Assignment_06 - STAT 689

Philip Anderson; panders2@tamu.edu

2/14/2018

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

Read in our data

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

```
##  [1] "obs"      "age"      "sbp21"    "sbp22"    "sbp31"    "sbp32"
##  [7] "smoker"   "cholest2" "cholest3" "chd"
```

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.

```
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.

```
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
```

Question 5

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

```
# making the smoker variable a factor, and ensuring that the reference level is 0
semi_mod <- mgcv::gam(lsbp ~ relevel(factor(smoker), ref='0') +
                        s(lcholest, k=23, bs="cr")
                      , data=fram2
                      , method="REML"
                     )
summary(semi_mod)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## lsbp ~ relevel(factor(smoker), ref = "0") + s(lcholest, k = 23,
##     bs = "cr")
##
## Parametric coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          4.39713    0.01100 399.735  < 2e-16
## relevel(factor(smoker), ref = "0")1 -0.03799    0.01251  -3.036  0.00244
##
## (Intercept)                         ***
## relevel(factor(smoker), ref = "0")1 **
## ---
## 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

```
plot(semi_mod)
```