Assignment 08 - STAT 689

Philip Anderson; panders2@tamu.edu

2/26/2018

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

Part 1

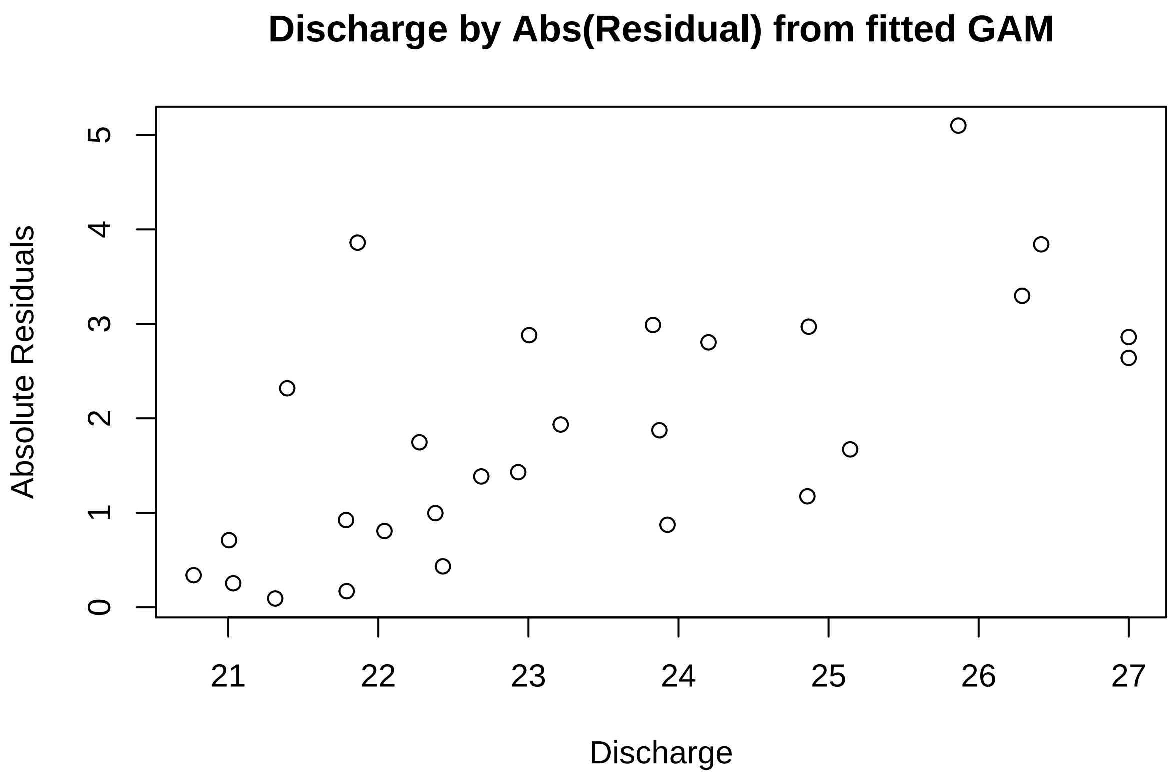
```
# read in the modified salinity dataset
salinity <- read.csv('/Users/panders2/Documents/schools/tamu/stat_689/homework/semiparametric-regression
names(salinity) <- tolower(names(salinity))
str(salinity)
```

```
## 'data.frame':    28 obs. of  4 variables:
##  $ salinity       : num  7.6 7.7 4.3 5.9 5 6.5 8.3 8.2 13.2 12.6 ...
##  $ lagged.salinity: num  8.2 7.6 4.6 4.3 5.9 5 6.5 8.3 10.1 13.2 ...
##  $ trend          : int  4 5 0 1 2 3 4 5 0 1 ...
##  $ discharge      : num  23 23.9 26.4 24.9 27 ...
```

Question 1

Fit a GAM where salinity is regressed on a smooth function of discharge. Plot discharge against the absolute residuals from this fit.

```
gam_mod <- mgcv::gam(salinity ~ s(discharge)
                   , data=salinity
                   )
abs_resid <- abs(gam_mod$residuals)
plot(salinity$discharge, abs_resid
     , main="Discharge by Abs(Residual) from fitted GAM"
     , xlab="Discharge"
     , ylab="Absolute Residuals"
     )
```



Note that we appear to have a positive trend. This reflects evidence of hetereoscedacticity.

Question 2

Calculate the ratio of the maximum predicted absolute residual to the minimum predicted absolute residual.

```
cat("\nMaximum Value: ", max(abs_resid))
```

```
##
## Maximum Value:  5.098749
```

```
cat("\nMinimum Value: ", min(abs_resid))
```

```
##
## Minimum Value:  0.09310846
```

```
cat("\nRatio of max to min: ", max(abs_resid) / min(abs_resid))
```

```
##
## Ratio of max to min:  54.7614
```

Question 3

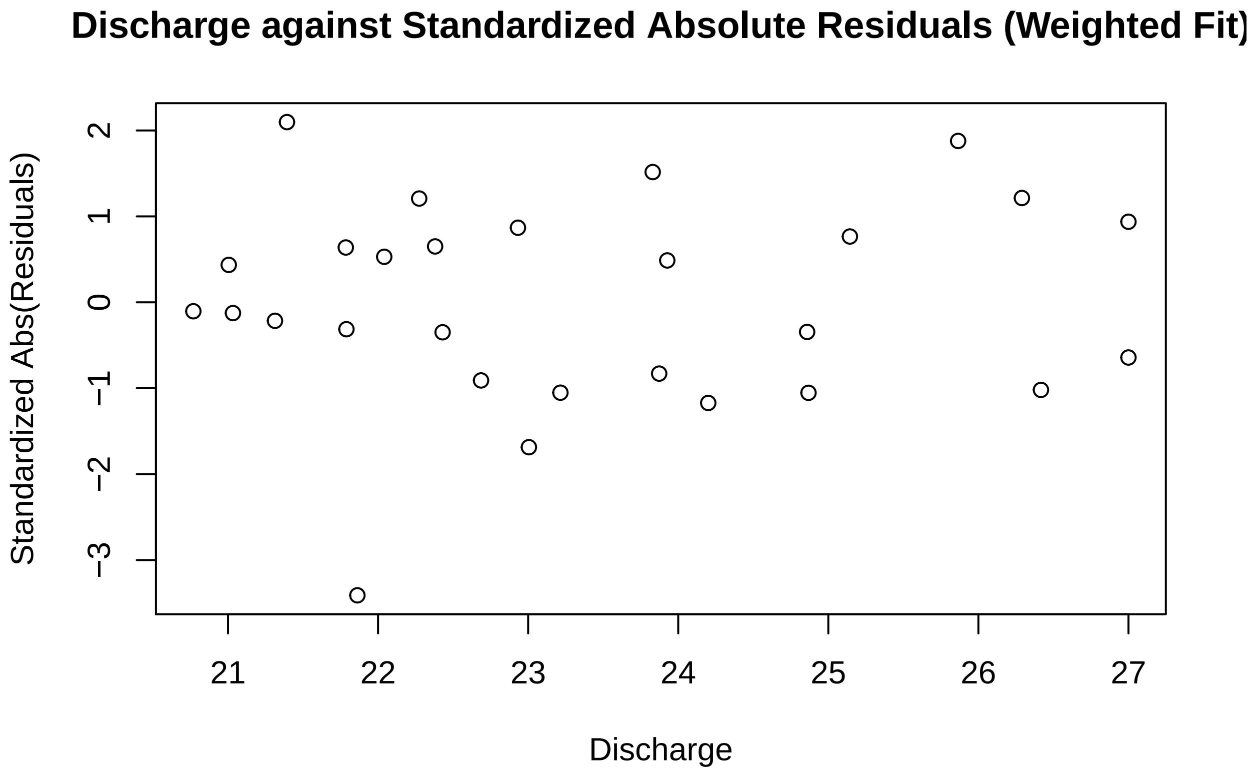
The ratio found in Question 2 is quite high (definitely higher than our heuristic of 3), indicating that we have evidence of heteroscedacticity. This will make the calculation of pointwise confidence intervals more involved, and render naive estimates incorrect.

Question 4

Conduct a weighted fit of the data. Plot discharge against the absolute residuals of the new model divided by the estimated standard deviations.

```r
gam_mod_resid <- mgcv::gam(abs_resid ~ s(salinity$discharge))
weight <- 1/(fitted(gam_mod_resid)^2)
gam_mod_weight <- mgcv::gam(salinity ~ s(discharge)
                            , data=salinity
                            , weight=weight
                            )
```

```r
plot(salinity$discharge, gam_mod_weight$residuals/fitted(gam_mod_resid)
     , main="Discharge against Standardized Absolute Residuals (Weighted Fit)"
     , xlab="Discharge"
     , ylab="Standardized Abs(Residuals)"
     )
```



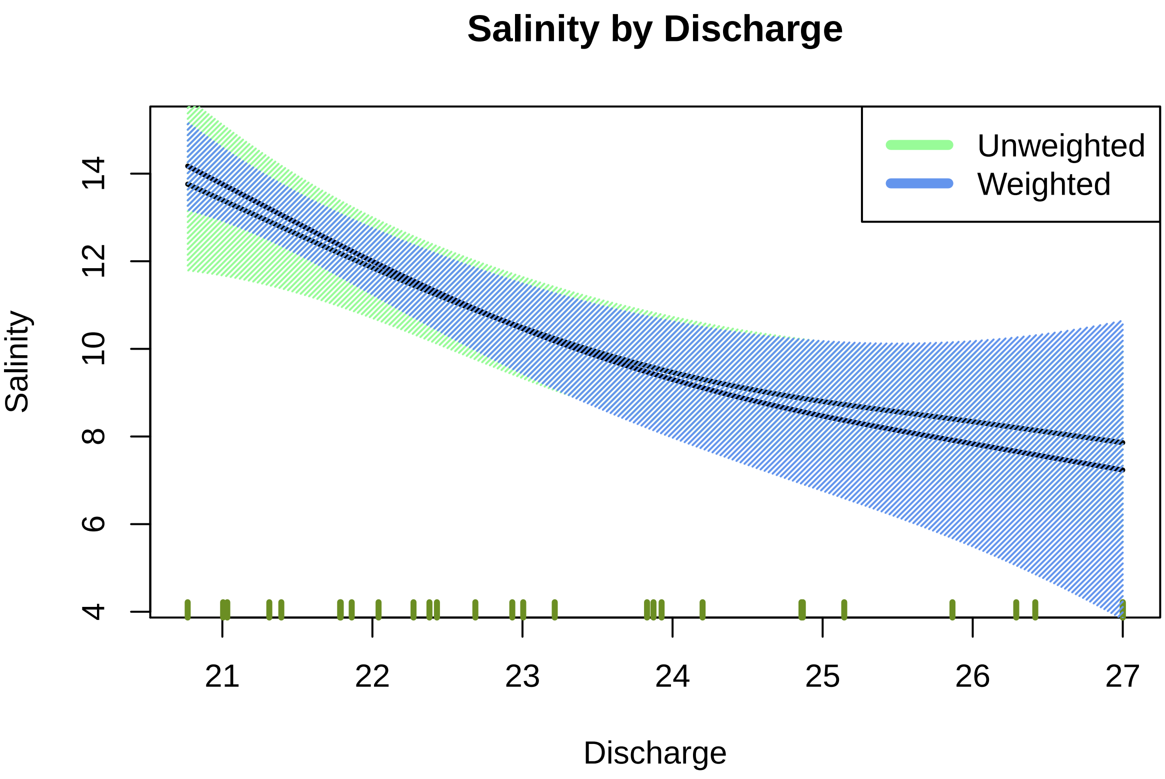
Question 5

Compute the 95% confidence intervals for the unweighted and weighted fits.

```r
x_var <- seq(from=min(salinity$discharge), to=max(salinity$discharge)
             , len=1000)
# predictions
unweight_pred <- predict(gam_mod
                         , newdata=data.frame(discharge=x_var)
                         , se.fit=TRUE)
```

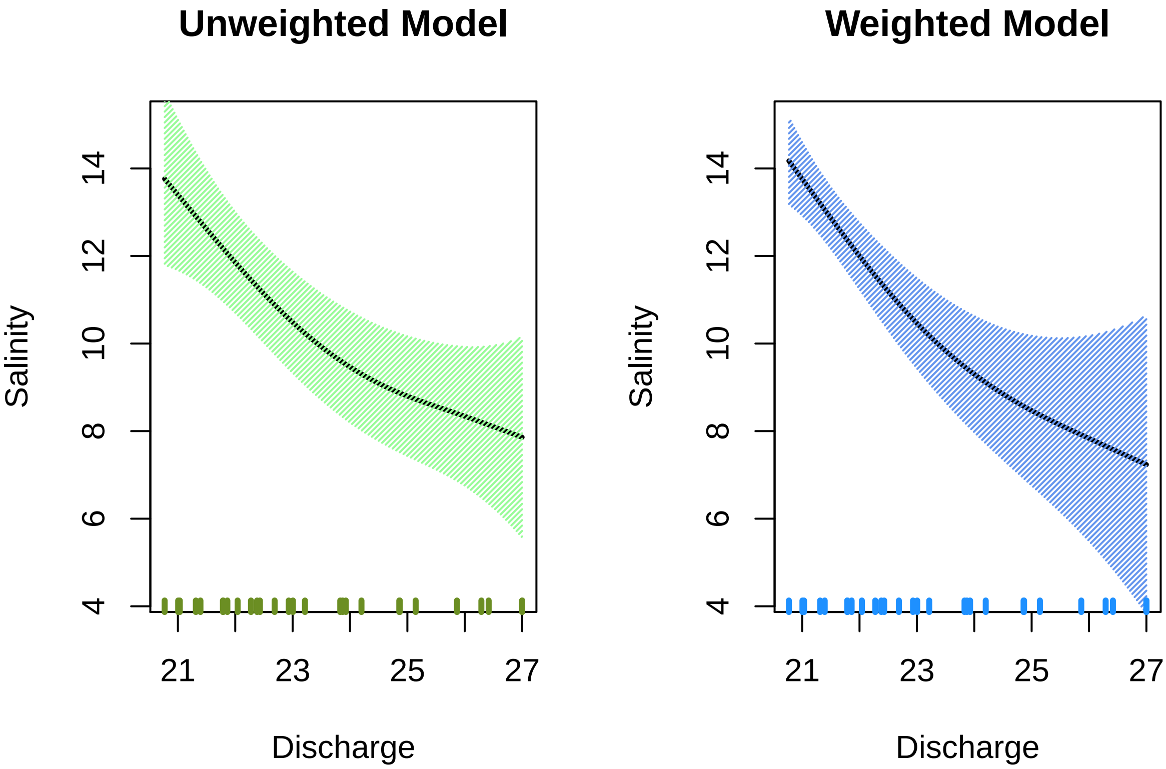
```
weight_pred <- predict(gam_mod_weight
                         , newdata=data.frame(discharge=x_var)
                         , se.fit=TRUE)

# confidence bounds
unweight_lower <- unweight_pred$fit - (1.96 * unweight_pred$se.fit)
unweight_upper <- unweight_pred$fit + (1.96 * unweight_pred$se.fit)

weight_lower <- weight_pred$fit - (1.96 * weight_pred$se.fit)
weight_upper <- weight_pred$fit + (1.96 * weight_pred$se.fit)
```

```
plot(salinity$discharge, salinity$salinity, type="n"
     , main="Salinity by Discharge"
     , xlab="Discharge"
     , ylab="Salinity"
     )

gam_plotter <- function(pred_obj, lower_bnd, upper_bnd, shader) {
  lines(x_var, pred_obj$fit, lwd=2.5)
  polygon(x=c(x_var, rev(x_var))
          , y=c(upper_bnd, rev(lower_bnd))
          , col=shader
          , border=NA
          , density=50
          )
}

# unweighted GAM first
gam_plotter(pred_obj=unweight_pred
            , lower_bnd=unweight_lower
            , upper_bnd=unweight_upper
            , shader="palegreen"
            )
rug(salinity$discharge, lwd=3, col="olivedrab")

# weighted GAM second
gam_plotter(pred_obj=weight_pred
            , lower_bnd=weight_lower
            , upper_bnd=weight_upper
            , shader="cornflowerblue"
            )
legend("topright"
       , c("Unweighted", "Weighted")
       , lty=c(1,1)
       , lwd=c(5,5)
       , col=c("palegreen", "cornflowerblue")
       )
```



I also found it helpful to see two side-by-side plots, due to the overlapping confidence bands in the above image.

```
par(mfrow=c(1,2))

plot(salinity$discharge, salinity$salinity, type="n"
     , main="Unweighted Model"
     , xlab="Discharge"
     , ylab="Salinity"
     )

# unweighted GAM first
gam_plotter(pred_obj=unweight_pred
            , lower_bnd=unweight_lower
            , upper_bnd=unweight_upper
            , shader="palegreen"
            )
rug(salinity$discharge, lwd=3, col="olivedrab")

plot(salinity$discharge, salinity$salinity, type="n"
     , main="Weighted Model"
     , xlab="Discharge"
     , ylab="Salinity"
     )

# weighted GAM second
gam_plotter(pred_obj=weight_pred
            , lower_bnd=weight_lower
            , upper_bnd=weight_upper
            , shader="cornflowerblue"
            )
rug(salinity$discharge, lwd=3, col="dodgerblue")
```



Question 6

Are there any systematic differences in the two plots?

It appears that the weighted fit has generally narrower confidence bands in the regions of x where we have more observations. In the regions where we do not have as many observations (x>25), it appears that the confidence bands are wider.

Part 2

Question 7

For the Framingham data, fit a factor-by-curve GAM where CHD is regressed on LSBP, LCholest, and age (all splines), but allow an interaction between s(age) and smoker.

```
# first, read in the data
fram <- read.csv("/Users/panders2/Documents/schools/tamu/stat_689/homework/semiparametric-regression/mi
names(fram) <- tolower(names(fram))
str(fram)
```

```
## 'data.frame':    1615 obs. of  12 variables:
##  $ obs     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age     : int  56 38 54 42 47 43 58 43 43 36 ...
##  $ sbp21   : int  110 118 140 122 136 138 108 130 116 150 ...
##  $ sbp22   : int  120 126 120 116 134 120 98 120 110 146 ...
##  $ sbp31   : int  136 120 125 115 130 118 112 112 108 122 ...
##  $ sbp32   : int  134 114 124 110 144 110 105 120 102 115 ...
##  $ smoker  : int  0 1 1 1 1 1 1 0 1 0 ...
##  $ cholest2: int  276 211 284 225 292 192 194 319 205 247 ...
##  $ cholest3: int  295 255 287 285 240 208 208 334 242 244 ...
##  $ chd     : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ lsbp    : num  4.32 4.24 4.35 4.19 4.45 ...
##  $ lcholest: num  5.65 5.45 5.65 5.54 5.58 ...
```

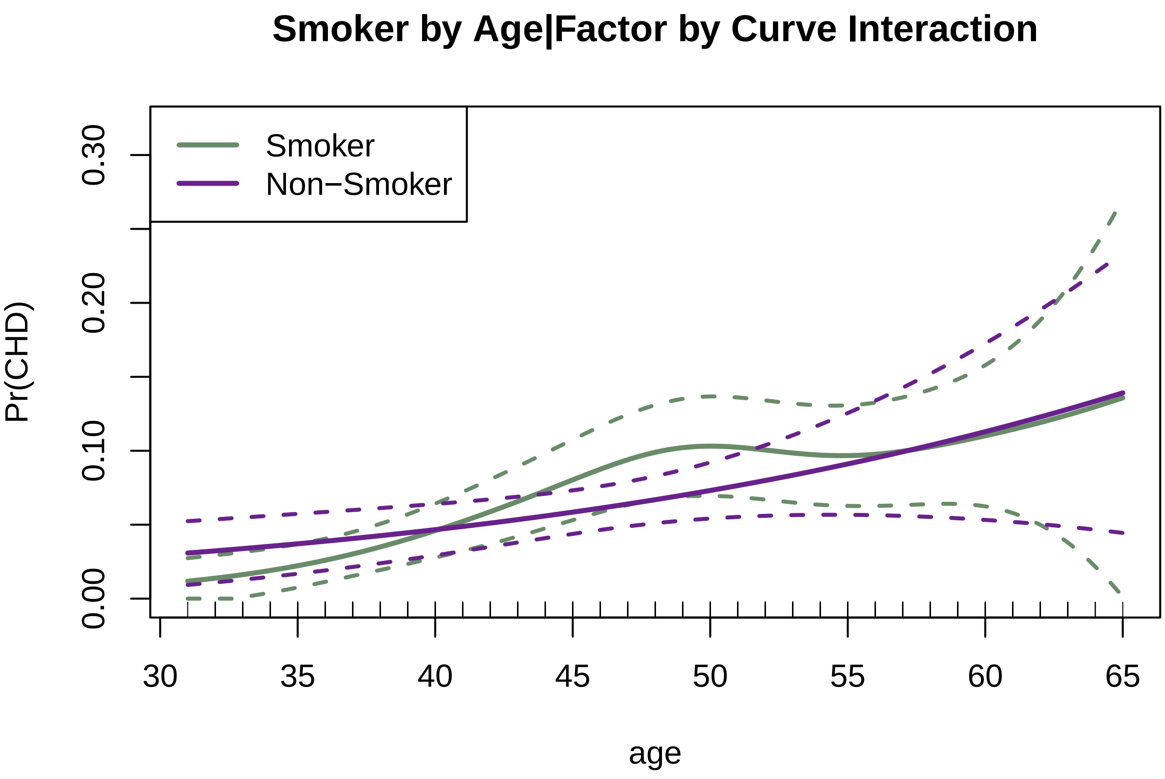
```
fram_mod <- mgcv::gam(chd ~ s(lsbp, bs="cr", k=23)
                       + s(lcholest, bs="cr", k=23)
                       + s(age, by=factor(smoker))
                      , family=binomial(link="logit")
                      , data=fram
                      )
```

```
ng <- 1001

lsbp_low <- min(fram$lsbp); lsbp_high <- max(fram$lsbp)
lsbpg <- seq(lsbp_low, lsbp_high, length=ng)

lcholest_low <- min(fram$lcholest); lcholest_high <- max(fram$lcholest)
lcholestg <- seq(lcholest_low, lcholest_high, length=ng)

age_low <- min(fram$age); age_high <- max(fram$age)
ageg <- seq(age_low, age_high, length=ng)

lsbp_avg <- rep(mean(fram$lsbp), ng)
lcholest_avg <- rep(mean(fram$lcholest), ng)
age_avg <- rep(mean(fram$age), ng)

smoke_n <- as.factor(rep('0', ng))
smoke_y <- as.factor(rep('1', ng))
```

```
# plot as a function of age
fAgeSmoke <- predict(fram_mod, type="response"
                     , newdata=data.frame(
                      lsbp=lsbp_avg
                      , lcholest=lcholest_avg
                      , age=ageg
                      , smoker=smoke_y
                      )
                     , se.fit=TRUE
                     )

fAgeNoSmoke <- predict(fram_mod, type="response"
                     , newdata=data.frame(
                      lsbp=lsbp_avg
                      , lcholest=lcholest_avg
                      , age=ageg
                      , smoker=smoke_n
                      )
                     , se.fit=TRUE
                     )
```

```
# SMOKER==1
prob_age_smoke_g <- fAgeSmoke$fit
sd_age_smoke_g <- fAgeSmoke$se.fit

lwr_age_smoke <- prob_age_smoke_g - 2*sd_age_smoke_g
upr_age_smoke <- prob_age_smoke_g + 2*sd_age_smoke_g
# trim
lwr_age_smoke[lwr_age_smoke<0] <- 0
upr_age_smoke[upr_age_smoke>1] <- 1

# SMOKER==0
prob_age_nosmoke_g <- fAgeNoSmoke$fit
sd_age_nosmoke_g <- fAgeNoSmoke$se.fit

lwr_age_nosmoke <- prob_age_nosmoke_g - 2*sd_age_nosmoke_g
upr_age_nosmoke <- prob_age_nosmoke_g + 2*sd_age_nosmoke_g
# trim
lwr_age_nosmoke[lwr_age_nosmoke<0] <- 0
upr_age_nosmoke[upr_age_nosmoke>1] <- 1
```

```
plot(0, typ="n", xlim=range(ageg), ylim=c(0, 0.32), xlab="age", ylab="Pr(CHD)"
     , main="Smoker by Age|Factor by Curve Interaction")
rug(fram$age, col="black", quiet=TRUE)
# smoker
lines(ageg,prob_age_smoke_g, lwd=2.5, col="darkseagreen4")
lines(ageg, lwr_age_smoke, lwd=2, lty=2, col="darkseagreen4")
lines(ageg, upr_age_smoke, lwd=2, lty=2, col="darkseagreen4")
# non-smoker
lines(ageg,prob_age_nosmoke_g, lwd=2.5, col="darkorchid4")
lines(ageg, lwr_age_nosmoke, lwd=2, lty=2, col="darkorchid4")
lines(ageg, upr_age_nosmoke, lwd=2, lty=2, col="darkorchid4")

legend("topleft"
       , c("Smoker", "Non-Smoker")
       , lty=c(1,1)
       , lwd=c(2.5,2.5)
       , col=c("darkseagreen4", "darkorchid4")
       )
```



Question 8

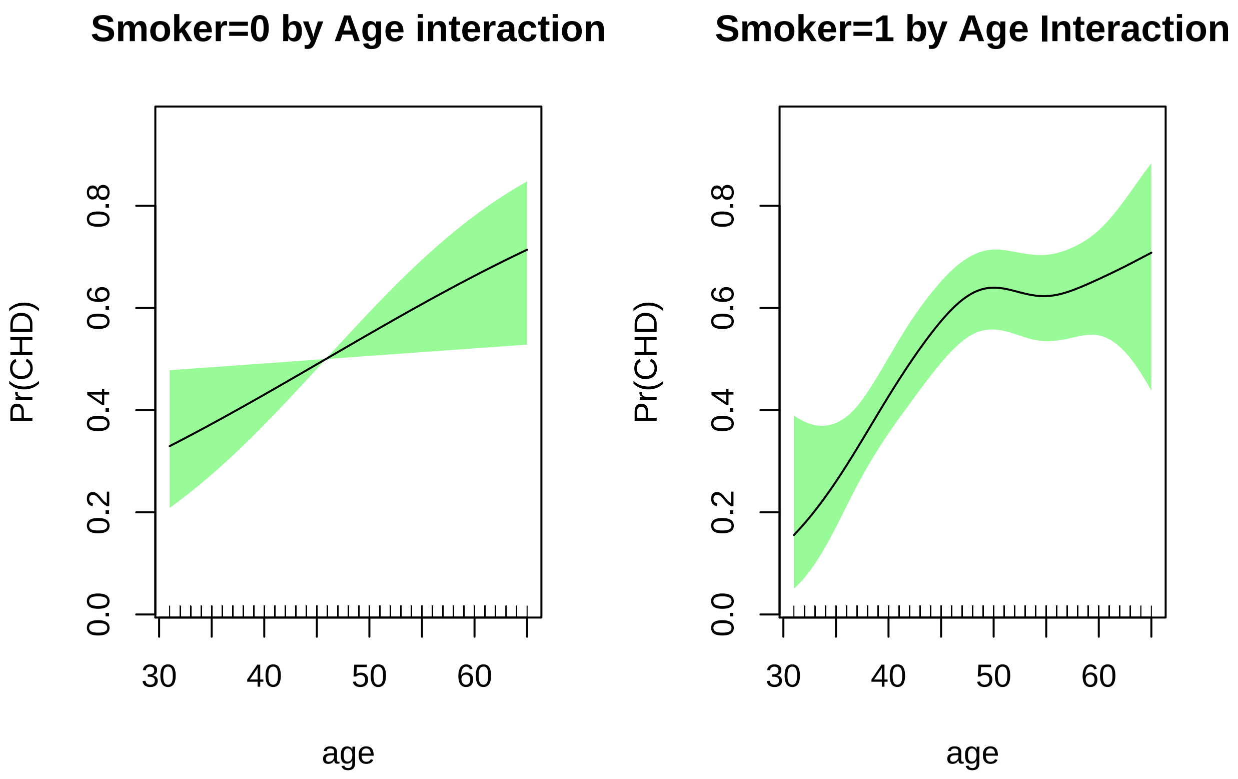
In Question 7, do you think there is any statistical evidence that the risk of developing CHD over 8 years depends on whether or not you are a smoker?

Based on the confidence intervals in the above graphic, it does not appear that we have statistical evidence that the risk of developing CHD depends on whether you are a smoker. In both cases, the trend is positive, but the lines overlap in several places, as the confidence bands. See the Appendix for relevant partial dependence plots that lead to a similar conclusion.

Appendix

Plot the fitted splines to age for smokers and non-smokers.

```
par(mfrow=c(1,2))
plot.gam(fram_mod, shade = TRUE, shade.col="palegreen", select=3
         , trans=plogis
         , ylab="Pr(CHD)"
         , main="Smoker=0 by Age interaction")
plot.gam(fram_mod, shade = TRUE, shade.col="palegreen", select=4
         , trans=plogis
         , ylab="Pr(CHD)"
         , main="Smoker=1 by Age Interaction"
         )
```



```r
par(mfrow=c(1,1))
```