BST 210 Homework 5

Wenjie Gu

```r
# import and clean dataset
library(haven)
framingham <- read_dta("framingham.dta")
framingham = na.omit(framingham[c('sex', 'bmi', 'age','agecat','death')])
framingham$sex = framingham$sex -1
```

Problem 1

1(a)

```r
# fit logistic regression model -  linear
bmi.mortality = glm(death~bmi, family = binomial(), data = framingham)
summary(bmi.mortality)
```

```
##
## Call:
## glm(formula = death ~ bmi, family = binomial(), data = framingham)
##
## 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
## AIC: 5666.4
##
## Number of Fisher Scoring iterations: 4
```

Intercept: The odds of death from any cause is estimated to be $e^{-1.949618} \approx 0.14233$ for individuals with bmi score equal zero. (Not sensible in real life)

Slope: The odds of death from any cause for an individual is estimated to be $e^{0.050932 \approx 1.0523}$ times higher for every 1-unit increase in his/her bmi value.

```r
# Calculate odds ratio for the effect of a 5-unit change in bmi
odds_ratio = exp(5* 0.050932)
cat("The odds ratio for the effect of a 5-unit change in bmi is",odds_ratio, '\n')
```

```
## The odds ratio for the effect of a 5-unit change in bmi is 1.290023
```

```r
lower_bound = exp(5* (0.050932 - 1.96*0.007686))
upper_bound = exp(5* (0.050932 + 1.96*0.007686))
```

```
sprintf("The 95 percent CI of OR for the effect of a 5-unit change in bmi is (%f, %f).",
        lower_bound, upper_bound)
```

```
## [1] "The 95 percent CI of OR for the effect of a 5-unit change in bmi is (1.196424, 1.390944)."
```

1(b)

```
# fit logistic regression model - linear and quadratic bmi
bmi.quad.mortality = glm(death ~ bmi + I(bmi^2), family = binomial, data = framingham)
summary(bmi.quad.mortality)
```

```
## 
## Call:
## glm(formula = death ~ bmi + I(bmi^2), family = binomial, data = framingham)
## 
## 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  
## I(bmi^2)     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
## AIC: 5668.2
## 
## Number of Fisher Scoring iterations: 4
```

```
anova(bmi.mortality, bmi.quad.mortality, test = 'Chisq')
```

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

After including the quadratic term, the linear term becomes insignificant.

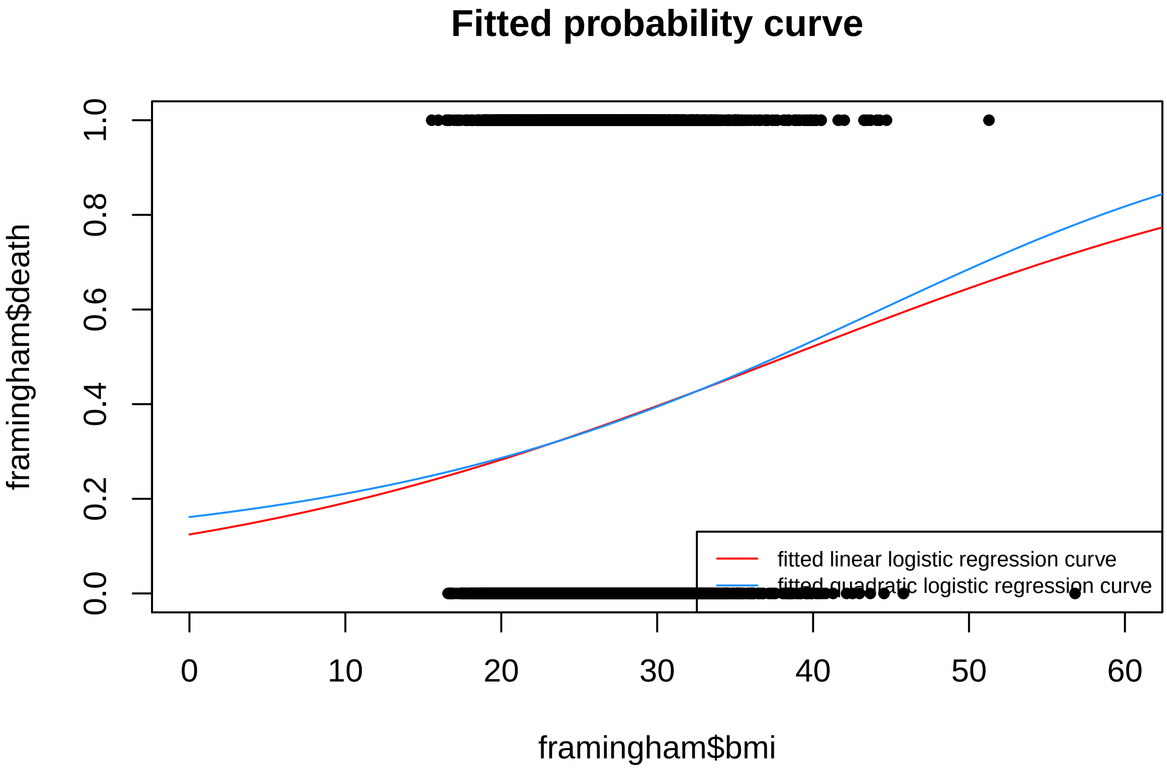
Since the Likelihood Ratio Test gives p = 0.6936, we fail to reject the null hypothesis where the linear model is sufficient. Therefore, it is not necessary to include the quadratic term.

```
bmi_range = range(framingham$bmi)
bmi_range
```

```
## [1] 15.54 56.80
```

```
xweight = seq(0,100,0.01)
yweight1 = predict(bmi.mortality,list(bmi = xweight), type = "response")
```

```
yweight2 = predict(bmi.quad.mortality, list(bmi = xweight), type = "response")
plot(framingham$death ~ framingham$bmi, col = "black", pch = 20, xlim = c(0,60),
     main = "Fitted probability curve")
lines(xweight,yweight1, col = "red")
lines(xweight, yweight2, col = "dodgerblue")
legend("bottomright", c("fitted linear logistic regression curve", "fitted quadratic logistic regressio
       col = c("red", "dodgerblue") ,lty = "solid", cex = 0.7)
```



With bmi <20, the model with quadratic bmi term will give a slightly higher mortality probability than the model with only the linear term for a fixed bmi value. For bmi in the approximate range of (20, 35), the two models give similar prediction (the curves overlap). For bmi >35, the model with quadratic bmi term will again give higher prediction result for a fixed bmi than the linear one. However, the trend and shape of the fitted curves is similar.

1(c)

```
odds = function(bmi){
  log_odds = -1.6476413 + 0.0287998 * bmi + 0.0003947 * bmi^2
  odds = exp(log_odds)
  return (odds)
}
oddsratio1 = odds(25)/odds(20)
oddsratio2 = odds(35)/odds(30)
```

```
cat("The odds ratio for a 5-unit increase in BMI (comparing 25 to 20) is:", oddsratio1, '\n')
```

```
## The odds ratio for a 5-unit increase in BMI (comparing 25 to 20) is: 1.262137
```

```
cat("The odds ratio for a 5-unit increase in BMI (comparing 35 to 30) is:", oddsratio2)
```

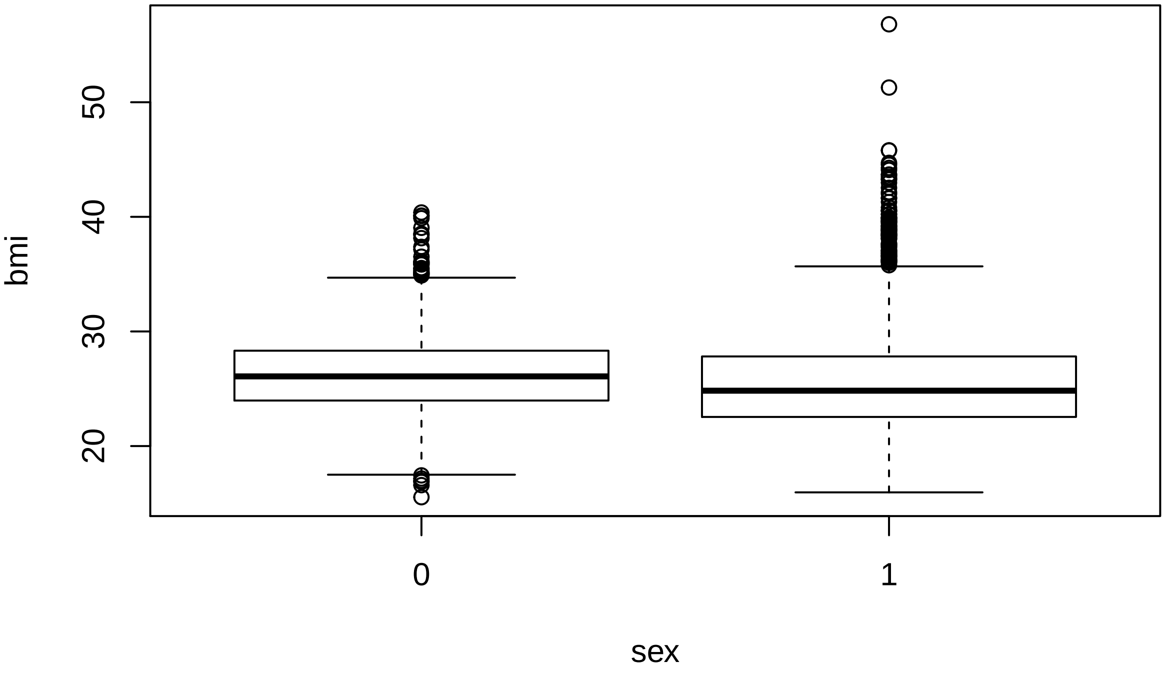
```
## The odds ratio for a 5-unit increase in BMI (comparing 35 to 30) is: 1.31295
```

1(d)

```
# Two sample t-test comparing the average bmi of males and females
t.test(bmi~sex, data = framingham)
```

```
##
##  Welch Two Sample t-test
##
## data:  bmi by sex
## 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:
##  0.3416388 0.8117584
## sample estimates:
## mean in group 0 mean in group 1
##        26.16958        25.59288
```

```
boxplot(bmi~sex, data = framingham)
```



By performing a t-test comparing the average bmi of males and females, we get a p-value less than 0.05 (p = 1.56e-6). There is evidence suggesting the association between participant sex and bmi.

```
# check if sex is a confounder
bmi.sex.mortality = glm(death~ bmi + sex, family = binomial, data = framingham)
summary(bmi.sex.mortality)
```

```
##
## Call:
## glm(formula = death ~ bmi + sex, family = binomial, data = framingham)
##
## 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 ***
## sex         -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
## AIC: 5567
##
## Number of Fisher Scoring iterations: 4
```

```
summary(bmi.mortality)
```

```
##
## Call:
## glm(formula = death ~ bmi, family = binomial(), data = framingham)
##
## 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
## AIC: 5666.4
##
## Number of Fisher Scoring iterations: 4
```

Coefficient for bmi changes from 0.050932 to 0.047439 (% of change: -6.8%), which is less than 10%. Therefore sex is not a confounder.

```
# check if sex is an effect modifier
bmi.sexint.mortality = glm(death~bmi+bmi*sex, family = binomial, data = framingham)
summary(bmi.sexint.mortality)
```

```
##
## Call:
## glm(formula = death ~ bmi + bmi * sex, family = binomial, data = framingham)
##
## 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|)
## (Intercept) -0.510089   0.355224  -1.436 0.151013
## bmi          0.009139   0.013453   0.679 0.496927
## sex         -2.148473   0.436929  -4.917 8.78e-07 ***
## bmi:sex      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
## AIC: 5556.9
##
## Number of Fisher Scoring iterations: 4
```

The interaction between bmi and sex has significant coefficient ($p = 0.000502$), therefore sex is an effect modifier of the effect of continuous BMI on mortality.

For males, the odds of mortality is estimated to be $e^{0.009139} \approx 1.00918$ times higher for every 1-unit increase in bmi value. For females, the odds of mortality is estimated to be $e^{0.009139+0.057502 \approx 1.0689}$ times higher for every 1-unit increase in bmi value.

1(e)

```
age.cont.mortality = glm(death~age, family = binomial, data = framingham)
age.cat.factor.mortality = glm(death~as.factor(agecat), family = binomial, data = framingham)
age.cat.cont.mortality= glm(death~agecat, family = binomial, data = framingham)
```

```
anova(age.cat.cont.mortality,age.cat.factor.mortality, 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      4413     4966.7
## 2      4411     4960.9  2    5.857  0.05348 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(age.cont.mortality)
```

```
## [1] 4885.81
```

```
AIC(age.cat.cont.mortality)
```

```
## [1] 4970.728
```

```
AIC(age.cat.factor.mortality)
```

```
## [1] 4968.871
```

```
BIC(age.cont.mortality)
```

```
## [1] 4898.596
```

```
BIC(age.cat.cont.mortality)
```

```
## [1] 4983.514
```

```
BIC(age.cat.factor.mortality)
```

```
## [1] 4994.442
```

```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
library(LogisticDx)
fitted.cont.results = ifelse(fitted(age.cont.mortality) > 0.5,1,0)
fitted.cat.cont.results = ifelse(fitted(age.cat.cont.mortality) > 0.5,1,0)
fitted.cat.factor.results = ifelse(fitted(age.cat.factor.mortality) > 0.5,1,0)

hoslem.test(framingham$death,fitted(age.cont.mortality) ,g=10)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  framingham$death, fitted(age.cont.mortality)
## X-squared = 15.761, df = 8, p-value = 0.04593
```

```
chisq.test(framingham$death,fitted(age.cat.cont.mortality))
```

```
##
##  Pearson's Chi-squared test
##
## data:  framingham$death and fitted(age.cat.cont.mortality)
## X-squared = 730.31, df = 3, p-value < 2.2e-16
```

```
chisq.test(framingham$death,fitted(age.cat.factor.mortality))
```

```
##
##  Pearson's Chi-squared test
##
## data:  framingham$death and fitted(age.cat.factor.mortality)
## X-squared = 730.31, df = 3, p-value < 2.2e-16
```

Of the three models, the continuous age model has the lowest AIC and BIC scores. After assessing for the goodness of fits of the three models, we found that all the three models give significant Hosmer-Lemeshow/Pearson Chi-squared statistics, indicating that none of the fits are adequate. However, if we have to select one “best” model, according to AIC and BIC values, the preferred one should be the continuous age model.

1(f)

```
bmi.agecat.mortality = glm(death~bmi + agecat, family=binomial, data = framingham)
bmi.agecatint.mortality = glm(death ~ bmi + agecat*bmi, family = binomial, data = framingham)
summary(bmi.mortality)
```

```
##
## Call:
## glm(formula = death ~ bmi, family = binomial(), data = framingham)
##
```

```
## 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
## AIC: 5666.4
##
## Number of Fisher Scoring iterations: 4
```

```
summary(bmi.agecat.mortality)
```

```
##
## Call:
## glm(formula = death ~ bmi + agecat, family = binomial, data = framingham)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7173  -0.7638  -0.6538   0.9208   2.3058
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.082978   0.242284 -16.852  < 2e-16 ***
## bmi          0.029577   0.008413   3.516 0.000439 ***
## agecat       1.009916   0.041566  24.296  < 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: 4954.4  on 4412  degrees of freedom
## AIC: 4960.4
##
## Number of Fisher Scoring iterations: 4
```

```
summary(bmi.agecatint.mortality)
```

```
##
## Call:
## glm(formula = death ~ bmi + agecat * bmi, family = binomial,
##     data = framingham)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7362  -0.7597  -0.6559   0.9238   2.2916
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.931320   0.754444  -5.211 1.88e-07 ***
## bmi          0.023704   0.028935   0.819 0.412674
## agecat       0.955220   0.261044   3.659 0.000253 ***
## bmi:agecat   0.002111   0.009950   0.212 0.832001
## ---
## 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: 4954.3  on 4411  degrees of freedom
## AIC: 4962.3
##
## Number of Fisher Scoring iterations: 4
```

The coefficient of bmi changes from 0.050932 to 0.029577 (-36% change). Thus age category is a confounder for the effect of bmi on mortality. Whereas, the interaction between age category and bmi is not significant, indicating that age category is not an effect modifier for the effect of bmi on mortality.