Homework 7

Skylar Liu

2024-02-24

Homework #7, Stat 660, Spring 2024, Due Saturday, February 25

1. Fit an ordinary logistic regression with response = CHD, and predictor = Age. Show the summary table.

```
# clear workspace
rm(list = ls())
# set the seed
set.seed(382957)
library(HRW)
library(mgcv)
```

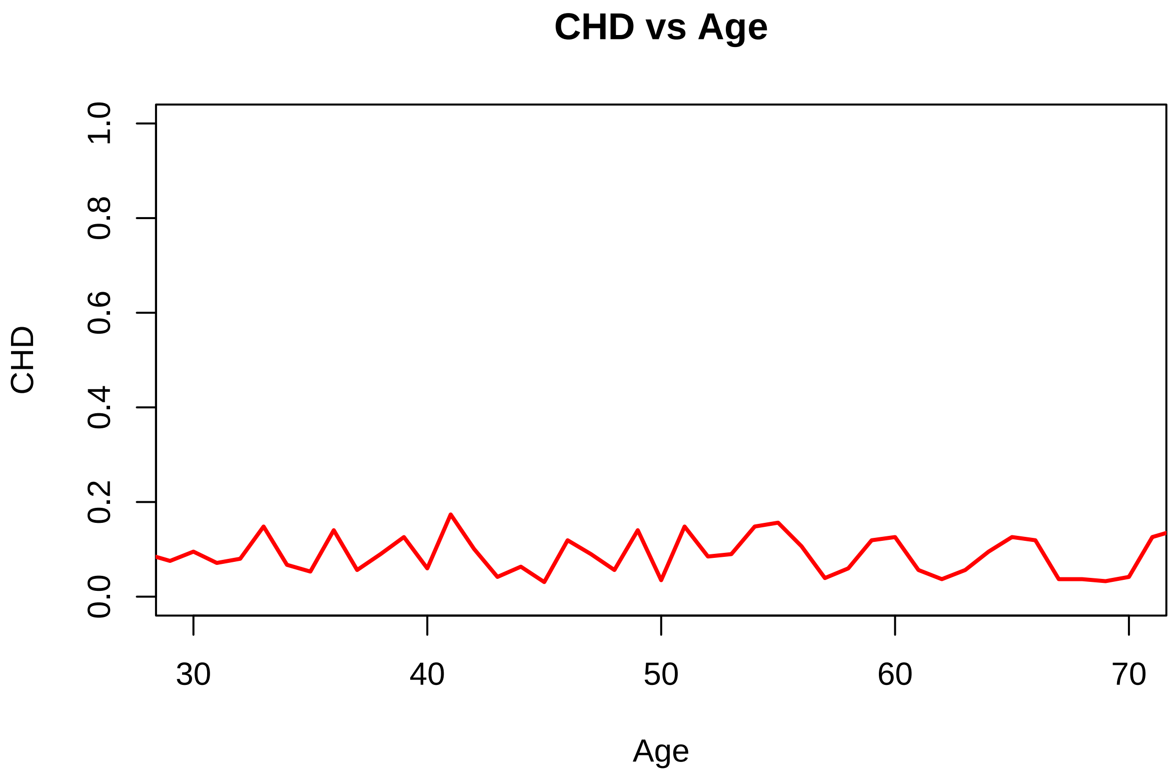
```
## Loading required package: nlme

## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

```
# import data
init = read.csv("~/660 - Flexible Regression/Homework/Homework6/Framingham.csv")
SBP      = rowMeans(init[, c(3:6)])
LSBP     = log(SBP - 50)
cholest  = rowMeans(init[, c(8:9)])
Lcholest = log(cholest)
data     = cbind(init[, c("CHD", "Age", "Smoker")], LSBP, Lcholest)

# fit an ordinary logistic regression
data$CHD <- as.numeric(as.character(data$CHD))

framlog <- mgcv::gam(CHD ~ Age, family = binomial(link="logit"), data = data)
summary(framlog)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## CHD ~ Age
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.46397    0.55630  -9.822  < 2e-16 ***
## Age          0.06298    0.01103   5.711 1.12e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.0183   Deviance explained = 3.81%
## UBRE = -0.46475  Scale est. = 1         n = 1615
```

2. In Question 1, display the fit without the data points. Just plot the model object. We know that when plotting the model object, mgcv::gam ignored the intercept.

```r
#Model predictor
gampred <- predict(framlog, type = "response")

# Plot the log regression model object
plot(data$Age, data$CHD, type="n",xlim = c(30,70), ylim = c(0, 1),xlab="Age", ylab="CHD", main = "CHD v
lines(gampred, lwd=2, col = "red")
```



3. In Question 1, is the fit statistically significant?

The deviance explained for the ordinary logistic regression model is small (3.81%), suggesting that the model is a good fit. The p-values for Age is also statistically significant, with 1.12x10ˆ-8 much smaller than 0.05.

4. In Question 1, test whether the fit is linear or quadratic versus the need to do a semiparametric fit, i.e., a spline in age

```r
# Fit a quadratic fit
framquad = mgcv::gam(CHD ~ poly(Age,degree=2), family = binomial(link="logit"),
                      data = data)

# Fit a semiparametric fit
framparam = mgcv::gam(CHD ~ s(Age,bs="cr"), family = binomial(link="logit"),
                      data = data)

# Test linear vs semiparametric
anova(framlog,framparam,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: CHD ~ Age
## Model 2: CHD ~ s(Age, bs = "cr")
##   Resid. Df Resid. Dev    Df Deviance Pr(>Chi)
## 1      1613     860.43
## 2      1611     853.55 2.046   6.8824  0.03356 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Test quadratic vs semiparametric
anova(framquad,framparam,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: CHD ~ poly(Age, degree = 2)
## Model 2: CHD ~ s(Age, bs = "cr")
##   Resid. Df Resid. Dev    Df Deviance Pr(>Chi)
## 1      1612     855.16
## 2      1611     853.55 1.046   1.6114   0.2153
```

A spline is only necessary in the linear fit, with a statistically significant p-value of 0.033 which is less than 0.05. The p-value for the quadratic fit is not statistically significant at 0.2153, suggesting we don't need the spline for this case.

5. Fit a logistic gam with all the predictors but only LSBP modeled as a spline.

```r
# Fit a logistic game with all predictors, LSBP as spline
framall <- mgcv::gam(CHD ~ Age + Smoker + Lcholest + s(LSBP,bs="cr"),
                     family = binomial(link="logit"),
                     data = data)
summary(framall)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## CHD ~ Age + Smoker + Lcholest + s(LSBP, bs = "cr")
##
```

```
## Parametric coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -20.33917    3.28985  -6.182 6.31e-10 ***
## Age           0.05673    0.01190   4.767 1.87e-06 ***
## Smoker        0.60476    0.25094   2.410    0.016 *
## Lcholest      2.67952    0.57842   4.632 3.61e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Approximate significance of smooth terms:
##           edf Ref.df Chi.sq  p-value
## s(LSBP) 1.738    2.2  16.06 0.000537 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## R-sq.(adj) =  0.0432   Deviance explained = 9.17%
## UBRE = -0.48979  Scale est. = 1         n = 1615
```

a. Quote the p-values for all of the predictors. P-values: Age: 1.87e-06 Smoker: 0.016 Lcholest: 3.61e-06 LSBP: 0.000537

b. Answer whether the fit suggests that LSBP should be modeled as a spline. Remember, you need to do an ANOVA for this will the null model having everything modeled as ordinary logistic regression.

```r
# Fit a logistic game with all predictors
framall2 <- mgcv::gam(CHD ~ Age + Smoker + Lcholest + LSBP,
                      family = binomial(link="logit"),
                      data = data)

# Test LSBP with or without spline
anova(framall2,framall,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: CHD ~ Age + Smoker + Lcholest + LSBP
## Model 2: CHD ~ Age + Smoker + Lcholest + s(LSBP, bs = "cr")
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
## 1    1610.0     814.76
## 2    1608.8     812.52 1.2004   2.2489   0.1691
```

LSBP does not need to be modeled as a spline; when tested against everything modeled as ordinary logistic regression, the p-value is 0.1691 which is not statistically significant at the 0.05 level.

6. Fit a logistic gam with LSBP, Lcholest and age modeled as splines. Quote the p-values for all predictors. Tell me which of the spline terms seem like they are worth modeling as a spline. Remember, you need to do an ANOVA for this will the null model having everything modeled as ordinary linear logistic regression, but use mgcv::gam.

```r
# Fit a logistic gam with LSBP, Lcholest, and age as splines
framsplines <- mgcv::gam(CHD ~ s(LSBP,bs="cr") + s(Lcholest,bs="cr") + s(Age,bs="cr") + Smoker,
                     family = binomial(link="logit"),
                     data = data)
summary(framsplines)
```

```
## 
## Family: binomial 
## Link function: logit 
## 
## Formula:
## CHD ~ s(LSBP, bs = "cr") + s(Lcholest, bs = "cr") + s(Age, bs = "cr") + 
##     Smoker
## 
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -3.2512     0.2457 -13.230   <2e-16 ***
## Smoker        0.5925     0.2507   2.363   0.0181 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Approximate significance of smooth terms:
##               edf Ref.df Chi.sq  p-value    
## s(LSBP)     1.630  2.051  16.21 0.000353 ***
## s(Lcholest) 1.001  1.002  20.62 6.60e-06 ***
## s(Age)      2.179  2.725  23.32 4.45e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## R-sq.(adj) =  0.0446   Deviance explained =  9.7%
## UBRE = -0.4914  Scale est. = 1         n = 1615
```

```r
# Compare to ordinary linear logistic regression
anova(framall2,framsplines,test="Chisq")
```

```
## Analysis of Deviance Table
## 
## Model 1: CHD ~ Age + Smoker + Lcholest + LSBP
## Model 2: CHD ~ s(LSBP, bs = "cr") + s(Lcholest, bs = "cr") + s(Age, bs = "cr") + 
##     Smoker
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)  
## 1    1610.0     814.76                           
## 2    1607.2     807.76 2.7776   7.0003  0.06072 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Test each spline to see which is significant
framLSBP <- mgcv::gam(CHD ~ s(LSBP,bs="cr") + Lcholest + Age + Smoker,
                     family = binomial(link="logit"),
                     data = data)
framLcholest <- mgcv::gam(CHD ~ LSBP + s(Lcholest,bs="cr") + Age + Smoker,
                     family = binomial(link="logit"),
                     data = data)
framAge <- mgcv::gam(CHD ~ LSBP + Lcholest + s(Age,bs="cr") + Smoker,
                     family = binomial(link="logit"),
                     data = data)

# Compare to ordinary linear logistic regression
anova(framall2,framLSBP,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: CHD ~ Age + Smoker + Lcholest + LSBP
## Model 2: CHD ~ s(LSBP, bs = "cr") + Lcholest + Age + Smoker
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
## 1    1610.0     814.76
## 2    1608.8     812.52 1.2004   2.2489   0.1691
```

```
anova(framall2,framLcholest,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: CHD ~ Age + Smoker + Lcholest + LSBP
## Model 2: CHD ~ LSBP + s(Lcholest, bs = "cr") + Age + Smoker
##   Resid. Df Resid. Dev        Df  Deviance Pr(>Chi)
## 1      1610     814.76
## 2      1610     814.76 0.0068931 0.0057096  0.01806 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(framall2,framAge,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: CHD ~ Age + Smoker + Lcholest + LSBP
## Model 2: CHD ~ LSBP + Lcholest + s(Age, bs = "cr") + Smoker
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
## 1    1610.0     814.76
## 2    1608.2     809.40 1.7966   5.3606  0.05629 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-values: Smoker: 0.0181 LSBP: 0.000353 Lcholest: 6.60e-06 Age: 4.45e-05

Lcholest is the only statistically significant spline term, with a p-value of 0.018. LSBP has a p-value of 0.169 and Age has a p-value of 0.056, which are not statistically significant at the 0.05 level.

7. This is an open-ended question with no absolutely correct answer. It will not be graded. At some point in life, you are going to have to summarize your results to people who do not care for the details of the analysis. So, having done an exhaustive and exhausting analysis, write a paragraph free of technical jargon about what things you think might be important in predicting who is at higher risk of getting coronary heart disease.

There are many possible variables that may contribute to a hightened risk of getting coronoary heart disease. Based on our recent analysis, I have analyzed Age, Cholesterol, Smoking status, and Systolic blood pressure. From the data provided, I have observed cholesterol and age to be the greatest predictors of getting coronary heart disease.