HW #6

Hillary Miller

October 31, 2017

BST 210 HOMEWORK #6

Due 8:00 AM, Wednesday, November 8, 2017

For this assignment, you are welcome to work with one or two colleagues (maximum of three people working together) and turn in this assignment together, or you can work alone, your choice. (This is not the second project yet, but just a regular homework.)

1. Consider again the Framingham Heart Study data set. Suppose we are interested in looking at a three-level outcome for incidence of either coronary heart disease (with the subject still alive) or death, compared to a reference group of subjects who neither died nor had coronary heart disease in the follow-up period. Thus, we want to restrict the analysis to exclude subjects with prevalent coronary heart disease (prevchd = 1), and create a three-level outcome consisting of:

1 = no death or coronary heart disease in the follow-up period (reference category)

2 = coronary heart disease in the follow-up period, but the subject remained alive

3 = death from any cause in the follow-up period.

Thus, you will need to use prevchd, anychd, and death to create the outcome variable and sample to use. Using this sample, we will explore some multinomial and ordinal logistic regression models, using participant sex and continuous age as predictor variables. It might be easiest to recode sex to be an indicator for female (i.e., = 1 for female, = 0 for male). There should be 4,240 observations if you are using the Framingham dataset, with no one missing outcome, age, or sex.

```
library(foreign)
library(nnet)
library(haven)
library(dplyr)
```

Hillary Miller
BST 210 HW 6

```
##
## 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

framingham <- read_dta("C:/Users/millerhillaryv/Desktop/HSPH/BST210/BST 210 l
ab/Lab Week 6/framingham.dta")

dat <- framingham

dat <- dat %>% mutate(sex= factor((sex-1))) %>% as.data.frame()
dat <- dat %>% select(sex, age, death, prevchd, anychd) %>%filter(prevchd==0)
dat <- dat[complete.cases(dat), ]

dat <- mutate(dat, outcome = ifelse(death==1,3,
                                    ifelse(anychd==1,2,1)))
```

Fit four multinomial logistic regression models using age alone, sex alone, both age and sex, and finally age, sex, and their interaction, and answer the following questions:

```
mod.1 = multinom(outcome ~ age, data = dat)

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

mod.2 = multinom(outcome ~ sex, data = dat)

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

mod.3 = multinom(outcome ~ age + sex, data = dat)

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

```
mod.4 = multinom(outcome ~ age + sex + age*sex, data = dat)

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

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))
  }
}
summ.MNfit(mod.1)

##
## 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

summ.MNfit(mod.2)

##
## 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
## sex1         -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
## sex1         -0.678 0.068 0.000 0.508  0.444  0.581
```

```
summ.MNfit(mod.3)
```

```
##
## Level 2 vs. Level 1
##              betaHat    se pval   RRR RRR.lo RRR.up
## (Intercept)  -2.703 0.314    0 0.067  0.036  0.124
## age           0.029 0.006    0 1.030  1.017  1.043
## sex1         -0.558 0.103    0 0.572  0.468  0.700
##
## Level 3 vs. Level 1
##              betaHat    se pval   RRR RRR.lo RRR.up
## (Intercept)  -6.224 0.248    0 0.002  0.001  0.003
## age           0.122 0.005    0 1.130  1.119  1.141
## sex1         -0.891 0.077    0 0.410  0.353  0.477
```

```
summ.MNfit(mod.4)
```

```
##
## Level 2 vs. Level 1
##              betaHat    se  pval   RRR RRR.lo RRR.up
## (Intercept)  -2.006 0.474 0.000 0.134  0.053  0.340
## age           0.014 0.010 0.149 1.015  0.995  1.035
## sex1         -1.849 0.640 0.004 0.157  0.045  0.551
## age:sex1      0.027 0.013 0.042 1.027  1.001  1.054
##
## Level 3 vs. Level 1
##              betaHat    se  pval   RRR RRR.lo RRR.up
## (Intercept)  -6.500 0.371 0.000 0.002  0.001  0.003
## age           0.128 0.007 0.000 1.136  1.120  1.153
## sex1         -0.322 0.505 0.525 0.725  0.269  1.952
## age:sex1     -0.011 0.010 0.280 0.989  0.971  1.009
```

(a) For the model with age alone, calculate and graph the fitted probabilities for each category as a function of age. Briefly interpret your graph. Also, what is the estimated relative risk ratio and 95% CI for the effect of 10 years of age when comparing outcome 2 to outcome 1? What is the estimated relative risk ratio and 95% CI for the effect of 10 years of age when comparing outcome 3 to outcome 1? A little harder: What is the relative risk ratio and 95% CI for the effect of 10 years of age when comparing outcome 3 to outcome 2?

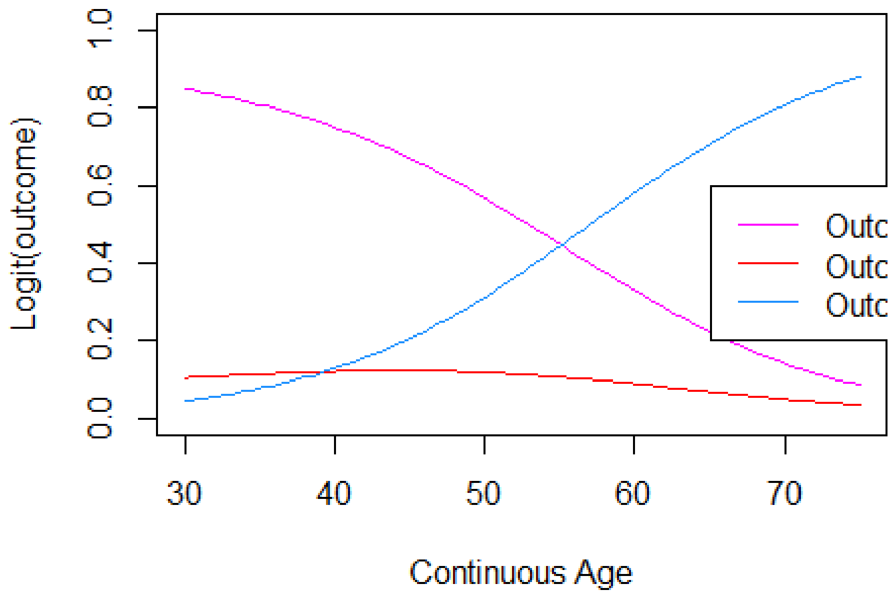
We estimate that the risk ratio of having outcome 2 (coronary heart disease in the follow-up period, but the subject remained alive) to having outcome 1 (no death or coronary heart disease in the follow-up period (reference category)) for a population given of age = k+10 is 1.29 times this risk ratio for a population of age k, with 95% Confidence Interval [1.14,1.47].

We estimate that the risk ratio of having outcome 3 (death from any cause in the follow-up period) to having outcome 1 (no death or coronary heart disease in the

follow-up period (reference category)) for a population given of age = k+10 is 3.21 times this risk ratio for a population of age k, with 95% Confidence Interval [2.92,3.52].

We estimate that the risk ratio of having outcome 3 (death from any cause in the follow-up period) to having outcome 2 (coronary heart disease in the follow-up period, but the subject remained alive) for a population given of age = k+10 is 2.48 times this risk ratio for a population of age k, with 95% Confidence Interval [2.17,2.83].

```
summary(mod.1)

## Call:
## multinom(formula = outcome ~ age, data = dat)
##
## Coefficients:
##   (Intercept)        age
## 2   -2.859125 0.02578039
## 3   -6.424430 0.11653522
##
## Std. Errors:
##   (Intercept)         age
## 2   0.3118915 0.006387847
## 3   0.2443119 0.004697477
##
## Residual Deviance: 7163.616
## AIC: 7171.616

summ.MNfit(mod.1)

##
## 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

curve(1/(1+ exp(coef(mod.1)[1,1]+(coef(mod.1)[1,2]*x)) + exp(coef(mod.1)[2,1]
+(coef(mod.1)[2,2]*x))), col="magenta", xlab="Continuous Age", ylab="Logit(ou
tcome)", xlim=c(30,75), ylim=c(0,1))

curve(exp(coef(mod.1)[1,1]+(.026*x))/(1+ exp(coef(mod.1)[1,1]+(coef(mod.1)[1,
2]*x)) + exp(coef(mod.1)[2,1]+(coef(mod.1)[2,2]*x))), col="red", add=T, ylim=
c(0,1), xlim=c(30,75))
curve(exp(coef(mod.1)[2,1]+(coef(mod.1)[2,2]*x))/(1+ exp(coef(mod.1)[1,1]+(co
ef(mod.1)[1,2]*x)) + exp(coef(mod.1)[2,1]+(coef(mod.1)[2,2]*x))), col="dodger
blue", add=T, ylim=c(0,1), xlim=c(30,75))
legend(65,.6, legend=c("Outcome 1","Outcome 2", "Outcome 3"),lty=c(1,1),col=c
("magenta","red", "dodgerblue"))
```



what is the estimated relative risk ratio and 95% CI for the effect of 10 years of age when comparing outcome 2 to outcome 1?

$$exp(\alpha_2 + \beta_{21}(x + 10))/\exp(\alpha_2 + \beta_{21}(x)) = \exp(\beta_{2_1} * 10)$$

```
##relative ridk ratio 2 to 1
exp(coef(mod.1)[1,1]+(coef(mod.1)[1,2]*11))/exp(coef(mod.1)[1,1]+(coef(mod.1)
[1,2]*1))

## [1] 1.294085

exp(coef(mod.1)[1,2]*10)

## [1] 1.294085

##95% CI
exp(10*(coef(mod.1)[1,2]-(1.96*(summary(mod.1)$standard.errors[1,2]))))

## [1] 1.141796

exp(10*(coef(mod.1)[1,2]+(1.96*(summary(mod.1)$standard.errors[1,2]))))

## [1] 1.466686
```

Hillary Miller
BST 210 HW 6

What is the estimated relative risk ratio and 95% CI for the effect of 10 years of age when comparing outcome 3 to outcome 1?

$$\exp(\alpha_3 + \beta_{31}(x + 10))/\exp(\alpha_3 + \beta_{31}(x)) = \exp(\beta_{31} * 10)$$

```
#relative risk ratio 3  to 1
exp(coef(mod.1)[2,1]+(coef(mod.1)[2,2]*11))/exp(coef(mod.1)[2,1]+(coef(mod.1)
[2,2]*1))

## [1] 3.207052

exp(coef(mod.1)[2,2]*10)

## [1] 3.207052

#CI
 exp(10*(coef(mod.1)[2,2]-(1.96*(summary(mod.1)$standard.errors[2,2]))))

## [1] 2.924963

exp(10*(coef(mod.1)[2,2]+(1.96*(summary(mod.1)$standard.errors[2,2]))))

## [1] 3.516347
```

A little harder: What is the relative risk ratio and 95% CI for the effect of 10 years of age when comparing outcome 3 to outcome 2?

```
##relative risk ratio 3 to 2
exp(10*(coef(mod.1)[2,2]-coef(mod.1)[1,2]))

## [1] 2.478239



#must find the covariance

vcov(mod.1)

##which we find to be 7.955112 * 10-6

##CI

exp(10*((coef(mod.1)[2,2]-coef(mod.1)[1,2])-(1.96*(sqrt(0.006387847^2 + 0.004697477^
2- (2*.000007955112))))))

## [1] 2.166762

exp(10*((coef(mod.1)[2,2]-coef(mod.1)[1,2])+(1.96*(sqrt(0.006387847^2 + 0.004697477
^2-(2*.000007955112))))))

## [1] 2.834493
```

Hillary Miller
BST 210 HW 6

(b) For the model with sex alone, confirm that the fitted probabilities match those of an outcome × sex tabulation exactly. Also confirm that the estimated relative risk ratios for sex from your model match the relative risk ratios from the tabulation. Note that this would only occur with a "saturated model" like when you only have a single dichotomous predictor as here – this will not happen for continuous covariates, say.

As demonstrated below, the fitted probabilities match those of an outcome x sex tabulation exactly (comparing table outcome the fitted probabilities of the model). Also, the estimated relative risk ratios for sex match the relative risk ratios from the tabulation. This can be seen from the tabulation below. (Relative Risk Ratio from the model was found using exp(B). This is true for saturated models.

```
#mod.2 = multinom(outcome ~ sex, data = dat)

# Look at how fitted values compare to observed proportions
table(dat$outcome, dat$sex)

## 
##        0    1
##   1  855 1515
##   2  225  239
##   3  740  666

prop.table(table(dat$outcome, dat$sex), margin = 2)

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

# Female fitted probs.
fitted(mod.2)[dat$sex == 1,][1,]

##         1         2         3 
## 0.6260492 0.0987825 0.2751683

# Male fitted probs.
fitted(mod.2)[dat$sex == 0,][1,]

##         1         2         3 
## 0.4697938 0.1236231 0.4065831

#confirming the estimated relative risk ratios are the same
summ.MNfit(mod.2)

## 
## 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
## sex1         -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
## sex1         -0.678 0.068 0.000 0.508  0.444  0.581

##RRR level 2 to level 1 for Females (sex=1) vs. Males (sex=0)
(239/225)/(1515/855)

## [1] 0.5994719

###RRR level 3 to level 1 for Females (sex=1) vs. Males (sex=0)
(666/740)/(1515/855)

## [1] 0.5079208

###RRR level 3 to level 2 for Females (sex=1) vs. Males (sex=0)
(666/740)/(239/225)

## [1] 0.8472803

#RRR comparing 3 to 2 for the fitted model
exp(-0.6775408--0.5114517)

## [1] 0.8469708
```

(c) Use a LRT to decide between the models including both age and sex and the model including age, sex, and their interaction. What do you conclude? Are there any other models you might recommend fitting next?

Comparing the model including both age and sex and the model including age, sex and their interaction, the full model is preferred. We can reject the null that the reduced model is sufficient (P=.028). To further explore the relationship between the outcomes, I would potentially break the age varialbe up into categories, in order to see if different age groups have different outcomes that are statistically signficant. You could also test to see if quadrtic age is significant (done below). From the results of this anova test, you can see that the model with quadratic age is preferred (P=.00006) If so, this could lead to a model that is easier to interpret.

```
#LRT will compare model 3 (reduced) and model 4 (full)
anova(mod.4, mod.3, 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
```

Formal hypothesis: (Ho: the reduced model, with age and sex, is sufficient) versus (H1: the full model, with age, sex and the interaction, is preferred)

```
mod.5 = multinom(outcome ~ age + sex + age*sex + I(age^2), data = dat)

## # weights:  18 (10 variable)
## initial  value 4658.116104
## iter  10 value 3525.599975
## final  value 3498.514289
## converged

anova(mod.4, mod.5, test="Chisq")

##                               Model Resid. df Resid. Dev   Test    Df
## 1             age + sex + age * sex      8472   7011.826           NA
## 2 age + sex + age * sex + I(age^2)      8470   6997.029 1 vs 2     2
##   LR stat.      Pr(Chi)
## 1       NA           NA
## 2 14.79787 0.0006119026
```

2. Now, fit four ordinal logistic regression models using age alone, sex alone, both age and sex, and finally age, sex, and their interaction, and answer the following questions:

```
###1    =  no death or coronary heart disease in the follow-up period (refere
nce category)
###2    =  coronary heart disease in the follow-up period, but the subject re
mained alive
###3    =  death from any cause in the follow-up period.



# NOTE:
# We will get the negative of the coefficients and same "intercept" as in Sta
ta if we set reverse = FALSE(default)

library(VGAM)

## Loading required package: stats4

## Loading required package: splines

mod2.1 = vglm(outcome ~ age,
             cumulative(parallel=TRUE, reverse=FALSE), data=dat)
summary(mod2.1)

##
## Call:
## vglm(formula = outcome ~ age, family = cumulative(parallel = TRUE,
##     reverse = FALSE), data = dat)
##
##
## Pearson residuals:
##                      Min      1Q Median     3Q   Max
## logit(P[Y<=1]) -3.733 -0.4974 0.4423 0.6986 2.036
## logit(P[Y<=2]) -2.979 -0.7561 0.2742 0.4269 2.827
##
## 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
##
## Number of linear predictors:  2
##
## Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])
##
## Residual deviance: 7208.382 on 8477 degrees of freedom
##
## Log-likelihood: -3604.191 on 8477 degrees of freedom
##
## Number of iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##       age
## 0.9054247
```

```
mod2.2 = vglm(outcome ~ sex,
             cumulative(parallel=TRUE, reverse=FALSE), data=dat)
summary(mod2.2)
```

```
##
## Call:
## vglm(formula = outcome ~ sex, family = cumulative(parallel = TRUE,
##     reverse = FALSE), data = dat)
##
##
## Pearson residuals:
##                   Min      1Q Median     3Q    Max
## logit(P[Y<=1]) -2.359 -0.6873 0.6896 0.6896 0.9474
## logit(P[Y<=2]) -1.483 -1.0803 0.3546 0.4672 1.8868
##
## 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 ***
## sex1           0.61834    0.06081  10.168  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  2
```

```
##
## Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])
##
## Residual deviance: 7810.012 on 8477 degrees of freedom
##
## Log-likelihood: -3905.006 on 8477 degrees of freedom
##
## Number of iterations: 3
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##     sex1
## 1.855841
```

```
mod2.3 = vglm(outcome ~ age + sex,
             cumulative(parallel=TRUE, reverse=FALSE), data=dat)
summary(mod2.3)
```

```
##
## Call:
## vglm(formula = outcome ~ age + sex, family = cumulative(parallel = TRUE,
##     reverse = FALSE), data = dat)
##
##
## Pearson residuals:
##                    Min      1Q Median     3Q   Max
## logit(P[Y<=1]) -4.294 -0.4848 0.3992 0.6809 2.295
## logit(P[Y<=2]) -3.756 -0.7178 0.2618 0.4173 3.056
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1   4.984813   0.205544   24.25   <2e-16 ***
## (Intercept):2   5.545466   0.209178   26.51   <2e-16 ***
## age            -0.104617   0.004084  -25.61   <2e-16 ***
## sex1            0.801887   0.065794   12.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  2
##
## Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])
##
## Residual deviance: 7056.827 on 8476 degrees of freedom
##
## Log-likelihood: -3528.413 on 8476 degrees of freedom
##
## Number of iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
```

```
##
## Exponentiated coefficients:
##       age      sex1
## 0.9006692 2.2297449
```

```
mod2.4 = vglm(outcome ~ age + sex + age*sex,
              cumulative(parallel=TRUE, reverse=FALSE), data=dat)
summary(mod2.4)
```

```
##
## Call:
## vglm(formula = outcome ~ age + sex + age * sex, family = cumulative(parall
el = TRUE,
##     reverse = FALSE), data = dat)
##
##
## Pearson residuals:
##                    Min      1Q Median     3Q   Max
## logit(P[Y<=1]) -4.155 -0.4837 0.4094 0.6834 2.398
## logit(P[Y<=2]) -3.617 -0.7150 0.2630 0.4147 3.177
##
## 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 ***
## sex1           0.315679   0.415731   0.759    0.448
## age:sex1       0.009686   0.008155   1.188    0.235
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  2
##
## Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])
##
## Residual deviance: 7055.417 on 8475 degrees of freedom
##
## Log-likelihood: -3527.708 on 8475 degrees of freedom
##
## Number of iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##       age      sex1  age:sex1
## 0.8957957 1.3711896 1.0097330
```

Hillary Miller
BST 210 HW 6

(a) For the model with age alone, what is the estimated odds ratio and 95% CI for the effect of 10 years of age when comparing outcome 3 vs. outcome 1 and 2 (combined)? Also, what is the estimated odds ratio and 95% CI for the effects of 10 years of age when comparing outcome 2 and 3 (combined) vs. outcome 1?

Due to the proportional odds assumption, these two values are equivalent. The B1 coefficient is the log Odds Ratio for being in Outcome 3 vs. combined outcome 2 and 1 for a 1 unit increase in age (or a k+1 vs. a k value of age). It is also the log Odds Ratio for being in combined outcome 3 and 2 vs. outcome 1 for a k+1 age compared to someone of k age.

We estimate that the odds of having outcome 3 (death from any cause in the follow-up period) to having combined outcome 1 (no death or coronary heart disease in the follow-up period (reference category)) and 2 (coronary heart disease in the follow-up period, but the subject remained alive) for a population given of age = k+10 is 2.70 times the odds ratio for a population of age k, with 95% Confidence Interval [2.498,2.920].

We estimate that the odds of having outcome 2 (coronary heart disease in the follow-up period, but the subject remained alive) and outcome 3 (death from any cause in the follow-up period) combined to having outcome 1 (no death or coronary heart disease in the follow-up period (reference category)) for a population given of age = k+10 is 2.70 times the odds ratio for a population of age k, with 95% Confidence Interval [2.498,2.920].

```
summary(mod2.1)

## 
## Call:
## vglm(formula = outcome ~ age, family = cumulative(parallel = TRUE, 
##     reverse = FALSE), data = dat)
## 
## 
## Pearson residuals:
##                   Min      1Q Median     3Q   Max
## logit(P[Y<=1]) -3.733 -0.4974 0.4423 0.6986 2.036
## logit(P[Y<=2]) -2.979 -0.7561 0.2742 0.4269 2.827
## 
## 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
## 
## Number of linear predictors:  2 
## 
## Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])
```

```
##
## Residual deviance: 7208.382 on 8477 degrees of freedom
##
## Log-likelihood: -3604.191 on 8477 degrees of freedom
##
## Number of iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##       age
## 0.9054247

#odds ratio 3 to combined 1+2 AND odds ratio for combined outcome 2 and 3 vs.
outcome 1
exp(0.099351*10)

## [1] 2.700697

#CI
exp(10*(0.099351-(1.96*0.003977)))

## [1] 2.498176

exp(10*(0.099351+(1.96*0.003977)))

## [1] 2.919637
```

(b) For the model with age alone, is the proportional odds assumption satisfied or rejected? Let's explore this further by these additional looks at the data: First, create a binary outcome variable that equals 1 when you are in category 3 and equals 0 when you are in category 1 or 2. Run a logistic regression model using age to predict this binary outcome. Second, create a new binary outcome variable that equals 1 when you are in category 2 or 3 and equals 0 when you are in category 1. Again, run a logistic regression model using age to predict this new binary outcome. If the proportional odds assumption holds, we would expect that the two beta coefficients for age in these two models would be close to each another. What happens in this example? (Do the CI's for the age beta coefficients overlap or not?) Given this comparison of the beta coefficients, do you believe the proportional odds model assumption holds or not for the ordinal logistic regression model with age alone?

From the initial exploration, we can use an approximate likelihood ratio test between an ordinal model with the proportional odds assumption and a generalized ordinal model without the proportional odds assumption. It does not appear the proportional odds assumption holds (P<<.001). It seems we should reject the null hypothesis and conclude the proportional odds assumption is not an appropriate assumption for this data set. With further exploration and creating new variables, we find the following:

From the new models, holding age constant, the log odds ratio of having outcome 3 (death from any cause in the follow-up period) to having combined outcome 1 (no

death or coronary heart disease in the follow-up period (reference category)) and 2 (coronary heart disease in the follow-up period, but the subject remained alive) is .112 , with 95% Confidence Interval [0.103,0.121]. Alternatively, the log odds ratio of having combined outcome 3 (death from any cause in the follow-up period) and outcome 2 (coronary heart disease in the follow-up period, but the subject remained alive) to outcome 1 (no death or coronary heart disease in the follow-up period (reference category)) is .091, with 95% Confidence Interval [0.083,0.099].

Given this comparison of the beta coefficient, it does not appear the proportional odds model assumption holds for the ordinal logistic regression model with age alone.

```
#testing proportionality of odds assumptions
fit.po = vglm(outcome ~ age,
              cumulative(parallel=TRUE, reverse=T), data=dat)
fit.npo = vglm(outcome ~ age,
               cumulative(parallel=FALSE, reverse=T), data=dat)
pchisq(deviance(fit.po)-deviance(fit.npo),
       df=df.residual(fit.po)-df.residual(fit.npo),lower.tail=F)

## [1] 1.025796e-10

##First, create a binary outcome variable that equals 1 when you are in categ
ory 3 and equals 0 when you are in category 1 or 2.
dat <- mutate(dat, outcome.3 = ifelse(outcome==3,1,0))

##Run a logistic regression model using age to predict this binary outcome.
mod2.b1 = glm(outcome.3 ~ age, family="binomial"(link="logit"), data=dat)
summary(mod2.b1)

##
## Call:
## glm(formula = outcome.3 ~ age, family = binomial(link = "logit"),
##     data = dat)
##
## 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

#Second, create a new binary outcome variable that equals 1 when you are in c
ategory 2 or 3 and equals 0 when you are in category 1.
dat <- mutate(dat, outcome.2.3 = ifelse(outcome==3,1,
                                        ifelse(outcome==2,1,0)))

##Again, run a logistic regression model using age to predict this new binary
outcome.
mod2.b2 <- glm(outcome.2.3 ~ age, family="binomial"(link="logit"), data=dat)
summary(mod2.b2)

##
## Call:
## glm(formula = outcome.2.3 ~ age, family = binomial(link = "logit"),
##     data = dat)
##
## 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

##If the proportional odds assumption holds, we would expect that the two bet
a coefficients for age in these two models would be close to each another. Wh
at happens in this example?  (Do the CI’s for the age beta coefficients overl
ap or not?)  Given this comparison of the beta coefficients, do you believe t
he proportional odds model assumption holds or not for the ordinal logistic r
egression model with age alone?


##comparing the log odds ratios
coef(mod2.b1)[2]
```

```
##       age
## 0.1118963

coef(mod2.b2)[2]

##        age
## 0.09120769

##CI of binary with 3 against 2 and 1
((coef(mod2.b1)[2])-(1.96*coef(summary(mod2.b1))[2,2]))

##       age
## 0.1030298

((coef(mod2.b1)[2])+(1.96*coef(summary(mod2.b1))[2,2]))

##       age
## 0.1207629

##CI of binary with 3 and 2 against 1
((coef(mod2.b2)[2])-(1.96*coef(summary(mod2.b2))[2,2]))

##        age
## 0.08315125

((coef(mod2.b2)[2])+(1.96*coef(summary(mod2.b2))[2,2]))

##        age
## 0.09926412
```

(c) Now focus on the model with sex alone. Is the proportional odds assumption satisfied or rejected? Let's explore this further by these additional looks at the data: Consider again the outcome × sex tabulation as above. With the ordinal model, we need to tabulate category 1 vs. 2 and 3 (combined) and category 1 and 2 (combined) vs. category 3. If the proportional odds assumption is satisfied, we should feel comfortable with a common odds ratio estimate for these two categorizations. What are the associated odds ratio estimates for sex to predict these categorizations based on hand calculations? How do these compare with the ordinal logistic regression-based odds ratio estimate for sex? Given your comparison of these odds ratio estimates, do you believe the proportional odds model assumption holds or not for the ordinal logistic regression model with sex alone? (One could also perform the pair of logistic regressions as in 2 (b) with sex as the only predictor and compare the beta coefficients for sex in these two models. Try that if you like.) Finally, is this ordinal logistic regression model saturated or not? Defend your answer.

Based on hand calculations, below, it does not appear the hand calculations of the odds ratios match exactly, but they are very close. The hand calculated OR for outcome 1 vs. 2 and 3 is .529, while the OR for outcome 1 and 2 vs. 3 is .554. Comparing to the ordinal logistic regression-based odds ratio of 0.539, which is between the between the two hand calculations.

Hillary Miller
BST 210 HW 6

From further exploration, we can use an approximate likelihood ratio test between an ordinal model with the proportional odds assumption and a generalized ordinal model without the proportional odds assumption. It does appear the proportional odds assumption holds (P=.275). It seems we should fail to reject the null hypothesis and conclude the proportional odds assumption is an appropriate assumption for this data set. With further exploration and creating new variables, we find the following:

From the new models, holding gender constant, the odds ratio of having outcome 3 (death from any cause in the follow-up period) to having combined outcome 1 (no death or coronary heart disease in the follow-up period (reference category)) and 2 (coronary heart disease in the follow-up period, but the subject remained alive) is 0.554, with 95% Confidence Interval [.487,.631]. Alternatively, the odds ratio of having combined outcome 3 (death from any cause in the follow-up period) and outcome 2 (coronary heart disease in the follow-up period, but the subject remained alive) to outcome 1 (no death or coronary heart disease in the follow-up period (reference category)) is .529, with 95% Confidence Interval [.468,.599]. As you can see, these estimated Odds Ratios overlap.

Given this comparison of the beta coefficient, it does appear the proportional odds model assumption holds for the ordinal logistic regression model with age alone. This model is not a saturated model. While a binary predictor, one can look at the fitted probabilities from the ordinal model and compare with the observed proportions in the data and see that they are close, but not exactly the same.

```
table(dat$outcome, dat$sex)

## 
##        0    1
##   1  855 1515
##   2  225  239
##   3  740  666

##OR for category 1 vs. 2 and 3
((239+666)/(225+740))/(1515/855)

## [1] 0.5292669

##OR for category 1 and 2 vs. 3
(666/740)/((1515+239)/(855+225))

## [1] 0.5541619

##OR from model
exp(-0.61834)

## [1] 0.5388382

#testing proportionality of odds assumptions
fit.po = vglm(outcome ~ sex,
              cumulative(parallel=TRUE, reverse=T), data=dat)
fit.npo = vglm(outcome ~ sex,
               cumulative(parallel=FALSE, reverse=T), data=dat)
pchisq(deviance(fit.po)-deviance(fit.npo),
       df=df.residual(fit.po)-df.residual(fit.npo),lower.tail=F)

## [1] 0.2747365

table(dat$outcome, dat$sex)

##
##        0    1
##   1  855 1515
##   2  225  239
##   3  740  666

prop.table(table(dat$outcome, dat$sex), margin = 2)

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

# Male fitted probs.
fitted(mod2.2)[dat$sex == 0,][1,]

##         1         2         3
## 0.4726259 0.1175958 0.4097783

# Female fitted probs.
fitted(mod2.2)[dat$sex == 1,][1,]

##         1         2         3
## 0.6245092 0.1032375 0.2722533

#confirming the estimated odds ratios are the same
summary(mod2.2)

##
## Call:
## vglm(formula = outcome ~ sex, family = cumulative(parallel = TRUE,
##     reverse = FALSE), data = dat)
##
##
## Pearson residuals:
##                   Min      1Q Median     3Q    Max
## logit(P[Y<=1]) -2.359 -0.6873 0.6896 0.6896 0.9474
## logit(P[Y<=2]) -1.483 -1.0803 0.3546 0.4672 1.8868
##
## 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 ***
## sex1           0.61834    0.06081  10.168  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Number of linear predictors:  2 
## 
## Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])
## 
## Residual deviance: 7810.012 on 8477 degrees of freedom
## 
## Log-likelihood: -3905.006 on 8477 degrees of freedom
## 
## Number of iterations: 3 
## 
## No Hauck-Donner effect found in any of the estimates
## 
## Exponentiated coefficients:
##     sex1 
## 1.855841
```

```
##Again, run a logistic regression model using age to compare the new binary outcomes.
mod2.c1 <- glm(outcome.3 ~ sex, family="binomial"(link="logit"), data=dat)
summary(mod2.c1)
```

```
## 
## Call:
## glm(formula = outcome.3 ~ sex, family = binomial(link = "logit"), 
##     data = dat)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.0216  -1.0216  -0.8023   1.3416   1.6064  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -0.37807    0.04772  -7.922 2.33e-15 ***
## sex1        -0.59030    0.06595  -8.951  < 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: 5306.9  on 4238  degrees of freedom
## AIC: 5310.9
## 
## Number of Fisher Scoring iterations: 4
```

```
mod2.c2 <- glm(outcome.2.3 ~ sex, family="binomial"(link="logit"), data=dat)
summary(mod2.c2)
```

```
##
## Call:
## glm(formula = outcome.2.3 ~ sex, family = binomial(link = "logit"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2292  -0.9678  -0.9678   1.1265   1.4026
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.12103    0.04697   2.577  0.00997 **
## sex1        -0.63626    0.06302 -10.097  < 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: 5715.8  on 4238  degrees of freedom
## AIC: 5719.8
##
## Number of Fisher Scoring iterations: 4
```

```
##comparing the odds ratios
exp(coef(mod2.c1)[2])
```

```
##      sex1
## 0.5541619
```

```
exp(coef(mod2.c2)[2])
```

```
##      sex1
## 0.5292669
```

```
##CI of binary with 3 against 2 and 1
exp(((coef(mod2.c1)[2])-(1.96*coef(summary(mod2.c1))[2,2])))
```

```
##      sex1
## 0.4869699
```

```
exp(((coef(mod2.c1)[2])+(1.96*coef(summary(mod2.c1))[2,2])))
```

```
##     sex1
## 0.630625
```

```
##CI of binary with 3 and 2 against 1
exp(((coef(mod2.c2)[2])-(1.96*coef(summary(mod2.c2))[2,2])))
```

```
##      sex1
## 0.4677731

exp(((coef(mod2.c2)[2])+(1.96*coef(summary(mod2.c2))[2,2])))

##      sex1
## 0.5988448
```

(d) Do we have any evidence that the age × sex interaction is needed for ordinal logistic regression modeling? Why or why not?

No, this interaction is not needed. Based on the approximate likelihood ratio test (P=.235) and the Wald test (P=.235) comparing the null, Ho (reduced model with age + sex) and the alternative, Ha (full mdoel with age + sex + ageXsex), we fail to reject the null hypothesis test that the reduced model is preferred.

```
#testing proportionality of odds assumptions
pchisq(deviance(mod2.3)-deviance(mod2.4), df=1, lower.tail=F)

## [1] 0.235037

summary(mod2.4)

##
## Call:
## vglm(formula = outcome ~ age + sex + age * sex, family = cumulative(parall
el = TRUE,
##     reverse = FALSE), data = dat)
##
##
## Pearson residuals:
##                    Min      1Q Median     3Q   Max
## logit(P[Y<=1]) -4.155 -0.4837 0.4094 0.6834 2.398
## logit(P[Y<=2]) -3.617 -0.7150 0.2630 0.4147 3.177
##
## 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 ***
## sex1           0.315679   0.415731   0.759    0.448
## age:sex1       0.009686   0.008155   1.188    0.235
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  2
##
## Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])
##
## Residual deviance: 7055.417 on 8475 degrees of freedom
##
```

```
## Log-likelihood: -3527.708 on 8475 degrees of freedom
##
## Number of iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##       age      sex1   age:sex1
## 0.8957957 1.3711896 1.0097330
```

(e) Finally, assess whether or not the proportional odds assumption holds for the model including both main effects of age and sex (but not their interaction). Based on the results of this analysis, what would be your recommendations for model choices if you wanted to include continuous age in the modeling? Would you recommend using ordinal or multinomial logistic regression? Is there anything else you might recommend?

From the approximate LRT, it does not appear the proportional odds assumption holds. Therefore, we should not use the proportional odds ordinal model. We could do a multinomial regression model or we could do a generalized ordinal model. The generalized ordinal model would make the most sense, since the outcomes truly represent an ordinal response.

I did a further exploration using logistic regression from the variables created in part (b) and compared the beta coefficients and the confidence intervals to confirm my conclusions from the approximate LRT. Looking at the age coefficient for the model with outcome 3 vs. 2 and 1 and then outcome 2 and 3 vs. 1, Odds Ratios and 95% Confidence Intervals are 1.12 [1.11,1.34] and 1.10 [1.09,1.109], respectively. These do not overlap. Because of this, I would recommend using the multinomial logistic regression model rather than the ordinal model.

One additional possibility would be to look more closely at the age variable and see if cut points may exist within the data that could be used to create categorical age data. However, if we want to keep continuous age in the model, this is not ideal.

```
#testing proportionality of odds assumptions
fit.po = vglm(outcome ~ age + sex,
              cumulative(parallel=TRUE, reverse=T), data=dat)
fit.npo = vglm(outcome ~ age + sex,
               cumulative(parallel=FALSE, reverse=T), data=dat)
pchisq(deviance(fit.po)-deviance(fit.npo),
       df=df.residual(fit.po)-df.residual(fit.npo),lower.tail=F)

## [1] 1.423149e-09

##Again, run a logistic regression model using age to compare the new binary
outcomes.
mod2.e1 <- glm(outcome.3 ~ sex + age, family="binomial"(link="logit"), data=d
at)
summary(mod2.e1)
```

```
##
## Call:
## glm(formula = outcome.3 ~ sex + age, family = binomial(link = "logit"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8465  -0.8203  -0.5320   0.9470   2.4767
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.206782   0.240004  -25.86   <2e-16 ***
## sex1        -0.784918   0.073881  -10.62   <2e-16 ***
## age          0.116833   0.004654   25.10   <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: 4537.3  on 4237  degrees of freedom
## AIC: 4543.3
##
## Number of Fisher Scoring iterations: 4
```

```
mod2.e2 <- glm(outcome.2.3 ~ sex + age, family="binomial"(link="logit"), data
=dat)
summary(mod2.e2)
```

```
##
## Call:
## glm(formula = outcome.2.3 ~ sex + age, family = binomial(link = "logit"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9131  -0.9553  -0.6266   1.0472   2.1063
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.570532   0.211564  -21.60   <2e-16 ***
## sex1        -0.788923   0.068650  -11.49   <2e-16 ***
## age          0.095779   0.004226   22.66   <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: 5121.1  on 4237  degrees of freedom
## AIC: 5127.1
##
## Number of Fisher Scoring iterations: 4
```

```
##comparing the odds ratios
exp(coef(mod2.e1)[3])
```

```
##      age
## 1.123932
```

```
exp(coef(mod2.e2)[3])
```

```
##      age
## 1.100516
```

```
##CI of binary with 3 against 2 and 1
exp(((coef(mod2.e1)[3])-(1.96*coef(summary(mod2.e1))[3,2])))
```

```
##      age
## 1.113726
```

```
exp(((coef(mod2.e1)[3])+(1.96*coef(summary(mod2.e1))[3,2])))
```

```
##      age
## 1.134231
```

```
##CI of binary with 3 and 2 against 1
exp(((coef(mod2.e2)[3])-(1.96*coef(summary(mod2.e2))[3,2])))
```

```
##      age
## 1.091437
```

```
exp(((coef(mod2.e2)[3])+(1.96*coef(summary(mod2.e2))[3,2])))
```

```
##     age
## 1.10967
```