Homework 14; STAT 689

Philip Anderson; panders2@tamu.edu

4/30/2018

```
rm(list=ls())
```

```
# bring in data
sim <- read.csv("/Users/panders2/Documents/schools/tamu/stat_689/homework/semiparametric-regression/mis
names(sim) <- tolower(names(sim))
str(sim)
```

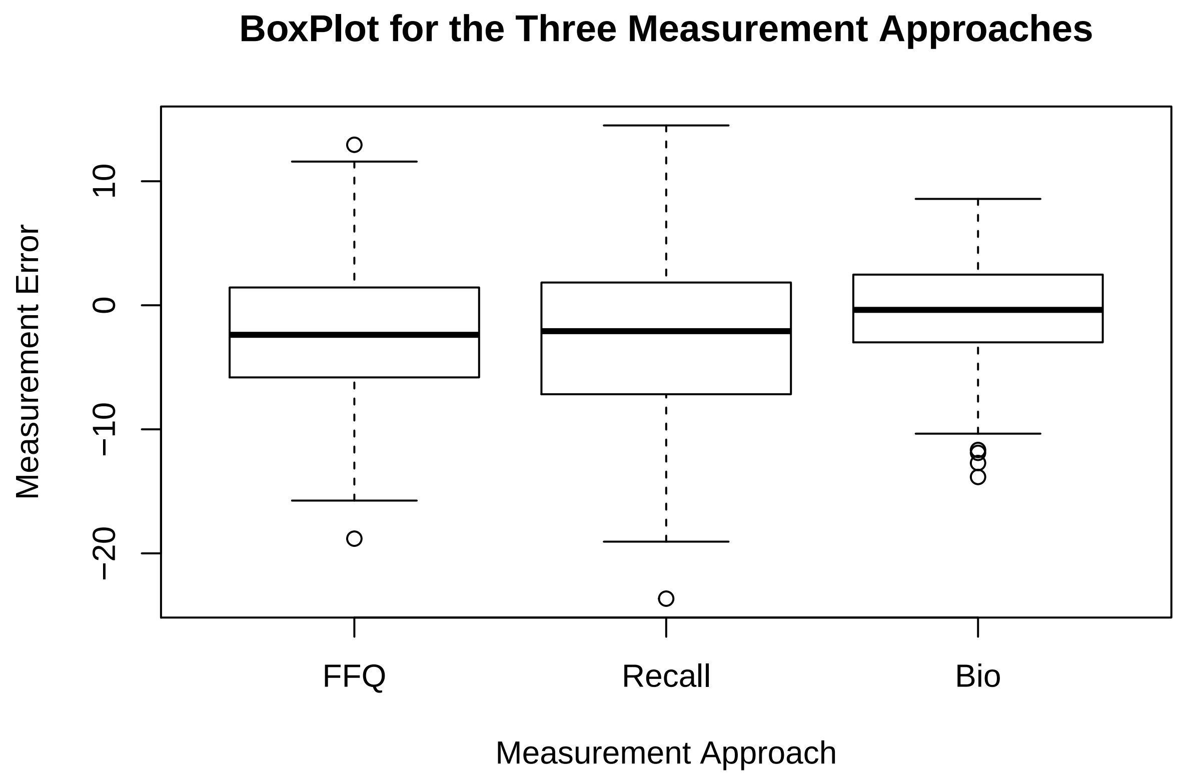
```
## 'data.frame':    446 obs. of  11 variables:
##  $ id       : int  1 1 2 2 3 3 4 4 5 5 ...
##  $ meas     : int  1 2 1 2 1 2 1 2 1 2 ...
##  $ age      : int  49 49 62 62 46 46 51 51 69 69 ...
##  $ bmi      : num  31.3 31.3 21 21 19.1 ...
##  $ truth    : num  27.9 27.9 23.1 23.1 26.5 ...
##  $ ffq      : num  36.5 46.5 26.5 20.9 23.5 ...
##  $ recall   : num  30.5 38.8 25.2 16.2 23.3 ...
##  $ bio      : num  26.3 20.5 21.9 17.6 29.6 ...
##  $ avgffq   : num  41.5 41.5 23.7 23.7 25.6 ...
##  $ avgrecall: num  34.7 34.7 20.7 20.7 23.2 ...
##  $ avgbio   : num  23.4 23.4 19.8 19.8 31.3 ...
```

Question 1

Compare the true %CFP to the average %CFP estimated by each instrument (avgFFQ, avgRecall, avgBio); make a boxplot containing the errors of the three instruments.

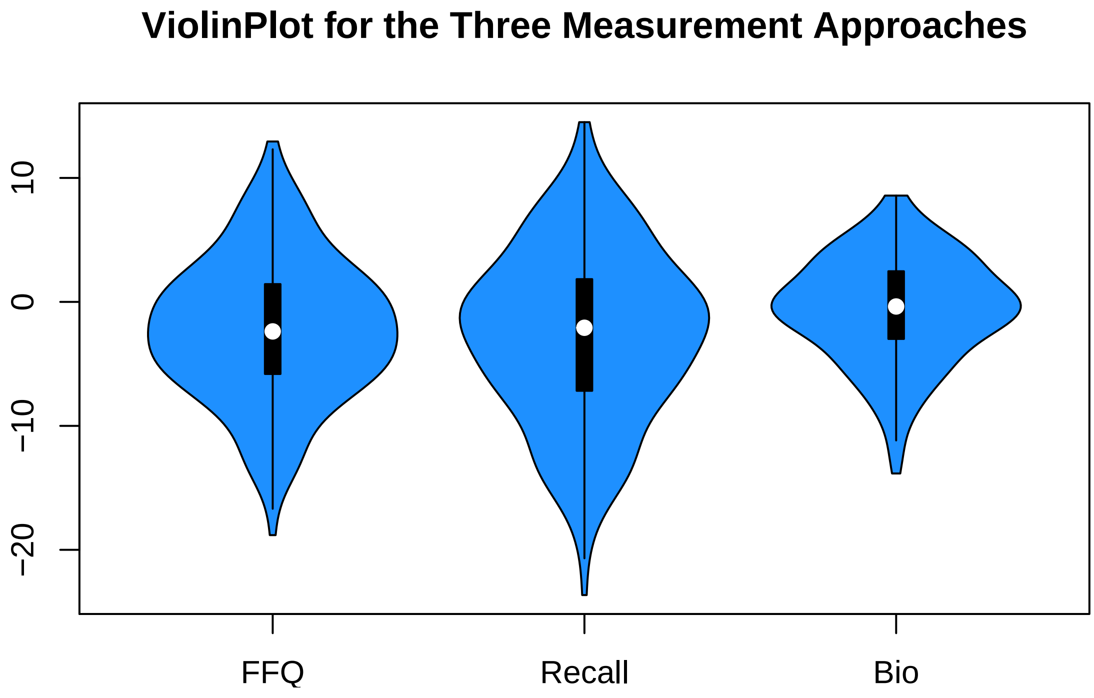
```
# note that there is some repitition in this data.
# For more accurate variance representations within the boxplot, I will
# deduplicate the table so we only have one 'average' measurement per id number
# this will halve the number of records in the table
dedup <- sim %>%
  dplyr::select(truth, avgffq, avgrecall, avgbio) %>%
  unique()

ffq_error <- dedup$truth - dedup$avgffq
recall_error <- dedup$truth - dedup$avgrecall
bio_error <- dedup$truth - dedup$avgbio
```

```
boxplot(ffq_error, recall_error, bio_error
        , names = c("FFQ", "Recall", "Bio")
        , xlab="Measurement Approach"
        , ylab="Measurement Error"
        , main="BoxPlot for the Three Measurement Approaches"
        )
```



I am also going to include a violin plot becaue I find the distribution aspect useful.

```
vioplot::vioplot(ffq_error, recall_error, bio_error, col="dodgerblue"
                 , names = c("FFQ", "Recall", "Bio")
                    )
title("ViolinPlot for the Three Measurement Approaches")
```



Based on the above graphics, I believe that the Biomarker measurement approach is the best of the three.

Each approach is very similar in terms of mean error, but the Biomarker approach has decidedly lower variance than the other two. To me, this gives some indication that it could be more reliable than the other approaches.

Question 2

Fit two random intercept spline models with the responses being the biomarkers, and the predictors being FFQ and 24hr summary.

```
mod_one <- mgcv::gamm(bio ~ s(ffq)
                     , random=list(id= ~ 1)
                     , data=sim
                      )
summary(mod_one$gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## bio ~ s(ffq)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.4184     0.3825    74.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##        edf Ref.df    F  p-value
## s(ffq)   1      1 19.5 1.26e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0571
##   Scale est. = 37.497    n = 446
```

```
mod_two <- mgcv::gamm(bio ~ s(recall)
                     , random=list(id= ~ 1)
                     , data=sim
                      )
summary(mod_two$gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## bio ~ s(recall)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.4184     0.3946   72.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df     F p-value
## s(recall)   1      1 4.177  0.0416 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0181
##   Scale est. = 37.559    n = 446
```

From the above output, it appears that both FFQ and 24hr recall are significant predictors of the number of Biomarkers, but FFQ is significant at the <1% level, while 24hr recall is significant at the 5% level.

Question 3

What are the between and within standard deviations of the two fits?

```
summary(mod_one$lme)
```

```
## Linear mixed-effects model fit by maximum likelihood
##  Data: strip.offset(mf)
##       AIC      BIC    logLik
##   3015.17 3035.672 -1502.585
##
## Random effects:
##  Formula: ~Xr - 1 | g
##  Structure: pdIdnot
##                  Xr1          Xr2          Xr3          Xr4          Xr5
## StdDev: 0.0005797775 0.0005797775 0.0005797775 0.0005797775 0.0005797775
##                  Xr6          Xr7          Xr8
## StdDev: 0.0005797775 0.0005797775 0.0005797775
##
##  Formula: ~1 | id %in% g
##         (Intercept) Residual
## StdDev:    3.715664  6.12351
##
## Fixed effects: y ~ X - 1
##                  Value Std.Error  DF  t-value p-value
## X(Intercept) 28.418408 0.3829405 222 74.21103       0
## Xs(ffq)Fx1    1.583594 0.3590627 222  4.41036       0
##  Correlation:
##            X(Int)
## Xs(ffq)Fx1 0
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.07916343 -0.63274713 -0.06217181  0.49384144  3.29921648
##
## Number of Observations: 446
## Number of Groups:
##         g id %in% g
##         1       223
```

```
summary(mod_two$lme)
```

```
## Linear mixed-effects model fit by maximum likelihood
##  Data: strip.offset(mf)
##        AIC      BIC    logLik
##   3029.412 3049.914 -1509.706
##
## Random effects:
##  Formula: ~Xr - 1 | g
##  Structure: pdIdnot
##                  Xr1         Xr2         Xr3         Xr4         Xr5
## StdDev: 0.001912287 0.001912287 0.001912287 0.001912287 0.001912287
##                  Xr6         Xr7         Xr8
## StdDev: 0.001912287 0.001912287 0.001912287
##
##  Formula: ~1 | id %in% g
##         (Intercept) Residual
## StdDev:    3.983204  6.12851
##
## Fixed effects: y ~ X - 1
##                    Value Std.Error  DF  t-value p-value
## X(Intercept)  28.418408 0.3950438 222 71.93736  0.0000
## Xs(recall)Fx1  0.699748 0.3427683 222  2.04146  0.0424
##  Correlation:
##               X(Int)
## Xs(recall)Fx1 0
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.02684777 -0.61720800 -0.08559678  0.49586058  3.12583469
##
## Number of Observations: 446
## Number of Groups:
##         g id %in% g
##         1       223
```

The between-group standard deviation for FFQ is 0.0005797775, while the between-group standard deviation for 24hr is 0.001912287; 24hr has a standard deviation that is over 3 times larger than that of FFQ.

The within-group standard deviation for FFQ is 3.715664, while the within-group standard deviation for the 24hr is 3.983204.

Question 4

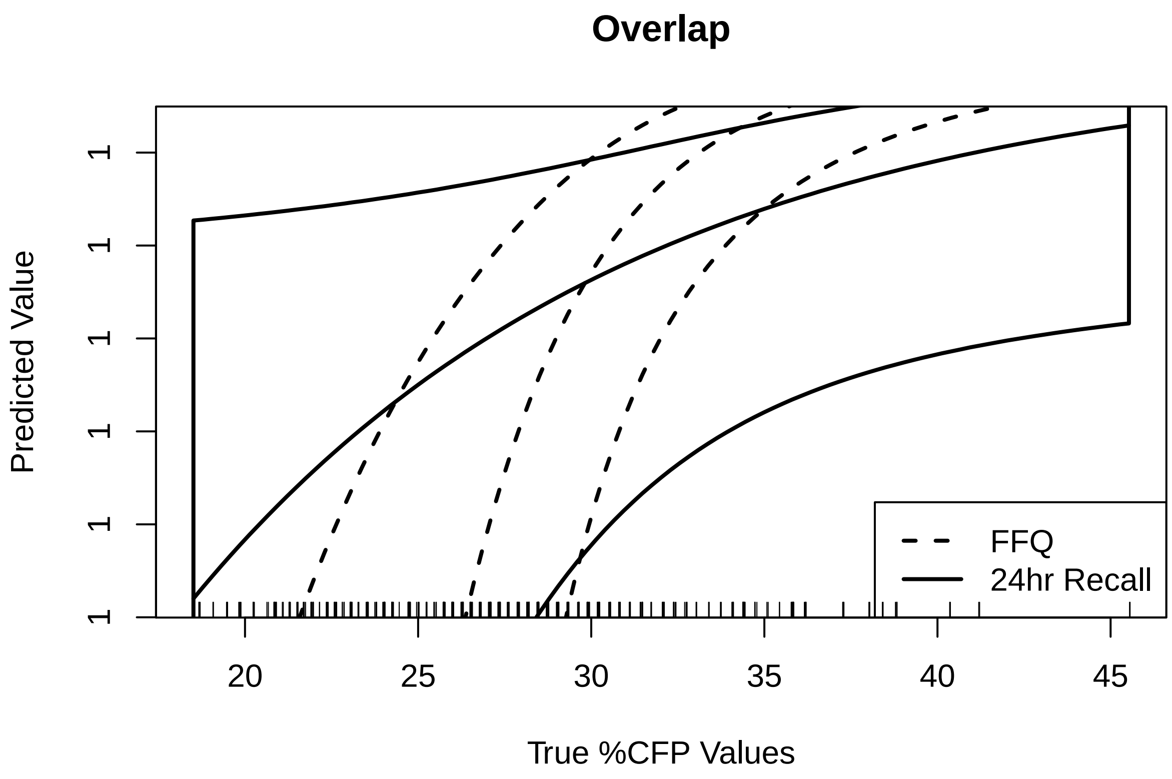
Display the fitted curves for both the *FFQ* and *24hr* in one graph. Fit them separately and then plot on the same graph.

```
ng <- 101

cfpg <- seq(from=min(sim$truth)
            , to=max(sim$truth)
            , length=ng
            )
```

```
newDataDF_one <- as.data.frame(cfpg)

names(newDataDF_one) <- c("ffq")
newDataList_one <- as.list(newDataDF_one)

predObj_one <- predict(mod_one$gam, newDataList_one, se=T)

muHatg_one <- 1/(1+exp(-predObj_one$fit))
aa_one     <- predObj_one$fit + 2*predObj_one$se
bb_one     <- predObj_one$fit - 2*predObj_one$se
lowergg_one <- 1 / (1 + exp(-bb_one))
uppergg_one <- 1 / (1 + exp(-aa_one))
```

```
newDataDF_two <- as.data.frame(cfpg)

names(newDataDF_two) <- c("recall")
newDataList_two <- as.list(newDataDF_two)

predObj_two <- predict(mod_two$gam, newDataList_two, se=T)

muHatg_two <- 1/(1+exp(-predObj_two$fit))
aa_two     <- predObj_two$fit + 2*predObj_two$se
bb_two     <- predObj_two$fit - 2*predObj_two$se
lowergg_two <- 1 / (1 + exp(-bb_two))
uppergg_two <- 1 / (1 + exp(-aa_two))
```

```
plot(cfpg, muHatg_two, type="n"
     , xlab="True %CFP Values"
     , ylab="Predicted Value"
     , main="Overlap"
     )
# ffq
polygon(c(cfpg, rev(cfpg)), c(lowergg_one, rev(uppergg_one))
         , border=T
         , lty=2
         , lwd=2
        )
lines(cfpg, muHatg_one, col="black", lty=2, lwd=2)

# 24hr recall
polygon(c(cfpg, rev(cfpg)), c(lowergg_two, rev(uppergg_two))
       , border=T
       , lwd=2
        )
lines(cfpg, muHatg_two, col="black", lwd=2)
rug(jitter(sim$truth))
legend("bottomright", c("FFQ", "24hr Recall"), lty=c(2,1), lwd=c(2,2))
```



Question 5

In the models from *Question 2*, add *age* and *BMI* as linear predictors. Which are statistically significant?

```
mod_five_one <- mgcv::gamm(bio ~ s(ffq) + age + bmi
                           , random=list(id = ~ 1)
                           , data=sim
                           )
summary(mod_five_one$gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## bio ~ s(ffq) + age + bmi
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.304895   3.178694  10.163   <2e-16 ***
## age          0.006005   0.048444   0.124    0.901
## bmi         -0.187327   0.076744  -2.441    0.015 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##         edf Ref.df     F  p-value
## s(ffq)   1      1 18.69 1.89e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0686
##   Scale est. = 37.48     n = 446
```

```
mod_five_two <- mgcv::gamm(bio ~ s(recall) + age + bmi
                           , random=list(id = ~ 1)
                           , data=sim
                           )
summary(mod_five_two$gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## bio ~ s(recall) + age + bmi
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.15471    3.21788  10.614  < 2e-16 ***
## age         -0.01833    0.04917  -0.373  0.70950
## bmi         -0.21254    0.07866  -2.702  0.00716 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df     F p-value
## s(recall) 1.708  1.708 2.448  0.0534 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0408
##   Scale est. = 37.421    n = 446
```

In the random-intercept model of *Biomarkers* on *FFQ*, *BMI* is significant, along with the smoothed *FFQ* term.

In the random-intercept model of *Biomarkers* on *24hr*, *BMI* is highly significant, but the smoothed *24hr* term has lost its significance, when compared with the previously fit model.