Homework 4, Due October 27 at 11:45am

Zach White

Provide clearly documented code in an R markdown document where appropriate. Whenever possible, provide statistical evidence to support your answers.

1. Heart Failure Severity

Using the Duke Databank for Cardiovascular Diseases data in cathhw4.Rdata, consider the outcome of heart failure severity (0=no heart failure, 4=most severe) as a function of age group, gender, and smoking history. Note the variable definitions:

CHFSEV (severity of congestive heart failure, 0=None, 1-4 correspond to New York Heart Association classifications I (cardiac disease does not limit normal activity though symptoms may present with exertion), II (slight limitation of ordinary activity though could have symptoms with light activity such as walking more than 2 blocks or climbing more than one flight of stairs), III (patient comfortable at rest but marked limitations to physical activity), and IV (patient has symptoms even at rest that increase with any physical activity))

GENDER (1=female, 0=male)

AGE_G (categorized as 1=18-24; 2=25-29; 3=30-34; 4=35-39; 5=40-44; 6=45-49; 7=50-54; 8=55-59; 9=60-64; 10=65-69; 11=70-74; 12=75-79; 13=>=80)

HXSMOKE (1=history of smoking, 0=no history of smoking)

1a. Descriptive Statistics

Explore relationships among the four variables using graphical displays and describe your findings.

Solution

```
load("cathhw4.Rdata")
cath = fac.cath = num.cath = cathhw4
cath$cath1.CHFSEV = fac.cath$cath1.CHFSEV = num.cath$cath1.CHFSEV = factor(cath$cath1.CHFSEV)
cath$cath1.GENDER = fac.cath$cath1.GENDER = num.cath$cath1.GENDER = factor(cath$cath1.GENDER)
cath$cath1.HXSMOKE = fac.cath$cath1.HXSMOKE = num.cath$cath1.HXSMOKE = factor(cath$cath1.HXSMOKE)
sum(is.na(cath))
```

```
## [1] 0
```

```
# Test this out a little bit more
fac.cath$cath1.AGE_G = factor(cath$cath1.AGE_G)

# The visualizations for two factors.

# Numeric Age, Factor CHFSEV
age.mid.vec = c(21,27,32,37,42,47,52,57,62,67,72,77,85)
for(i in 1:13){
  num.cath$cath1.AGE_G[num.cath$cath1.AGE_G == i]  = age.mid.vec[i]
}
mid.age = quantile(age.mid.vec,.5)
num.cath$centered.age = num.cath$cath1.AGE_G - mid.age
```

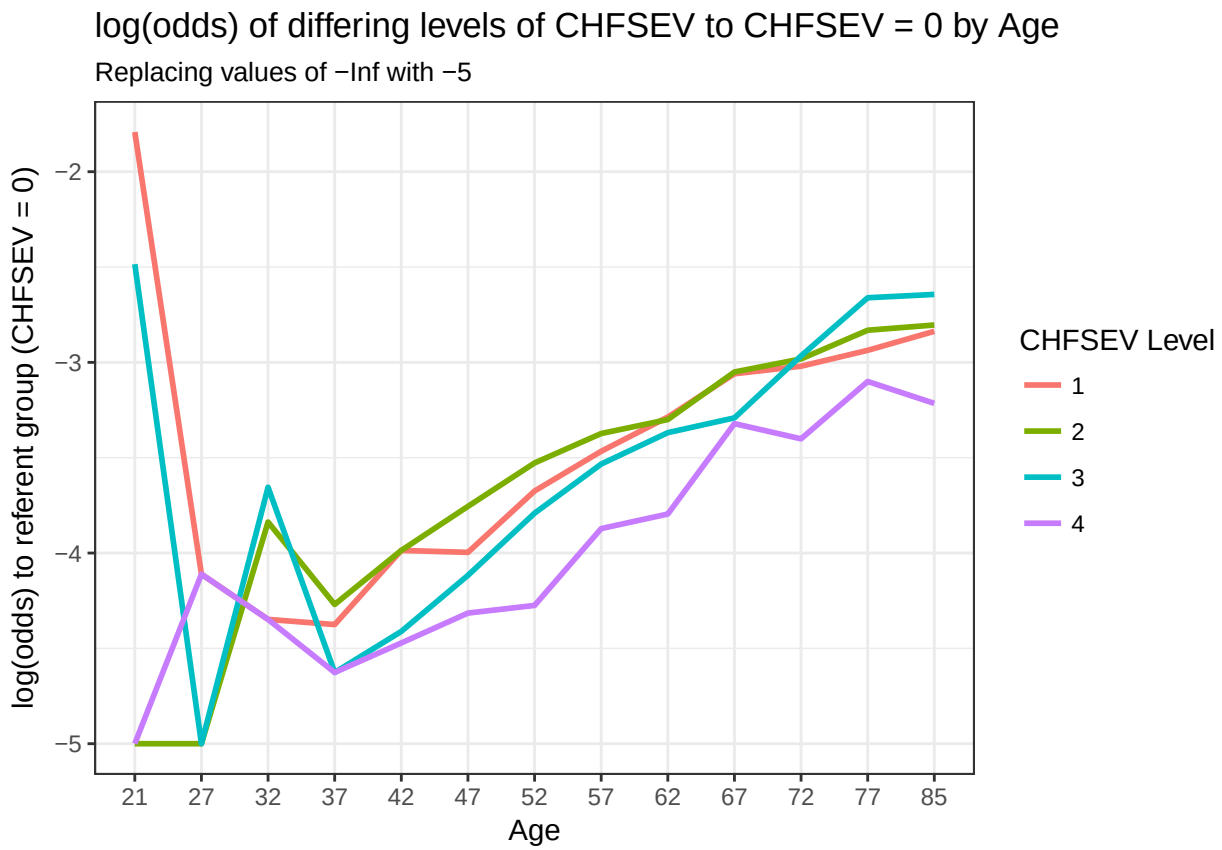
```r
# What about the scaled age?
with(num.cath,do.call(rbind,
                      tapply(cath1.AGE_G,cath1.CHFSEV, function(x) c(M = mean(x), SD = sd(x)))))
```

```
##          M       SD
## 0 61.53113 12.03533
## 1 65.57279 11.11957
## 2 65.39568 11.14692
## 3 66.93508 11.21947
## 4 66.22647 10.97068
```

```r
num.cath$age10 = num.cath$cath1.AGE_G / 10

prop.tab.chfsev.age = log.odds.tab = table.chfsev.age = as.matrix(table(num.cath$cath1.CHFSEV, num.cath$
colnames(table.chfsev.age) = age.mid.vec
for(i in 1:5){
  prop.tab.chfsev.age[i,] = table.chfsev.age[i,] / colSums(table.chfsev.age)
  if(i >= 2){
    log.odds.tab[i,] = log(prop.tab.chfsev.age[i,] / prop.tab.chfsev.age[1,])
  }
}
log.odds.tab = log.odds.tab[-1,]
log.odds.df = as.data.frame(log.odds.tab)
log.odds.df[log.odds.df == -Inf] = -5

table(num.cath$cath1.CHFSEV,num.cath$cath1.AGE_G)
```

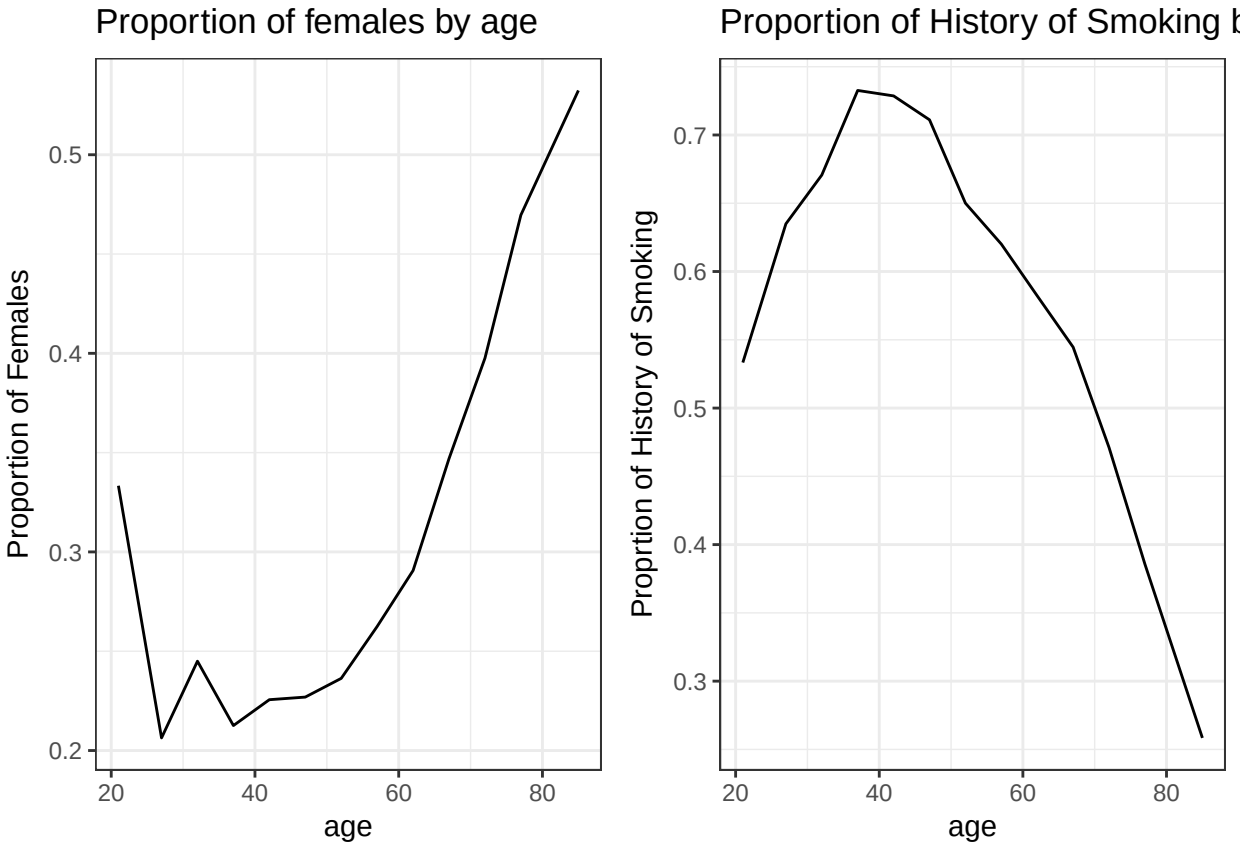
```
##
##       21   27   32   37   42   47   52   57   62   67   72   77   85
##   0   12   61  232  715 1400 2394 3234 3938 4094 4542 3689 2376 1519
##   1    2    1    3    9   26   44   82  123  153  213  180  126   89
##   2    0    0    5   10   26   56   95  135  151  215  187  140   92
##   3    1    0    6    7   17   39   73  115  141  169  190  166  108
##   4    0    1    3    7   16   32   45   82   92  164  123  107   61
```

```r
ggplot(log.odds.df, aes(x = Var2, y = Freq, group = Var1, color = Var1)) + geom_line(size=1) +
  theme_bw() +
  labs(x = "Age",
       y = "log(odds) to referent group (CHFSEV = 0)",
       color = "CHFSEV Level",
       title = "log(odds) of differing levels of CHFSEV to CHFSEV = 0 by Age",
       subtitle = "Replacing values of -Inf with -5")
```



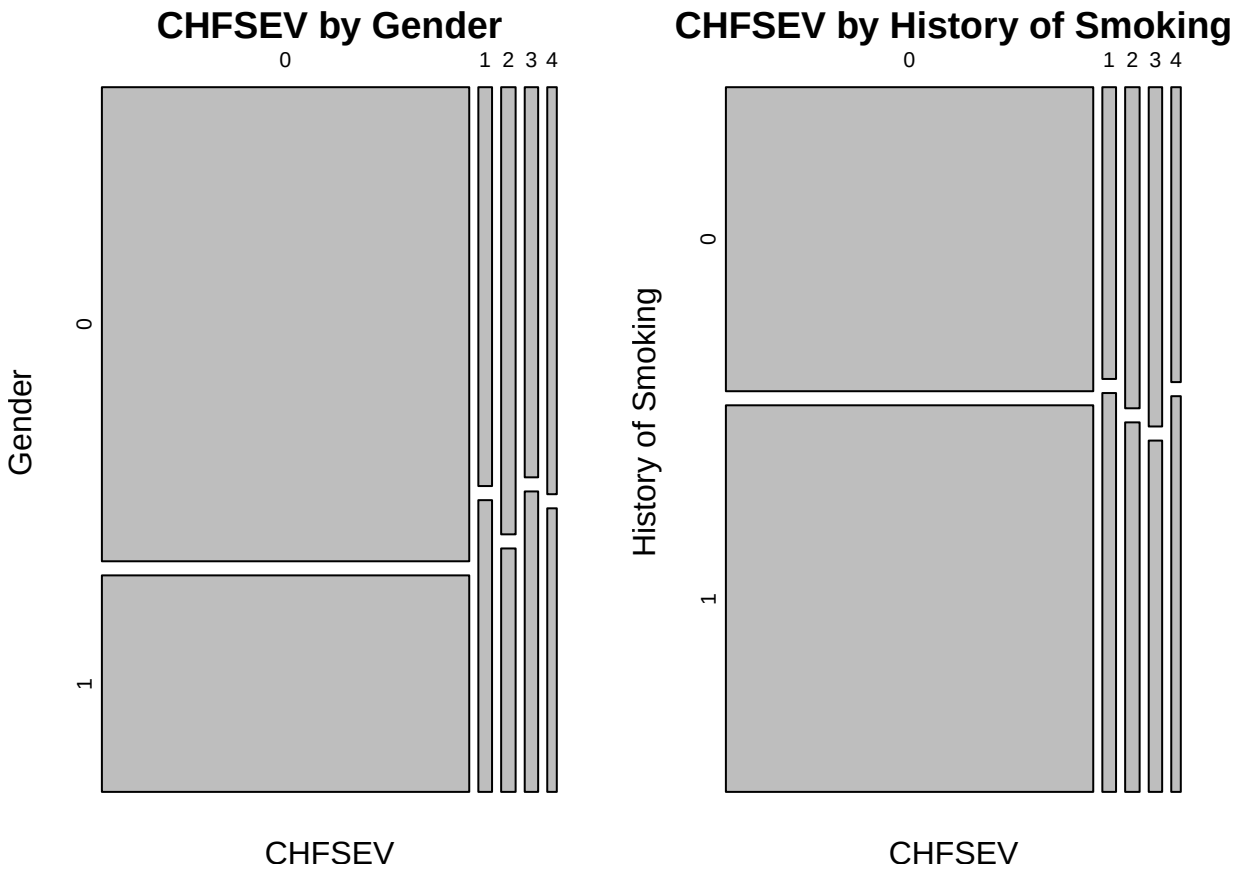
For this plot, it seems like the log odds to the referent group is pretty clearly linearly related, except for some sparse categories in the younger categories. This seems like a good sign that we will be able to model age as a simple linear term.

```
age.gender.hx.df = data_frame()
for(i in seq_along(age.mid.vec)){
  df = data_frame(
    age = age.mid.vec[i],
    gender.age = mean(num.cath$cath1.GENDER[num.cath$cath1.AGE_G == age.mid.vec[i]] == 1),
    hx.age = mean(num.cath$cath1.HXSMOKE[num.cath$cath1.AGE_G == age.mid.vec[[i]]] == 1)
  )
  age.gender.hx.df = bind_rows(age.gender.hx.df,df)
}
gender.by.age = ggplot(age.gender.hx.df,aes(x = age, y = gender.age)) + geom_line() +
  theme_bw() +
  labs(title = "Proportion of females by age",
       y = "Proportion of Females")

smoking.by.age = ggplot(age.gender.hx.df,aes(x = age, y = hx.age)) + geom_line() +
  theme_bw() +
  labs(title = "Proportion of History of Smoking by Age",
       y = "Proprtion of History of Smoking")

grid.arrange(gender.by.age,smoking.by.age, ncol = 2)
```



The plots are meant to explore the relationships between the covariates, which might be useful as far as the interactions are concerned, but not necessarily considering that interactions are more about how the covariates influence each other with respect to the response. However, given that we are supposed to explore the relationships between all four variables, we explore these plots.

```
par(mfrow = c(1,2), mar = c(1.8,1.8,1.3,1.3))
mosaicplot(table(num.cath$cath1.CHFSEV,num.cath$cath1.GENDER), main = "CHFSEV by Gender",
           ylab = "Gender",
           xlab = "CHFSEV")
mosaicplot(table(num.cath$cath1.CHFSEV, num.cath$cath1.HXSMOKE), main = "CHFSEV by History of Smoking",
           ylab = "History of Smoking",
           xlab = "CHFSEV")
```



It is clear that looking at these mosaic plots shows that for every level of the covariates, the most common CHFSEV level is 0, which is to be expected. Also, it is important to note that it seems like for females the lines for the different levels of CHFSEV are higher, which seems to indicate that the proportion of females in those higher levels is higher than males in the baseline group of CHFSEV = 0. This story is less clear when analyzing smoking.

1b. Baseline Category Logistic Model

Fit a baseline category logistic model to the heart failure severity data, with the baseline category defined as CHFSEV=0, and main effects of the predictors age group, gender, and smoking history. Use the functional form of age group that provides the best fit to the data and assess whether any interaction terms are needed. Provide a clearly-labeled table that gives the OR's and 95% CI's relating the predictors to heart failure severity and interpret all estimates that are statistically significant at the $\alpha = 0.05$ level.

Solution

```
fac.cath$cath1.AGE_G = relevel(fac.cath$cath1.AGE_G, ref = "13")
fac.mod = multinom(cath1.CHFSEV ~ cath1.AGE_G + cath1.HXSMOKE + cath1.GENDER, data = fac.cath)
summary(fac.mod)$deviance
num.mod = multinom(cath1.CHFSEV ~ centered.age + cath1.HXSMOKE + cath1.GENDER, data = num.cath)
summary(num.mod)$deviance
poly.mod = multinom(cath1.CHFSEV ~ poly(centered.age,2) + cath1.HXSMOKE + cath1.GENDER, data = num.cath
summary(poly.mod)$deviance

# What about interactions? If we add interactions, then no way will we do the factor ones.
```

```
# Change referent group
fac.mod.int = multinom(cath1.CHFSEV ~ . + .:., data = fac.cath)
sum.fac.int = summary(fac.mod.int)
#dim(sum.fac.int$coefficients)[2]
z.fac.int = sum.fac.int$coefficients / sum.fac.int$standard.errors
p.fac.int = (1 - pnorm(abs(z.fac.int),0,1))*2
## If we do this one, eliminate the interaction between cath1.AGE_G:cath1.HXSMOKE


poly.mod.int = multinom(cath1.CHFSEV ~ poly(centered.age,2) + cath1.HXSMOKE + cath1.GENDER + poly(cente
                        poly(centered.age,2):cath1.GENDER + cath1.HXSMOKE:cath1.GENDER, data = num.cat
sum.poly.int = summary(poly.mod.int)
#sum.poly.int$deviance
#dim(sum.poly.int$coefficients)[2]

#sum.fac.int$deviance - sum.poly.int$deviance
```

```
anova(fac.mod, num.mod, poly.mod, test = "Chisq")
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: cath1.CHFSEV
##                                                  Model Resid. df
## 1          centered.age + cath1.HXSMOKE + cath1.GENDER    128520
## 2 poly(centered.age, 2) + cath1.HXSMOKE + cath1.GENDER    128516
## 3           cath1.AGE_G + cath1.HXSMOKE + cath1.GENDER    128476
##   Resid. Dev   Test    Df LR stat.    Pr(Chi)
## 1   34054.49
## 2   34044.31 1 vs 2     4 10.17768 0.03753879
## 3   33996.93 2 vs 3    40 47.38273 0.19682212
```

Analyzing these values, it seems like centered continuous centered age, continuous centered age and polynomial term for the centered age, and factor for age are all comparable. There doesn't seem to be a clear improvement between these different models, and thus for interpretability sake, we will treat age as continuous and then center them so the coefficients for our model are reasonable.

```
# Centered age with all the interactions.
num.mod.int = multinom(cath1.CHFSEV ~ centered.age + cath1.HXSMOKE + cath1.GENDER + centered.age:cath1.
                        centered.age:cath1.GENDER + cath1.HXSMOKE:cath1.GENDER, data = num.cath)
sum.num.int = summary(num.mod.int)
#dim(sum.num.int$coefficients)[2]
z.num.int = sum.num.int$coefficients / sum.num.int$standard.errors
p.num.int = (1 - pnorm(abs(z.num.int),0,1))* 2

# Leave out centered age:Gender
wo.gend.age.int = multinom(cath1.CHFSEV ~ centered.age + cath1.HXSMOKE + cath1.GENDER + centered.age:ca
                       cath1.HXSMOKE:cath1.GENDER, data = num.cath)
sum.wo.gend.age.int = summary(wo.gend.age.int)
```

```
p.num.int
```

```
##   (Intercept) centered.age cath1.HXSMOKE1 cath1.GENDER1
## 1           0 3.320566e-06   0.0004521822  7.410541e-10
## 2           0 5.127258e-07   0.3343031120  5.221571e-01
## 3           0 1.258086e-08   0.5931684324  4.437990e-05
## 4           0 4.483117e-06   0.2743378191  5.758403e-04
```

```
##   centered.age:cath1.HXSMOKE1 centered.age:cath1.GENDER1
## 1                   0.088512957                  0.1329412
## 2                   0.603724719                  0.4778979
## 3                   0.004960231                  0.4629554
## 4                   0.027226370                  0.1960144
##   cath1.HXSMOKE1:cath1.GENDER1
## 1                   6.847362e-05
## 2                   1.837702e-01
## 3                   1.068465e-01
## 4                   3.719538e-01
```

```
anova(num.mod.int,wo.gend.age.int, test = "Chisq")
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: cath1.CHFSEV
##
## 1                              centered.age + cath1.HXSMOKE + cath1.GENDER + centered.age:cath1.HXSMOI
## 2 centered.age + cath1.HXSMOKE + cath1.GENDER + centered.age:cath1.HXSMOKE + centered.age:cath1.GENDI
##   Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
## 1    128512   34019.67
## 2    128508   34014.78 1 vs 2     4 4.888199 0.2989625
```

We first test all of the interactions. I test the significance of the coefficient to see if there are any that are universally insignificant. Doing so, it seems clear that leaving out the interaction between age and gender does not harm our model significantly.

```
z.wo.gend.age.int = summary(wo.gend.age.int)$coefficients / summary(wo.gend.age.int)$standard.errors
p.wo = (1 - pnorm(abs(z.wo.gend.age.int),0,1))*2
```

Although it sems like the history of smoking is not significant for most referent groups, when comparing CHFSEV = 1 to CHFSEV = 0, then it is clearly significant. We will keep it in our model for this reason. Also, the interaction between history of smoking and gender describes a similar story.

```
OR.mat = exp(coef(wo.gend.age.int))
conf.array = exp(confint(wo.gend.age.int))

cvfsev1 = cbind(OR.mat[1,],conf.array[,,1])
cvfsev2 = cbind(OR.mat[2,],conf.array[,,2])
cvfsev3 = cbind(OR.mat[3,],conf.array[,,3])
cvfsev4 = cbind(OR.mat[4,],conf.array[,,4])

rownames(cvfsev1) = rownames(cvfsev2) = rownames(cvfsev3) = rownames(cvfsev4) = c("Referent Group","Cen
colnames(cvfsev1) = colnames(cvfsev2) = colnames(cvfsev3) = colnames(cvfsev4) = c("OR","Lower","Upper")

kable(cvfsev1, digits = 4, format = "latex", caption = "Coefficients comparing cath1.CHFSEV = 1 to cathi

kable(cvfsev2, digits = 4, format = "latex", caption = "Coefficients comparing cath1.CHFSEV = 1 to cathi

kable(cvfsev3, digits = 4, format = "latex", caption = "Coefficients comparing cath1.CHFSEV = 1 to cathi

kable(cvfsev4, digits = 4, format = "latex", caption = "Coefficients comparing cath1.CHFSEV = 1 to cathi
```

It is important to note that for all of these interpretations, the referent group is a a male between 50-54 with no history of smoking.

Holding all else constant an increase of one year in age is associated with an odds ratio of $e^{\beta_{\text{c.age}}+\beta_{\text{Smoking:Age}}} = e^{0.021+.0131I(\text{H.Smoking})}$ (1.021 if not a smoker 1.03 if a smoker) of being in the CHFSEV = 1 group instead of

Table 1: Coefficients comparing cath1.CHFSEV = 1 to cath1.CHFSEV = 0

	OR	Lower	Upper
Referent Group	0.0182	0.0150	0.0220
Centered Age	1.0213	1.0125	1.0302
Hist. Smoke	1.4672	1.1680	1.8430
Gender	2.0381	1.6691	2.4886
Age:Smoking	1.0133	1.0019	1.0248
Smoking:Gender	0.5949	0.4540	0.7795

Table 2: Coefficients comparing cath1.CHFSEV = 1 to cath1.CHFSEV = 0

	OR	Lower	Upper
Referent Group	0.0262	0.0221	0.0309
Centered Age	1.0271	1.0188	1.0354
Hist. Smoke	1.1279	0.9184	1.3852
Gender	1.1763	0.9818	1.4093
Age:Smoking	1.0020	0.9911	1.0130
Smoking:Gender	0.8134	0.6237	1.0608

Table 3: Coefficients comparing cath1.CHFSEV = 1 to cath1.CHFSEV = 0

	OR	Lower	Upper
Referent Group	0.0211	0.0177	0.0252
Centered Age	1.0283	1.0199	1.0368
Hist. Smoke	0.9210	0.7325	1.1582
Gender	1.6311	1.3565	1.9611
Age:Smoking	1.0183	1.0066	1.0300
Smoking:Gender	0.8152	0.6251	1.0632

Table 4: Coefficients comparing cath1.CHFSEV = 1 to cath1.CHFSEV = 0

	OR	Lower	Upper
Referent Group	0.0143	0.0115	0.0178
Centered Age	1.0251	1.0146	1.0356
Hist. Smoke	1.1291	0.8621	1.4787
Gender	1.5196	1.2056	1.9154
Age:Smoking	1.0179	1.0041	1.0318
Smoking:Gender	0.8957	0.6544	1.2260

0. Thus the change associated with age also depends on the status of history of smoking.

Holding all else constant, a smoker is associated with an odds-ratio of $e^{\beta_{\text{Smoking}}+\beta_{\text{Smoking:Age}}x_{\text{c.age}}+\beta_{\text{Smoking:Gender}}I(\text{GEnder = female})} = e^{.38+0.131x_{\text{c.age}}-.51I(\text{Gender = Female})}$. Thus the change associated with smoking depends on the age and gender of the individual.

Holding all else constant, females are associated with an odds-ratio of $e^{\beta_{\text{Gender}}+\beta_{\text{Gender:Smoking}}I(\text{Smoker})} = e^{.71+-0.519I(\text{Smoker})} = 1.21$ if a smoker and 2.03 if not a smoker. The change associated with smoking depends on the history of smoking of the individual.

Note these interpretations are for comparing the $\text{logit}(\frac{\pi_{\text{CHFSEV = 1}}}{\pi_{\text{CHFSEV = 0}}})$. The interpretations for all the other comparisons of the groups with the baseline will be similar, albeit different estimates and confidence intervals, For this reason, I do not interpret the other ones

1c. Prediction in Baseline Category Logistic Model

Using your preferred linear predictor from 1b, find and report in a table the predicted probabilities of each level of heart failure severity from the baseline category logistic model for (i) a 25 year-old woman who has never smoked, (ii) a 25 year-old woman who smokes, (iii) a 25 year-old man who has never smoked, (iv) a 25 year-old man who smokes, (v) an 80 year-old woman who has never smoked, (vi) an 80 year-old woman who smokes, (vii) an 80 year-old man who has never smoked, (viii) an 80 year-old man who smokes.

Solution

```
pred_df = data_frame(
  cath1.GENDER = factor(c(1,1,0,0,1,1,0,0)),
  cath1.HXSMOKE = factor(c(0,1,0,1,0,1,0,1)),
  centered.age = c(27,27,27,27,85,85,85,85) - mid.age
)

predict(wo.gend.age.int,pred_df, "probs")
```

```
##           0          1          2           3           4
## 1 0.9376773 0.02047127 0.01480115 0.016067353 0.010982927
## 2 0.9585246 0.01313300 0.01320597 0.007845229 0.007291170
## 3 0.9593755 0.01027690 0.01287408 0.010078857 0.007394644
## 4 0.9639079 0.01089254 0.01387974 0.005933248 0.005386618
## 5 0.7786745 0.05779493 0.05783496 0.067378203 0.038317391
## 6 0.7245047 0.07254772 0.05272814 0.085520335 0.064699139
## 7 0.8438865 0.03073265 0.05328489 0.044769189 0.027326750
## 8 0.7615964 0.06289849 0.05793022 0.067609501 0.049965355
```

These predictions make sense considering the data.

1d. Proportional Odds Model

Evaluate whether the proportional odds assumption is satisfied for any of the predictors, using the same functional forms as in 1b. Provide test results to support your conclusions.

Solution

```
mpo = vglm(ordered(num.cath$cath1.CHFSEV) ~ centered.age + cath1.HXSMOKE + cath1.GENDER + centered.age:
                        cath1.HXSMOKE:cath1.GENDER, data = num.cath, family = cumulative(parallel = TRU
sum.wo.gend.age.int = summary(mpo)
```

```
just.age.par = vglm(ordered(num.cath$cath1.CHFSEV) ~ centered.age + cath1.HXSMOKE +
                      cath1.GENDER + centered.age:cath1.HXSMOKE +
                      cath1.HXSMOKE:cath1.GENDER, data = num.cath, family =
                      cumulative(parallel = FALSE~cath1.HXSMOKE + cath1.GENDER))
sum.just.age.par = summary(just.age.par)

age.gender.par =vglm(ordered(num.cath$cath1.CHFSEV) ~ centered.age + cath1.HXSMOKE +
                      cath1.GENDER + centered.age:cath1.HXSMOKE +
                      cath1.HXSMOKE:cath1.GENDER, data = num.cath, family =
                      cumulative(parallel = FALSE~cath1.HXSMOKE))
sum.age.gender.par = summary(age.gender.par)
```

1e. Interpretation of Estimates in Proportional Odds Model

Regardless of your conclusions in part 1c, use the same functional forms as in 1b and provide a clearly-labeled table that gives the OR's and 95% CI's relating the predictors to heart failure severity in the (fully) proportional odds model. Interpret all estimates that are statistically significant at the $\alpha = 0.05$ level.

Solution

```
or.mat = cbind(exp(-coef(mpo)), exp(-confint(mpo))[,c(2,1)])
colnames(or.mat) = c("OR","Lower","Upper")
or.mat
```

```
##                                       OR      Lower      Upper
## (Intercept):1                 0.08026468 0.07291608 0.08835387
## (Intercept):2                 0.05636892 0.05110171 0.06217904
## (Intercept):3                 0.03311533 0.02988767 0.03669156
## (Intercept):4                 0.01321932 0.01176257 0.01485647
## centered.age                  1.02564739 1.02099800 1.03031795
## cath1.HXSMOKE1                1.13718543 1.01001551 1.28036716
## cath1.GENDER1                 1.53478924 1.38525955 1.70045967
## centered.age:cath1.HXSMOKE1   1.01200702 1.00581640 1.01823574
## cath1.HXSMOKE1:cath1.GENDER1 0.78070119 0.67548215 0.90231008
```

These intercepts are large because under the proportional odds model, we are dealing with the logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3]), and logit(P[Y<=4]). Since the most likely state is CHFSEV = 0 by a lot, it makes sense that the intercepts are as follows because as we get to the more extreme ordered end (CHFSEV = 4), the logit(P[Y<=4]) is extremely high.

Under the proportional-odds model, all of the coefficients are significant, and the odds-ratios represent the following:

Holding all else constant, for every year somone lives, they have $e^{\beta_{\text{c.age}}+\beta_{\text{c.age:smoking}}I(\text{Smoking})} = e^{(0.025+0.012I(\text{Smoking}))} = 1.037$ (if a smoker) or 1.025 (if not a smoker) times the odds of being in CHFSEV = 4 versus CHFSEV = 0,1,2, or 3. We can make the same comparison of CHFSEV = 3,4 versus CHFSEV = 0,1,2, and so on until CHFSEV = 1,2,3,4 versus CHFSEV = 0.

Holding all else constant, people who smoke are associated with an odds ratio of $e^{\beta_{\text{smoke}}I(\text{Smoking})+\beta_{\text{c.age:smoking}}x_{\text{Age}}+\beta_{\text{smoke:Gender}}I(\text{Fer}}$ $e^{0.13+0.012x_{\text{Age}}-.25I(\text{Female})}$ when comparing CHFSEV = 4 versus CHFSEV = 0,1,2, or 3. We can make the same comparison of CHFSEV = 3,4 versus CHFSEV = 0,1,2, and so on until CHFSEV = 1,2,3,4 versus CHFSEV = 0. Thus it is clear that the effect of smoking depends both on age and gender.

Holding all else constant, women are associated with an odds ratio of $e^{\beta_{\text{Gender}}+\beta_{\text{Gender:smoking}}I(\text{Smoking})} = e^{(0.428-0.25I(\text{Smoking}))} = 1.523$ (if not a smoker) or 1.19 (if a smoker) odds of being CHFSEV = 4 versus

CHFSEV = 0,1,2, or 3. We can make the same comparison of CHFSEV = 3,4 versus CHFSEV = 0,1,2, and so on until CHFSEV = 1,2,3,4 versus CHFSEV = 0.

1f. Prediction in Proportional Odds Model

Using your preferred linear predictor from 1b, find and report in a table the predicted probabilities of each level of heart failure severity from the proportional odds model for (i) a 25 year-old woman who has never smoked, (ii) a 25 year-old woman who smokes, (iii) a 25 year-old man who has never smoked, (iv) a 25 year-old man who smokes, (v) an 80 year-old woman who has never smoked, (vi) an 80 year-old woman who smokes, (vii) an 80 year-old man who has never smoked, (viii) an 80 year-old man who smokes.

Solution

```
predictvglm(mpo,pred_df,"response")
```

```
##           0          1          2           3           4
## 1 0.9386088 0.01747422 0.01764080 0.015618718 0.010657430
## 2 0.9586925 0.01193653 0.01190462 0.010419990 0.007046327
## 3 0.9591259 0.01181509 0.01178044 0.010308774 0.006969787
## 4 0.9652885 0.01007947 0.01001276 0.008731661 0.005887565
## 5 0.7787402 0.05491421 0.06142164 0.060221284 0.044702617
## 6 0.7277972 0.06418030 0.07434284 0.075655815 0.058023873
## 7 0.8437941 0.04115395 0.04409403 0.041370870 0.029587085
## 8 0.7621141 0.05808791 0.06570906 0.065194215 0.048894705
```

Although there are differences, these are quite similar to the responses for the previous predictions.

2. Agresti 8.34 (Separate Logistic Regressions)

Solution

```
fish = c(7,4,16,3,2,13,3,0,3,8,2,0,13,9,3,8)
invertebrates = c(1,0,3,0,2,7,9,1,7,6,4,1,10,0,9,1)
reptile = c(0,0,2,1,0,6,1,0,1,6,1,0,0,0,1,0)
bird = c(0,1,2,2,0,0,0,1,0,3,1,0,2,1,0,0)
other = c(5,2,3,3,1,0,2,0,1,5,4,0,2,2,1,1)
lake = c(rep("Hancock",4),rep("Oklawaha",4),rep("Trafford",4),rep("George",4))
gender = rep(c(rep("Male",2),rep("Female",2)),4)
size = rep(c(1,0),8)

# Playing aroun with some stuff
all.data = data.frame(
  Freq = c(fish,invertebrates,reptile,bird,other),
  species = c(rep("fish",16),rep("invertebrates",16),rep("reptile",16),rep("bird",16),rep("other",16)),
  lake.vec = rep(lake,5),
  gender.vec = rep(gender,5),
  size.vec = rep(size,5)
)

test.frame = expand.dft(all.data)
test.frame$species = relevel(test.frame$species, ref = "fish")
```

```
all.mod = multinom(species ~ size.vec + lake.vec, data = test.frame)
```

```
## # weights:  30 (20 variable)
## initial  value 352.466903
## iter  10 value 271.607785
## iter  20 value 270.046051
## final  value 270.040140
## converged
```

```
summary(all.mod)
```

```
## Call:
## multinom(formula = species ~ size.vec + lake.vec, data = test.frame)
##
## Coefficients:
##               (Intercept)   size.vec lake.vecHancock lake.vecOklawaha
## bird            -2.093358 -0.6306329       0.6954256     -0.652622721
## invertebrates   -1.549021  1.4581457      -1.6581178      0.937237973
## other           -1.904343  0.3315514       0.8263115      0.005792737
## reptile         -3.314512 -0.3512702       1.2428408      2.458913302
##               lake.vecTrafford
## bird                 1.088098
## invertebrates        1.122002
## other                1.516461
## reptile              2.935262
##
## Std. Errors:
##               (Intercept)  size.vec lake.vecHancock lake.vecOklawaha
## bird            0.6622972 0.6424863       0.7813123        1.2020025
## invertebrates   0.4249185 0.3959418       0.6128466        0.4719035
## other           0.5258313 0.4482504       0.5575446        0.7765655
## reptile         1.0530577 0.5800207       1.1854031        1.1181000
##               lake.vecTrafford
## bird                 0.8417085
## invertebrates        0.4905122
## other                0.6214371
## reptile              1.1163844
##
## Residual Deviance: 540.0803
## AIC: 580.0803
```

```
res.mat = rbind(coef(all.mod)["invertebrates",],
summary(all.mod)$standard.errors["invertebrates",])
rownames(res.mat) = c("Estimate","SE")
#res.mat

fi.data = test.frame %>% filter(species %in% c("fish","invertebrates"))
fi.mod = glm(species ~ size.vec + lake.vec, data = fi.data,
             family = binomial())
sum.log  = summary(fi.mod)
#sum.log$coefficients[,1:2]

comp.mat= cbind(t(res.mat),sum.log$coefficients[,1:2])
colnames(comp.mat) = c("m.Est","m.SE","log.Est","log.SE")
```

```
comp.mat
```

```
##                       m.Est      m.SE   log.Est    log.SE
## (Intercept)      -1.549021 0.4249185 -1.693577 0.4498302
## size.vec          1.458146 0.3959418  1.659802 0.4256290
## lake.vecHancock  -1.658118 0.6128466 -1.778682 0.6189471
## lake.vecOklawaha  0.937238 0.4719035  1.052417 0.4948307
## lake.vecTrafford  1.122002 0.4905122  1.217881 0.5165468
```

Clearly, when we do a logistic regression under this paradigm, it is similar to the baseline multinomial model. In this case, the first two columns are the estimate and standard error for the multinomial model and the last two columns are the estimate and standard error for the logstic regression model. Both the estimates and standard errors are quite similar to each other.