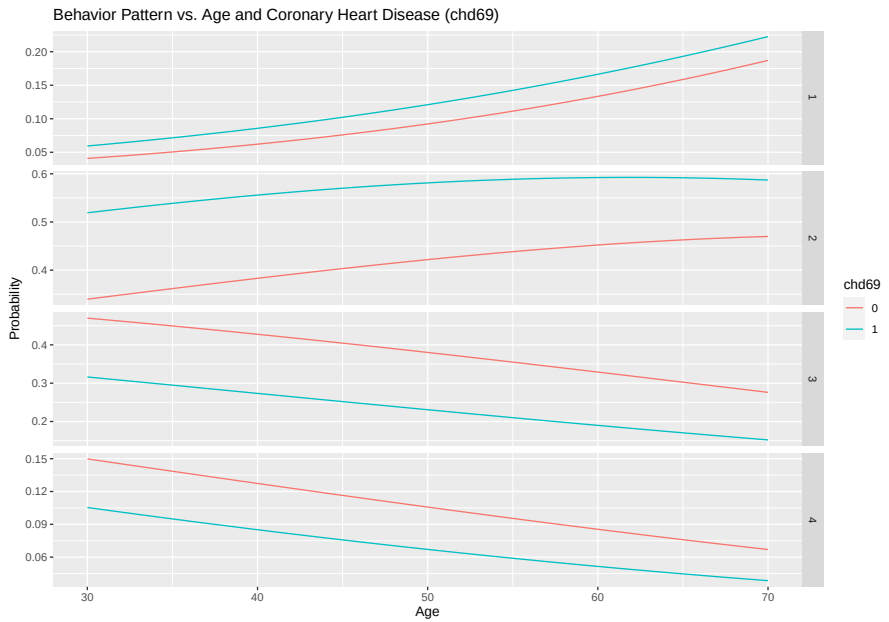
```
## p value for the coefficients
1 - pnorm(abs(beta/beta.se))
```

```
##    (Intercept)         age0       chd691
## 2 4.319943e-08 5.790182e-03 0.4094863045
## 3 1.634692e-12 1.055377e-05 0.0005591098
## 4 5.686254e-06 4.297988e-05 0.0100251449
```

Plot and Understand Model Specification

GGplot Logistic regression for nominal outcome

```
xx <- data.frame(chd69 = rep(as.factor(c(0, 1)), each = 41), age0 = rep(c(30:70),2))
pp.xx <- cbind(xx, predict(model, type = "probs", newdata =xx, se = TRUE)) #
library("reshape")
lpp <- melt(pp.xx, id.vars = c("age0","chd69"), value.name = "probability")
library("ggplot2")
ggplot(lpp, aes(x = age0, y = value, colour = chd69)) + geom_line() + facet_grid(variable ~., scales = "free")+ labs(lines= "Nitrogen",
       x = "Age",
       y = "Probability",
       title = "Behavior Pattern vs. Age and Coronary Heart Disease (chd69)")
```



In class exercise

Use the wcgs data and plot Figure 1.6 but substitue variable chd69 with arcus0.

Take home exercise

Use the wcgs data, choose your own outcome and predictors, conduct a Logit Model for Nominal Responses and make two plots: one from model with main effects only and the other from model with interaction.

Class over