Homework 12

Skylar Liu

2024-04-06

Homework #12, Stat 660, Spring 2024 Due 11:59 PM Central, April 10

Consult the directory OPEN Data, and get the files OPEN_sim.csv and OPEN_README.pdf. The README file will tell you about the data, so look at it carefully for definitions. You will be comparing different estimates of the percentage of calories from protein (%CFP) for women. This is simulated data but based on actual data that I cannot share. There is an ID number there that makes this a repeated measures problem.

```
rm(list = ls())
set.seed(382957)
options(repos = list(CRAN="http://cran.rstudio.com/"))

# import data
data = read.csv("~/660 - Flexible Regression/Homework/Homework12/OPEN_sim_2019.csv")
```

1. Compare the true %CFP to the average %CFP estimated by each instrument (AvgFFQ, AvgRecall, and AvgBio). Make a boxplot comparing the three instruments, where you take their differences with the truth. Describe why you think one of the instruments is best.

```
# Variable creation
truth  = data$Truth
FFQ    = data$FFQ
recall = data$Recall
bio    = data$Bio
ID     = data$ID

# Average %CFP for each instrument
mean(truth)
```

```
## [1] 27.80081
```

```
mean(FFQ)
```

```
## [1] 30.02108
```

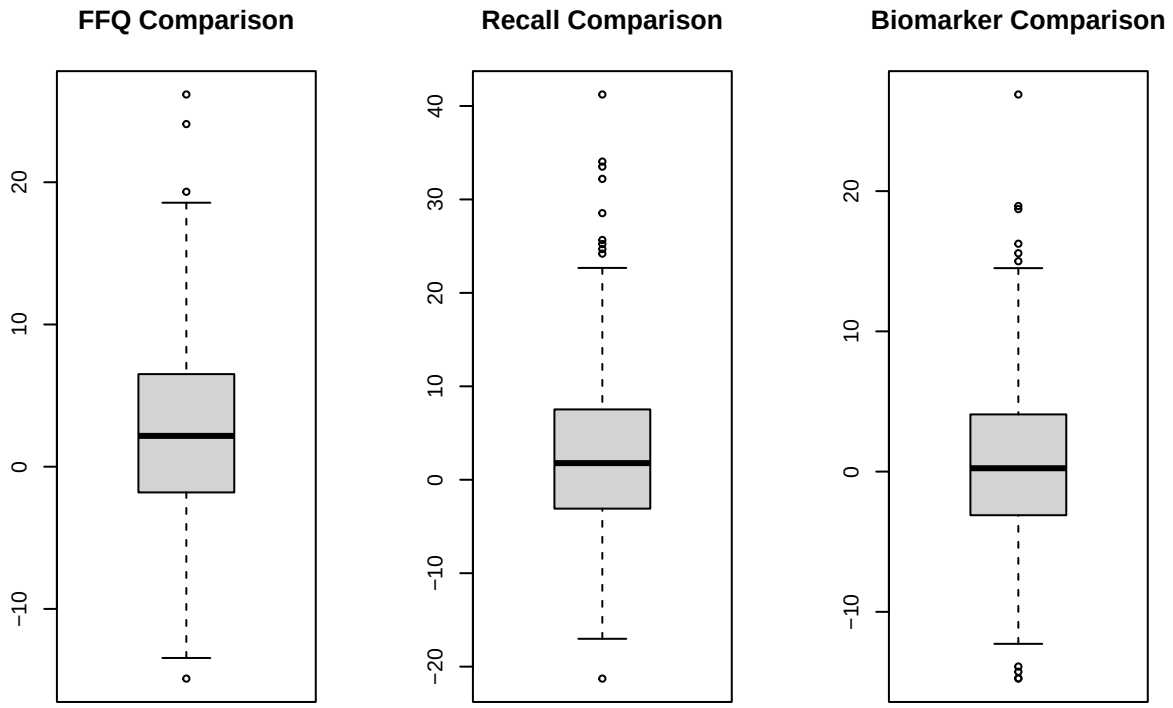
```
mean(recall)
```

```
## [1] 30.5241
```

```r
mean(bio)
```

```
## [1] 28.41841
```

```r
# Boxplots
par(mfrow = c(1,3))
boxplot(FFQ - truth, main = "FFQ Comparison")
boxplot(recall - truth, main = "Recall Comparison")
boxplot(bio - truth, main = "Biomarker Comparison")
```



> The Biomarker estimate is the best instrument, as it has the smallest variability and median difference close to 0. 24 Hour Recall is the worst instrument since it has almost twice the amount of variability as the other two. FFQ is a close second to biomarker, but has slightly larger variation and a median farther from 0.

2. Fit a random intercept spline model with the response being the biomarker (Bio in the data set), and the predictors being the FFQ and the 24HR, both modeled as splines. >a. Display the summaries of the fit. Tell us which, if any, is a statistically significant predictor of the biomarker.

```r
library(gamm4)
```

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
## Loading required package: mgcv

## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:lme4':
##
##     lmList

## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.

## This is gamm4 0.2-6
```

```
# Random intercept spline model
fitbio = gamm4(bio ~ s(I(FFQ),k=4,bs="cr") + s(I(Recall),k=4,bs="cr"),
               random =  ~(1|ID), data = data)
summary(fitbio)
```

```
##     Length Class   Mode
## mer  1     lmerMod S4
## gam 32     gam     list
```

```
summary(fitbio$gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## bio ~ s(I(FFQ), k = 4, bs = "cr") + s(I(Recall), k = 4, bs = "cr")
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.4184     0.3795   74.88   <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(I(FFQ))    1.000  1.000 15.648 8.92e-05 ***
## s(I(Recall)) 1.706  1.706  2.234    0.218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0657
## lmer.REML =   2997  Scale est. = 37.956    n = 446
```

FFQ is a statistically significant predictor of biomarker, with a p-value of 8.92x10^-5. Recall is not statistically significant, with a p-value of 0.218 which is greater than alpha=0.05.

3. What are the between and within standard deviations of the fit? This is a single fit.

```
summary(fitbio$mer)
```

```
## Linear mixed model fit by REML ['lmerMod']
##
## REML criterion at convergence: 2997
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1027 -0.6352 -0.0472  0.4661  3.2617
##
## Random effects:
##  Groups   Name         Variance Std.Dev.
##  ID       (Intercept)  13.14    3.6250
##  Xr.0     s(I(Recall))  0.50    0.7071
##  Xr       s(I(FFQ))     0.00    0.0000
##  Residual              37.96    6.1608
## Number of obs: 446, groups:  ID, 223; Xr.0, 2; Xr, 2
##
## Fixed effects:
##                   Estimate Std. Error t value
## X(Intercept)      28.4184     0.3795  74.881
## Xs(I(FFQ))Fx1      8.0888     2.0448   3.956
## Xs(I(Recall))Fx1   1.1889     2.0770   0.572
##
## Correlation of Fixed Effects:
##             X(Int) X(I(FF
## X(I(FFQ))F1  0.000
## Xs(I(Rc))F1  0.000 -0.105
```

The between standard deviation is 3.625 while the within standard deviation is 6.161.

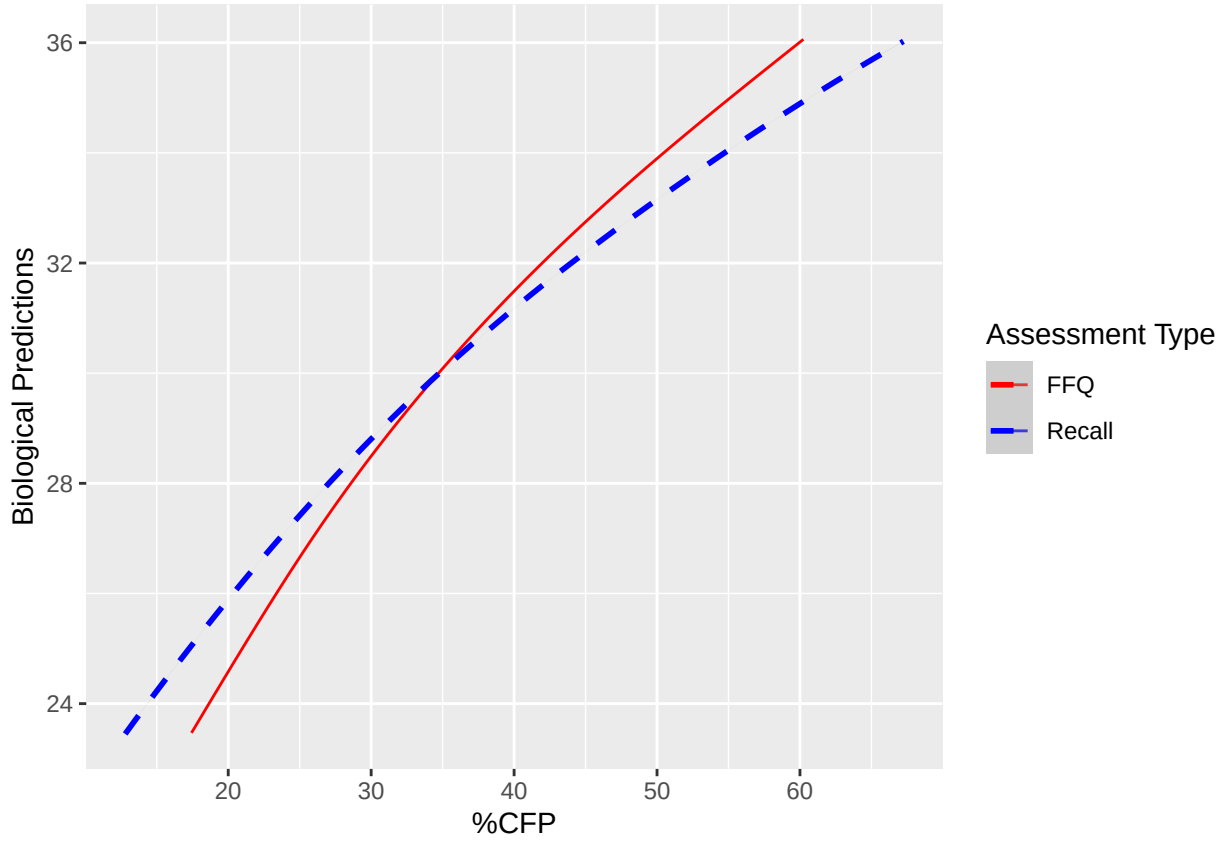
4. Display the fitted curves for both the FFQ and the 24HR in one graph. I suggest that you set up a grid of %CFP as your x-values and use that grid when predicting responses from your model.

Hint: If your gamm or gamm4 object is called fit, to make predictions you must type predict(fit$gam, ...).

```
# Setup
newdata <- data.frame(FFQ = seq(min(data$FFQ), max(data$FFQ), length.out = 446),
                      Recall = seq(min(data$Recall), max(data$Recall), length.out = 446))
newdata$pred <- predict(fitbio$gam, newdata = newdata, type = "response")

# Plot fixed effects function against num.weeks with 95% confidence intervals
library(ggplot2)
ggplot(newdata, aes(x = FFQ, y = pred, color = "FFQ")) +
  geom_line() +
  geom_smooth(aes(x = Recall, y = pred, color = "Recall"), linetype = "dashed") +
  labs(x = "%CFP", y = "Biological Predictions", color = "Assessment Type") +
  scale_color_manual(values = c("red", "blue"), labels = c("FFQ", "Recall"))
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



5. In the model from question (2), add Age and BMI as linear predictors. Are either statistically significant predictors?

```
fitnew = gamm4(bio ~ s(I(FFQ),k=4,bs="cr") + s(I(Recall),k=4,bs="cr") + Age + BMI,
               random =  ~(1|ID), data = data)
summary(fitnew$gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## bio ~ s(I(FFQ), k = 4, bs = "cr") + s(I(Recall), k = 4, bs = "cr") +
##     Age + BMI
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.06057    3.15630  10.158   <2e-16 ***
## Age          0.01422    0.04825   0.295   0.7684
## BMI         -0.19575    0.07620  -2.569   0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                edf Ref.df      F  p-value
## s(I(FFQ))    1.000  1.000 14.789 0.000139 ***
```

```
## s(I(Recall)) 1.819  1.819  2.694 0.143841
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0799
## lmer.REML = 2997.8  Scale est. = 37.996    n = 446
```

BMI is statistically significant at the 0.05 level with a p-value of 0.0105. Age is not statistically significant since the p-value of 0.768 is greater than an alpha of 0.05.

6. I have no idea if a random intercept model is sufficient. With the Indiana data , we were able to use gamm4 to test whether we needed a random intercept model or a random function model. In these data, see if the biomarker (not the truth) needs a random function model or a random intercept model. >a. Of course, you can test this in gamm4.

```
# random function model
fitrand = gamm4(Bio ~ s(FFQ, bs="cr")
                + s(FFQ, factor(ID), bs="re", xt=list(bs="cr"))
                + s(Recall, bs="cr")
                + s(Recall, factor(ID), bs="re", xt=list(bs="cr"))
                + Age + BMI, data = data)

# test whether random functio is needed
anova(fitrand$mer,fitnew$mer)
```

```
## refitting model(s) with ML (instead of REML)

## Data: NULL
## Models:
## fitnew$mer: NULL
## fitrand$mer: NULL
##             npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## fitnew$mer     9 3014.6 3051.5 -1498.3   2996.6
## fitrand$mer   10 3017.7 3058.7 -1498.9   2997.7     0  1          1
```

The p-value from the ANOVA is essentially 1, indicating that a random function model is not necessary.

7. Run a random intercept logistic spline regression with Y = the indicator that Bio < 27.5, X = FFQ, and Z = (Age, BMI). Here Y is the binary response, X enters the model as a spline with a random intercept, and Z are linear predictors.

```
# Create Y variable for Bio < 27.5
data$Y = ifelse(data$Bio < 27.5, 1, 0)

# Random intercept logistic spline regression
logfit = gamm4(Y ~ s(FFQ, k=4, bs="cr") + Age + BMI,
                random= ~(1|ID),
                family = binomial, data = data)
```

8. Which among X and Z are statistically significant predictors?

```r
summary(logfit$gam)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## Y ~ s(FFQ, k = 4, bs = "cr") + Age + BMI
##
## Parametric coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.1688205  0.9689782  -1.206   0.2277
## Age         -0.0002577  0.0147692  -0.017   0.9861
## BMI          0.0472401  0.0232713   2.030   0.0424 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##        edf Ref.df Chi.sq  p-value
## s(FFQ)   1      1  11.53 0.000683 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0308
## glmer.ML = 476.86  Scale est. = 1         n = 446
```

At the alpha=0.05 level, BMI (p-value of 0.0424) and FFQ (p-value of 0.000683) are significant.

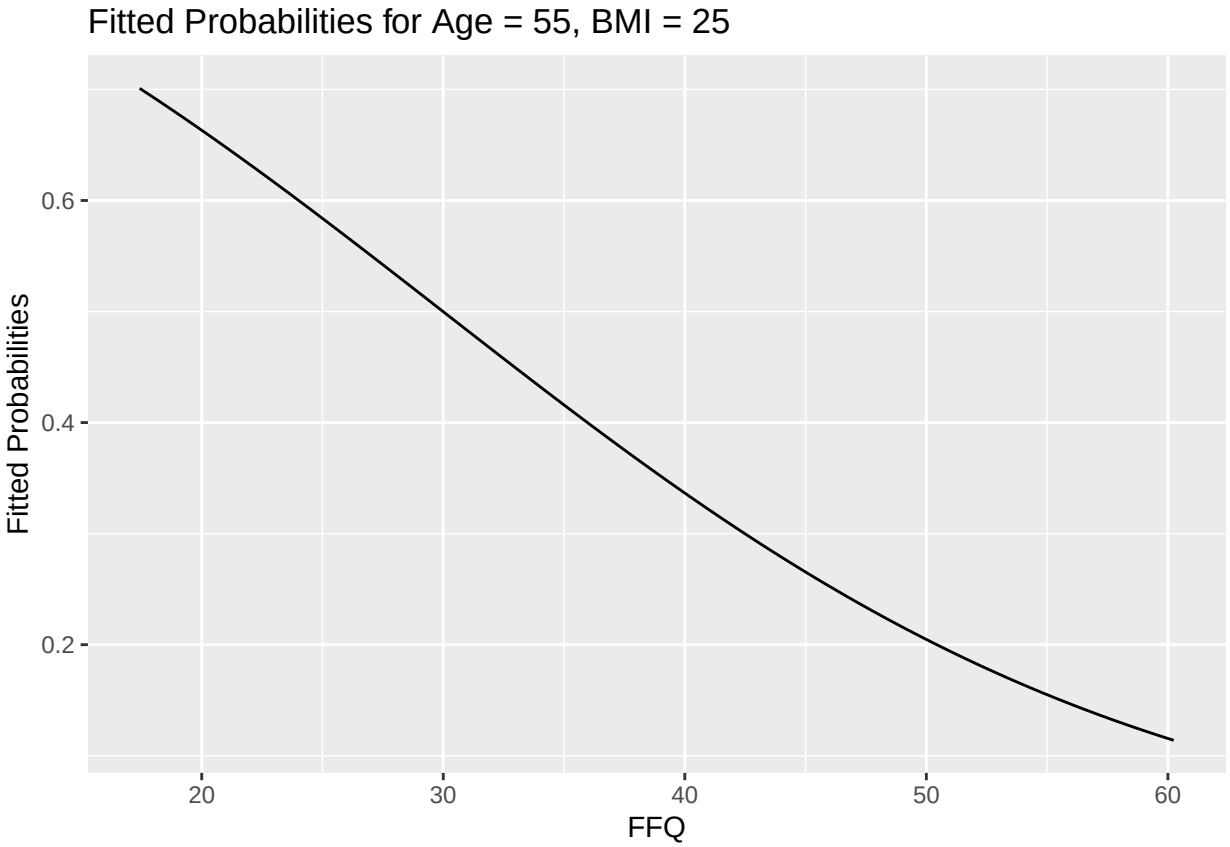
9. From the model in (7), graph the fitted probabilities for people who are 55 years old and whose BMI = 25.

```r
# New dataframe
probdata <- data.frame(FFQ = seq(min(data$FFQ), max(data$FFQ), length.out = length(FFQ)),
                       Age = 55,
                       BMI = 25)

# Predicted values
predictprobs <- predict(logfit$gam, newdata = probdata, type = "response")

# Plotted dataframe
plottedvalues <- data.frame(FFQ = probdata$FFQ, Fitted_Probability = predictprobs)

# Plot fitted values
ggplot(plottedvalues, aes(x = FFQ, y = Fitted_Probability)) +
  geom_line() +
  labs(x = "FFQ", y = "Fitted Probabilities", title = "Fitted Probabilities for Age = 55, BMI = 25")
```


Fitted Probabilities for Age = 55, BMI = 25
Fitted Probabilities
0.6
0.4
0.2
20
30
40
50
60
FFQ