Assignment 07 - STAT 689

Philip Anderson; panders2@tamu.edu

2/26/2018

```
library("mgcv")
```

```
# read in our data set
fram <- read.csv("/Users/panders2/Documents/schools/tamu/stat_689/homework/semiparametric-regression/mis
names(fram) <- tolower(names(fram))
```

Our response variable Y, equals chd; our predictors are lsbp, lcholest, age, and smoking status.

Question 1

Fit a logistic gam with only LSBP modeled as a spline.

```
logit1 <- mgcv::gam(chd ~ s(lsbp, k=23, bs="cr") + lcholest + age + smoker
                    , family=binomial(link="logit")
                    , data=fram
                    )
summary(logit1)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## chd ~ s(lsbp, k = 23, bs = "cr") + lcholest + age + smoker
##
## Parametric coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -20.33972    3.28988  -6.183 6.31e-10 ***
## lcholest      2.67958    0.57843   4.633 3.61e-06 ***
## age           0.05673    0.01190   4.767 1.87e-06 ***
## smoker        0.60475    0.25094   2.410    0.016 *
## ---
## 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) 1.748   2.22  16.07 0.000547 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0433   Deviance explained = 9.17%
## UBRE = -0.48979  Scale est. = 1         n = 1615
```

P-values for 4 predictors are as follows:

```
pval_smry <- function(gam_mod) {
  gam_smry <- summary(gam_mod)
  print("Parametric Term p-values:")
```

```
  print(gam_smry$pTerms.pv)
  cat("\nSmoothed Term p-values:\n")
  cat(gam_smry$s.pv)

}

pval_smry(logit1)
```

```
## [1] "Parametric Term p-values:"
##     lcholest          age       smoker
## 3.612535e-06 1.869438e-06 1.595596e-02
##
## Smoothed Term p-values:
## 0.0005469875
```

Note from the output of the summary statement that we are not explaining a great deal of the deviance present within the data, so our model has some room for improvement. Nonetheless, the p-value on the smoothed term suggests that *LSBP* should be modeled as a spline.

Question 2

Fit a logisitc gam with only Lcholest modeled as a spline.

```
logit2 <- mgcv::gam(chd ~ lsbp + s(lcholest, k=23, bs="cr") + age + smoker
                    , family=binomial(link="logit")
                    , data=fram
                    )
summary(logit2)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## chd ~ lsbp + s(lcholest, k = 23, bs = "cr") + age + smoker
##
## Parametric coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -13.04593    1.85195  -7.044 1.86e-12 ***
## lsbp          1.66355    0.42118   3.950 7.82e-05 ***
## age           0.05580    0.01188   4.697 2.64e-06 ***
## smoker        0.60900    0.25134   2.423   0.0154 *
## ---
## 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(lcholest) 1.003  1.006  22.48 2.23e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.042   Deviance explained = 8.92%
## UBRE = -0.48931  Scale est. = 1         n = 1615
```

```
pval_smry(logit2)
```

```
## [1] "Parametric Term p-values:"
##         lsbp          age       smoker
## 7.823087e-05 2.642826e-06 1.539144e-02
##
## Smoothed Term p-values:
## 2.225305e-06
```

It looks as though the *lcholest* variable should be modeled within a spline function.

Question 3

Fit a logistic gam with only age modeled as a spline.

```
logit3 <- mgcv::gam(chd ~ lsbp + lcholest + s(age, k=23, bs="cr") + smoker
                    , family=binomial(link="logit")
                    , data=fram
                    )
summary(logit3)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## chd ~ lsbp + lcholest + s(age, k = 23, bs = "cr") + smoker
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -25.1600     3.6525  -6.888 5.64e-12 ***
## lsbp          1.6861     0.4214   4.001 6.30e-05 ***
## lcholest      2.6873     0.5815   4.622 3.81e-06 ***
## smoker        0.5955     0.2510   2.372   0.0177 *
## ---
## 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(age) 2.255   2.83  22.88 4.63e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0435   Deviance explained = 9.52%
## UBRE = -0.4911  Scale est. = 1         n = 1615
```

```
pval_smry(logit3)
```

```
## [1] "Parametric Term p-values:"
##         lsbp     lcholest       smoker
## 6.302367e-05 3.808389e-06 1.768211e-02
##
## Smoothed Term p-values:
## 4.626735e-05
```

From the above p-values, it appears as though the *age* term should be smoothed as well.

Question 4

Fit the model with all continuous terms modeled as splines; and smoking_status as-is.

```
logit4 <- mgcv::gam(chd ~ s(lsbp, k=23, bs="cr") +
                      s(lcholest, k=23, bs="cr") +
                      s(age, k=23, bs="cr") + smoker
                    , family=binomial(link="logit")
                    , data=fram
                    )
summary(logit4)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## chd ~ s(lsbp, k = 23, bs = "cr") + s(lcholest, k = 23, bs = "cr") +
##     s(age, k = 23, 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)     1.638  2.065  16.23 0.000361 ***
## s(lcholest) 1.002  1.004  20.62 5.74e-06 ***
## s(age)      2.194  2.754  23.32 3.70e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0446   Deviance explained = 9.71%
## UBRE = -0.4914  Scale est. = 1         n = 1615
```

```
pval_smry(logit4)
```

```
## [1] "Parametric Term p-values:"
##    smoker
## 0.0181234
##
## Smoothed Term p-values:
## 0.0003607555 5.73582e-06 3.701926e-05
```

From the above summaries, it appears that all of the spline terms we put into the fourth logistic regression are appropriate; this model has the highest %Deviance explained out of any of the models we have seen so far.

Question 5

Refit the model from Question 4, with only the terms that you picked as being significant under spline transformation. I elected to model all continuous terms as splines, so I will simply repeat the model from question 4.

```
logit5 <- mgcv::gam(chd ~ s(lsbp, k=23, bs="cr") +
                      s(lcholest, k=23, bs="cr") +
                      s(age, k=23, bs="cr") + smoker
                    , family=binomial(link="logit")
                    , data=fram
                    )
```

Question 6

Test whether the model from Question 5 differs from the model where nothing is modeled as a spline.

```
# the null model will be a standard logistic regression
logit_null <- mgcv::gam(chd ~ lsbp + lcholest + age + smoker
                        , family=binomial(link="logit")
                        , data=fram
                      )
```

```
# now, conduct our test
(aov_tab <- anova(logit5, logit_null, test="Chisq"))
```

```
## Analysis of Deviance Table
##
## Model 1: chd ~ s(lsbp, k = 23, bs = "cr") + s(lcholest, k = 23, bs = "cr") +
##     s(age, k = 23, bs = "cr") + smoker
## Model 2: chd ~ lsbp + lcholest + age + smoker
##   Resid. Df Resid. Dev      Df Deviance Pr(>Chi)
## 1    1607.2     807.72
## 2    1610.0     814.76 -2.8228  -7.0493  0.06152 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print("P-value from test given by:")
```

```
## [1] "P-value from test given by:"
```

```
print(aov_tab$`Pr(>Chi)`)
```

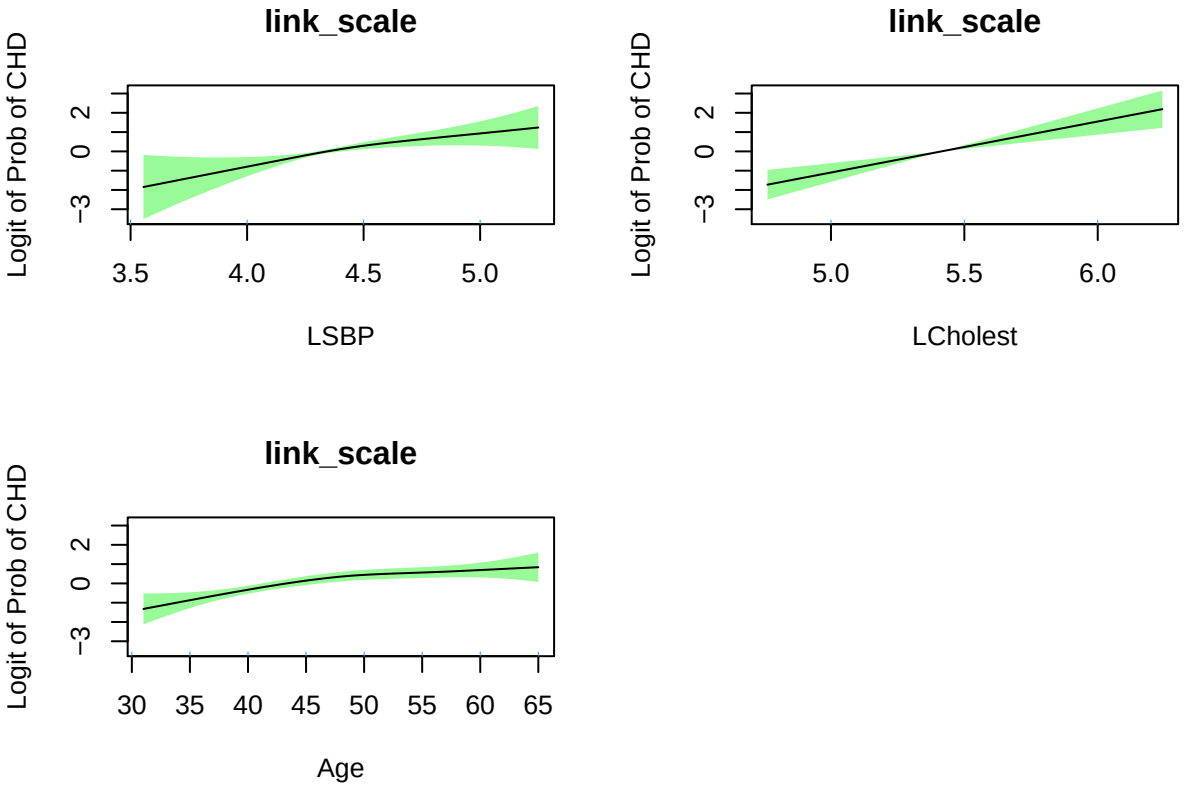
```
## [1]         NA 0.06151563
```

The model with 3 smoothed terms is not significantly different from the model that contains no smoothed terms.

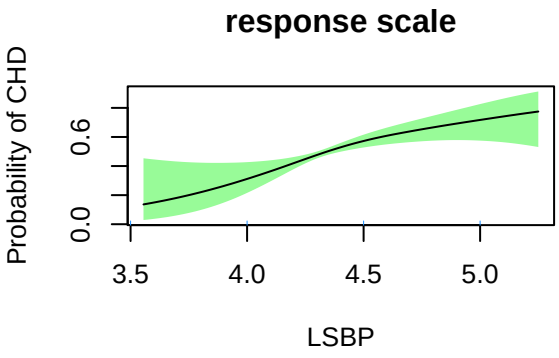
Question 7

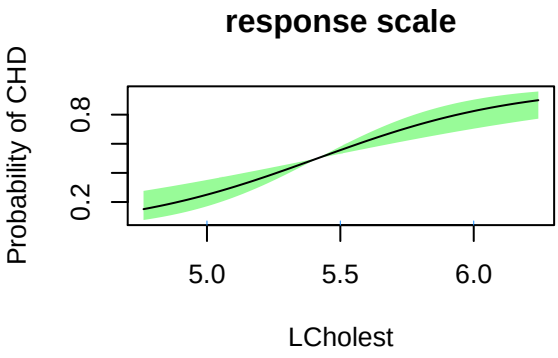
For the model that passed Question 5, plot the smooth fits. Note that I will produce plots in both the model link and response scales.

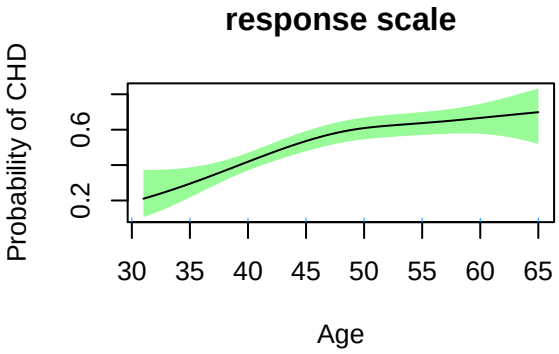
```r
color_plot_ls <- function(mod_obj,var_name, var_idx, x_label) {

  plot(mod_obj, shade=TRUE, shade.col="palegreen"
       , select=var_idx, ylab="Logit of Prob of CHD",
       xlab=x_label, main="link_scale", rug=FALSE)
  rug(fram$var_name, col="dodgerblue", quiet=TRUE)
}
par(mfrow=c(2,2))
color_plot_ls(mod_obj=logit5, var_name=lsbp, var_idx=1, x_label="LSBP")
color_plot_ls(mod_obj=logit5, var_name=lcholest, var_idx=2, x_label="LCholest")
color_plot_ls(mod_obj=logit5, var_name=age, var_idx=3, x_label="Age")
par(mfrow=c(1,1))
```




```r
color_plot_rs <- function(mod_obj, var_name, var_idx, x_label){

  plot(mod_obj,shade = TRUE,shade.col = "palegreen",
       trans = plogis, scale = FALSE, select = var_idx,
       ylab = "Probability of CHD",
       xlab=x_label,main = "response scale",rug = FALSE)
  rug(fram$var_name,col = "dodgerblue",quiet = TRUE)

}

par(mfrow=c(2,2))
color_plot_rs(mod_obj=logit5, var_name=lsbp, var_idx=1, x_label="LSBP")
color_plot_rs(mod_obj=logit5, var_name=lcholest, var_idx=2, x_label="LCholest")
color_plot_rs(mod_obj=logit5, var_name=age, var_idx=3, x_label="Age")
par(mfrow=c(1,1))
```







Question 8

For the regressions in questions 1-5, state the p-value of the smoking variable and state the odds ratio of it.

```
smok <- function(mod_obj, mod_name) {
  cat(mod_name)
  smry <- summary(mod_obj)
  cat("\nsmoker term p-value: ")
  cat(smry$p.pv['smoker'])
  cat("\nodds of CHD for a smoker vs. non-smoker: ")
  cat(exp(smry$p.coeff['smoker']))
  cat("\n----------\n")
}

smok(mod_obj=logit1, mod_name="Model from Question 1")
```

```
## Model from Question 1
## smoker term p-value: 0.01595596
## odds of CHD for a smoker vs. non-smoker: 1.830799
## ----------
```

```
smok(mod_obj=logit2, mod_name="Model from Question 2")
```

```
## Model from Question 2
## smoker term p-value: 0.01539144
## odds of CHD for a smoker vs. non-smoker: 1.83859
## ----------
```

```
smok(mod_obj=logit3, mod_name="Model from Question 3")
```

```
## Model from Question 3
## smoker term p-value: 0.01768211
## odds of CHD for a smoker vs. non-smoker: 1.813936
## ----------
```

```
smok(mod_obj=logit4, mod_name="Model from Question 4")
```

```
## Model from Question 4
## smoker term p-value: 0.0181234
## odds of CHD for a smoker vs. non-smoker: 1.808479
## ----------
```

```
smok(mod_obj=logit5, mod_name="Model from Question 5")
```

```
## Model from Question 5
## smoker term p-value: 0.0181234
## odds of CHD for a smoker vs. non-smoker: 1.808479
## ----------
```

There is not a dramatic difference between the models fit in Questions 1-5. This implies that the impact of smoking_status on a CHD does not differ much in the presence of different smoothing terms.

Question 9

For the model in Question 4 (all continuous terms modeled as splines), find a 95% confidence interval for the odds ratio on the smoker variable.

```
logit4_smry <- summary(logit4)
# grab the coefficient
beta_smok <- logit4_smry$p.coeff['smoker']
# grab the coefficient standard error
se_beta_smok <- logit4_smry$se['smoker']

lower <- exp(beta_smok - 1.96*se_beta_smok) # lower bound
upper <- exp(beta_smok + 1.96*se_beta_smok) # upper bound

cat("Odds ratio given by : ",exp(beta_smok), "\n")
```

```
## Odds ratio given by :  1.808479
```

```
cat("Odds ratio 95% confidence interval given by: ", lower, upper, "\n")
```

```
## Odds ratio 95% confidence interval given by:  1.106348 2.956211
```

The confidence interval just barely misses a value of 1.

Question 10

For the model in Question 5, pick any one of the variables modeled as a spline; state the odds ratio for the highest level of that variable to the lowest level.

I will conduct the analysis on the LSBP variable. From the fit object generated in Question 5, we can grab the max and min values of the spline coefficient. These will represent the log odds ratio of the highest and

lowest values of the smoothed lsbp term. We can then exponentiate them and take their ratio to determine the difference in odds ratio of pr(CHD) for the highest to lowest level of smoothed(LSBP).

```
lsbp_low <- logit5$coefficients['s(lsbp).1']
lsbp_high <- logit5$coefficients['s(lsbp).22']

or <- (exp(lsbp_high) / exp(lsbp_low))

print("Our odds ratio of interest is: ")
```

```
## [1] "Our odds ratio of interest is: "
```

```
print(or)
```

```
## s(lsbp).22
##   7.749413
```

Question 11

From the analysis conducted, it appears that the probability of CHD diagnosis is positively associated with blood pressure, cholesterol levels, age, and smoking status. The relationship between the continuous variables (blood pressure, cholesterol, age) and the CHD probability is linear, which means that if we increase any of these by a single unit, we will increase our CHD probability by a set amount, regardless of if we have high or low values of these inputs. For example, if I am age 40, and move to age 41, I will see the same probability increase in CHD diagnosis, as if I were 60 and moving to age 61, in the presence of the other variables.