Stat 526 **Assignment #11**

Nonparametric Regression and Additive Models

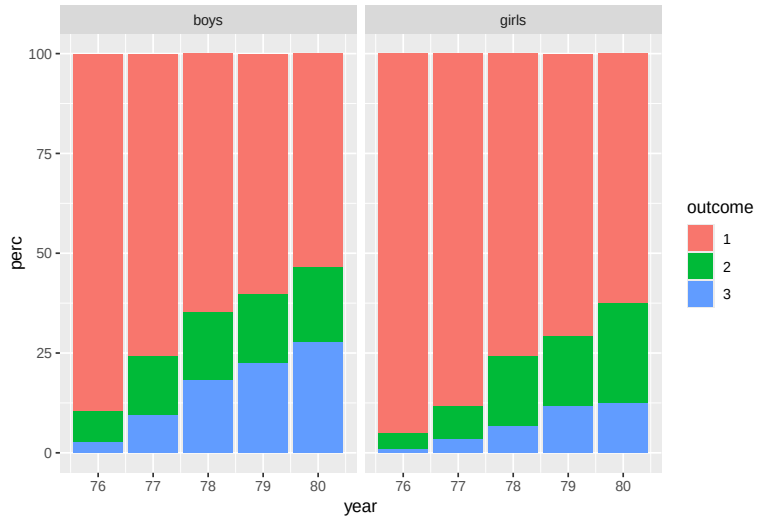
READING - Faraway Chapters 14 and 15

0. (5 pts) Name

1. (7 pts) Faraway Chapter 13 Exercise 2

a) Since the number remains constant over the years, I decided to plot the percent of individuals in each category. That way we can directly compare the boys and girls. The response categories are 1=never used, 2=used no more than once a month, and 3=used more than once a month. Although the book says there were 117 boys, the data set has counts for only 116.

In both cases, we see a steady increase in the percent of 3's and decline in 1's. For the girls there is also an increase in 2's but for the boys it is rather constant, especially for the last three to four years.



b) Getting the data into long form is required in order to fit the GLMMs. Below is the summary of the model fit with sex and year (numeric) and their interaction in the model. We could consider a random coefficients model, but I just included ID as a random effect here.

```
Generalized linear mixed model fit by maximum likelihood (Adaptive
  Gauss-Hermite Quadrature, nAGQ = 20) [glmerMod]

Formula: outcome1 ~ factor(sex) * yr + (1 | id)

     AIC      BIC   logLik deviance df.resid
  1004.4   1029.7   -497.2    994.4     1175

Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 8.182    2.86
Number of obs: 1180, groups:  id, 236
```

```
Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -4.3953     0.5267  -8.345  < 2e-16 ***
factor(sex)2     -1.5934     0.7341  -2.171    0.030 *
yr                0.8645     0.1127   7.671  1.7e-14 ***
factor(sex)2:yr   0.1553     0.1623   0.957    0.339
```

Girls are less likely to use marijuana than boys ($\hat{\beta} = -1.6$, $P = 0.03$). This can also be seen in the percentage bar chart above. The percent of blue and green in each year is always larger for the boys than for girls. There is a strong year effect but it does not look like this linear rate varies across sex.

c) If we remove sex from the model, we also remove the interaction. The following summarizes the test of whether we can remove sex from the model. The results suggest we need to leave sex in the model.

```
> anova(mod1,mod2)

Models:
mod2: outcome1 ~ yr + (1 | id)
mod1: outcome1 ~ factor(sex) * yr + (1 | id)
     Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
mod2  3 1006.7 1021.9 -500.35  1000.71
mod1  5 1004.4 1029.7 -497.18   994.35 6.3575      2    0.04164 *
```

d) If we treat year as a factor variable and also include the interaction, we run into some warnings about convergence. This is not surprising given that this model would be the saturated GLM model. If we drop the interaction term, the model converges. Here are those results.

```
Generalized linear mixed model fit by maximum likelihood (Adaptive
  Gauss-Hermite Quadrature, nAGQ = 20) [glmerMod]

Formula: outcome1 ~ factor(sex) + year + (1 | id)

     AIC      BIC   logLik deviance df.resid
   997.3   1032.8   -491.6    983.3     1173

Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 8.456    2.908
Number of obs: 1180, groups:  id, 236

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.3823     0.5053  -8.673  < 2e-16 ***
factor(sex)2  -1.0642     0.4583  -2.322   0.0202 *
yearY77        1.7245     0.4122   4.184 2.87e-05 ***
yearY78        3.0575     0.4235   7.220 5.19e-13 ***
yearY79        3.5087     0.4320   8.122 4.58e-16 ***
yearY80        4.2049     0.4494   9.357  < 2e-16 ***
```

The difference between boys and girls is still there. As far as the year effects, we see they are increasing over time, which we should expect. The fit of this model has BIC 1032.8. The BIC for the model treating year as numeric is 1023.6. This suggest it is reasonable to treat year as numeric.

e) Finally, the GEE fit is shown below. I did not include the interaction and treated year as numeric.

```
 Coefficients:
             Estimate  Std.err    Wald Pr(>|W|)
(Intercept)  -2.28017  0.21715 110.262   <2e-16 ***
factor(sex)2 -0.48838  0.23373   4.366   0.0367 *
yr            0.47172  0.04435 113.134   <2e-16 ***

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)   0.9571 0.08275

Correlation: Structure = exchangeable  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha   0.4143 0.05796
```

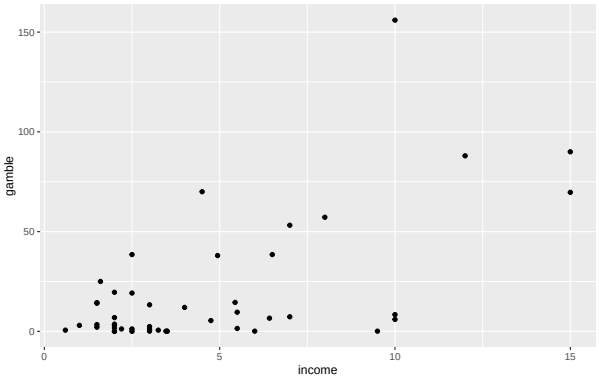
We see that girls still have a lower probability of using marijuana but the effect is not as strong. This is expected as the mean of the marginal distribution will be pulled more out into the tail compared to the median.

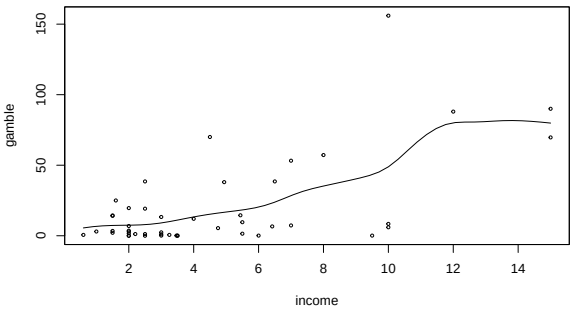
2. (7 pts) A client is interested in predicting the number of Canada thistle seeds that germinate under different conditions. She collects seeds from 20 different locations and randomly assigns a proportion of them from each location to each of the conditions. Because she randomly chose these locations, she is debating between a GEE and GLMM analysis. Describe the differences between these two approaches and how she should choose between them.

The GEE analysis will focus on the population mean and the GLMM analysis will focus on the population median (conditional mean). If she is interested in developing a model to predict the proportion at a given site, the GLMM model would be more appropriate. If interested is a general prediction over sites, the GEE would be acceptable.

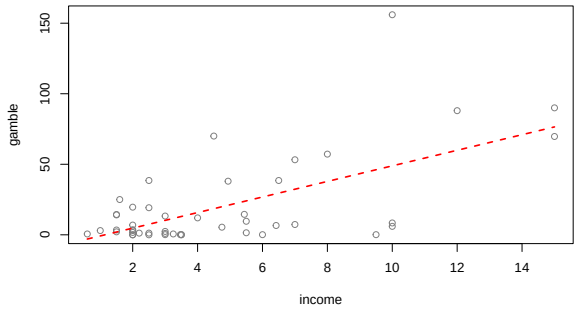
3. (7 pts) Faraway Chapter 14 Exercise 1

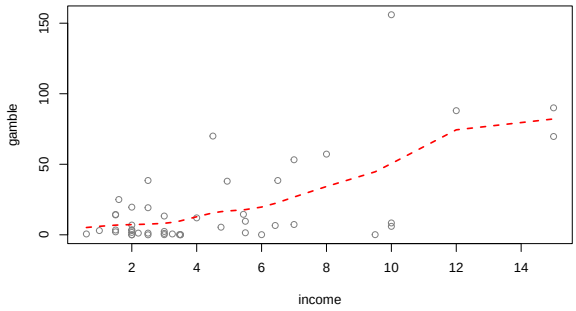
a) A plot of the data is shown below on the left. Clearly the spread in gambling increases with income.
b) When fitting a kernel smoother to the data, the resulting curve is somewhat S-shaped.



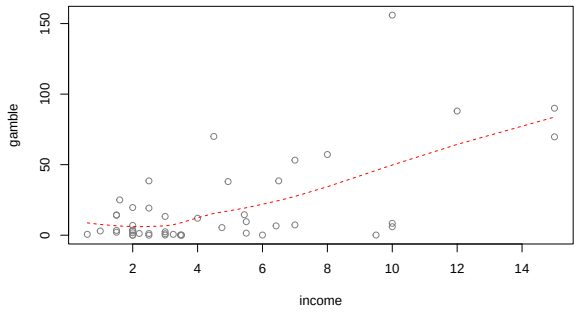


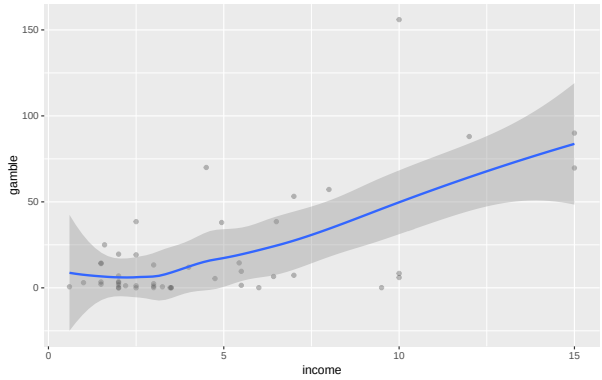
c) When we use a smoothing spline, the effective degrees of freedom is 2 and the resulting curve looks very linear. When the df is increased to 6, we begin to see more of the S-shape again. I do not think the automatic choice is effective here.





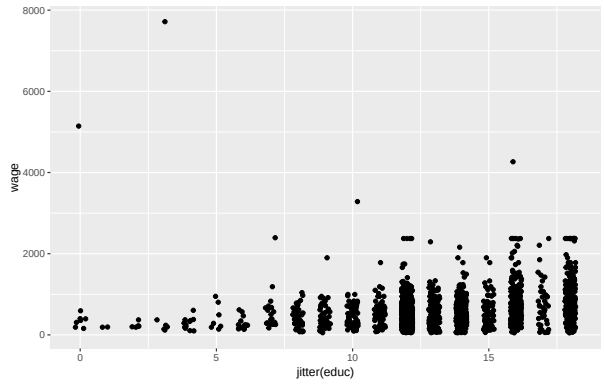
d) The loess fit is shown below. It suggests a slight curvature. e) When you add a confidence band, we see a linear model may be reasonable. The only reason one may consider something else is that the gambling amount cannot be negative.

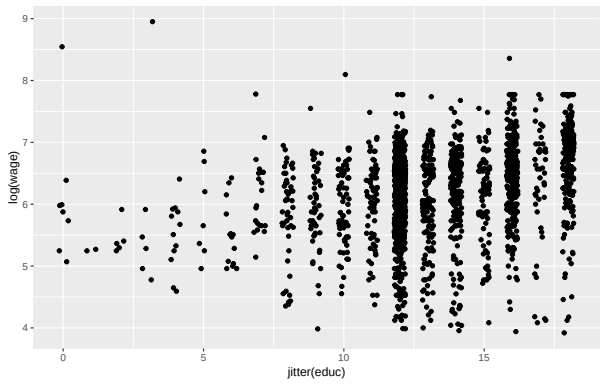




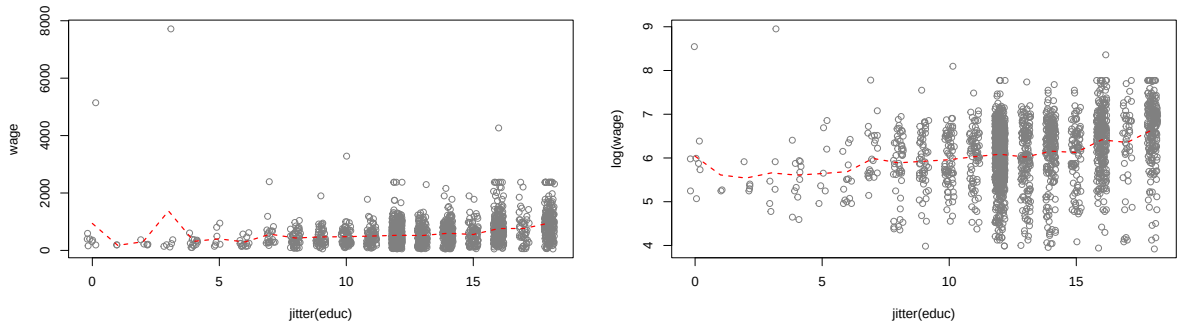
4. (8 pts) Faraway Chapter 14 Exercise 2

a) Education runs from 0 to 18 years. To more clearly see all the incomes, I added jitter around the years of education. There is still some overplotting but we clearly see that a majority of the sample had at least 8 years of education. When using the log of income, we can more clearly see that the most common years are 12 (high school), 13, 14, 16 (college) and 18 and there is a positive correlation between the two variables.

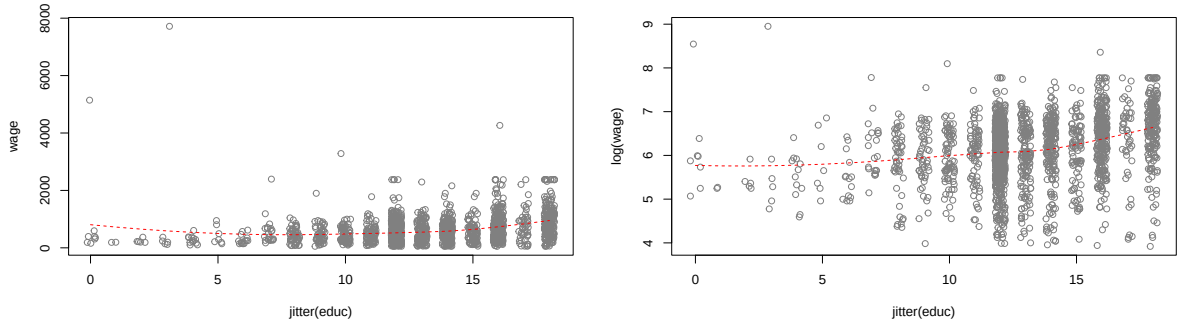




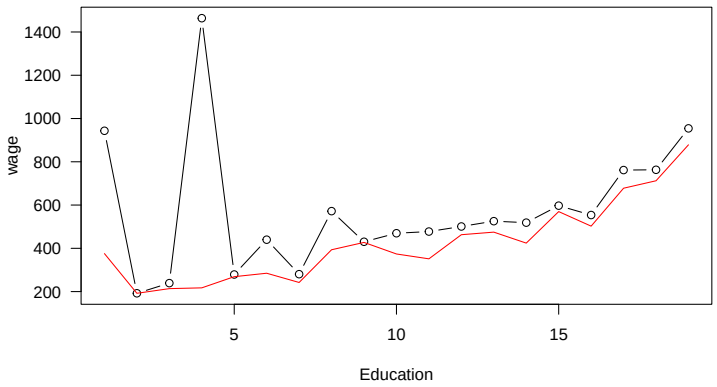
b) The effective degrees of freedom are roughly 18.5. The fit appears too sensitive to some of the outlying values. After about 5 years, the fit looks quite linear. When fit on the log scale, the default choice looks more reasonable but maybe a little too wiggly.



c) The default loess fit is strongly influenced by some early large values which makes the plot look somewhat quadratic. I do not think this is a good fit. The fit to log income on other hand looks very good.

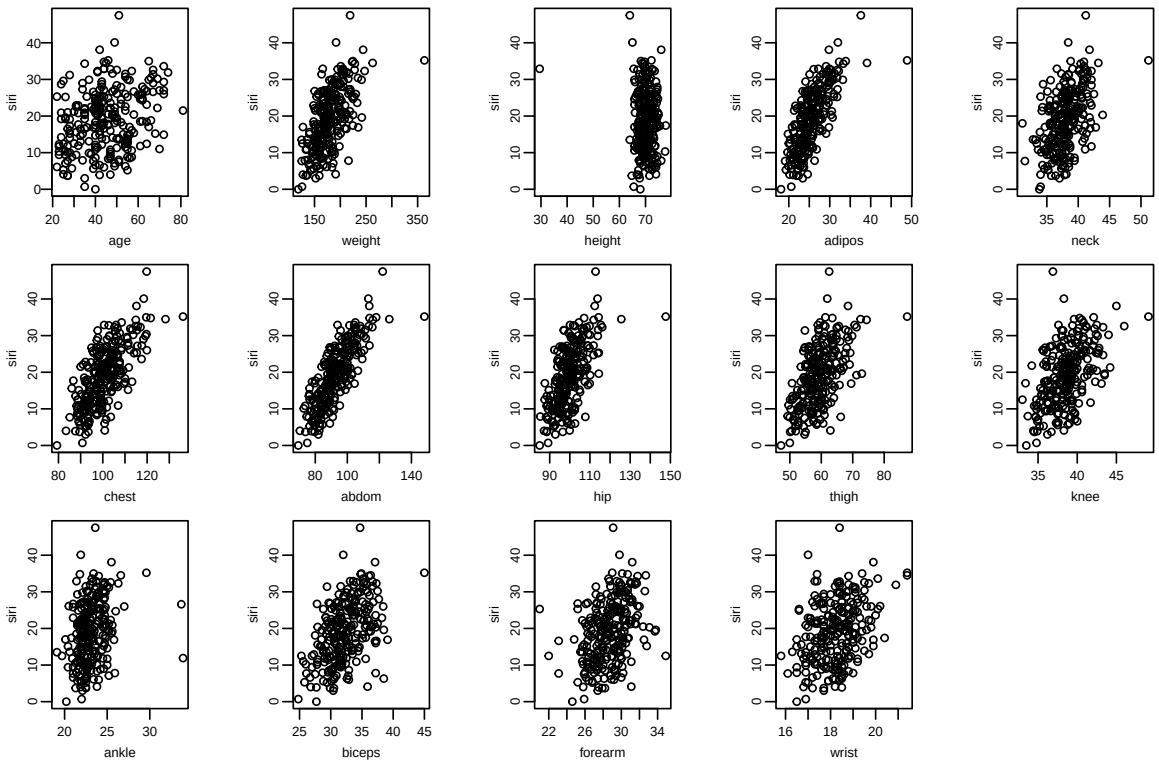


d) Because of the large outlying values in the early years, the median does a much better describing the trend. The trend suggests fitting a a linear or quadratic model to log(y) and then back-transforming using $\exp$.

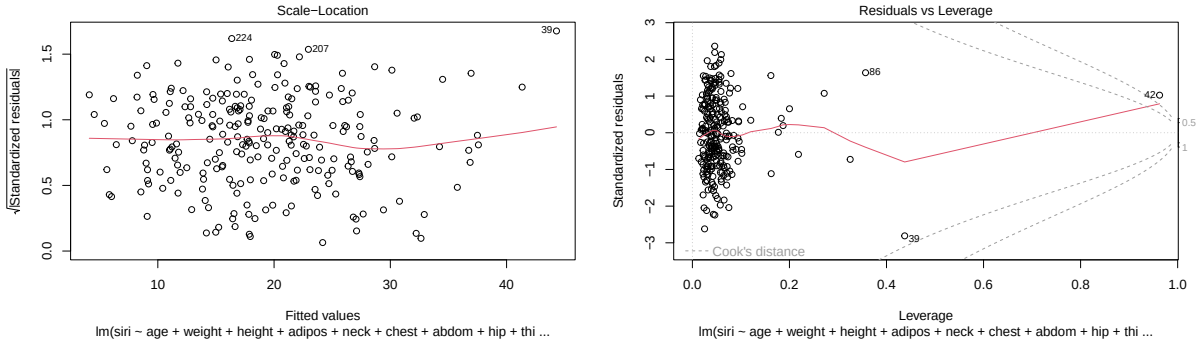


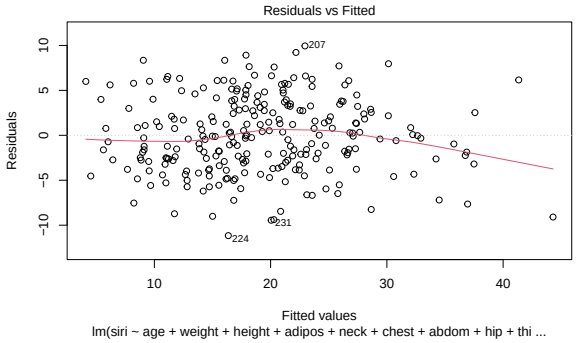
5. (8 pts) Faraway Chapter 15 Exercise 1

a) The scatterplots in general show a positive relationship. The on exception is height, which has an outlier but otherwise very little association.



b) In fitting the linear model, we find one observation (Case 42) that is reasonably influential. This is the very short individual (less than 3 ft tall) and the value may be a typo. Removing that case leaves Case 39 near the border of being overly influential. Case 39 has very large weight, adipos, neck, abdomin, thigh, biceps, hip, and knee. I'll take this individual out but it is debatable whether to do this. After that, the remaining values appear to behave well.





Removing insignificant variables using the step and drop1 functions, my final model is

```
Call: lm(formula = siri ~ age + adipos + chest + abdom + wrist, data = fatr)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.45616    6.10750  -0.893 0.372548
age          0.08740    0.02301   3.799 0.000184 ***
adipos       0.56174    0.23768   2.363 0.018891 *
chest       -0.22329    0.09406  -2.374 0.018374 *
abdom        0.75948    0.07708   9.854  < 2e-16 ***
wrist       -2.26525    0.39321  -5.761 2.51e-08 ***

Residual standard error: 4.279 on 244 degrees of freedom
Multiple R-squared:  0.7391,Adjusted R-squared:  0.7338
F-statistic: 138.3 on 5 and 244 DF,  p-value: < 2.2e-16
```

It has an R^2 of almost 74% with abdominen being the most signficant.

c) Fitting an additive model to the full data set is shown below prior to insignificant values being removed. The key difference is that hip is significant and adipos chest are not.

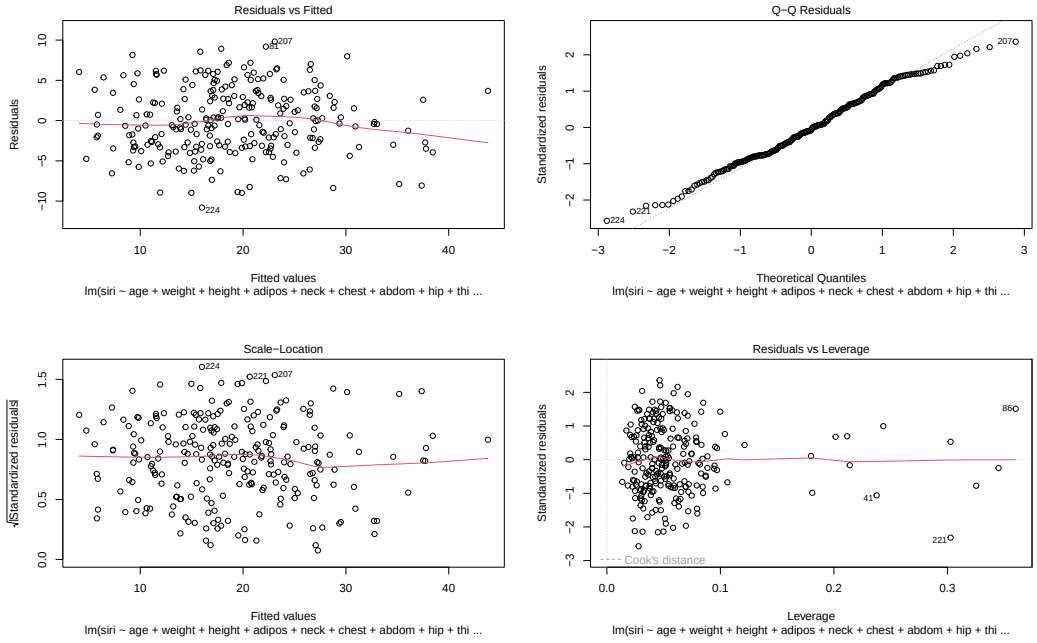
```
Formula:
siri ~ s(age) + s(weight) + s(height) + s(adipos) + s(neck) +
    s(chest) + s(abdom) + s(hip) + s(thigh) + s(knee) + s(ankle) +
    s(biceps) + s(forearm) + s(wrist)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.1508     0.2432   78.75   <2e-16 ***

Approximate significance of smooth terms:
             edf Ref.df      F  p-value
s(age)     5.470  6.620  2.163 0.042241 *
s(weight)  1.000  1.000  0.224 0.636773
s(height)  1.000  1.000  0.132 0.716415
s(adipos)  2.355  3.052  0.870 0.488354
s(neck)    1.669  2.103  2.520 0.101446
s(chest)   1.000  1.000  0.465 0.496235
s(abdom)   6.693  7.545 14.803  < 2e-16 ***
s(hip)     7.470  8.287  2.582 0.007360 **
s(thigh)   1.000  1.000  0.962 0.327671
s(knee)    1.452  1.785  0.357 0.578946
s(ankle)   2.843  3.503  0.863 0.598646
s(biceps)  4.746  5.763  2.427 0.047747 *
s(forearm) 1.000  1.000  1.935 0.165612
s(wrist)   1.768  2.219  7.617 0.000396 ***

R-sq.(adj) =  0.787   Deviance explained = 82.1%
GCV = 17.756  Scale est. = 14.904    n = 252
```

d) The model identifies the same influential cases but the impact is reduced because of the relaxation of linearity. In other words, the transformation is able to adjust and fit these cases well thereby reducing their influence on the model. Below are several of the transformation plots. The most influential variable is again abdomen.



e) The additive model is using all the cases while the linear regression model involved eliminating two unusual cases. Whether we include or exclude those them, the explained variation is about 82%, which is much better.

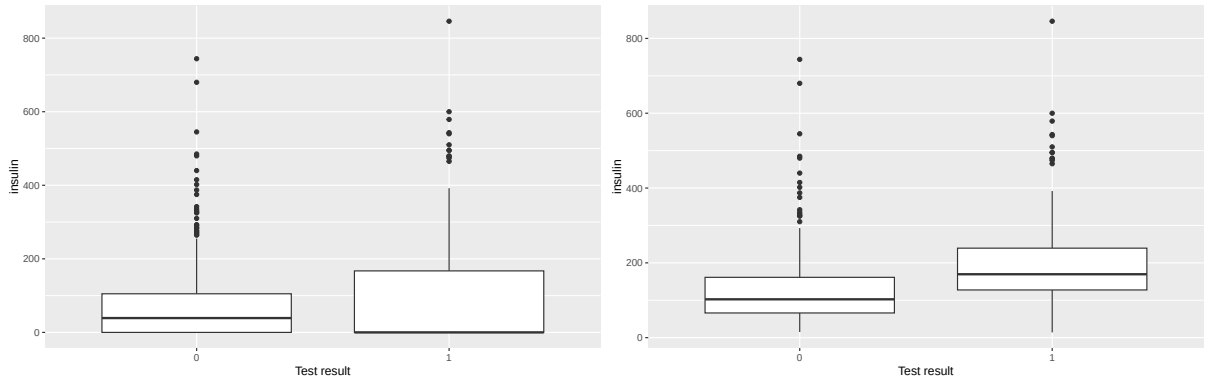
f) Regardless of whether all cases are used or not, the two most nonlinear transformations are biceps and hip. The test for hip using all the cases is shown below. It should not be made linear. If anything a quadratic relationship may be more appropraite.

```
 > anova(mod1,mod1a,test="F")

  Resid. Df Resid. Dev      Df Deviance      F  Pr(>F)
1    205.12     3152.8
2    213.49     3433.4 -8.3676  -280.63 2.2502 0.02329 *
```

6. (8 pts) Faraway Chapter 15 Exercise 3

a) For many of the predictors, 0's are impossible so I replaced the 0's with NA. This was done for insulin, diastolic, bmi, glucose, diabetes, age and triceps. Note that a 0 for pregnant can be possible. The plots below are for insulin before and after this switch. You can see that there is a more obvious shift in insulin values for the two responses.



The cases that are complete are those without any NAs. There are 393 of them. I took a random sample of 100 complete cases and used the others for training. The following model results:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.711809   1.350832  -7.190 6.50e-13 ***
age          0.038369   0.022774   1.685   0.0920 .
diabetes     1.016323   0.462683   2.197   0.0281 *
bmi          0.059708   0.029770   2.006   0.0449 *
insulin     -0.001338   0.001585  -0.844   0.3987
triceps      0.012411   0.019543   0.635   0.5254
diastolic   -0.003051   0.013438  -0.227   0.8204
glucose      0.040522   0.006946   5.834 5.42e-09 ***
pregnant     0.091280   0.066345   1.376   0.1689

  (375 observations deleted due to missingness)
```

This model correctly predicts 68 of the 78 0's and 16 of the 22 1's in the test sample. However, 375 of the cases are not used in the training sample due to missingness.

When I step down through models be removing the least significant, I end up with the following model:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.803923   0.688768 -12.782  < 2e-16 ***
diabetes     0.929599   0.296107   3.139  0.00169 **
bmi          0.089136   0.014569   6.118 9.47e-10 ***
glucose      0.033887   0.003364  10.074  < 2e-16 ***
pregnant     0.138111   0.027349   5.050 4.42e-07 ***

  (11 observations deleted due to missingness)
```

This model uses all but 11 cases in the training set but its prediction is quite similar predicting 66 of the 78 0's correctly and 17 of the 22 1's.

When I fit a GAM usng all the predictors, I end up with a model that predicts 65 of the 78 0's and 16 of the 22 1's. The reduced GAM is

```
Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.7715     0.1077  -7.162 7.94e-13 ***

Approximate significance of smooth terms:
             edf Ref.df Chi.sq  p-value
s(age)     3.163  3.938  41.97 1.88e-08 ***
```

```
       s(diabetes) 2.302  2.889  16.77 0.000983 ***
       s(glucose)  4.897  5.906 108.14  < 2e-16 ***

       R-sq.(adj) =  0.328   Deviance explained = 28.7%
       UBRE = -0.028103  Scale est. = 1         n = 668
```

This model predicts 66 of the 78 of the 0's correctly and 16 of the 1's. These results are all basically the same. To better assess the models, it would be good to consider some sort of cross-validation were the complete cases were split into say 5 folds and this procedure used for each of the folds.

```
 ############## Problem 1 ################
library(faraway)
head(potuse,n=30)

## Creating summary data set:  3 outcomes for each of 5 years by sex
  sex = c(rep(rep(c("boys","girls"),3),5))
  outcome = factor(c(rep(c(1,1,2,2,3,3),5)))
  year = c(rep(76,6),rep(77,6),rep(78,6),rep(79,6),rep(80,6))

  cnt = c(c(xtabs(count~sex+year.76,potuse)),c(xtabs(count~sex+year.77,potuse)),
        c(xtabs(count~sex+year.78,potuse)),c(xtabs(count~sex+year.79,potuse)),
        c(xtabs(count~sex+year.80,potuse)))
  library(tigerstats)
  perc = c(c(rowPerc(xtabs(count~sex+year.76,potuse))[,c(1:3)]),c(rowPerc(xtabs(count~sex+year.77,potuse))[,c(1:3)]),
         c(rowPerc(xtabs(count~sex+year.78,potuse))[,c(1:3)]),c(rowPerc(xtabs(count~sex+year.79,potuse))[,c(1:3)]),
         c(rowPerc(xtabs(count~sex+year.80,potuse))[,c(1:3)]))
  year = c(rep(76,6),rep(77,6),rep(78,6),rep(79,6),rep(80,6))

  parta = data.frame(year,cnt,sex,outcome,perc)

## Creating the bar graphs
library(ggplot2)
ggplot(data=parta, aes(x=year, y=cnt, fill=outcome))+geom_bar(stat="identity")+facet_grid(~sex)
ggplot(data=parta, aes(x=year, y=perc, fill=outcome))+geom_bar(stat="identity")+facet_grid(~sex)

## Getting the data into long format
  ##Filter out the zero counts
  potuse1 = filter(potuse, count!=0)
  potuse1= data.matrix(potuse1)
  ##There are 68 non-zero sequences (39 for sex=1 and 29 for sex=2)
  ##There are 116+120=236 total individuals
  ##Now create individual row for each subject
  partb = matrix(0,nrow=236,ncol=7)
  cnt=1
  for(i in 1:length(potuse1[,1])){
    for(j in 1:potuse1[i,7]){
      partb[cnt,] = c(cnt,potuse1[i,c(1:6)])
      cnt=cnt+1
    }
  }
  partb = data.frame(partb)
  ##Now convert wide table into tall table using plyr and tidyr
  library(plyr)
  partb = rename(partb,c("X1"="id","X2"="sex","X3"="Y76","X4"="Y77","X5"="Y78","X6"="Y79","X7"="Y80"))
  library(tidyr)
  partb_long <- gather(partb, year, outcome, Y76:Y80, factor_key=TRUE)
  partb_long = partb_long[with(partb_long,order(id,year)),]

  partb_long$yr = rep(1:5,236)
  partb_long$outcome1 = ifelse(partb_long$outcome==1,0,1)

 library(lme4)
mod1 = glmer(outcome1~factor(sex)*yr+(1 | id), nAGQ=20, family="binomial", partb_long)
mod1a = glmer(outcome1~factor(sex)+yr+(1 | id), nAGQ=20, family="binomial", partb_long)
mod2 = glmer(outcome1~yr+(1 | id), nAGQ=20,family="binomial", partb_long)
```

```
summary(mod1)
anova(mod1,mod2)

mod3 = glmer(outcome1~factor(sex)*year+(1 | id), nAGQ=20,family="binomial", partb_long)
mod3a = glmer(outcome1~factor(sex)+year+(1 | id), nAGQ=20,family="binomial", partb_long)
summary(mod3a)

library(geepack)
mod4 = geeglm(outcome1~factor(sex)+yr, family="binomial", id=id,
              corstr="exchangeable",partb_long)



########### Exercise #3
library(faraway)
library(ggplot2)
ggplot(teengamb, aes(x=income, y=gamble))+geom_point()
library(sm)
with(teengamb, sm.regression(income,gamble,h=h.select(income,gamble)))

smooth.spline(teengamb$income,teengamb$gamble)

with(teengamb, {
  plot(gamble~income,col=gray(0.5))
  lines(smooth.spline(income,gamble),lty=2,col="red",lwd=2)
})

with(teengamb, {
  plot(gamble~income,col=gray(0.5))
  lines(smooth.spline(income,gamble,df=6),lty=2,col="red",lwd=2)
})

with(teengamb, {
  plot(gamble~income,col=gray(0.5))
  f = loess(gamble~income)
  ford = order(f$x)
  lines(f$x[ford],f$fitted[ford],lty=2,col="red")
})

ggplot(teengamb, aes(x=income, y=gamble))+ geom_point(alpha=0.25) +
  geom_smooth(method="loess", span=0.75)

########### Exercise #4

ggplot(uswages, aes(x=jitter(educ), y=wage))+geom_point()
ggplot(uswages, aes(x=jitter(educ), y=log(wage)))+geom_point()

smooth.spline(uswages$educ,uswages$wage)
with(uswages, {
  plot(wage~jitter(educ),col=gray(0.5))
  lines(smooth.spline(educ,wage),lty=2,col="red",lwd=1.5)
})

smooth.spline(uswages$educ,log(uswages$wage))
with(uswages, {
  plot(log(wage)~jitter(educ),col=gray(0.5))
  lines(smooth.spline(educ,log(wage)),lty=2,col="red",lwd=1.5)
})


with(uswages, {
  plot(wage~jitter(educ),col=gray(0.5))
  f = loess(wage~educ)
  ford = order(f$x)
  lines(f$x[ford],f$fitted[ford],lty=2,col="red")
})
```

```
with(uswages, {
  plot(log(wage)~jitter(educ),col=gray(0.5))
  f = loess(log(wage)~educ)
  ford = order(f$x)
  lines(f$x[ford],f$fitted[ford],lty=2,col="red")
})

wagemean = sapply(split(uswages$wage,uswages$educ),mean)
wagemedian = sapply(split(uswages$wage,uswages$educ),median)

plot(wagemean,type="b",las=1,ylab="wage",xlab="Education")
lines(wagemedian,col="red")

####Exercise #5

head(fat)

par(mfrow=c(3,5),mar=c(2.75,4.5,0.5,0.5),
    mgp=c(1.5,.5,0),cex.lab=0.8,cex.axis=0.8)
attach(fat)
plot(age,siri,cex=0.9,las=1)
plot(weight,siri,cex=0.9)
plot(height,siri,cex=0.9)
plot(adipos,siri,cex=0.9)
plot(neck,siri,cex=0.9)
plot(chest,siri,cex=0.9)
plot(abdom,siri,cex=0.9)
plot(hip,siri,cex=0.9)
plot(thigh,siri,cex=0.9)
plot(knee,siri,cex=0.9)
plot(ankle,siri,cex=0.9)
plot(biceps,siri,cex=0.9)
plot(forearm,siri,cex=0.9)
plot(wrist,siri,cex=0.9)

par(mfrow=c(1,1),mar=c(5.1,4.1,4.1,2.1),
    mgp=c(3,1,0),cex.lab=1,cex.axis=1)

mod1 = lm(siri~age+weight+height+adipos+neck+chest+abdom+hip+thigh+knee+ankle+biceps+forearm+wrist,fat)
plot(mod1)

fatr = fat[-c(39,42),]
mod1 = lm(siri~age+weight+height+adipos+neck+chest+abdom+hip+thigh+knee+ankle+biceps+forearm+wrist,fatr)
plot(mod1)

mod2 = step(mod1,fatr)
drop1(mod2,test="F")
mod2a = lm(siri~age+adipos+neck+chest+abdom+hip+wrist,fatr)
drop1(mod2a,test="F")
mod2a = lm(siri~age+adipos+chest+abdom+hip+wrist,fatr)
drop1(mod2a,test="F")
mod2a = lm(siri~age+adipos+chest+abdom+wrist,fatr)
drop1(mod2a,test="F")

summary(mod2a)

library(mgcv)
library(faraway)
mod1 = gam(siri~s(age)+s(weight)+s(height)+s(adipos)+s(neck)+
           s(chest)+s(abdom)+s(hip)+s(thigh)+s(knee)+s(ankle)+
           s(biceps)+s(forearm)+s(wrist),data=fat)
summary(mod1)

plot(mod1,residuals=TRUE)

plot(residuals(mod1)~predict(mod1),xlab="Predicted",ylab="Residual")

mod1a = gam(siri~s(age)+s(weight)+s(height)+s(adipos)+s(neck)+
```

```
            s(chest)+s(abdom)+hip+s(thigh)+s(knee)+s(ankle)+
            s(biceps)+s(forearm)+s(wrist),data=fat)

anova(mod1,mod1a,test="F")

####Exercise #6

library(faraway)
attach(pima)
str(pima)
library(ggplot2)
ggplot(pima, aes(y=insulin, x=as.factor(test))) + geom_boxplot() + labs(x="Test result")

detach(pima)
pima1 = pima
pima1$insulin[pima1$insulin==0] = NA
attach(pima1)
ggplot(pima1, aes(y=insulin, x=as.factor(test))) + geom_boxplot() + labs(x="Test result")

pima1$diastolic[pima1$diastolic==0] = NA
pima1$triceps[pima1$triceps==0] = NA
pima1$bmi[pima1$bmi==0] = NA
pima1$diabetes[pima1$diabetes==0] = NA
pima1$age[pima1$age==0] = NA

pima1c = na.omit(pima1)
pima1i = pima1[!complete.cases(pima1),]

set.seed(612)
bc = sample(nrow(pima1c),100)
pima1test = pima1c[bc,]
head(pima1test)
pima1training = rbind(pima1c[-bc,],pima1i)

mod2 = glm(test~age+diabetes+bmi+insulin+triceps+diastolic+glucose+pregnant,family=binomial,pima1training)
summary(mod2)

bc1 = predict(mod2,pima1test,type="response")
bc1p = as.numeric(bc1 > 0.5)

xtabs(~bc1p+pima1test$test)

## Remove diastolic
mod2r = glm(test~age+diabetes+bmi+glucose+pregnant,family=binomial,pima1)
summary(mod2r)

## Remove age
mod2r = glm(test~diabetes+bmi+glucose+pregnant,family=binomial,pima1)
summary(mod2r)

bc1 = predict(mod2r,pima1test,type="response")
bc1p = as.numeric(bc1 > 0.5)

xtabs(~bc1p+pima1test$test)

###Fitting a gam
mod1 = gam(test~s(age)+s(diabetes)+s(bmi)+s(insulin)+s(triceps)+
          s(diastolic)+s(glucose)+s(pregnant),family=binomial,pima1training)

bc1 = predict(mod1,pima1test,type="response")
bc1p = as.numeric(bc1 > 0.5)

xtabs(~bc1p+pima1test$test)

summary(mod1)
mod1r = gam(test~s(age)+s(diabetes)+s(glucose),family=binomial,pima1training)
summary(mod1r)
```

```
bc1 = predict(mod1r,pima1test,type="response")
bc1p = as.numeric(bc1 > 0.5)

xtabs(~bc1p+pima1test$test)
```