Homework 10

Skylar Liu

2024-03-17

Homework #10, Stat 660, Spring 2024, Due Tuesday March 19 by 11:59PM

1. We have talked about the spinal bone mineral density data, with the random intercept case. I stated that I think, or perhaps hope, that the id numbers should not have to be 1,2,3... I also think/hope that the idnumbers do not need to be ordered. It is generally the case that data sets are "cleaned" and there is a code book that converts the actual ids to 1,2,3,...,n. This helps de-identify data and helps preserve anonymity.

    a. Test this out in the spinal bone mineral density data, by defining a new variable, femSBMD$idnum2 = 2 * femSBMD$idnum.

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

# load libraries
library(HRW)
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
```

```r
# import data
femSBMD = read.csv("~/660 - Flexible Regression/Homework/Homework10/femSBMD.csv")

# add new variable
femSBMD$idnum2 = 2*femSBMD$idnum
```

b. Then rerun the gamm4::gamm4 given in class to see if you get the same results. I think/hope you will.

```r
# fit from class
fitclass <- gamm4(spnbmd ~ s(age,k=10,bs="cr") + black
                + hispanic + white,
                random= ~(1|idnum),data = femSBMD)

fitLclass <- gamm4(spnbmd ~ I(age) + black + hispanic + white,
                random= ~(1|idnum),data = femSBMD)

# fit with new variable
fit <- gamm4(spnbmd ~ s(age,k=10,bs="cr") + black
                + hispanic + white,
                random= ~(1|idnum2),data = femSBMD)

fitL <- gamm4(spnbmd ~ I(age) + black + hispanic + white,
                random= ~(1|idnum2),data = femSBMD)

# comparison
anova(fitclass$mer,fit$mer)
```

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

## Data: NULL
## Models:
## fitclass$mer: NULL
## fit$mer: NULL
##               npar     AIC     BIC logLik deviance Chisq Df Pr(>Chisq)
## fitclass$mer    8 -2404.4 -2365.1 1210.2  -2420.4
## fit$mer         8 -2404.4 -2365.1 1210.2  -2420.4     0  0
```

```r
anova(fitLclass$mer,fitL$mer)
```

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

## Data: NULL
## Models:
## fitLclass$mer: NULL
## fitL$mer: NULL
##                npar     AIC     BIC logLik deviance Chisq Df Pr(>Chisq)
## fitLclass$mer    7 -2030.8 -1996.4 1022.4  -2044.8
## fitL$mer         7 -2030.8 -1996.4 1022.4  -2044.8     0  0
```

The p-value from the anova for comparing both the fit and fitL from class compared to with the new variable is 0, showing there is no difference between these models.

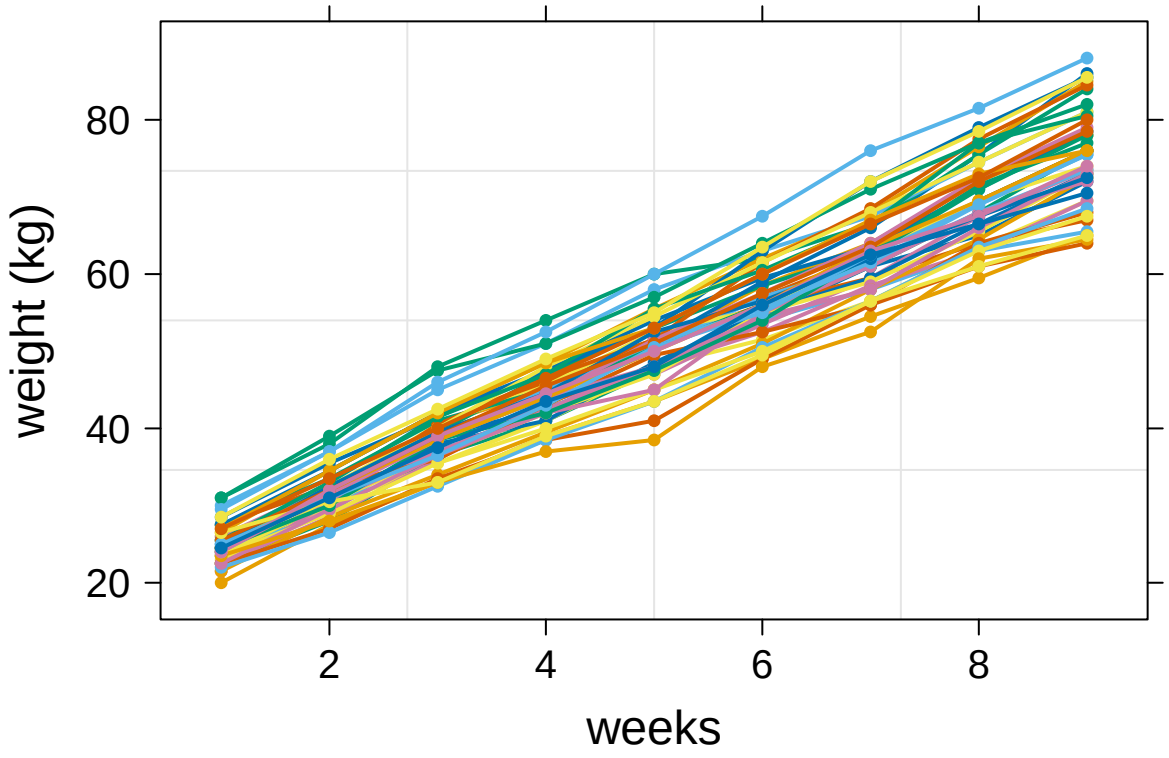
2. Get the data set pigWeights.csv from Canvas. The variable weight is the response, and the variable num.weeks is the date of the repeated measures.

    a. Display the lattice plot from library(lattice). Use the example from the spinal bone mineral density data to do this. I covered this in class, but in this case there is only 1 population and no ANCOVA.

```
# import data
pigWeights = read.csv("~/660 - Flexible Regression/Homework/Homework10/pigWeights.csv")

# load packages
library(lattice)

# lattice plot
pigWeightsvis <- xyplot(weight ~ num.weeks,
                        group = id.num,as.table = TRUE,
                        data = pigWeights,
                        strip = strip.custom(par.strip.text
                                             = list(cex = 1.5)),
                        par.settings = list(layout.heights
                                            =list(strip=1.6)),
                        scales = list(cex = 1.25),
                        xlab = list("weeks",cex = 1.5),
                        ylab = list(expression(paste(
                          "weight (kg)")),
                          cex = 1.5),
                        panel = function(x,y,subscripts,groups)
                        {
                          panel.grid()
                          panel.superpose(x,y,subscripts,groups,
                                          type = "b",pch = 16,lwd = 2)
                        })
plot(pigWeightsvis)
```



b. Looking at the data, do you think a random intercept model holds for these data? Why or why not? You might want to look at Lecture 15 where I described the means and variances of a random intercept model. It is a subjective call, but just answer it.

Each line seems to have the same shape and linear slope with minimal changes of subject to subject variance, so I believe a random intercept model holds for this data.

c. Fit the random intercept model with num_weeks modeled as a spline. Do a summary and show your results. Show the between-person variance of the intercept and the within-person variance of the random errors. You may use either mgcv::gamm or gamm4::gamm4. They should be similar because gamm and gamm4 are theoretically justified in this family=gaussian case

```
# random intercept gamm
fit <- gamm4(weight ~ s(num.weeks,k=9,bs="cr"),
             random= ~(1|id.num),data = pigWeights)
summary(fit$mer)
```

```
## Linear mixed model fit by REML ['lmerMod']
##
## REML criterion at convergence: 2027
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.8105 -0.5407  0.0069  0.4755  3.9437
##
## Random effects:
##  Groups   Name         Variance  Std.Dev.
##  id.num   (Intercept)  15.152336 3.89260
##  Xr       s(num.weeks)  0.003797 0.06162
##  Residual               4.297895 2.07314
## Number of obs: 432, groups:  id.num, 48; Xr, 7
##
## Fixed effects:
##                   Estimate Std. Error t value
## X(Intercept)       50.4051     0.5706   88.33
## Xs(num.weeks)Fx1   48.1016     0.2992  160.75
##
## Correlation of Fixed Effects:
##             X(Int)
## Xs(nm.wk)F1 0.000
```

The between-person variance of the intercept is 15.152 while the within-person variance of the random errors is 4.298.

d. Using anova() in gamm4::gamm4, to test whether a spline is needed as compared to a linear and a quadratic effect.

```
# test whether a spline is needed
fitL <- gamm4(weight ~ I(num.weeks),
              random= ~(1|id.num),data = pigWeights)
fitQ <- gamm4(weight ~ I(num.weeks) + I(num.weeks^2),
              random= ~(1|id.num),data = pigWeights)

anova(fit$mer,fitL$mer)
```

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

## Data: NULL
## Models:
## fitL$mer: NULL
## fit$mer: NULL
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fitL$mer    4 2037.8 2054.1 -1014.9   2029.8
## fit$mer     5 2037.1 2057.5 -1013.6   2027.1 2.7395  1     0.0979 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit$mer,fitQ$mer)
```

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

## Data: NULL
## Models:
## fit$mer: NULL
## fitQ$mer: NULL
```

```
##          npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## fit$mer     5 2037.1 2057.5 -1013.6   2027.1
## fitQ$mer    5 2039.1 2059.4 -1014.5   2029.1     0  0
```

Compared to both linear and quadratic fits, a spline is necessary only for the quadratic fit as the p-value is very small (~0). A spline may not be necessary compared to the linear fit sincce the p-value is slightly above an alpha of 0.05 (p-value = 0.098)

e. Compare the quadratic and linear fits as well.

```
# comparing linear vs quadratic fit
anova(fitQ$mer,fitL$mer)
```

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

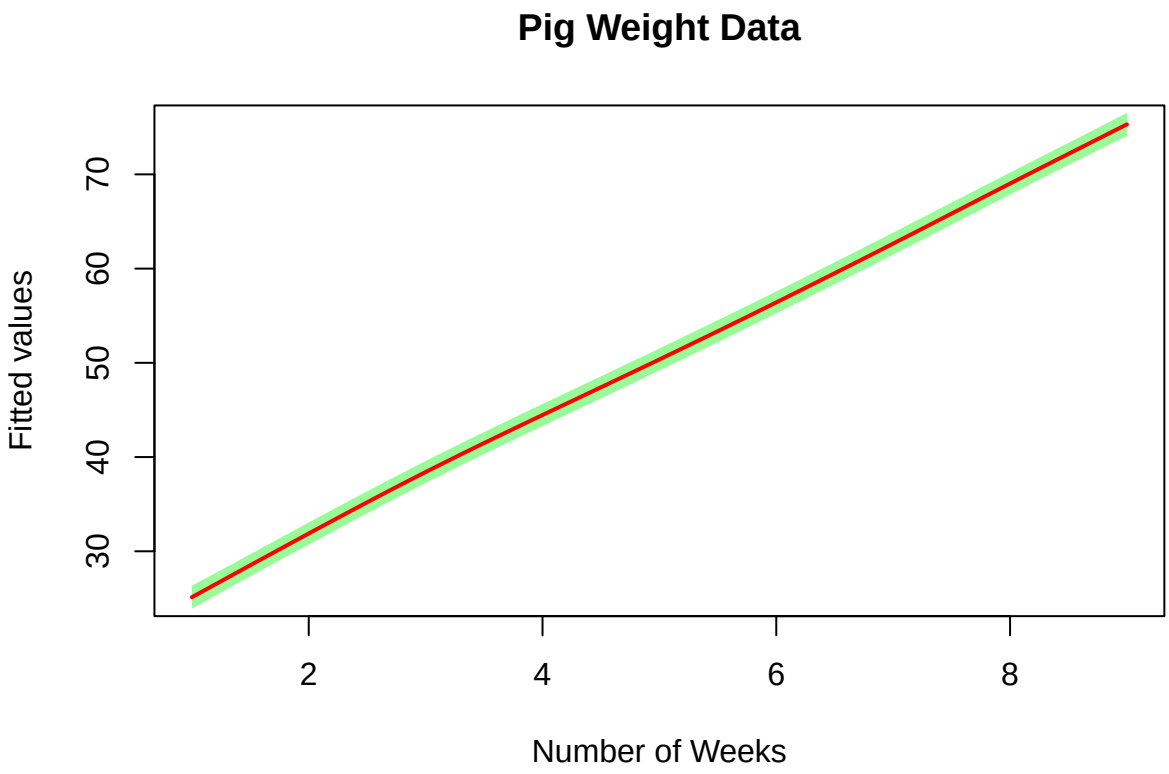
```
## Data: NULL
## Models:
## fitL$mer: NULL
## fitQ$mer: NULL
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fitL$mer    4 2037.8 2054.1 -1014.9   2029.8
## fitQ$mer    5 2039.1 2059.4 -1014.5   2029.1 0.7488  1     0.3869
```

The resulting p-value is 0.39 when comparing a quadratic fit to a linear fit, suggesting that the linear fit is the better fit.

f. As in the spinal bone mineral density data, plot the fixed effects function against num.weeks, and include a pointwise 95% confidence interval for it.

```
# setup predictions
ng <- 432
num.weeks = seq(min(pigWeights$num.weeks) , max(pigWeights$num.weeks),length=ng)
pred <- predict(fit$gam,newdata=data.frame(num.weeks = num.weeks),se.fit=TRUE)
lowdirg <- pred$fit - qnorm(0.975) * pred$se.fit
uppdirg <- pred$fit + qnorm(0.975) * pred$se.fit
ymin = min(min(pred$fit))
ymax = max(max(pred$fit))

# Plot fixed effects function against num.weeks with 95% confidence intervals
plot(0, type = "n",
     xlab = "Number of Weeks",
     ylab = "Fitted values",
     main = "Pig Weight Data",
xlim = c(min(num.weeks), max(num.weeks)), ylim = c(ymin, ymax))
polygon(c(num.weeks, rev(num.weeks)), c(lowdirg, rev(uppdirg)), col = "palegreen",
border = FALSE)
lines(num.weeks, pred$fit, col = "red", lwd = 2)
```



g. Since you have already computed var(U) and var(epsilon), what is the estimated within-person correlation for this model?

```
# run the fit summary again
summary(fit$mer)
```

```
## Linear mixed model fit by REML ['lmerMod']
##
## REML criterion at convergence: 2027
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8105 -0.5407  0.0069  0.4755  3.9437
##
## Random effects:
##  Groups   Name         Variance  Std.Dev.
##  id.num   (Intercept)  15.152336 3.89260
##  Xr       s(num.weeks)  0.003797 0.06162
##  Residual               4.297895 2.07314
## Number of obs: 432, groups:  id.num, 48; Xr, 7
##
## Fixed effects:
##                  Estimate Std. Error t value
## X(Intercept)      50.4051     0.5706   88.33
## Xs(num.weeks)Fx1  48.1016     0.2992  160.75
```

```
##
## Correlation of Fixed Effects:
##             X(Int)
## Xs(nm.wk)F1 0.000
```

```
# within person correlation calculation
15.152336/(15.152336 + 4.297895)
```

```
## [1] 0.7790312
```

The within-person correlation for this model is 0.779.