Assignment 08 - STAT 689

Philip Anderson; panders2@tamu.edu

2/26/2018

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

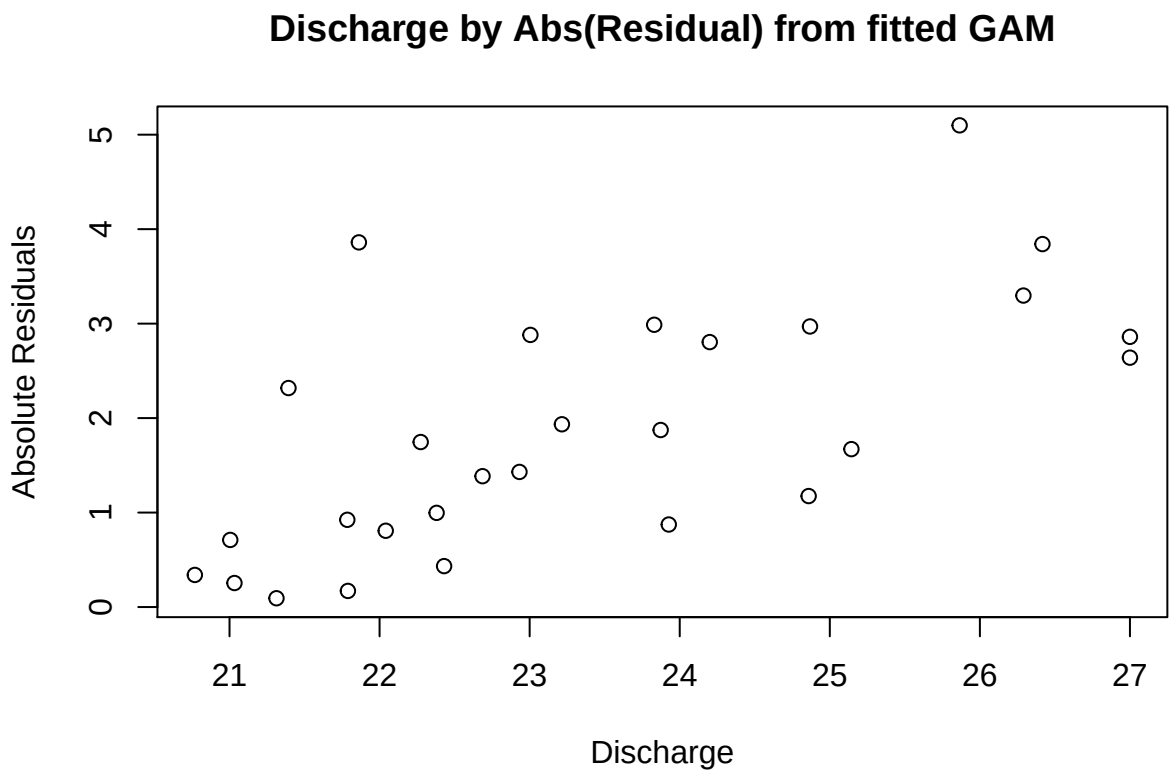
Part 1

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

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.

```
# following the solutions from hw05
x1 <- fitted(gam_mod)
y1 <- abs(gam_mod$residuals)
gam_mod_abs <- mgcv::gam(y1 ~ s(x1))

cat("\nMaximum Value: ", max(abs(fitted(gam_mod_abs))))
```

```
##
## Maximum Value:  3.09951
```

```
cat("\nMinimum Value: ", min(abs(fitted(gam_mod_abs))))
```

```
##
## Minimum Value:  0.4075468
```

```
cat("\nRatio of max to min: ", max(abs(fitted(gam_mod_abs))) / min(abs(fitted(gam_mod_abs))))
```

```
##
## Ratio of max to min:  7.605286
```

Question 3

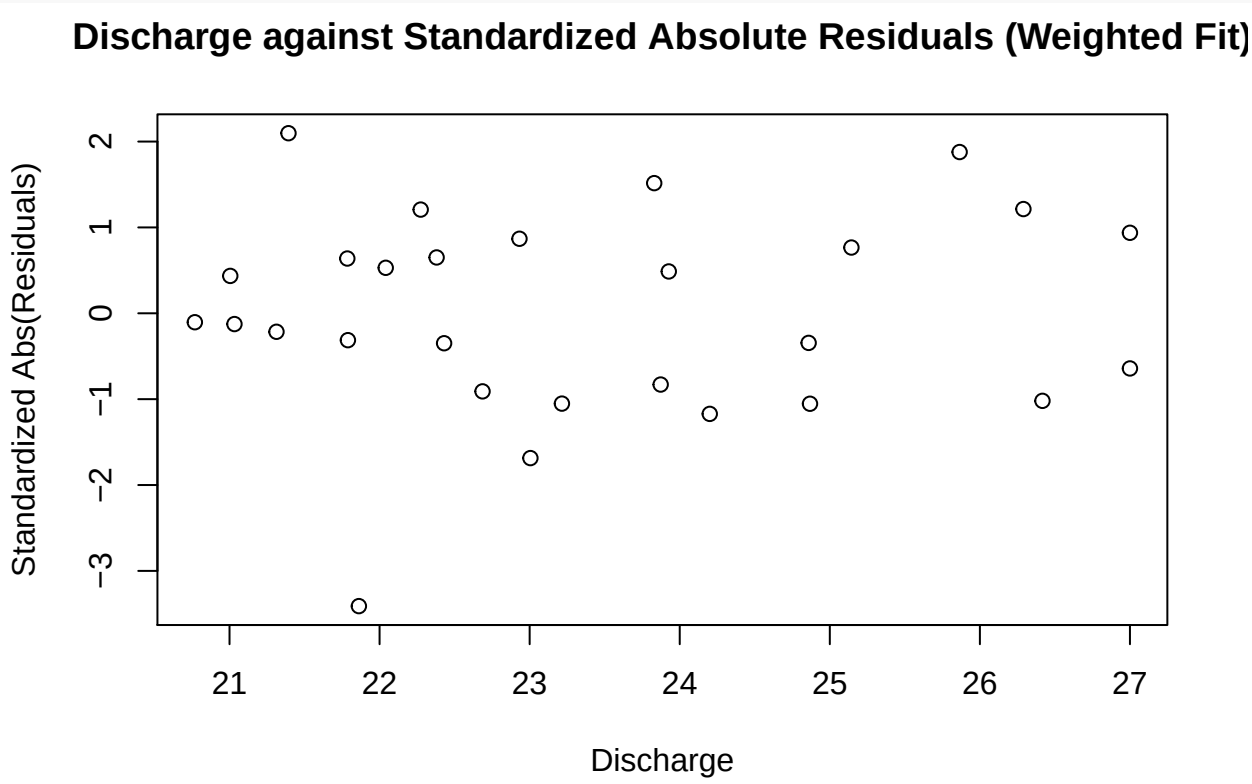
The ratio found in Question 2 is higher than our heuristic of 3, indicating that we have evidence of heteroscedacticity. This will make the calculation of pointwise confidence intervals more involved, and render naive estimates invalid.

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.

```
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
                            )
```

```
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)"
     )
```



The data points above no longer appear to display any sort of meaningful trend. This indicates that we may have corrected for the hetereoscedacticity we saw earlier.

Question 5

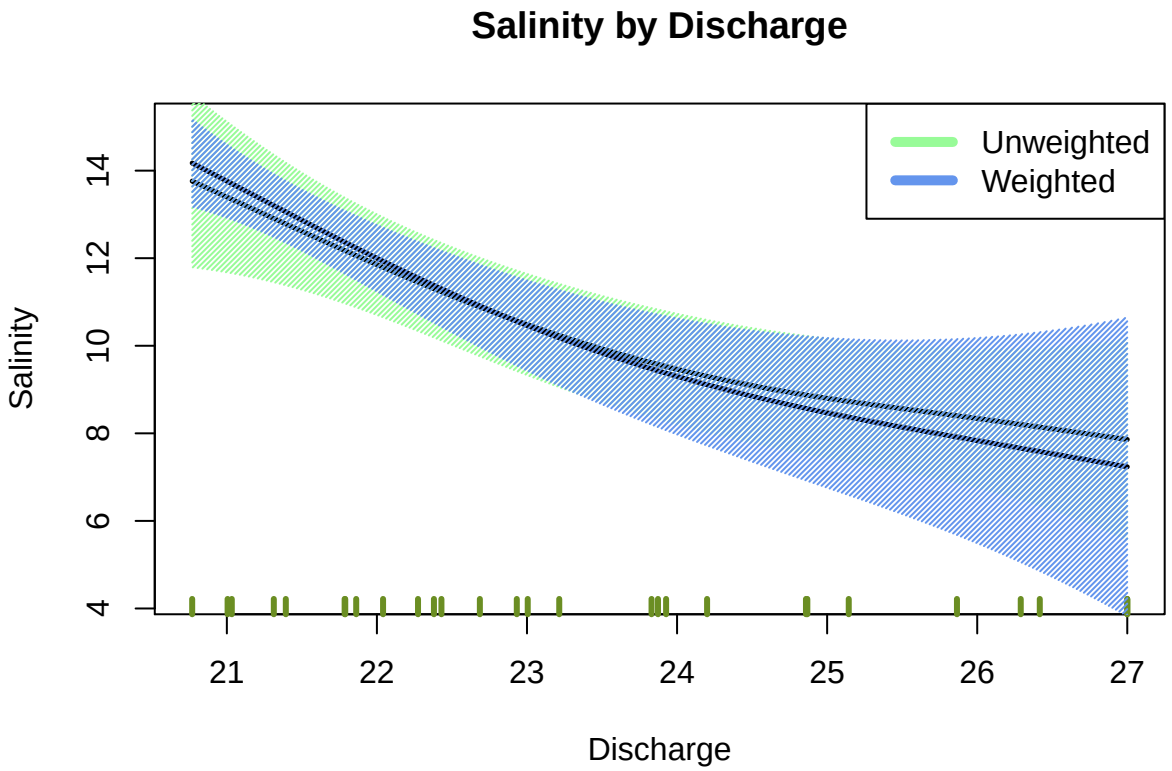
Compute the 95% confidence intervals for the unweighted and weighted fits.

```
# range variable
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)
```

The below will plot the weighted and unweighted fits with confidence bounds within the same graphic.

```
# plot wireframe
plot(salinity$discharge, salinity$salinity, type="n"
     , main="Salinity by Discharge"
     , xlab="Discharge"
     , ylab="Salinity"
     )
# line with CI function
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. These are located within the Appendix.

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 for the weighted fit.

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.

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

```r
# model fit
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
                       )
```

Now that the model has been fit, begin the lengthy plotting process.

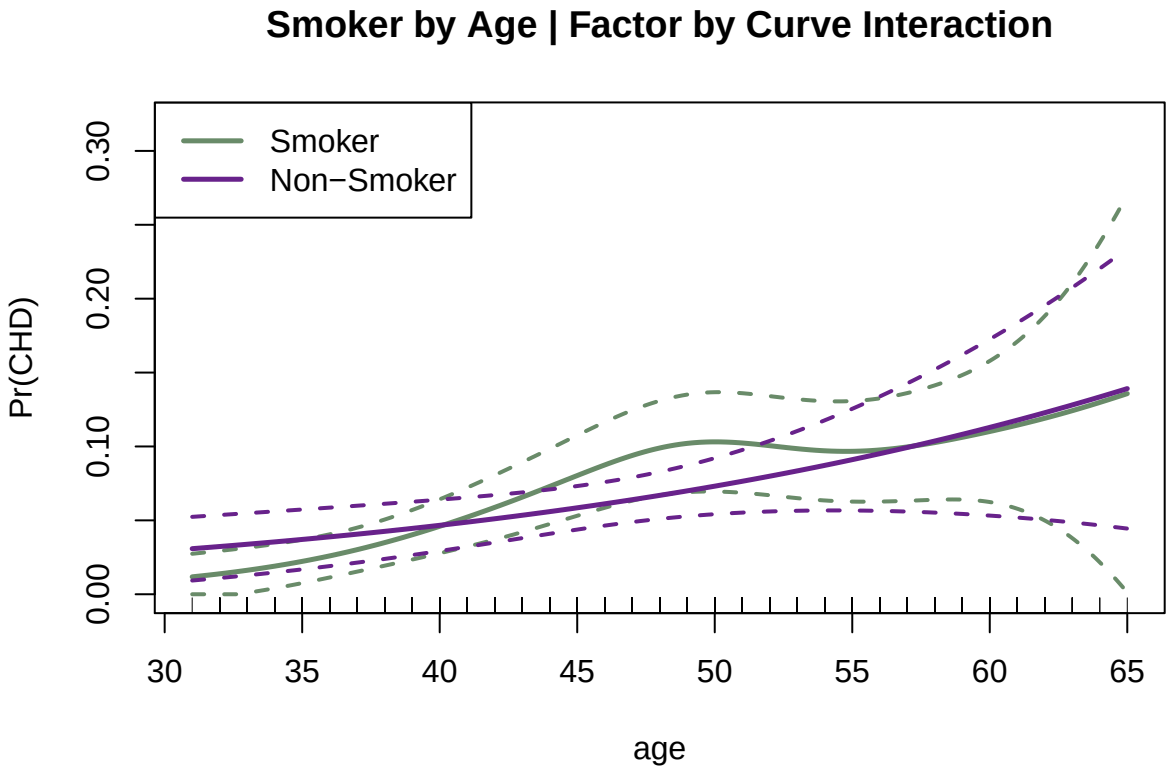
```r
# graphing parameters
ng <- 1001 # grid elements

# ranges for specific variables
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)

# mean values for conditional variable evaluation
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))
```

```r
# grab prediction values for graphing
# we will be plotting 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
```

```
                      )
```

```
# develop our core fit line values and associated confidence bounds
# 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
```

```
# finally, we plot
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?

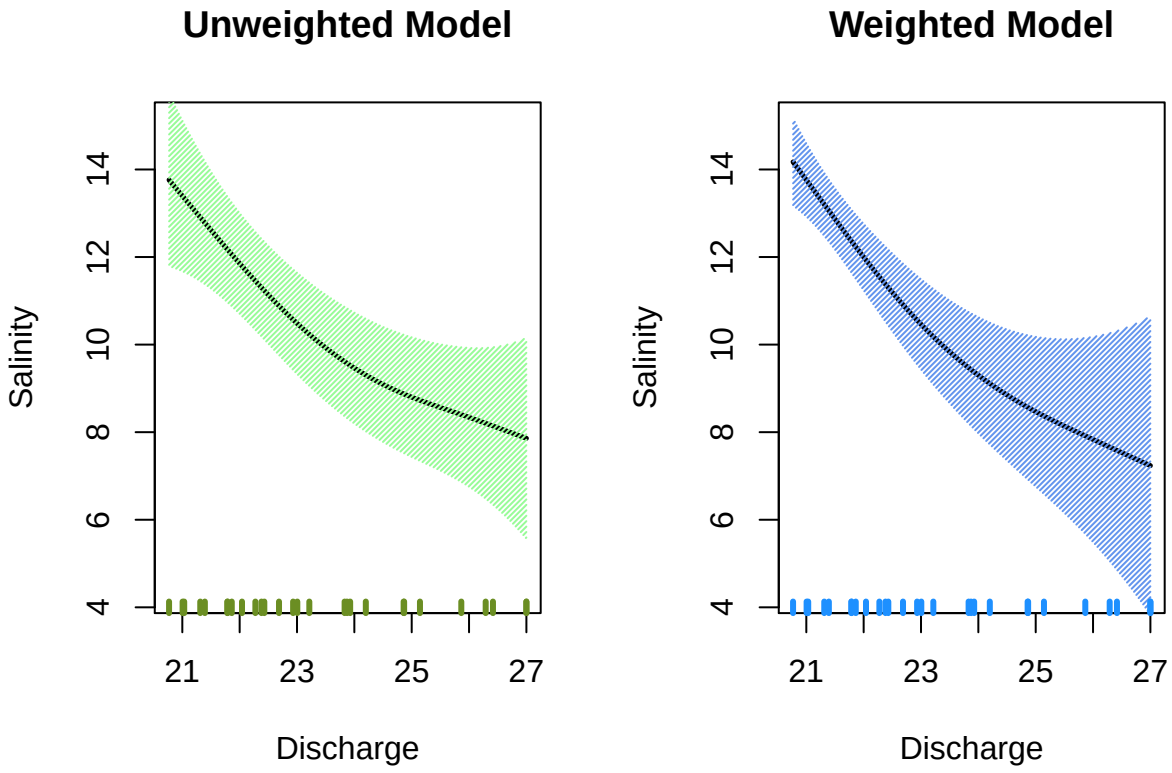
Based on the confidence intervals in the above graphic, it does not appear that we have statistical evidence that the risk of developing CHD depends on whether you are a smoker. In both cases, the trend is positive, but the lines overlap in several places, as do the confidence bands. See the Appendix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

Appendix

Question 5 Expanded

```
# similar code as above, just split out.
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")
```

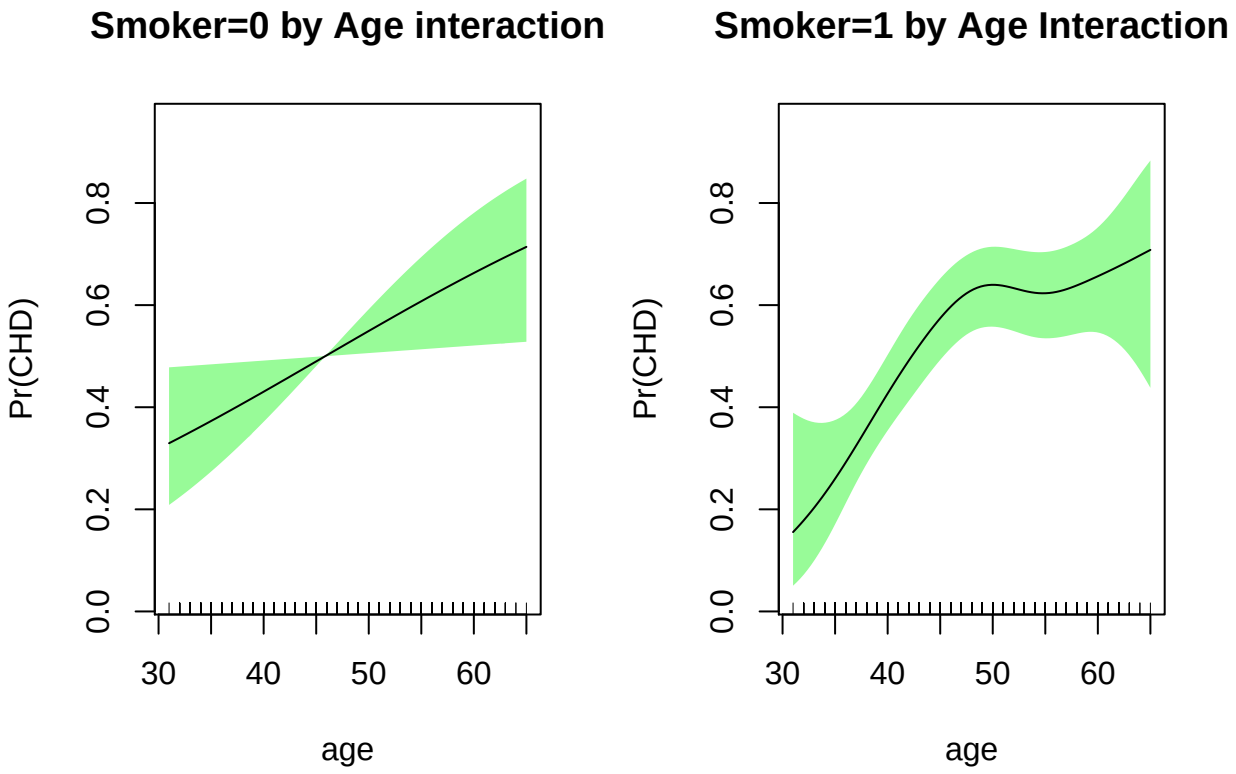


Question 8 Expanded

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"
         )
```




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