Homework 8

Sarah Adilijiang

Problem 2

(a)

The GLM model for binary response here is:
likelihood: $P(chd_i = y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i},\ y_i = 0,1$
linear predictor: $\eta_i = \beta_0 + \beta_1 sbp_i + \beta_2 tobacco_i + \beta_3 ldl_i + \beta_4 1_{famhist_i=Present} + \beta_5 obesity_i + \beta_6 alcohol_i + \beta_7 age_i$
link function (logit): $\eta_i = \log\frac{p_i}{1-p_i}$

```
mod = glm(chd~., family=binomial, SAheart)
coef(mod)
```

```
##    (Intercept)            sbp        tobacco            ldl famhistPresent
##  -4.1295996883   0.0057606767   0.0795256305   0.1847793334   0.9391854851
##        obesity        alcohol            age
##  -0.0345434340   0.0006065017   0.0425412093
```

```
# significance of each predictor
drop1(mod, test="Chi")
```

```
## Single term deletions
##
## Model:
## chd ~ sbp + tobacco + ldl + famhist + obesity + alcohol + age
##         Df Deviance    AIC     LRT  Pr(>Chi)
## <none>       483.17 499.17
## sbp      1   484.22 498.22  1.0492 0.3056944
## tobacco  1   493.05 507.05  9.8796 0.0016712 **
## ldl      1   494.09 508.09 10.9197 0.0009515 ***
## famhist  1   500.89 514.89 17.7110 2.571e-05 ***
## obesity  1   484.61 498.61  1.4352 0.2309253
## alcohol  1   483.19 497.19  0.0185 0.8917985
## age      1   501.51 515.51 18.3397 1.848e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Discussion:

The predictors **tobacco**, **ldl** and **age** have significant positive effects on odds of the response **chd**, and the present of family history of heart desease will significantly increase the odds of the response **chd** while controlling other predictors.

However, the predictors **sbp**, **obesity** and **alcohol** are not significant at 5% significance level, though **sbp** and **alcohol** has positive effect and **obesity** has negative effect on odds of the response **chd**.

Next, we test if there is a simpler model that fits the data well.

From the chi-square test above, we see that the predictor **alcohol** has a p-value = 0.8917985, which is the most insignificant term in the model, so we drop this term first and then continue to drop the next one. We repeat this process until we find all the predictors are significant at 5% significance level.

```
mod1 = glm(chd~sbp+tobacco+ldl+famhist+obesity+age, family=binomial, SAheart)
drop1(mod1, test="Chi")
```

```
## Single term deletions
##
## Model:
## chd ~ sbp + tobacco + ldl + famhist + obesity + age
##         Df Deviance    AIC     LRT  Pr(>Chi)
## <none>       483.19 497.19
## sbp      1   484.30 496.30  1.1042 0.2933437
## tobacco  1   493.62 505.62 10.4227 0.0012448 **
## ldl      1   494.12 506.12 10.9268 0.0009478 ***
## famhist  1   501.07 513.07 17.8752 2.359e-05 ***
## obesity  1   484.63 496.63  1.4358 0.2308160
## age      1   501.54 513.54 18.3522 1.836e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
mod2 = glm(chd~sbp+tobacco+ldl+famhist+age, family=binomial, SAheart)
drop1(mod2, test="Chi")
```

```
## Single term deletions
##
## Model:
## chd ~ sbp + tobacco + ldl + famhist + age
##         Df Deviance    AIC     LRT  Pr(>Chi)
## <none>       484.63 496.63
## sbp      1   485.44 495.44  0.8155  0.366499
## tobacco  1   495.17 505.17 10.5439  0.001166 **
## ldl      1   494.21 504.21  9.5866  0.001960 **
## famhist  1   502.19 512.19 17.5585 2.786e-05 ***
## age      1   502.17 512.17 17.5385 2.815e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
mod3 = glm(chd~tobacco+ldl+famhist+age, family=binomial, SAheart)
drop1(mod3, test="Chi")
```

```
## Single term deletions
##
## Model:
## chd ~ tobacco + ldl + famhist + age
##         Df Deviance    AIC     LRT  Pr(>Chi)
## <none>       485.44 495.44
## tobacco  1   496.18 504.18 10.7364  0.001050 **
## ldl      1   495.39 503.39  9.9415  0.001616 **
## famhist  1   502.82 510.82 17.3808 3.059e-05 ***
## age      1   507.24 515.24 21.7987 3.028e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# compare with full model
1 - pchisq(mod3$deviance-mod$deviance, mod3$df.residual-mod$df.residual)
```

```
## [1] 0.5183256
```

Answer:

Now all the predictors are significant at 5% significance level, and the p-value of the difference-in-deviance chi-square test is $0.5183256 > 0.5$, so we do not reject the smaller null model, i.e. the smaller model is preferred.

Therefore, we get the final smaller model that fits the data better than the full model:

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 tobacco_i + \beta_2 ldl_i + \beta_3 1_{famhist_i=Present} + \beta_4 age_i$$

(b)

```r
# new GLM model with Natural Cubic Splines
library(splines)
SAheart$sbp_NS = ns(SAheart$sbp, df=4)
SAheart$tobacco_NS = ns(SAheart$tobacco, df=4)
SAheart$ldl_NS = ns(SAheart$ldl, df=4)
SAheart$obesity_NS = ns(SAheart$obesity, df=4)
SAheart$alcohol_NS = ns(SAheart$alcohol, df=4)
SAheart$age_NS = ns(SAheart$age, df=4)

mod_NS = glm(chd~famhist+sbp_NS+tobacco_NS+ldl_NS+obesity_NS+alcohol_NS+age_NS,
          family=binomial, SAheart)

# significance of each predictor
drop1(mod_NS, test="Chi")
```

```
## Single term deletions
##
## Model:
## chd ~ famhist + sbp_NS + tobacco_NS + ldl_NS + obesity_NS + alcohol_NS +
##     age_NS
##            Df Deviance    AIC     LRT  Pr(>Chi)
## <none>          457.63 509.63
## famhist     1   478.76 528.76 21.1319 4.287e-06 ***
## sbp_NS      4   466.77 510.77  9.1429 0.0576257 .
## tobacco_NS  4   469.61 513.61 11.9753 0.0175355 *
## ldl_NS      4   470.90 514.90 13.2710 0.0100249 *
## obesity_NS  4   465.41 509.41  7.7749 0.1001811
## alcohol_NS  4   458.09 502.09  0.4562 0.9776262
## age_NS      4   480.37 524.37 22.7414 0.0001426 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# compare this Natural Cubic Spline GLM full model with the simple linear GLM full model
1 - pchisq(mod$deviance - mod_NS$deviance, mod$df.residual - mod_NS$df.residual)
```

```
## [1] 0.1106942
```

Answer:

There are 7 predictors in the original dataset, among which 6 ones are continuous predictors. Now we produce 4 B-spline bases for each of the continuous predictors, so now there will be $1 + 4 \times 6 = 25$ predictors in total (except the intercept) in the Natural Cubic Spline GLM model.

The Natural Cubic Spline GLM model for binary response here is:
likelihood: $P(chd_i = y_i | p_i) = p_i^{y_i}(1-p_i)^{1-y_i},\ y_i = 0, 1$
model:

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 1_{famhist_i=Present} + \sum_{j=1}^{4} (\beta_{2j} sbp_{ij} + \beta_{3j} tobacco_{ij} + \beta_{4j} ldl_{ij} + \beta_{5j} obesity_{ij} + \beta_{6j} alcohol_{ij} + \beta_{7j} age_{ij})$$

Comparing this Natural Cubic Spline GLM full model with the original simple linear GLM full model, we see that again the predictors **famhist**, **tobacco**, **ldl** and **age** are significant predictors at 5% significance level, while **sbp**, **obesity** and **alcohol** are not.

Besides, the p-value of the likelihood-based chi-square test for comparing these two models is 0.1106942 > 0.05, so we do not reject the null model (the smaller simple linear GLM full model). Therefore, the Natural Cubic Spline GLM full model is not significant comparing with the simple linear GLM full model at 5% significane level.

Next, we still use the chi-square test to test if there is a simpler Natural Cubic Spline GLM model that fits the data well.

From the chi-square test above, we see that the 4 B-spline predictors of **alcohol_NS** has a p-value = 0.9776262, which is the most insignificant term in the model, so we drop this term first and then continue to drop the next one. We repeat this process until we find all the predictors are significant at 5% significance level.

```
mod_NS1 = glm(chd~famhist+sbp_NS+tobacco_NS+ldl_NS+obesity_NS+age_NS,
              family=binomial, SAheart)
drop1(mod_NS1, test="Chi")
```

```
## Single term deletions
##
## Model:
## chd ~ famhist + sbp_NS + tobacco_NS + ldl_NS + obesity_NS + age_NS
##            Df Deviance    AIC     LRT  Pr(>Chi)
## <none>          458.09 502.09
## famhist     1   479.44 521.44 21.3562 3.814e-06 ***
## sbp_NS      4   467.16 503.16  9.0762  0.059223 .
## tobacco_NS  4   470.48 506.48 12.3873  0.014692 *
## ldl_NS      4   472.39 508.39 14.3065  0.006378 **
## obesity_NS  4   466.24 502.24  8.1471  0.086336 .
## age_NS      4   481.86 517.86 23.7682 8.889e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod_NS2 = glm(chd~famhist+sbp_NS+tobacco_NS+ldl_NS+age_NS,
              family=binomial, SAheart)
drop1(mod_NS2, test="Chi")
```

```
## Single term deletions
##
## Model:
## chd ~ famhist + sbp_NS + tobacco_NS + ldl_NS + age_NS
##            Df Deviance    AIC     LRT  Pr(>Chi)
## <none>          466.24 502.24
## famhist     1   486.90 520.90 20.6624 5.478e-06 ***
## sbp_NS      4   474.75 502.75  8.5112 0.0745476 .
## tobacco_NS  4   479.51 507.51 13.2730 0.0100163 *
## ldl_NS      4   477.73 505.73 11.4970 0.0215113 *
## age_NS      4   485.58 513.58 19.3463 0.0006719 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod_NS3 = glm(chd~famhist+tobacco_NS+ldl_NS+age_NS, family=binomial, SAheart)
drop1(mod_NS3, test="Chi")
```

```
## Single term deletions
```

```
##
## Model:
## chd ~ famhist + tobacco_NS + ldl_NS + age_NS
##            Df Deviance    AIC     LRT  Pr(>Chi)
## <none>          474.75 502.75
## famhist     1   493.09 519.09 18.3441 1.844e-05 ***
## tobacco_NS  4   489.69 509.69 14.9462 0.0048141 **
## ldl_NS      4   484.40 504.40  9.6577 0.0466056 *
## age_NS      4   496.34 516.34 21.5967 0.0002411 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# compare with Natural Cubic Spline GLM full model
1 - pchisq(mod_NS3$deviance-mod_NS$deviance, mod_NS3$df.residual-mod_NS$df.residual)
```

```
## [1] 0.1453413
```

Answer:

Now all the predictors are significant at 5% significance level, and the p-value of the difference-in-deviance chi-square test is $0.1453413 > 0.5$, so we do not reject the smaller null model, i.e. the smaller model is preferred.

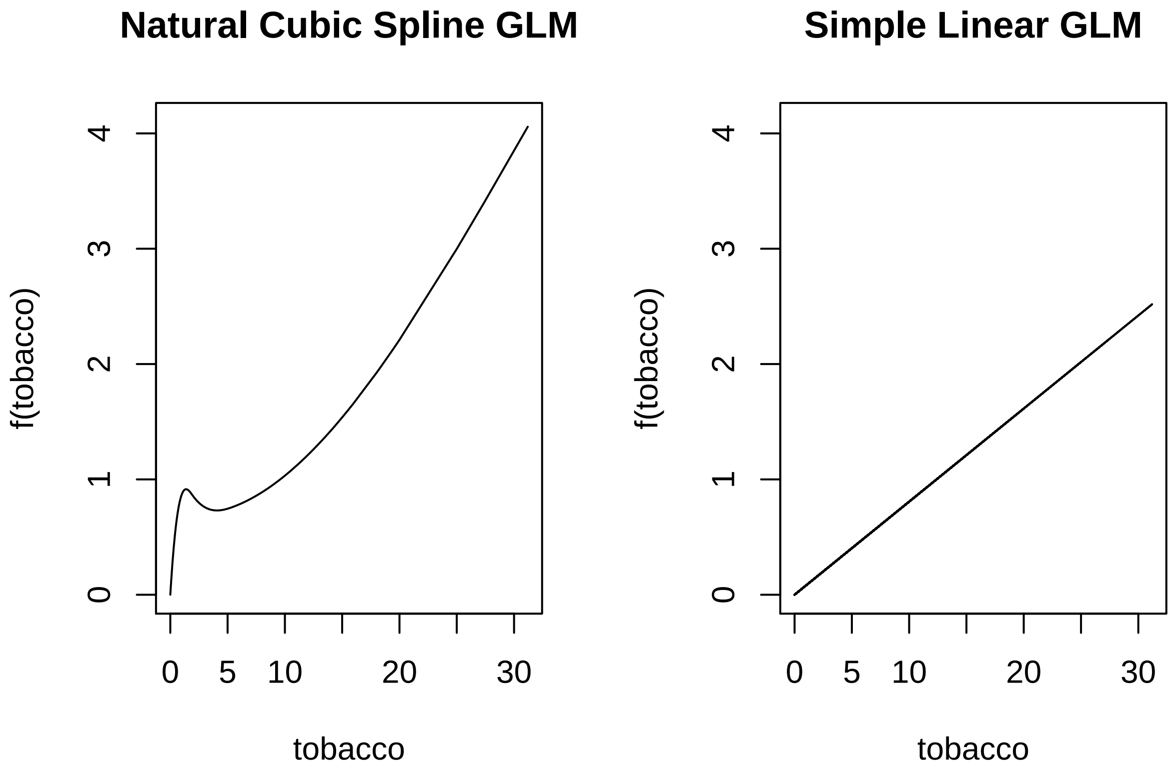
Therefore, we get the final smaller Natural Cubic Spline model that fits the data better than the full model:

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 1_{famhist_i=Present} + \sum_{j=1}^{4} (\beta_{2j} tobacco_{ij} + \beta_{3j} ldl_{ij} + \beta_{5j} age_{ij})$$

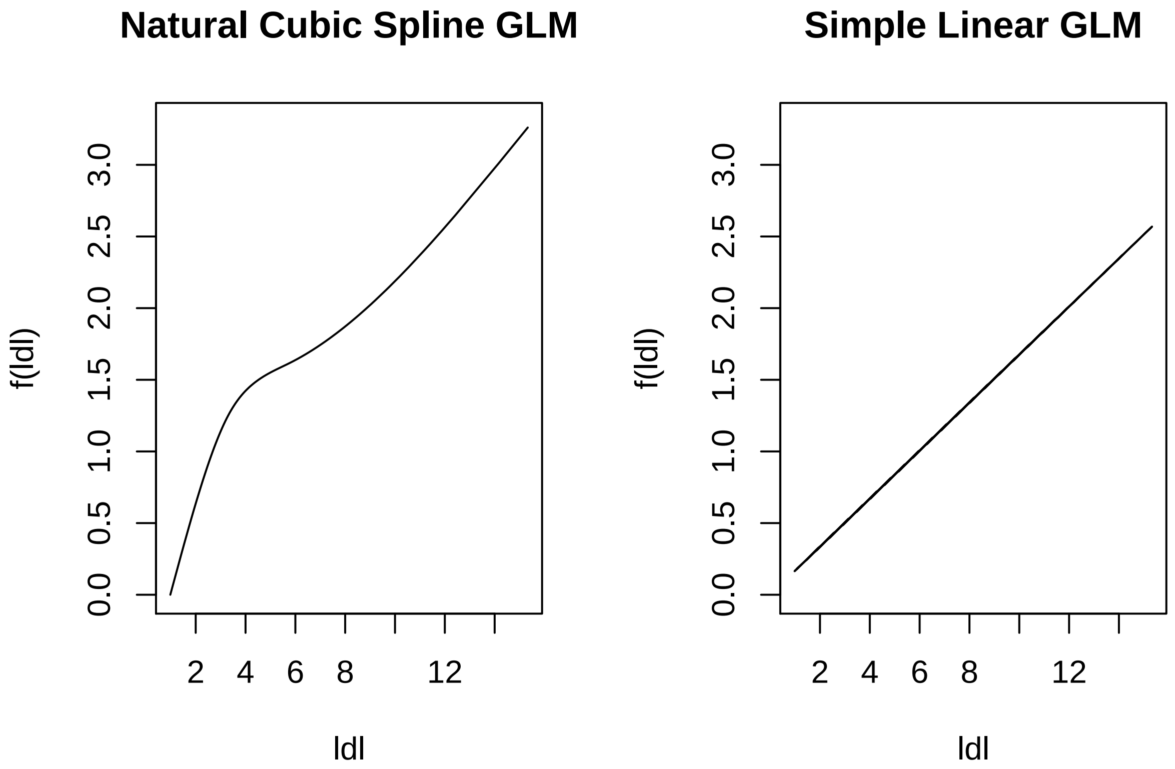
Notice that this model have the 4 B-spline blocks of the same corresponding variables in the final smaller model of simple linear GLM model.

Next, we plot the function $f(x_i) = \sum_{j=1}^{4} \beta_j x_{ij}$ for each of the 3 original continuous variables x_i (i.e. **tobacco**, **ldl**, and **age**) in the final smaller model.

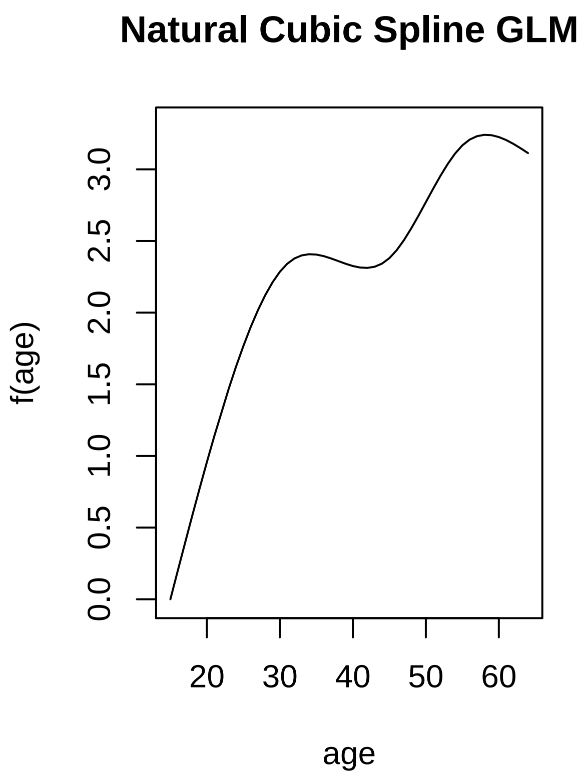
```r
# final Natural Cubic Spline GLM model
beta_NS = mod_NS3$coefficients
y_tobacco_NS = SAheart$tobacco_NS %*% beta_NS[3:6]
y_ldl_NS = SAheart$ldl_NS %*% beta_NS[7:10]
y_age_NS = SAheart$age_NS %*% beta_NS[11:14]

# final simple linear GLM model
beta = mod3$coefficients
y_tobacco = SAheart$tobacco * beta[2]
y_ldl = SAheart$ldl * beta[3]
y_age = SAheart$age * beta[5]

# plot function f(xi)'s while comparing with the simple linear GLM model
par(mfrow=c(1,2))
plot(y_tobacco_NS[order(SAheart$tobacco)] ~ SAheart$tobacco[order(SAheart$tobacco)], ylim=c(0,4.1),
     type="l", main="Natural Cubic Spline GLM", xlab="tobacco", ylab="f(tobacco)")
plot(y_tobacco ~ SAheart$tobacco, ylim=c(0,4.1),
     type="l", main="Simple Linear GLM", xlab="tobacco", ylab="f(tobacco)")
```

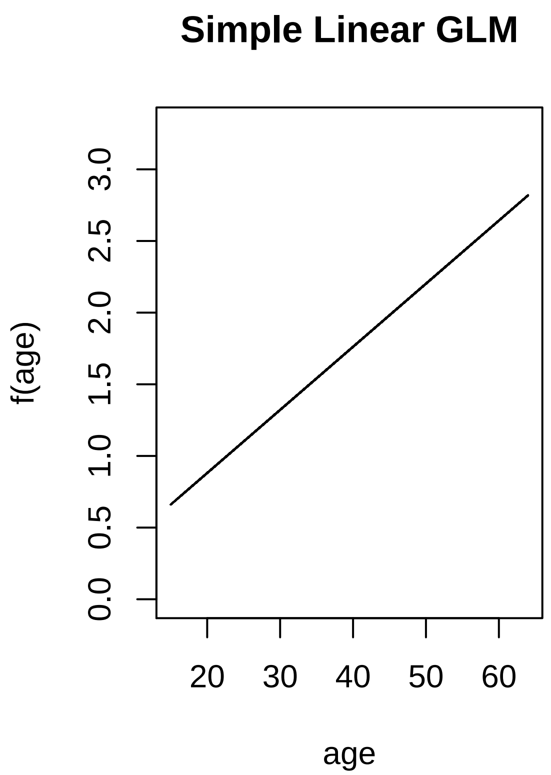


```
par(mfrow=c(1,2))
plot(y_ldl_NS[order(SAheart$ldl)] ~ SAheart$ldl[order(SAheart$ldl)], ylim=c(0,3.3),
     type="l", main="Natural Cubic Spline GLM", xlab="ldl", ylab="f(ldl)")
plot(y_ldl ~ SAheart$ldl, ylim=c(0,3.3),
     type="l", main="Simple Linear GLM", xlab="ldl", ylab="f(ldl)")
```



```
par(mfrow=c(1,2))
plot(y_age_NS[order(SAheart$age)] ~ SAheart$age[order(SAheart$age)], ylim=c(0,3.3),
     type="l", main="Natural Cubic Spline GLM", xlab="age", ylab="f(age)")
plot(y_age ~ SAheart$age, ylim=c(0,3.3),
     type="l", main="Simple Linear GLM", xlab="age", ylab="f(age)")
```


Natural Cubic Spline GLM


Simple Linear GLM

Answer:

Comparing the dependence on the variables between two models, we see that in the simple linear GLM model, all the dependence is linear positive, meaning that every unit change of one variable will make constant increase in the log-odds of the response while controlling for other variables.

However, the dependence on the variables in the Natural Cubic Spline GLM model is not linear (though having some linear parts), meaning that every unit change of one variable will make different changes in the log-odds of the response while controlling for other variables. And these changes are not always positive, having negative effect parts which depends on the range of the variables.

Therefore, the Natural Cubic Spline GLM model allows more flexible relationships between the log-odds of the response and the variables than the simple linear GLM model.