Homework 6

Skylar Liu

2024-02-09

Homework #6, Stat 660, Spring 2024, Due February 13

1. Fit a multiple linear regression of LSBP (Y) on Lcholest and smoker using mgcv::gam. Since the smoking variable is binary, this is an ordinary ANCOVA without an interaction. You will notice that the Rsquared is quite low. Produce a table of estimates, standard errors and p-values.

```
# 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 a multiple linear gam to the data
lm_LSBP = gam(LSBP ~ Lcholest + factor(Smoker), data = data)
summary(lm_LSBP)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## LSBP ~ Lcholest + factor(Smoker)
##
## Parametric coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.55569    0.17029  20.880  < 2e-16 ***
## Lcholest          0.15540    0.03140   4.949 8.22e-07 ***
## factor(Smoker)1 -0.03796    0.01251  -3.034  0.00246 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.0191   Deviance explained = 2.04%
## GCV = 0.044487  Scale est. = 0.044404  n = 1615
```

2. Just for fun (not graded but required) Do a little bit of a web search about whether smokers have higher or lower blood pressure than nonsmokers. Does the analysis in (1) agree? Just give a coherent answer.

   I found the general results to be that smokers have lower blood pressure than nonsmokers. This is consistent with the analysis in (1) since the estimate for "Smoker" is negative.

3. Just for fun (not graded but required) In Question 1, there is a subtle statistical interpretation of what I am asking you to do, because your analysis also includes the transformed cholesterol variable. I am curious if you can produce the correct terminology for what that subtle interpretation is. Please, less than 15 words. This is a test of your background, but I want to make sure you know how to report results appropriately.

   We interpret log transformations as percentages.

4. Do the same thing as in Question 1 but add an interaction between Lcholest and smoker.

```r
# Fit a multiple linear gam to the data with interaction
lm_interact = gam(LSBP ~ Lcholest + factor(Smoker) + Lcholest*factor(Smoker), data = data)
summary(lm_interact)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## LSBP ~ Lcholest + factor(Smoker) + Lcholest * factor(Smoker)
##
## Parametric coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.55075    0.33525  10.591   <2e-16 ***
## Lcholest                  0.15632    0.06191   2.525   0.0117 *
## factor(Smoker)1          -0.03130    0.38907  -0.080   0.9359
## Lcholest:factor(Smoker)1 -0.00123    0.07184  -0.017   0.9863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.0185   Deviance explained = 2.04%
## GCV = 0.044542  Scale est. = 0.044432  n = 1615
```

5. Now run a semiparametric regression using mgcv, one that is the semiparametric version of ANCOVA without an interaction.

```
# Fit a semiparametric multiple linear gam to the data
gam_LSBP = gam(LSBP ~ s(Lcholest, bs="cr", k=20) + factor(Smoker), method = "REML", data = data)
summary(gam_LSBP)
```

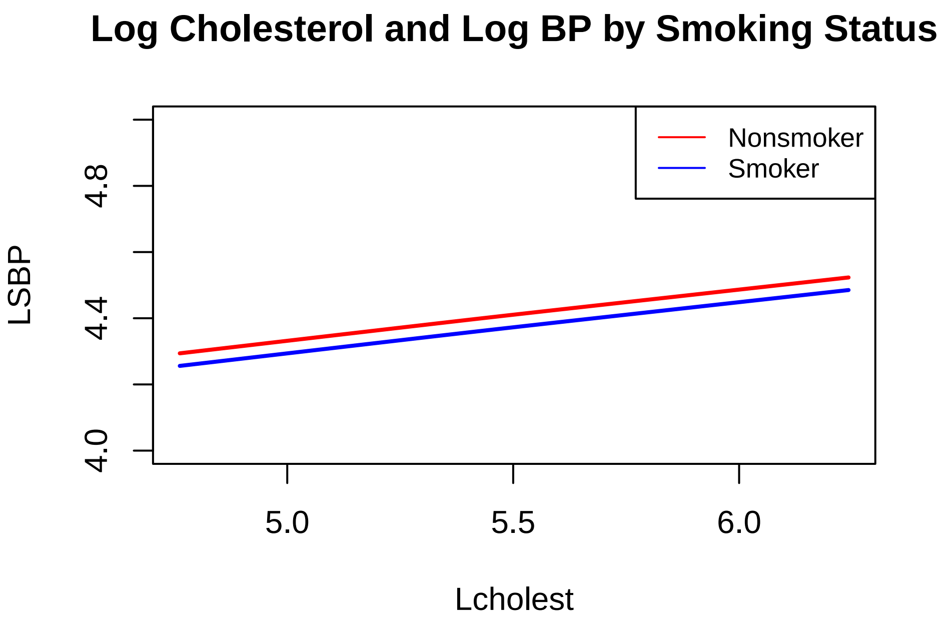
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## LSBP ~ s(Lcholest, bs = "cr", k = 20) + factor(Smoker)
##
## Parametric coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.39713    0.01100 399.734  < 2e-16 ***
## factor(Smoker)1 -0.03799    0.01251  -3.036  0.00244 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df     F  p-value
## s(Lcholest) 1.066  1.129 22.23 2.21e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0192   Deviance explained = 2.04%
## -REML = -214.81  Scale est. = 0.044402  n = 1615
```

6. In Question 5, display a plot of the two lines, but without the data.

```
# Non-Smoker
ng = 1615
xaxis = seq(min(Lcholest),max(Lcholest),length = ng)
yaxis = predict(gam_LSBP, newdata = data.frame(
  Lcholest = xaxis,
  Smoker = rep(0,ng)))
plot(xaxis, yaxis, type = "n",
     main = "Log Cholesterol and Log BP by Smoking Status",
     xlab = "Lcholest",
     ylab = "LSBP", ylim = c(4, 5))
lines(xaxis,yaxis, col='red', lwd=2)

# Smoker
xaxis = seq(min(Lcholest),max(Lcholest),length = ng)
yaxis = predict(gam_LSBP, newdata = data.frame(
  Lcholest = xaxis,
  Smoker = rep(1,ng)))
lines(xaxis,yaxis, col='blue', lwd=2)

legend("topright", legend = c("Nonsmoker", "Smoker"),
       col=c("red", "blue"),lty=1, cex=0.8)
```

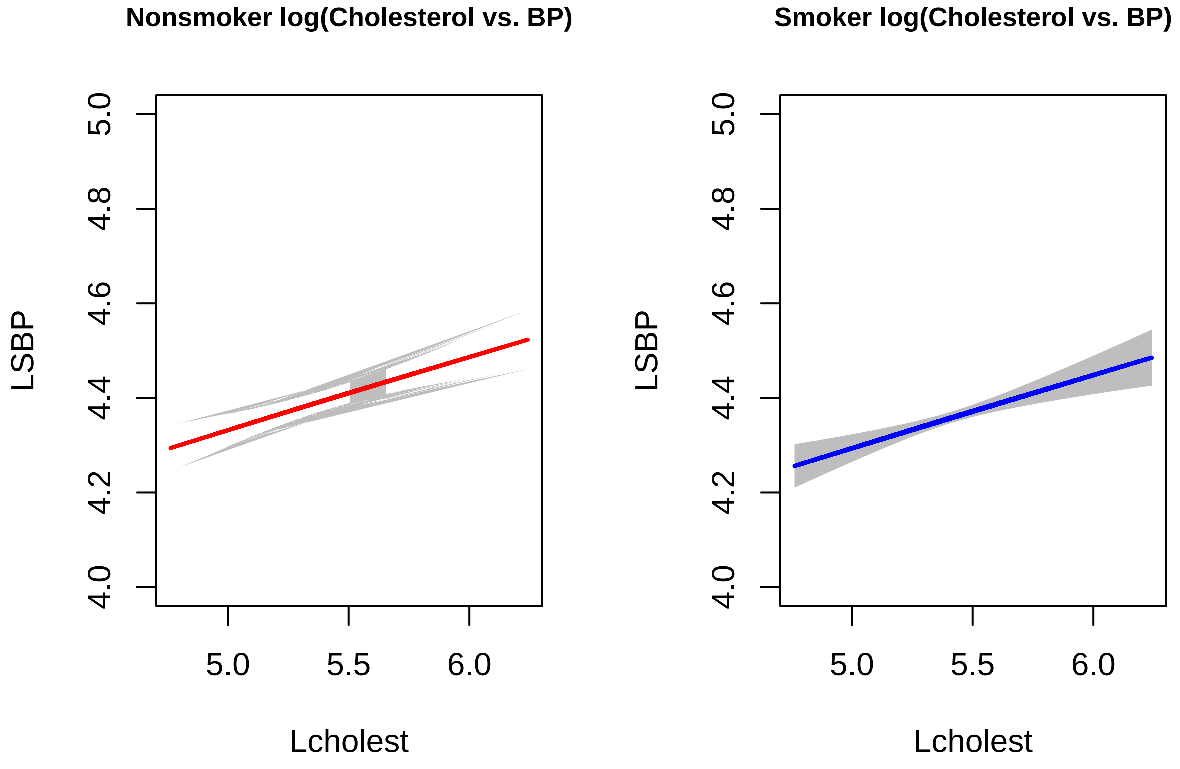


7. In Question 5, display a plot of the two lines but in one graph with 2 columns": par(mfrow(1,2) does this, I believe, and also show their pointwise 95% confidence intervals.

```
par(mfrow = c(1,2))
ord = sort(data$Lcholest, index.return = T)$ix

# Non-Smoker
ng = 1615
xaxis = seq(min(Lcholest),max(Lcholest),length = ng)
yaxis = predict(gam_LSBP, newdata = data.frame(
  Lcholest = Lcholest,
  Smoker = rep(0,ng)), se = T)
plot(data$Lcholest, data$LSBP, type = "n",
     main = "Nonsmoker log(Cholesterol vs. BP)",
     cex.main=0.8,
     xlab = "Lcholest",
     ylab = "LSBP", ylim = c(4, 5))
upper = yaxis$fit + (1.96 * yaxis$se.fit)
lower = yaxis$fit - (1.96 * yaxis$se.fit)
polygon(x=c(Lcholest, rev(Lcholest)), y=c(upper, rev(lower)),
        col="gray", border=NA)
lines(Lcholest, yaxis$fit, col='red', lwd=2)

# Smoker
xaxis = seq(min(Lcholest),max(Lcholest),length = ng)
yaxis = predict(gam_LSBP, newdata = data.frame(
  Lcholest = xaxis,
  Smoker = rep(1,ng)), se = T)
plot(data$Lcholest, data$LSBP, type = "n",
     main = "Smoker log(Cholesterol vs. BP)",
     cex.main=0.8,
     xlab = "Lcholest",
     ylab = "LSBP", ylim = c(4, 5))
upper = yaxis$fit + (1.96 * yaxis$se.fit)
lower = yaxis$fit - (1.96 * yaxis$se.fit)
polygon(x=c(xaxis, rev(xaxis)), y=c(upper, rev(lower)),
        col="gray", border=NA)
lines(xaxis[ord], yaxis$fit[ord], col='blue', lwd=2)
```



8. Run the semiparametric version of ANCOVA but with an interaction. Does it look like there is an interaction? Cite p-values and estimates

```
# Fit a semiparametric multiple linear gam to the data with an interaction
gam_interact = gam(LSBP ~ s(Lcholest, bs="cr", k=20, by = Smoker) + factor(Smoker),
               method = "REML", data = data)
summary(gam_interact)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## LSBP ~ s(Lcholest, bs = "cr", k = 20, by = Smoker) + factor(Smoker)
##
```

```
## Parametric coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       4.396790   0.011021 398.933   <2e-16 ***
## factor(Smoker)1 -0.009929   0.014125  -0.703    0.482    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df     F  p-value    
## s(Lcholest):Smoker   1  1.001 18.04 2.32e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Rank: 21/22
## R-sq.(adj) =  0.0153   Deviance explained = 1.65%
## -REML = -211.22  Scale est. = 0.04458   n = 1615
```

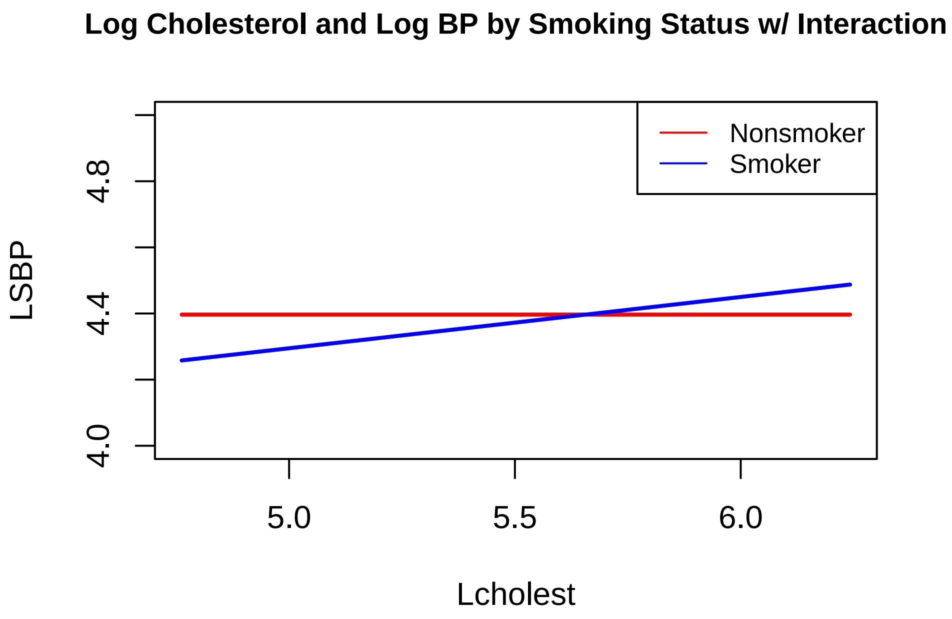
The interaction is significant at the 0.05 level, with a p-value of 2.32x10^-05.

9. In Question 8, display the fits but without the data points.

```r
# Non-Smoker
ng = 1615
xaxis = seq(min(Lcholest),max(Lcholest),length = ng)
yaxis = predict(gam_interact, newdata = data.frame(
  Lcholest = xaxis,
  Smoker = rep(0,ng)))
plot(xaxis, yaxis, type = "n",
     main = "Log Cholesterol and Log BP by Smoking Status w/ Interaction",
     cex.main=0.9,
     xlab = "Lcholest",
     ylab = "LSBP", ylim = c(4, 5))
lines(xaxis,yaxis, col='red', lwd=2)

# Smoker
xaxis = seq(min(Lcholest),max(Lcholest),length = ng)
yaxis = predict(gam_interact, newdata = data.frame(
  Lcholest = xaxis,
  Smoker = rep(1,ng)))
lines(xaxis,yaxis, col='blue', lwd=2)

legend("topright", legend = c("Nonsmoker", "Smoker"),
       col=c("red", "blue"),lty=1, cex=0.8)
```



10. What does having an interaction mean in the case when the factors are binary?

    There exists a relationship when the factor is one value vs the other value; i.e. the outcome depends on one factor value vs the other.

11. Now run an analysis of whether the two lines (smoker versus nonsmoker) are statistically significantly different when there is no interaction. You might remember that we contrasted the Srod district with the other Warsaw districts at one point. This can be your guide. Luckily, here smoker is a binary variable, so the reference population is the nonsmokers.

```r
# anova of no interaction model
anova(gam_LSBP)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## LSBP ~ s(Lcholest, bs = "cr", k = 20) + factor(Smoker)
##
## Parametric Terms:
##                df     F p-value
## factor(Smoker)  1 9.218 0.00244
##
## Approximate significance of smooth terms:
##               edf Ref.df     F  p-value
## s(Lcholest) 1.066  1.129 22.23 2.21e-06
```

The p-value for Smoker with no interaction is 0.00244, which is statistically significant at the 0.05 level.