

Assignment_06 - STAT 689

Philip Anderson; panders2@tamu.edu

2/14/2018

```
# import third-party modules
library("HRW")
library("tidyverse")
library("mgcv")
library("nlme")
```

Read in our data

```
fram <- read.csv("/Users/panders2/Documents/schools/tamu/stat_689/homework/semiparametric-regression/hw
names(fram) <- tolower(names(fram))
dim(fram)
```

```
## [1] 1615  10
```

Create new variables for the systolic blood pressure readings and the two cholesterol measurements.

```
# systolic blood pressure first
fram$lsbp <- log(((fram$sbp21 + fram$sbp22 + fram$sbp31 + fram$sbp32) / 4) - 50)
# cholesterol measurements second
fram$lcholest <- log((fram$cholest2 + fram$cholest3) / 2)
```

Keep only the variables that we will be working with directly and make sure everything seems reasonable.

```
fram2 <- fram %>%
  dplyr::select(chd, age, smoker, lsbp, lcholest)
head(fram2)
```

```
##   chd age smoker    lsbp lcholest
## 1   0  56      0 4.317488 5.654242
## 2   0  38      1 4.241327 5.451038
## 3   0  54      1 4.347047 5.654242
## 4   0  42      1 4.185860 5.541264
## 5   0  47      1 4.454347 5.583496
## 6   0  43      1 4.269697 5.298317
```

Question 1

Fit a multiple linear regression of LSBP on lcholest and smoker via “lm” function. Produce a estimates, standard errors, and p-values.

```
# fit model
lin_mod <- lm(lsbp ~ smoker + lcholest
              , data=fram2
              )
# produce summary
summary(lin_mod)
```

```
##
## Call:
```

```
## lm(formula = lsbp ~ smoker + lcholest, data = fram2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79148 -0.14009 -0.02043  0.10915  0.93289
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.55569    0.17029  20.880 < 2e-16 ***
## smoker        -0.03796    0.01251  -3.034  0.00246 **
## lcholest       0.15540    0.03140   4.949 8.22e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2107 on 1612 degrees of freedom
## Multiple R-squared:  0.02036,    Adjusted R-squared:  0.01915
## F-statistic: 16.75 on 2 and 1612 DF,  p-value: 6.299e-08
```

Question 2

Conduct web research on whether smokers have higher or lower blood pressure on average compared to non-smokers.

WebMD indicates that individuals who smoke tend to have higher blood pressure than those who do not. This is not consistent with my findings from Question 1, which indicate that participants who smoke have lower blood pressure than those who do not, conditional on our transformed cholesterol variable. There is nothing in the documentation to indicate that smoker is not encoded with '1' as the positive class. This is suspicious, and suggests that we need to check our data or revisit our model specification.

Question 3

The model produces the expectation of LSBP given smoking status *conditional on* cholesterol.

Question 4

Recreate the same model as in Question 1, but add in an interaction amongst the independent variables.

```
lin_mod2 <- lm(lsbp ~ smoker + lcholest + smoker:lcholest
              , data=fram2
              )
summary(lin_mod2)
```

```
##
## Call:
## lm(formula = lsbp ~ smoker + lcholest + smoker:lcholest, data = fram2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79151 -0.14039 -0.02046  0.10916  0.93280
```

```
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.55075    0.33525  10.591  <2e-16 ***
## smoker        -0.03130    0.38907  -0.080   0.9359
## lcholest       0.15632    0.06191   2.525   0.0117 *
## smoker:lcholest -0.00123    0.07184  -0.017   0.9863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2108 on 1611 degrees of freedom
## Multiple R-squared:  0.02036,    Adjusted R-squared:  0.01854
## F-statistic: 11.16 on 3 and 1611 DF,  p-value: 2.993e-07
```

The smoking indicator is still negatively associated with our blood pressure variable, but is no longer significant in the presence of the interaction term.

Question 5

Run a semiparametric ANCOVA with mgcv, the semiparametric version of an ANCOVA without an interaction.

```
semi_mod <- mgcv::gam(lsbp ~ factor(smoker) +
                      s(lcholest, k=23, bs="cr")
                      , data=fram2
                      , method="REML"
                      )
summary(semi_mod)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## lsbp ~ factor(smoker) + s(lcholest, k = 23, bs = "cr")
##
## Parametric coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.39713    0.01100 399.735  < 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) 1.064  1.126 22.27 1.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0192    Deviance explained = 2.04%
## -REML = -214.85    Scale est. = 0.044402    n = 1615
```

Question 6

For the data in part 5, display a plot of the two lines, but without the data

first, generate the data required for plotting

```
x_var <- seq(from=min(fram2$lcholest)
             , to=max(fram2$lcholest)
             , len=1000
             )
f_hat_smoker <- predict(semi_mod
                       , newdata=data.frame(
                         smoker=rep('0', 1000)
                         , lcholest=x_var
                       )
                       )

f_hat_nosmoke <- predict(semi_mod
                        , newdata=data.frame(
                          smoker=rep('1', 1000)
                          , lcholest=x_var
                        )
                        )

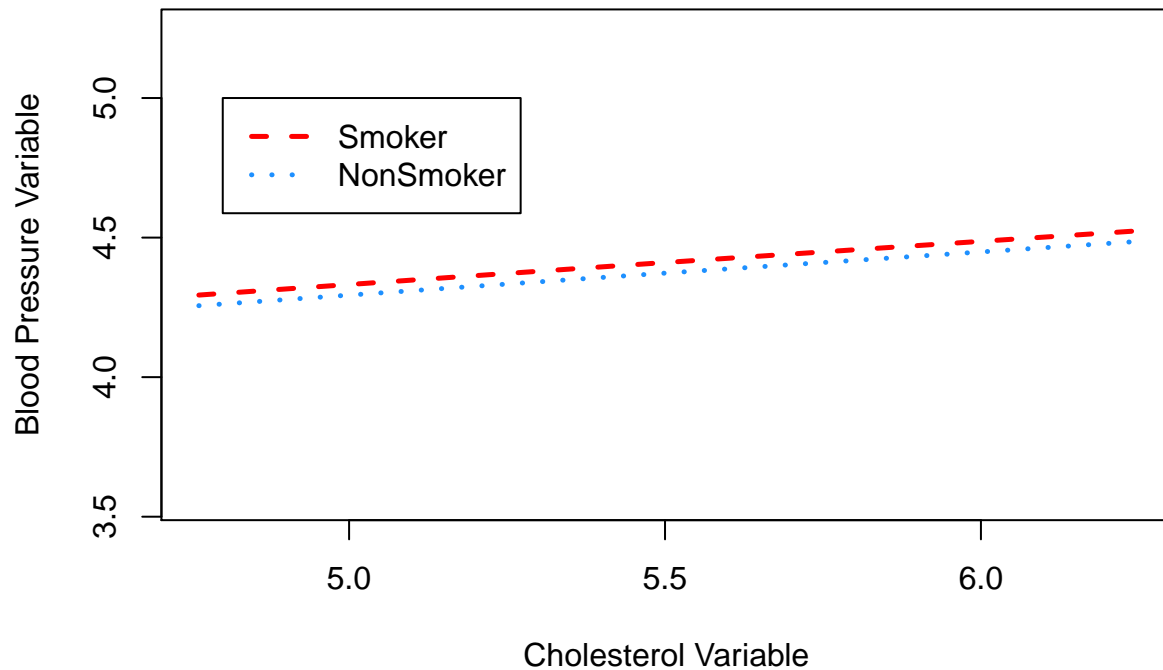
lineColors <- c("red", "dodgerblue")

plot(fram2$lcholest, fram2$lsbp, type="n"
     , xlab="Cholesterol Variable"
     , ylab = "Blood Pressure Variable"
     , main="Blood Presure by Cholesterol ANCOVA comparison"
     )

lines(x_var, f_hat_smoker, lty=2, lwd=2.5, col=lineColors[1])
lines(x_var, f_hat_nosmoke, lty=3, lwd=2.5, col=lineColors[2])

legend(4.8, 5
      , c("Smoker", "NonSmoker")
      , lty=c(2,3)
      , lwd=rep(2.5, 2)
      , col=c(lineColors[1], lineColors[2])
      )
```

Blood Pressure by Cholesterol ANCOVA comparison



Question 7

Skipping for now

```
# form confidence intervals for the regression lines
conf_int_smoke <- predict(semi_mod, newdata=data.frame(
  smoker=rep('1', 1000)
  , lcholest=x_var
  )
  , interval="confidence"
  , level=0.95
)
```

```
# the key is the polygon command in hw02
#https://stackoverflow.com/questions/14069629/plotting-confidence-intervals
glimpse(conf_int_smoke)
```

```
## num [1:1000(1d)] 4.26 4.26 4.26 4.26 4.26 ...
## - attr(*, "dimnames")=List of 1
## ..$ : chr [1:1000] "1" "2" "3" "4" ...
```

Question 8

Run the semiparametric version of ANCOVA but with an interaction.

```
semi_mod2 <- mgcv::gam(lsbp ~ factor(smoker)*lcholest +
  s(lcholest, k=23, bs="cr")
```

```

, data=fram2
# , method="REML"
)
summary(semi_mod2)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## lsbp ~ factor(smoker) * lcholest + s(lcholest, k = 23, bs = "cr")
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.3664545  0.0218837  16.746  <2e-16 ***
## factor(smoker)1 -0.0330055  0.3891394  -0.085    0.932
## lcholest        0.7444406  0.0044560 167.063  <2e-16 ***
## factor(smoker)1:lcholest -0.0009226  0.0718517  -0.013    0.990
## ---
## 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) 1.015  1.091 94.01  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 25/26
## R-sq.(adj) =  0.0186   Deviance explained = 2.05%
## GCV = 0.044542   Scale est. = 0.044429   n = 1615

```

From the above results, it does not appear that there is any sort of interaction present, as the p-value on the interaction term between *smoker* and *lcholest* is not significant.

Question 9

Display the fits of the above regressions, but without the data points.

```

ele_num <- 1000
x_vec <- seq(from=min(fram2$lcholest), to=max(fram2$lcholest), len=ele_num)
fHat_smoke <- predict(semi_mod2, newdata=data.frame(
  lcholest=x_vec
, smoker=rep('1', ele_num)
)
)

fHat_nosmoke <- predict(semi_mod2, newdata=data.frame(
  lcholest=x_vec
, smoker=rep('0', ele_num)
)
)

plot(fram2$lcholest, fram2$lsbp, type='n'
, xlab="Cholesterol"

```

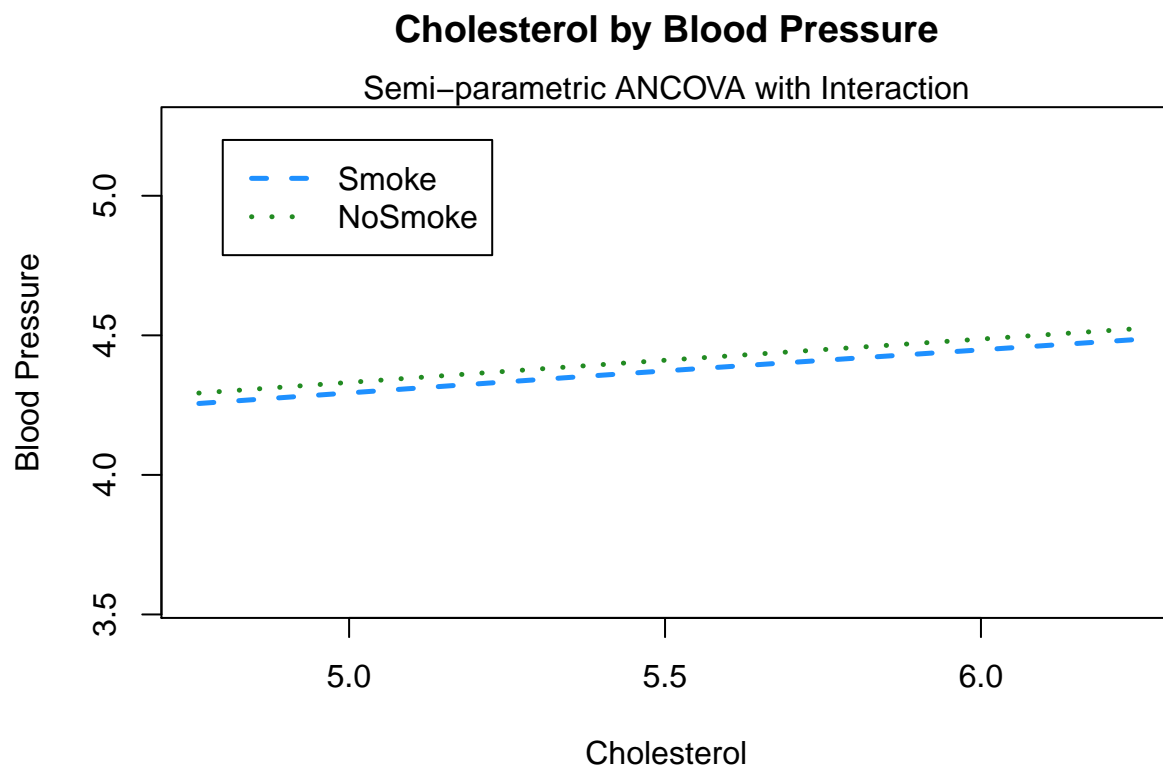
```

, ylab="Blood Pressure"
, main="Cholesterol by Blood Pressure"
)
mtext("Semi-parametric ANCOVA with Interaction")

col_vec <- c("dodgerblue", "forestgreen")
lines(x_vec, fHat_smoke, col=col_vec[1], lwd=2.5, lty=2)
lines(x_vec, fHat_nosmoke, col=col_vec[2], lwd=2.5, lty=3)

legend(4.8, 5.2, c("Smoke", "NoSmoke")
, col=col_vec
, lwd=rep(2.5, 2)
, lty=c(2,3)
)

```



Question 10

What does the interaction mean in the case when the factors are binary?

When we have a binary factor for our ANCOVA model, this indicates that the interaction term's coefficient is reflecting what happens to our outcome variable for that factor's non-reference class only.

Question 11

Run an analysis of whether our two regression lines are significantly different.

```

# First, fit the null model
contrast_mod1 <- mgcv::gam(lsbp ~ s(lcholest), data=fram2)

# indicator of the smoke variable taking positive class
smoke_ind <- as.numeric(fram2$smoker==1)
# now, fit the alternative model
contrast_mod2 <- mgcv::gam(lsbp ~ s(lcholest, smoke_ind), data=fram2)

anova(contrast_mod1, contrast_mod2, test="F")

```

```

## Analysis of Deviance Table
##
## Model 1: lsbp ~ s(lcholest)
## Model 2: lsbp ~ s(lcholest, smoke_ind)
##   Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
## 1      1613      71.988
## 2      1612      71.579  1  0.40861 9.2021 0.002456 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

It appears that there is a significant difference between the two fits, for smoker and non-smoker.

Question 12

Run native code using lme to fit LSBP vs. age.

```

x <- fram2$lcholest
y <- fram2$lsbp
numIntKnots <- 23
intKnots <- quantile(unique(x), seq(0, 1, length=numIntKnots+2))[-c(1, numIntKnots+2)]
a <- 1.01 * min(x) - 0.01*max(x)
b <- 1.01 * max(x) - 0.01*min(x)
Z <- HRW::ZOSull(x, range.x=c(a, b), intKnots=intKnots)
dummyID <- factor(rep(1, length(x)))
mm_fit <- nlme::lme(y ~ x, random=list(dummyID=pdIdent(~ -1+Z)))

```

Question 13

Display the fit without data points. Also provide a confidence band.

```

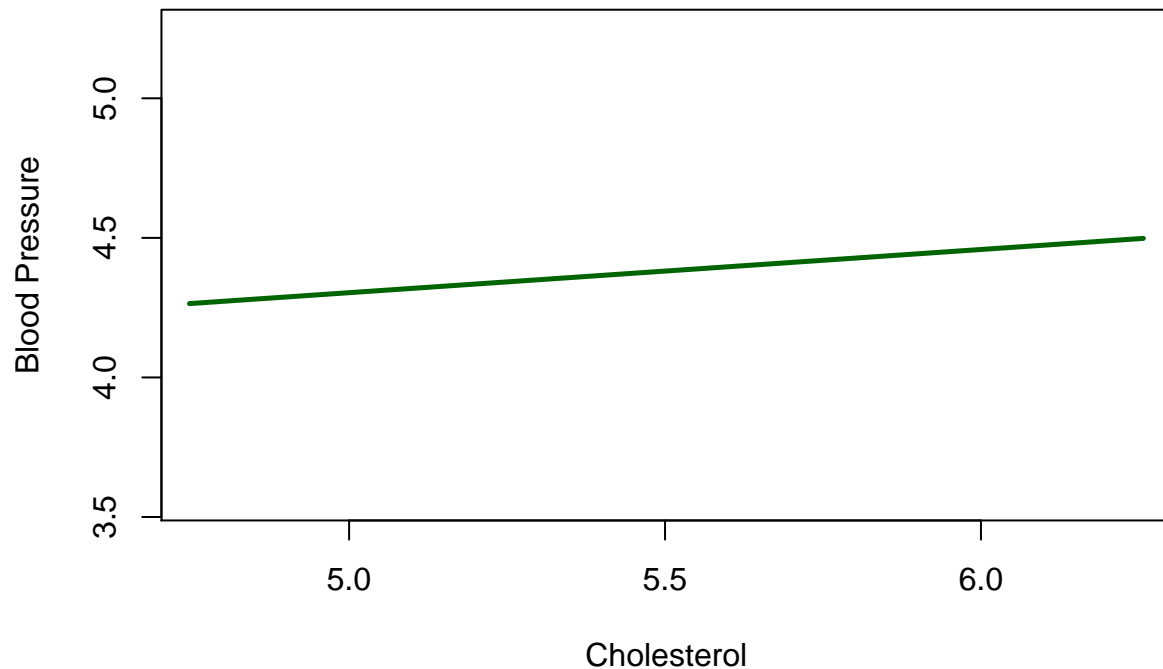
# extract our model coefficient
betaHat <- mm_fit$coefficients$fixed
uHat <- unlist(mm_fit$coefficients$random)
ng <- 1001
xg <- seq(a, b, length.out=ng)
Xg <- cbind(rep(1, ng), xg)
Zg <- HRW::ZOSull(xg, range.x=c(a,b), intKnots=intKnots)
fHatg <- as.vector(Xg %*% betaHat + Zg %*% uHat)
plot(x, y, type='n'
     , xlab="Cholesterol"
     , ylab="Blood Pressure"
     , main='Blood Pressure by Cholesterol Mixed Model Fit'

```



```
)  
lines(xg, fHatg, col="darkgreen", lwd=2.5)
```

Blood Pressure by Cholesterol Mixed Model Fit



```
# consider inclusion of se.fit=T option
```

Question 14

Is the fit of our model statistically significant?

```
anova(mm_fit)
```

```
##           numDF denDF  F-value p-value  
## (Intercept)      1  1613 690343.1 <.0001  
## x                1  1613   24.2  <.0001
```

The F-test indicates that the fit of our model is significant.

Question 15

Test whether the fit is linear or quadratic vs. the need to do a semiparametric fit.

```
mm_fit2 <- nlme::lme(y ~ x + x**2, random=list(dummyID=pdIdent(~ -1+Z)))  
summary(mm_fit2)
```

```
## Linear mixed-effects model fit by REML  
## Data: NULL  
##           AIC           BIC      logLik
```

```

##   -418.6971 -397.1536 213.3485
##
## Random effects:
## Formula: ~-1 + Z | dummyID
## Structure: Multiple of an Identity
##           Z1           Z2           Z3           Z4           Z5
## StdDev: 0.0002778356 0.0002778356 0.0002778356 0.0002778356 0.0002778356
##           Z6           Z7           Z8           Z9           Z10
## StdDev: 0.0002778356 0.0002778356 0.0002778356 0.0002778356 0.0002778356
##           Z11          Z12          Z13          Z14          Z15
## StdDev: 0.0002778356 0.0002778356 0.0002778356 0.0002778356 0.0002778356
##           Z16          Z17          Z18          Z19          Z20
## StdDev: 0.0002778356 0.0002778356 0.0002778356 0.0002778356 0.0002778356
##           Z21          Z22          Z23          Z24          Z25
## StdDev: 0.0002778356 0.0002778356 0.0002778356 0.0002778356 0.0002778356
##           Residual
## StdDev: 0.2112576
##
## Fixed effects: y ~ x + x^2
##           Value Std.Error   DF   t-value p-value
## (Intercept) 3.529742 0.17050870 1613 20.701240      0
## x           0.154779 0.03147687 1613  4.917213      0
## Correlation:
## (Intr)
## x -1
##
## Standardized Within-Group Residuals:
##           Min           Q1           Med           Q3           Max
## -3.7875892 -0.6517280 -0.1098031  0.5482766  4.3742402
##
## Number of Observations: 1615
## Number of Groups: 1

```