BST 210 HW #5

Hillary Miller

October 12, 2017

BST 210 HOMEWORK #5

Due 8:00 AM, Wednesday, October 18, 2017

(If you turn in the homework by 5:00 PM Tuesday, October 17, we will try to grade this before the midterm. Regardless, HW solutions will be posted late next week, ahead of the midterm.) Problems may be done in part using any computer package, but you must add in appropriate summaries and interpretation (not just the computer output). For this homework assignment, you are encouraged to interact with other students, but please submit your own solution set, providing responses in your own words.

Consider the Framingham Heart Study data set, that we used previously in a lab session. Here we focus on predicting "death from any cause" (mortality) over the 24 year period of follow-up, and focus on continuous BMI (body mass index), participant sex, and age at exam (or age category) as independent variables. The data set and a help file are available under "Datasets" at the bottom of the course Canvas home page.

```
library(foreign)
library(rms)
```

```
## Warning: package 'rms' was built under R version 3.4.2
```

```
## Loading required package: Hmisc
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
## Loading required package: SparseM
```

```
##
## Attaching package: 'SparseM'
```

Hillary Miller
BST HW #5

```
## The following object is masked from 'package:base':
##
##     backsolve

library(sandwich)

## Warning: package 'sandwich' was built under R version 3.4.2

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

dat <- framingham
```

a) Use logistic regression to assess the effects of (continuous) BMI on mortality. Briefly interpret your model. What are your conclusions? Also estimate an odds ratio and a 95% confidence interval for the effect of a 5-unit change in BMI.

Ultimately, I determined not to remove the missing data from this dataset. Removing observations with missing data resulted in removal of well over 10% of the data (closer to 50%), and since we do not at this time have effective ways of dealing with missing data, I opted to keep all records intact.

First, just to visualize the data:

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:Hmisc':
##
##     combine, src, summarize

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

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

Hillary Miller
BST HW #5

```
p <- dat %>%
  ggplot(aes(group=death, x=death, y=bmi))
p + geom_boxplot() + theme(legend.position = "none", title = element_text(col
or = "Brown")) + ylab("BMI") +
  xlab("Mortality") +
  ggtitle("Distribution of BMI in relation to Mortality")
```

```
## Warning: Removed 19 rows containing non-finite values (stat_boxplot).
```

The median for the BMI of the group that died is slightly higher than those who did not. both of the distribtuions have a positive skew with upper outliers.

Testing whether Ho: B1= 0 (BMI is not a significant predictor of death)

or

Ha: $\beta_1 \neq 0$ (BMI is a significant predictor of death)

```
lm.1 <- glm(death ~ bmi, family="binomial"(link="logit"), data=dat)
summary(lm.1)
```

```
##
## Call:
## glm(formula = death ~ bmi, family = binomial(link = "logit"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

Hillary Miller
BST HW #5

```
## -1.5951  -0.9305  -0.8644   1.3969   1.6919
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.949618   0.202824  -9.612  < 2e-16 ***
## bmi          0.050932   0.007686   6.627 3.42e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5706.7  on 4414  degrees of freedom
## Residual deviance: 5662.4  on 4413  degrees of freedom
##   (19 observations deleted due to missingness)
## AIC: 5666.4
##
## Number of Fisher Scoring iterations: 4
```

The fitted regression model is:

$$logit(death) = log(p/1-p) = \beta_0 + \beta_1(bmi)$$

Which in the linear form becomes:

$$p = exp(\beta_0 + \beta_1(bmi))/(1 + exp\beta_0 + \beta_1(bmi))$$

Here, the fitted model is

$$logit(death) = -1.949618 + 0.050932(bmi = 1,2,3\ldots)$$

```
exp(coef(lm.1)[2])
```

```
##      bmi
## 1.052252
```

Using the Wald test, we reject the null hypothesis at the α = 0.05 level, and conclude that the slope for BMI is siginificantly different from 0 (P<.01). In other words, we conclude that there exists a statistically significant association between risk of mortality and an increase in BMI.

(general conclusion) The odds of a subject with a bmi=k+1 is said to be

$$exp(0.050932) \approx 1.052252$$

more likely to die than those with a bmi=k (where k is any value of BMI). The estimated odds ratio associated with a 1 unit increase in BMI is 1.052.

For a subject with BMI 5 higher (k+5, where k is a BMI value):

```
##OR
exp(coef(lm.1)[2]*5)
```

```
##      bmi
## 1.290024

##CI
exp(5*((coef(lm.1)[2])-(1.96*coef(summary(lm.1))[2,2])))

##      bmi
## 1.196431

exp(5*((coef(lm.1)[2])+(1.96*coef(summary(lm.1))[2,2])))

##      bmi
## 1.390939
```

The odds of mortality within the study time frame for subjects of a particular BMI=k+5 are said to be exp(.0509*5)= 1.290023 times greater than the odds for a subject of the study with BMI=k (BMI 5 units lower). With 95% confidence, the odds ratio describing the association between BMI and mortality is between 1.196 and 1.391 (which does not include 1).

b) One way to assess possible nonlinear effects of BMI (on the logit scale) is to run a logistic regression model including (linear) BMI and (quadratic) BMI2 in the same model. Generate a BMI2 term, run models containing only the linear term and then including both the linear and quadratic terms, and determine if the quadratic term is needed or not. What happens to the linear effect when the quadratic term is included in the model? Also, graph the fitted probabilities from these two models overlaid on the same plot and (briefly) compare.

```
dat$bmi2 <- (dat$bmi)^2
##ran model with only linear term above, lm.1
lm.2 <- glm(death ~ bmi + bmi2, family="binomial"(link="logit"), data=dat)
summary(lm.2)

##
## Call:
## glm(formula = death ~ bmi + bmi2, family = binomial(link = "logit"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7384  -0.9279  -0.8654   1.4006   1.6679
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

Hillary Miller
BST HW #5

```
## (Intercept) -1.6476413  0.7949050  -2.073   0.0382 *
## bmi          0.0287998  0.0568957   0.506   0.6127
## bmi2         0.0003947  0.0010062   0.392   0.6949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5706.7  on 4414  degrees of freedom
## Residual deviance: 5662.2  on 4412  degrees of freedom
##   (19 observations deleted due to missingness)
## AIC: 5668.2
##
## Number of Fisher Scoring iterations: 4
```

```
anova(lm.1, lm.2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: death ~ bmi
## Model 2: death ~ bmi + bmi2
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4413     5662.4
## 2      4412     5662.2  1  0.15521   0.6936
```

So here, testing

$$Ho: \beta_2(bmi2) = 0$$

reduced model is sufficient

With the alternative

$$Ha: \beta_2(bmi2) \neq 0$$

full model is preferred. In this case, the linear model (reduced) is nested in the model with linear and quadratic BMI (full).

Based on the Likelihood Ratio Test, we fail to reject the null hypothesis, and conclude that the full model with quadratic BMI does not provide a significantly better fit than the reduced model (p =0.694).

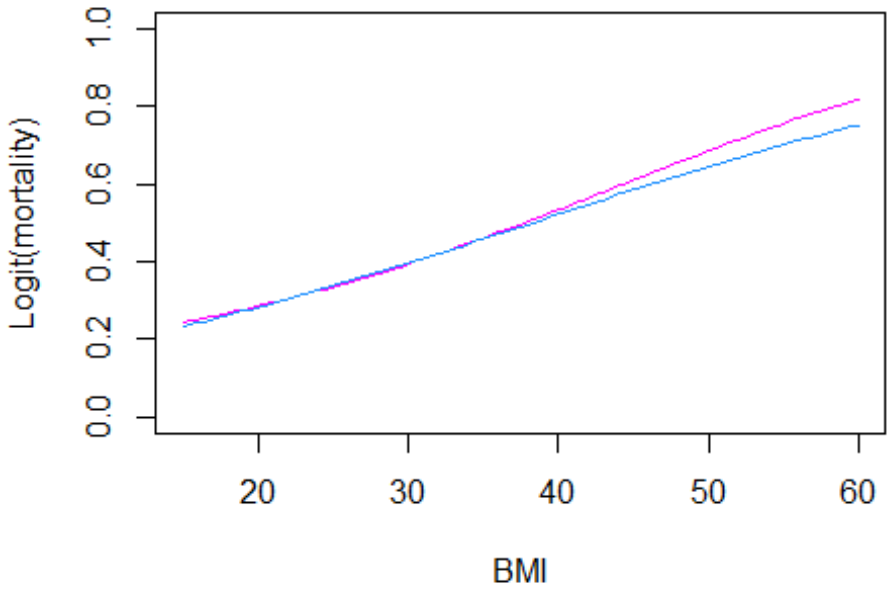
The intercept tells us nothing in either model, because these are continuous values. In the model with only linear continous BMI, the

$$\beta_1$$

coefficient is significant (P<.01). When the quadratic term of BMI is included in the model, it results in a decrease in the linear BMI coefficient and continous BMI is no longer a statistically signficant predictor of mortality (P= .613), and quadratic BMI is also not statistically significant (P=.6949).

Hillary Miller
BST HW #5

```
curve(exp(coef(lm.2)[1] + coef(lm.2)[2]*x +coef(lm.2)[3]*x^2)/(1+ exp(coef(lm
.2)[1] + coef(lm.2)[2]*x + coef(lm.2)[3]*x^2)), col="magenta", xlab="BMI", yl
ab="Logit(mortality)", xlim=c(15,60), ylim=c(0,1))
curve(exp(coef(lm.1)[1] +coef(lm.1)[2]*x)/(1 + exp(coef(lm.1)[1] +coef(lm.1)[
2]*x)), col="dodgerblue", add=T, xlim=c(15,60), ylim=c(0,1))
```



Looking at the two probabilities, the two models appear to predict almost the same outcome. However, with the model that includes quadratic BMI (magenta) makes BMI appear to have a slightly greater impact on mortality as BMI increases over 40 than the model with only linear BMI (dodgerblue). In general, adding quadratic BMI to the model does not give us more information about the prediction of mortality. As confirmed via the Anova test, we fail to reject the null, and favor the reduced model with only linear BMI.

c) For the model including both linear and quadratic BMI, estimate the odds ratio for a 5-unit increase in BMI (comparing 25 to 20) and for a 5-unit increase in BMI (comparing 35 to 30). (Because we have a quadratic BMI term in the model, these two odds ratio estimates should differ, because BMI is "interacting with itself".)

```
summary(lm.2)

##
## Call:
## glm(formula = death ~ bmi + bmi2, family = binomial(link = "logit"),
```

Hillary Miller
BST HW #5

```
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7384  -0.9279  -0.8654   1.4006   1.6679
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.6476413  0.7949050  -2.073   0.0382 *
## bmi          0.0287998  0.0568957   0.506   0.6127
## bmi2         0.0003947  0.0010062   0.392   0.6949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5706.7  on 4414  degrees of freedom
## Residual deviance: 5662.2  on 4412  degrees of freedom
##   (19 observations deleted due to missingness)
## AIC: 5668.2
##
## Number of Fisher Scoring iterations: 4
```

```
(25^2)-(20^2)
```

```
## [1] 225
```

```
(35^2)-(30^2)
```

```
## [1] 325
```

```
exp((5*coef(lm.2)[2]) + ((25^2-20^2)*(coef(lm.2)[3])))
```

```
##      bmi
## 1.262141
```

```
exp((5*coef(lm.2)[2]) + ((35^2-30^2)*(coef(lm.2)[3])))
```

```
##      bmi
## 1.312956
```

Among the population of subjects of a BMI of 25, the odds of mortality in the study period are estimated to be exp(0.02879985 + 2250.0003947)= 1.2621 greater than the odds among the population of subjects with a BMI of 20. (could also interpret as 26% increase in the odds of mortality)

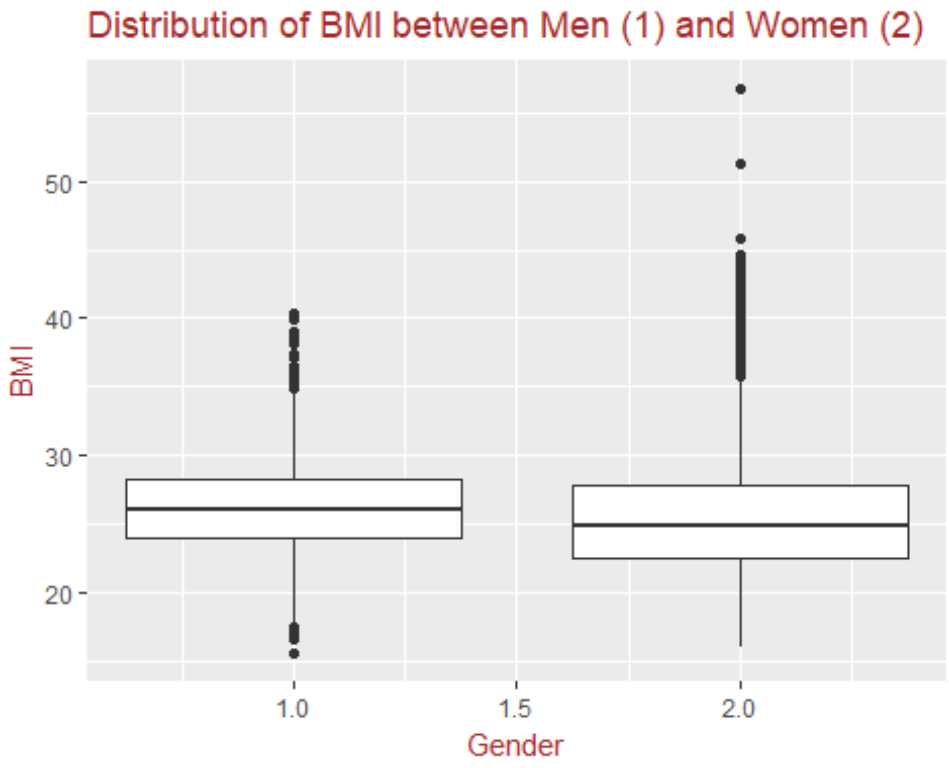
Among the population of subjects of a BMI of 35, the odds of mortality in the study period are estimated to be exp(0.02879985 + 3250.0003947)= 1.312956 greater than the odds among the population of subjects with a BMI of 30. (could also interpret as 31% increase in the odds of mortality for someone with a BMI of 35 vs. 30.)

Hillary Miller
BST HW #5

d) Go back to using only the linear BMI term. Perform some descriptive statistics or graphical display to assess the association between BMI and participant sex. Then perform an appropriate set of logistic regression analyses to determine whether or not sex is a confounder or an effect modifier of the effect of (continuous) BMI on mortality. What are your conclusions (in words) about the effect of BMI on mortality, considering the additional effects of sex? (Hint: It may be helpful to create a 0/1 indicator variable for sex, e.g., "Female = 1 for females, Female = 0 for males".)

```
library(dplyr)
p <- dat %>%
  ggplot(aes(group=sex, x=sex, y=bmi))
p <- p + geom_boxplot() + theme(legend.position = "none", title = element_tex
t(color = "Brown")) + ylab("BMI") +
  xlab("Gender") +
  ggtitle("Distribution of BMI between Men (1) and Women (2)")
p
```

```
## Warning: Removed 19 rows containing non-finite values (stat_boxplot).
```




```
dat <- dat %>% mutate(sex= factor((sex-1))) %>% as.data.frame()
```

T-test

Hillary Miller
BST HW #5

```
dat_f = subset(dat,sex==1)
dat_m =  subset(dat,sex==0)
t.test(dat_f$bmi, dat_m$bmi)
```

```
##
##  Welch Two Sample t-test
##
## data:  dat_f$bmi and dat_m$bmi
## t = -4.8099, df = 4403.8, p-value = 1.56e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8117584 -0.3416388
## sample estimates:
## mean of x mean of y
##  25.59288  26.16958
```

According to the boxplot above, the distribution of BMI is wider for females than for males. While the average BMI of females is lower than males, there are more upper outliers for the female group (wider distribution). There is a positive skew of BMI for women. Using a Welch sample T-test, we can conclude that the average BMI among women are signficantly different to the average BMI among men (p<.001). It is good to test for confounding here, since it is reasonable that gender is associated with both BMI and risk of death, but it is not a consequence of either.

```
summary(lm.1)
```

```
##
## Call:
## glm(formula = death ~ bmi, family = binomial(link = "logit"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5951  -0.9305  -0.8644   1.3969   1.6919
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.949618   0.202824  -9.612  < 2e-16 ***
## bmi          0.050932   0.007686   6.627 3.42e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5706.7  on 4414  degrees of freedom
## Residual deviance: 5662.4  on 4413  degrees of freedom
##   (19 observations deleted due to missingness)
## AIC: 5666.4
##
## Number of Fisher Scoring iterations: 4
```

```
lm.3 <- glm(death ~ bmi + sex, family=binomial(), data=dat)
summary(lm.3)
```

```
##
## Call:
## glm(formula = death ~ bmi + sex, family = binomial(), data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4118  -0.9672  -0.7758   1.2936   1.8009
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.513938   0.209687  -7.220 5.20e-13 ***
## bmi          0.047439   0.007798   6.084 1.18e-09 ***
## sex1        -0.644679   0.064299 -10.026  < 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: 5706.7  on 4414  degrees of freedom
## Residual deviance: 5561.0  on 4412  degrees of freedom
##   (19 observations deleted due to missingness)
## AIC: 5567
##
## Number of Fisher Scoring iterations: 4
```

Based on the 10% rule of thumb, we can conclude that sex is not a confounder of the effect of BMI on mortality, given the β coefficient changes by (0.051-0.047)/0.051 = 7.8%. However, this is still a relationship worth investigating since there is a change in the coefficient related to BMI that is between 5-10%.

```
lm.4 <- glm(death ~ bmi + sex + bmi*sex, family=binomial(), data=dat)
summary(lm.4)
```

```
##
## Call:
## glm(formula = death ~ bmi + sex + bmi * sex, family = binomial(),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6777  -1.0453  -0.7592   1.2933   1.8866
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

Hillary Miller
BST HW #5

```
## (Intercept) -0.510089   0.355224  -1.436 0.151013
## bmi          0.009139   0.013453   0.679 0.496927
## sex1        -2.148473   0.436929  -4.917 8.78e-07 ***
## bmi:sex1     0.057502   0.016525   3.480 0.000502 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5706.7  on 4414  degrees of freedom
## Residual deviance: 5548.9  on 4411  degrees of freedom
##   (19 observations deleted due to missingness)
## AIC: 5556.9
##
## Number of Fisher Scoring iterations: 4
```

```
exp((coef(lm.4)[2])-(1.96*coef(summary(lm.4))[2,2]))
```

```
##       bmi
## 0.9829185
```

```
exp((coef(lm.4)[2])+(1.96*coef(summary(lm.4))[2,2]))
```

```
##      bmi
## 1.036145
```

```
##males
exp(coef(lm.4)[2])
```

```
##      bmi
## 1.009181
```

```
##females
exp(coef(lm.4)[2]+coef(lm.4)[4])
```

```
##      bmi
## 1.068912
```

Based on the logistic regression analysis, the coefficient for sex by bmi is statistically significant (p = 0.0005), so sex is a meaningful effect modifier for BMI and the log odds of mortality. The relationship between BMI and the risk of death is different for males and females.

Among males, those with a one unit increase in BMI (k+1) are estimated to have a 1.009 times higher odds of mortality than for males with a BMI one unit lower (k). (could also interpret as .92% increase in the odds of mortality among males with a k+1 BMI)

Among females, those with a one unit increase in BMI (k+1) are estimated to have a 1.0689 times higher odds of mortality than for females with a BMI one unit lower (k).

Hillary Miller
BST HW #5

(With 95% confidence, the odds ratio describing the association between BMI and mortality among males is between 983 and 1.036 (so for males, since we cannot definitively say that there is a strong interaction between BMI and mortality). This does coincide with the results of the Wald test (P>.05).)

e) Now considering age and age category alone (not BMI or sex), compare models using (continuous) age, (ordinal) age category (i.e., age category used as a continuous covariate), and (categorical) age category. Which approach do you feel best models the effect of age on mortality? Justify your response. (It may be helpful to look at or plot fitted probabilities or run a hypothesis test.)

For continous age:

```
age.continuous <- glm(death~ age, family=binomial(), data=dat)
summary(age.continuous)
```

```
##
## Call:
## glm(formula = death ~ age, family = binomial(), data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8179  -0.8411  -0.5594   0.9656   2.3034
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.484358   0.230966  -28.07   <2e-16 ***
## age          0.114842   0.004392   26.15   <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: 5739.2  on 4433  degrees of freedom
## Residual deviance: 4906.3  on 4432  degrees of freedom
## AIC: 4910.3
##
## Number of Fisher Scoring iterations: 3
```

For ordinal categorical age (treated as continous):

```
age.ordinal <- glm(death ~ agecat, family=binomial(), data=dat)
summary(age.ordinal)
```

```
##
## Call:
## glm(formula = death ~ agecat, family = binomial(), data = dat)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5085  -0.6949  -0.6949   0.8791   2.1983
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.34747    0.11880  -28.18   <2e-16 ***
## agecat       1.02470    0.04125   24.84   <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: 5739.2  on 4433  degrees of freedom
## Residual deviance: 4992.6  on 4432  degrees of freedom
## AIC: 4996.6
##
## Number of Fisher Scoring iterations: 3
```

For categorical age:

```
age.category1 <- glm(death ~ as.factor(agecat), family=binomial(), data=dat)
summary(age.category1)
```

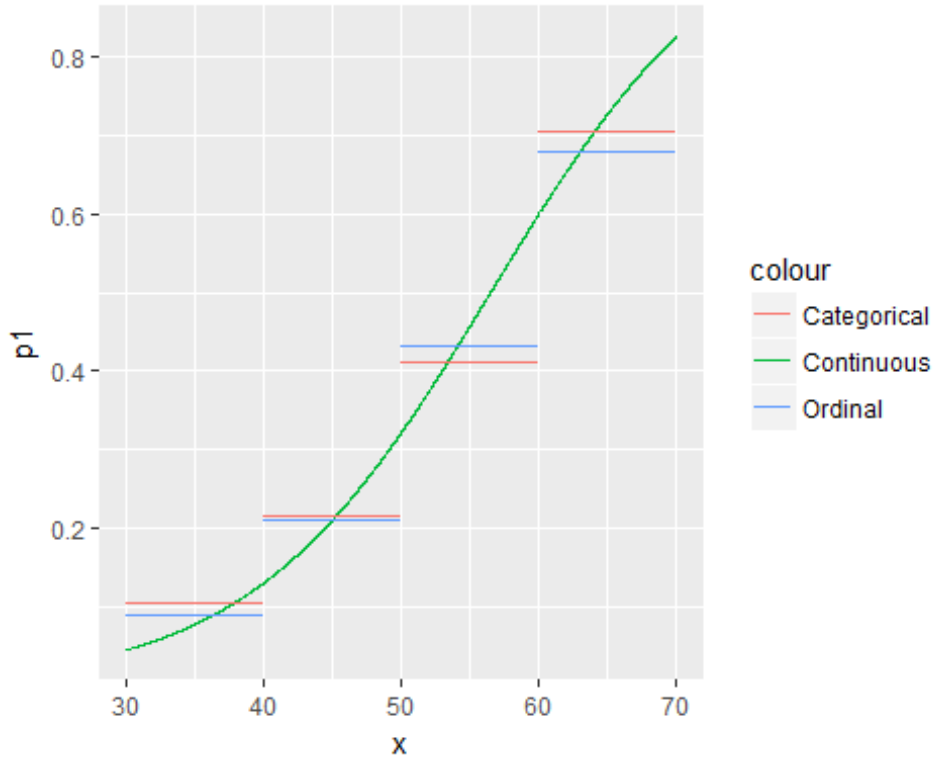
```
##
## Call:
## glm(formula = death ~ as.factor(agecat), family = binomial(),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5606  -0.6960  -0.6960   0.8377   2.1207
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.1371     0.1376 -15.525  < 2e-16 ***
## as.factor(agecat)2   0.8428     0.1498   5.625 1.85e-08 ***
## as.factor(agecat)3   1.7773     0.1480  12.010  < 2e-16 ***
## as.factor(agecat)4   3.0039     0.1583  18.972  < 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: 5739.2  on 4433  degrees of freedom
## Residual deviance: 4986.1  on 4430  degrees of freedom
## AIC: 4994.1
##
## Number of Fisher Scoring iterations: 4
```

```
AIC(age.ordinal, age.category1, age.continuous)
```

Hillary Miller
BST HW #5

```
##                df      AIC
## age.ordinal     2 4996.553
## age.category1   4 4994.062
## age.continuous  2 4910.299

BIC(age.ordinal, age.category1, age.continuous)

##                df      BIC
## age.ordinal     2 5009.347
## age.category1   4 5019.650
## age.continuous  2 4923.093

##age.ordinal, age.category1, age.continuous)

#Graph fitted probability
#fitted prob for continuous
x=seq(30,70,0.01)
p1=exp(coef(age.continuous)[1]+coef(age.continuous)[2]*x)/(1+exp(coef(age.con
tinuous)[1]+(coef(age.continuous)[2]*x)))
table(dat$agecat)

##
##    1    2    3    4
##  559 1692 1399  784

#fitted prob for ordinal probs
x2=c(1,2,3,4)
p2=exp(coef(age.ordinal)[1]+(coef(age.ordinal)[2]*x2))/(1+exp(coef(age.ordina
l)[1]+coef(age.ordinal)[2]*x2))
#fitted prob for categorical probs
p3=rep(0,4)
p3[1]=exp(coef(age.category1)[1])/(1+exp(coef(age.category1)[1]))
for(i in 2:4)
{
  p3[i]=exp(coef(age.category1)[1]+coef(age.category1)[i])/(1+exp(coef(age.ca
tegory1)[1]+coef(age.category1)[i]))
}

ggplot(data=as.data.frame(cbind(x,p1)),aes(x=x,y=p1))+geom_line(aes(color="Co
ntinuous"))+geom_segment(aes(x =30 , y = p2[1], xend = 40, yend = p2[1], colo
ur = "Ordinal"))+geom_segment(aes(x =40 , y = p2[2]-0.005, xend = 50, yend =
p2[2]-0.005,colour = "Ordinal"))+geom_segment(aes(x =50 , y = p2[3], xend = 6
0, yend = p2[3], colour = "Ordinal"))+geom_segment(aes(x =60 , y = p2[4], xen
d = 70, yend = p2[4], colour = "Ordinal"))+geom_segment(aes(x =30 , y = p3[1]
, xend = 40, yend = p3[1], colour = "Categorical"))+geom_segment(aes(x =40 ,
y = p3[2], xend = 50, yend = p3[2],colour = "Categorical"))+geom_segment(aes(
x =50 , y = p3[3], xend = 60, yend = p3[3], colour = "Categorical"))+geom_seg
ment(aes(x =60 , y = p3[4], xend = 70, yend = p3[4], colour = "Categorical"))
```

Hillary Miller
BST HW #5



Based on the results of the AIC/BIC evaluations (with continuous age having the lowest), and the plotted probabilities, the continuous age model best models the effect of age on mortality.

To compare categorical and ordinal age:

Formal hypothesis: (Ho: the reduced model, with ordinal age is sufficient) versus (H1: the full model, with categorical age is preferred)

```
anova(age.ordinal, age.category1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: death ~ agecat
## Model 2: death ~ as.factor(agecat)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      4432     4992.6
## 2      4430     4986.1  2   6.4915  0.03894 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hillary Miller
BST HW #5

Comparing ordinal and categorical models using a likelihood ratio test, we can confirm that the full model is preferred and reject the null hypothesis that the coefficients of the full (categorical) model are not significant (P=.039).

While the continuous is best overall for fitting the data (since it uses all of the data), it may be more difficult to interpret and to generalize in practice. While some information is lost in the categorical model, it is beneficial to simplify interpretations, specifically in a clinical setting. This can be viewed from the graph of the probabilities, above.

f) Perform an appropriate set of logistic regression analyses to determine whether or not age category is a confounder or an effect modifier of the possible effect of (continuous) BMI on mortality. What are your conclusions about the effect of BMI on mortality, considering the additional effects of age category?

```
summary(lm.1)
```

```
##
## Call:
## glm(formula = death ~ bmi, family = binomial(link = "logit"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5951  -0.9305  -0.8644   1.3969   1.6919
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.949618   0.202824  -9.612  < 2e-16 ***
## bmi          0.050932   0.007686   6.627 3.42e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5706.7  on 4414  degrees of freedom
## Residual deviance: 5662.4  on 4413  degrees of freedom
##   (19 observations deleted due to missingness)
## AIC: 5666.4
##
## Number of Fisher Scoring iterations: 4
```

```
lm.5 <- glm(death ~ bmi + as.factor(agecat), family=binomial(), data=dat)
summary(lm.5)
```

```
##
## Call:
```

Hillary Miller
BST HW #5

```
## glm(formula = death ~ bmi + as.factor(agecat), family = binomial(),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7736  -0.7675  -0.6547   0.8791   2.2352
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.91079    0.25486 -11.421  < 2e-16 ***
## bmi                 0.03025    0.00844   3.584 0.000339 ***
## as.factor(agecat)2  0.83656    0.15106   5.538 3.06e-08 ***
## as.factor(agecat)3  1.75306    0.14939  11.735  < 2e-16 ***
## as.factor(agecat)4  2.96480    0.15972  18.562  < 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: 5706.7  on 4414  degrees of freedom
## Residual deviance: 4948.0  on 4410  degrees of freedom
##   (19 observations deleted due to missingness)
## AIC: 4958
##
## Number of Fisher Scoring iterations: 4
```

Based on the 10% rule of thumb, we can conclude that categorical age is a confounder on the effect of BMI on mortality, given the β coefficient changes by (0.051-0.03025)/0.051 = 41%.

```
lm.6 <- glm(death ~ bmi + as.factor(agecat) + as.factor(agecat)*bmi, family=b
inomial(), data=dat)
summary(lm.6)
```

```
##
## Call:
## glm(formula = death ~ bmi + as.factor(agecat) + as.factor(agecat) *
##     bmi, family = binomial(), data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7561  -0.7788  -0.6487   0.8762   2.1551
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -2.351735   0.921596  -2.552   0.0107 *
## bmi                     0.008057   0.036348   0.222   0.8246
## as.factor(agecat)2      0.159868   0.994433   0.161   0.8723
## as.factor(agecat)3      1.182847   0.990614   1.194   0.2325
## as.factor(agecat)4      2.469458   1.041398   2.371   0.0177 *
```

```
## bmi:as.factor(agecat)2  0.026758   0.039064   0.685   0.4934
## bmi:as.factor(agecat)3  0.022613   0.038821   0.582   0.5602
## bmi:as.factor(agecat)4  0.019770   0.040633   0.487   0.6266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5706.7  on 4414  degrees of freedom
## Residual deviance: 4947.5  on 4407  degrees of freedom
##   (19 observations deleted due to missingness)
## AIC: 4963.5
##
## Number of Fisher Scoring iterations: 4
```

Based on the logistic regression analysis, the coefficient for categorical age by bmi is not statistically significant (p>0.05 for all age categories). Given this information, we do not need to calculate an OR for each individual age category.

```
exp(coef(lm.1)[2]) ## original model

##      bmi
## 1.052252

exp(coef(lm.5)[2]) ##model controlling for age

##      bmi
## 1.030708
```

When controlling for age, those with a one unit increase in BMI (k+1) are estimated to have a 1.031 times higher odds of mortality than for subjects with a BMI one unit lower (k). (could also interpret as 3.1% increase in the odds of mortality among those with k+1 BMI)