BST 210 Homework 6

Wenjie Gu

```r
library(foreign)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(nnet)
```

Problem 1

```r
# Data cleaning
library(haven)
framingham = read_dta("Data and Programs/framingham.dta")
framingham = framingham[framingham$prevchd == 0,]
```

```r
framingham$outcome = (framingham$death == 0 & framingham$anychd == 0)*1 +
  (framingham$death == 0 & framingham$anychd == 1)*2 + (framingham$death == 1)*3
framingham$prevchd = NULL
framingham$sex = framingham$sex -1
```

```r
summ.MNfit <- function(fit, digits=3){
  s <- summary(fit)
  for(i in 2:length(fit$lev))
  {
    ##
    cat("\nLevel", fit$lev[i], "vs. Level", fit$lev[1], "\n")
    ##
    betaHat <- s$coefficients[(i-1),]
    se <- s$standard.errors[(i-1),]
    zStat <- betaHat / se
    pval <- 2 * pnorm(abs(zStat), lower.tail=FALSE)
    ##
    RRR <- exp(betaHat)
    RRR.lo <- exp(betaHat - qnorm(0.975)*se)
    RRR.up <- exp(betaHat + qnorm(0.975)*se)
    ##
    results <- cbind(betaHat, se, pval, RRR, RRR.lo, RRR.up)
    print(round(results, digits=digits))
  }
}
```

```r
model.sex = multinom(outcome~sex, data = framingham)
```

```
## # weights:  9 (4 variable)
## initial  value 4658.116104
## iter  10 value 3904.409463
## iter  10 value 3904.409444
## final  value 3904.409444
## converged
```

```r
model.age = multinom(outcome~age, data = framingham)
```

```
## # weights:  9 (4 variable)
## initial  value 4658.116104
## iter  10 value 3581.808246
## iter  10 value 3581.808246
## final  value 3581.808246
## converged
```

```r
model.age.sex = multinom(outcome~age+sex, data = framingham)
```

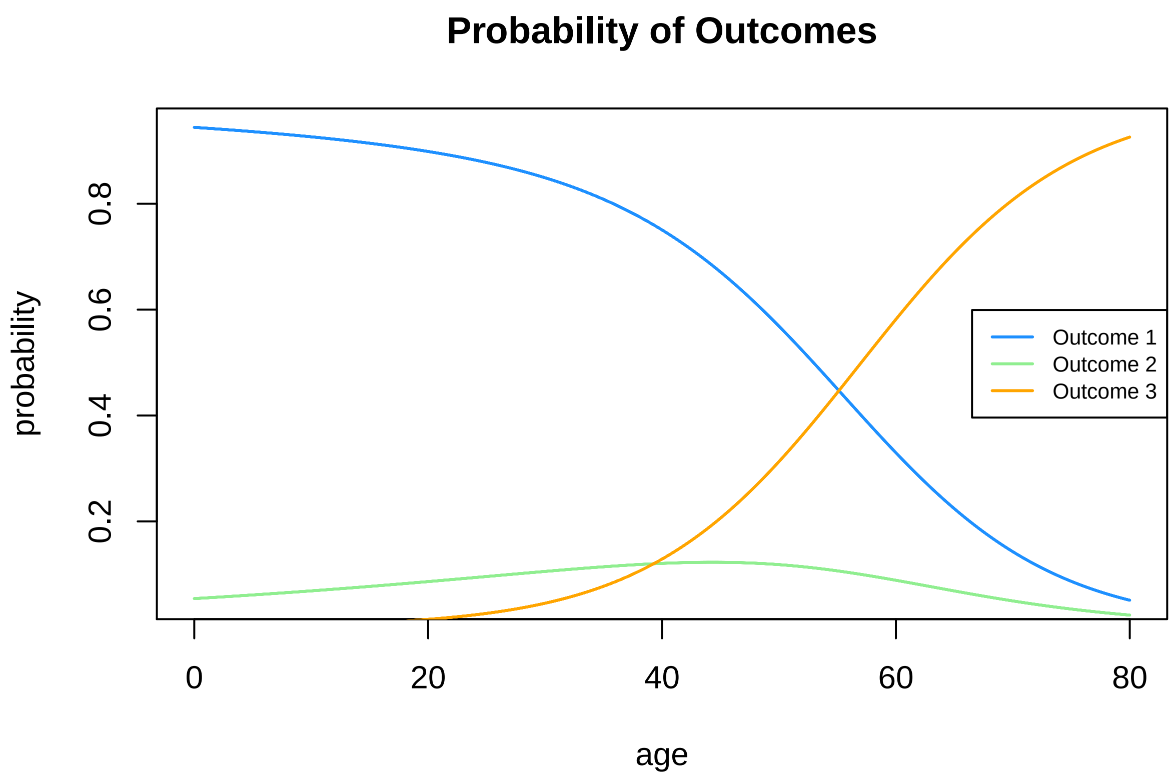
```
## # weights:  12 (6 variable)
## initial  value 4658.116104
## iter  10 value 3509.585533
## final  value 3509.484236
## converged
```

```r
model.age.sex.int = multinom(outcome~age+sex+age*sex, data = framingham)
```

```
## # weights:  15 (8 variable)
## initial  value 4658.116104
## iter  10 value 3508.375999
## final  value 3505.913226
## converged
```

1(a)

```r
age.seq = seq(0,80,0.01)
prob.model.age = predict(model.age,list(age = age.seq), type = 'probs')
par(mfrow = c(1,1))
plot(age.seq, prob.model.age[,1], cex = 0.05, main = "Probability of Outcomes",
     xlab ="age",ylab = "probability", col = "dodgerblue")
points(age.seq, prob.model.age[,2], cex = 0.05, col = "lightgreen")
points(age.seq, prob.model.age[,3], cex = 0.05, col = "orange")
legend("right",c("Outcome 1", "Outcome 2", "Outcome 3"),
       col = c("dodgerblue","lightgreen","orange"), lwd = 1.5, cex = 0.7)
```



As age goes up, the estimated probability of outcome 1d (no death or chronic heart disease in the follow-up period) decreases. The probability of outcome 2 (chronic heart disease, but remained alive) follows a parabolic pattern, and reaches maximum at around 40 yo. The probability of outcome 3 (death) gets larger as age goes up. According to the fitted probability curves, the probability of outcome 2 (CHD) is overall lower than outcome 1 (no chd or death) and outcome 3 (death). For people older than 55 years old, the prevalent outcome is death and for people younger than 55 years old, the prevalent outcome is no death or chd.

```
summ.MNfit(model.age)
```

```
##
## Level 2 vs. Level 1
##             betaHat    se pval   RRR RRR.lo RRR.up
## (Intercept)  -2.859 0.312    0 0.057  0.031  0.106
## age           0.026 0.006    0 1.026  1.013  1.039
##
## Level 3 vs. Level 1
##             betaHat    se pval   RRR RRR.lo RRR.up
## (Intercept)  -6.424 0.244    0 0.002  0.001  0.003
## age           0.117 0.005    0 1.124  1.113  1.134
```

```
confint(model.age)
```

```
## , , 2
##
##                   2.5 %      97.5 %
## (Intercept) -3.47042145 -2.24782917
## age          0.01326044  0.03830034
##
## , , 3
##
##                  2.5 %     97.5 %
```

```
## (Intercept) -6.9032729 -5.9455880
## age          0.1073283  0.1257421
```

```
vcov(model.age)
```

```
##                2:(Intercept)         2:age 3:(Intercept)         3:age
## 2:(Intercept)  0.0972763186 -1.965591e-03  0.0180569339 -3.734082e-04
## 2:age         -0.0019655907  4.080459e-05 -0.0003756535  7.955112e-06
## 3:(Intercept)  0.0180569339 -3.756535e-04  0.0596882838 -1.134172e-03
## 3:age         -0.0003734082  7.955112e-06 -0.0011341722  2.206629e-05
```

```
beta3_2_std = sqrt(4.080459e-05 + 2.206629e-05 - 2* 7.955112e-06)
beta3_2_std
```

```
## [1] 0.006852785
```

```
lower = exp(0.117*10 - 0.026*10- 1.96 * beta3_2_std *10)
upper = exp(0.117*10 - 0.026*10+ 1.96 * beta3_2_std *10)
```

```
sprintf("The estimated relative risk ratio of having outcome 2 to having outcome 1 for a population is %
        exp(0.026*10), exp(0.01326*10), exp(0.0383*10))
```

```
## [1] "The estimated relative risk ratio of having outcome 2 to having outcome 1 for a population is 1
```

```
sprintf("The estimated relative risk ratio of having outcome 3 to having outcome 1 for a population is %
        exp(0.117*10), exp(0.1073*10), exp(0.12574*10))
```

```
## [1] "The estimated relative risk ratio of having outcome 3 to having outcome 1 for a population is 3
```

```
sprintf("The estimated relative risk ratio of having outcome 3 to having outcome 2 for a population is %
        exp(0.117*10)/exp(0.026*10),lower, upper)
```

```
## [1] "The estimated relative risk ratio of having outcome 3 to having outcome 2 for a population is 2
```

1(b)

```
female = fitted(model.sex)[framingham$sex == 1,][1,]
male = fitted(model.sex)[framingham$sex == 0,][1,]
fitted_prob_table_sex = rbind(male,female)
outcome_sex_table = table(framingham$sex, framingham$outcome)
outcome_sex_table_prop = prop.table(outcome_sex_table, 1)
```

```
fitted_prob_table_sex
```

```
##                1         2         3
## male   0.4697938 0.1236231 0.4065831
## female 0.6260492 0.0987825 0.2751683
```

```
outcome_sex_table_prop
```

```
##
##              1          2          3
##   0 0.46978022 0.12362637 0.40659341
##   1 0.62603306 0.09876033 0.27520661
```

According to the tables above, we can confirm that for the model with sex alone, the fitted probabilities match the outcome-sex tabulation exactly.

```
summ.MNfit(model.sex)
```

```
##
## Level 2 vs. Level 1
##             betaHat    se pval   RRR RRR.lo RRR.up
## (Intercept)  -1.335 0.075    0 0.263  0.227  0.305
## sex          -0.511 0.102    0 0.600  0.491  0.733
##
## Level 3 vs. Level 1
##             betaHat    se  pval   RRR RRR.lo RRR.up
## (Intercept)  -0.145 0.050 0.004 0.865  0.784  0.955
## sex          -0.678 0.068 0.000 0.508  0.444  0.581
```

```r
outcome_sex_table_prop
```

```
##
##              1          2          3
##   0 0.46978022 0.12362637 0.40659341
##   1 0.62603306 0.09876033 0.27520661
```

RRR_21 = (P(Y=2|female)/P(Y=1|female))/(P(Y=2|male)/P(Y=1|male)) = (0.09876033/0.62603306)/(0.12362637/0.4697802
RRR_31 = (P(Y=3|female)/P(Y=1|female))/(P(Y=3|male)/P(Y=1|male)) = (0.27520661/0.62603306)/(0.40659341/0.4697802

```r
# calculated RRRs from the tabulation
RRR_21 = (0.09876033/0.62603306)/(0.12362637/0.46978022)
RRR_21
```

```
## [1] 0.599472
```

```r
RRR_31 = (0.27520661/0.62603306)/(0.40659341/0.46978022)
RRR_31
```

```
## [1] 0.5079208
```

According to the summary of the model with sex alone, relative risk ratio of outcome 2 to outcome 1 is 0.600, and relative risk ratio of outcome 3 to outcome 1 is 0.508. The calculated RRRs from the tabulation match the results.

1(c)

```r
anova(model.age.sex, model.age.sex.int, test = "Chisq")
```

```
##                   Model Resid. df Resid. Dev   Test    Df LR stat.
## 1             age + sex      8474   7018.968           NA       NA
## 2 age + sex + age * sex      8472   7011.826 1 vs 2     2  7.14202
##      Pr(Chi)
## 1         NA
## 2 0.02812743
```

The LRT statistic has a p-value = 0.028 ($p<0.05$). Therefore, we can reject the reduced model and conclude that the model including age,sex, and their interaction performs better than the one without interaction. We can consider fitting models with non-linear age terms in our next step.

Problem 2

```r
library(VGAM)
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```r
library(stats4)
library(splines)
ord.age = vglm(outcome~age,cumulative(parallel=TRUE, reverse=TRUE), data = framingham)
ord.sex = vglm(outcome~sex,cumulative(parallel=TRUE, reverse=TRUE), data = framingham)
ord.age.sex = vglm(outcome~age + sex,cumulative(parallel=TRUE, reverse=TRUE), data = framingham)
ord.age.sex.int = vglm(outcome~age + sex + age*sex,cumulative(parallel=TRUE, reverse=TRUE), data = fram
```

2(a)

```r
# model with age alone
summary(ord.age)
```

```
##
## Call:
## vglm(formula = outcome ~ age, family = cumulative(parallel = TRUE,
##     reverse = TRUE), data = framingham)
##
## Pearson residuals:
##                         Min      1Q  Median     3Q   Max
## logitlink(P[Y>=2]) -2.036 -0.6986 -0.4423 0.4974 3.733
## logitlink(P[Y>=3]) -2.827 -0.4269 -0.2742 0.7561 2.979
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -5.170906   0.202854  -25.49   <2e-16 ***
## (Intercept):2 -5.713239   0.206504  -27.67   <2e-16 ***
## age            0.099351   0.003977   24.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y>=2]), logitlink(P[Y>=3])
##
## Residual deviance: 7208.382 on 8477 degrees of freedom
##
## Log-likelihood: -3604.191 on 8477 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##      age
## 1.104454
```

```r
confint(ord.age)
```

```
##                   2.5 %     97.5 %
## (Intercept):1 -5.5684928 -4.7733183
## (Intercept):2 -6.1179799 -5.3084974
## age            0.0915555  0.1071468
```

```r
sprintf("The estimated odds ratio  for the effect of 10 years comparing outcome 3 vs. outcome 1 and 2 (
        exp(0.099351*10), exp(0.0915555*10), exp(0.1071468*10))
```

```
## [1] "The estimated odds ratio  for the effect of 10 years comparing outcome 3 vs. outcome 1 and 2 (co
```

```
sprintf("The estimated odds ratio  for the effect of 10 years comparing outcome 2 and 3(combined) vs. ou
        exp(0.099351*10), exp(0.0915555*10), exp(0.1071468*10))
```

```
## [1] "The estimated odds ratio  for the effect of 10 years comparing outcome 2 and 3(combined) vs. out
```

2(b)

```
framingham$outcome1 = 1*(framingham$outcome == 3)
framingham$outcome2 = 1*(framingham$outcome == 2 | framingham$outcome == 3)
log.outcome1 = glm(outcome1~age, family = binomial(), data = framingham)
log.outcome2 = glm(outcome2~age, family = binomial(), data = framingham)
summary(log.outcome1)
```

```
##
## Call:
## glm(formula = outcome1 ~ age, family = binomial(), data = framingham)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7719  -0.8255  -0.5540   0.9966   2.3034
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.384296   0.237005  -26.94   <2e-16 ***
## age          0.111896   0.004524   24.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5387.4  on 4239  degrees of freedom
## Residual deviance: 4652.9  on 4238  degrees of freedom
## AIC: 4656.9
##
## Number of Fisher Scoring iterations: 4
```

```
summary(log.outcome2)
```

```
##
## Call:
## glm(formula = outcome2 ~ age, family = binomial(), data = framingham)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8075  -0.9741  -0.6882   1.0812   1.9261
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.78576    0.20846  -22.96   <2e-16 ***
## age          0.09121    0.00411   22.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5818.8  on 4239  degrees of freedom
## Residual deviance: 5256.7  on 4238  degrees of freedom
## AIC: 5260.7
##
## Number of Fisher Scoring iterations: 4
```

```r
confint(log.outcome1)
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %     97.5 %
## (Intercept) -6.8535984 -5.9243745
## age          0.1031075  0.1208437
```

```r
confint(log.outcome2)
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %      97.5 %
## (Intercept) -5.19745098 -4.38019036
## age          0.08320882  0.09932408
```

The two beta coefficients for age in the two logistic regression models are 0.111896 and 0.09121 respectively, which are not close to each other. And the 95% CIs are (0.1031075, 0.1208437) and (0.08320882, 0.09932408) respectively, which do not overlap. Therefore, I suggest that proportional odds assumption doesn't hold for the ordinal logistic regression model with age alone.

2(c)

```r
summary(ord.sex)
```

```
##
## Call:
## vglm(formula = outcome ~ sex, family = cumulative(parallel = TRUE,
##     reverse = TRUE), data = framingham)
##
## Pearson residuals:
##                          Min      1Q  Median     3Q   Max
## logitlink(P[Y>=2]) -0.9474 -0.6896 -0.6896 0.6873 2.359
## logitlink(P[Y>=3]) -1.8868 -0.4672 -0.3546 1.0803 1.483
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  0.10961    0.04578   2.394   0.0166 *
## (Intercept):2 -0.36488    0.04612  -7.912 2.53e-15 ***
## sex           -0.61834    0.06081 -10.168  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y>=2]), logitlink(P[Y>=3])
##
## Residual deviance: 7810.012 on 8477 degrees of freedom
##
## Log-likelihood: -3905.006 on 8477 degrees of freedom
##
```

```
## Number of Fisher scoring iterations: 3
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##       sex
## 0.5388392
```

```r
# model with sex alone
tab1 = table(framingham$outcome1, framingham$sex)
tab2 = table(framingham$outcome2, framingham$sex)
tab1
```

```
##
##         0    1
##   0 1080 1754
##   1  740  666
```

```r
tab2
```

```
##
##         0    1
##   0  855 1515
##   1  965  905
```

```r
OR_12_3 = (666/1754)/(740/1080)
OR_12_3
```

```
## [1] 0.5541619
```

```r
OR_1_23 = (905/1515)/(965/855)
OR_1_23
```

```
## [1] 0.5292669
```

The associated odds ratio estimates are 0.554169 ((1,2) vs. 3) and 0.5292669 (1 vs (2,3)) are close to the ordinal logistic regrssion-based odds ratio estimate for sex: 0.5388392. Therefore, I suggest that the proportional odds model assumption holds for the ordinal logistic regrssion model with sex alone. However, in the ordinal logistic regrssion model, there are 2 covariate patterns (female or male) and 3 parameters, indicating that the ordinal logisitc regression model for sex alone is not saturated.

2(d)

```r
pchisq(deviance(ord.age.sex)-deviance(ord.age.sex.int),
       df = df.residual(ord.age.sex)-df.residual(ord.age.sex.int),lower.tail=F)
```

```
## [1] 0.235037
```

```r
summary(ord.age.sex.int)
```

```
##
## Call:
## vglm(formula = outcome ~ age + sex + age * sex, family = cumulative(parallel = TRUE,
##     reverse = TRUE), data = framingham)
##
## Pearson residuals:
##                        Min      1Q  Median     3Q   Max
## logitlink(P[Y>=2]) -2.398 -0.6834 -0.4094 0.4837 4.155
```

```
## logitlink(P[Y>=3]) -3.177 -0.4147 -0.2630 0.7150 3.617
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -5.250835   0.306883 -17.110   <2e-16 ***
## (Intercept):2 -5.811767   0.309567 -18.774   <2e-16 ***
## age            0.110043   0.006173  17.828   <2e-16 ***
## sex           -0.315679   0.415731  -0.759    0.448
## age:sex       -0.009686   0.008155  -1.188    0.235
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y>=2]), logitlink(P[Y>=3])
##
## Residual deviance: 7055.417 on 8475 degrees of freedom
##
## Log-likelihood: -3527.708 on 8475 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##       age       sex   age:sex
## 1.1163259 0.7292937 0.9903608
```

By comparing the ordinal logistic regrssion model with age $\times$ sex interaction and without age $\times$ sex interaction with a "likelihood ratio" test, we get p-value = 0.2305 (>0.05). Furthermore, according to Wald test, in ordinal logistic regrssion model with age $\times$ sex interaction, the interaction term is not significant. Therefore, we fail to reject the null hypothesis and conclude that the age $\times$ sex interaction is not necessary for ordinal logistic regrssion modeling.

2(e)

```r
ord.po = vglm(outcome~age+sex, family = cumulative(parallel = TRUE,
    reverse = TRUE), data = framingham)
ord.npo = vglm(outcome~age+sex,family = cumulative(parallel = FALSE,
    reverse = TRUE), data = framingham)
pchisq(deviance(ord.po)- deviance(ord.npo),
       df = df.residual(ord.po) - df.residual(ord.npo), lower.tail = F)
```

```
## [1] 1.423149e-09
```

By comparing the model with proportional odds assumption and the one without with a "likelihood ratio" test, we get p-value = 1.423149e-09 (<0.05). Therefore, we reject the null hypothesis and conclude that porportional odds assumption does not hold for the model including both effects of age and sex. I would recommend using multinomial logistic regression if we wanted to include continuous age in the modeling. On top of that, we can add quadratic term of age to the model and assess whether there is nonlinear relationship between age and outcome.