# 6. Analysis of Longitudinal Data

Hello!!



Welcome to the last of the weekly analysis entries! This week was interesting. Not only were there some major issues (more on those later), I also thought I’d experiment with the layout of the html some…

This week we will go back to building statistical models. Here is a quote I borrowed from the [course instructions](https://mooc.helsinki.fi/course/view.php?id=158#section-6):

*“The new challenge here is that the data may (and will) include two types of dependencies simultaneously: In addition to the more or less correlated variables that we have faced with all models and methods so far, the observations of the data will also be intercorrelated. Usually (in the above mentioned models and methods), we can (often pretty safely) assume that the observations are independent of each other. However, in longitudinal data this assumption seldom holds, because we have multiple observations or measurements of the same individuals. The concept of repeated measures highlights this phenomenon that is actually quite typical in many applications. Both types of dependencies must be taken into account; otherwise the models will be biased.”*

To cope with the type of setting described above, we will be applying **linear mixed effects models**.

Prior to creating this diary entry, I performed some data wrangling on two data sets, BPSR and RATS. Here is some information on the two datasets:

“40 male subjects were randomly assigned to one of two treatment groups and each subject was rated on the brief psychiatric rating scale (BPRS) measured before treatment began (week 0) and then at weekly intervals for eight weeks. The BPRS assesses the level of 18 symptom constructs such as hostility, suspiciousness, hallucinations and grandiosity; each of these is rated from one (not present) to seven (extremely severe). The scale is used to evaluate patients suspected of having schizophrenia” (from DataCamp).

RATS is from a “nutrition study conducted in three groups of rats (Crowder and Hand, 1990). The three groups were put on different diets, and each animal’s body weight (grams) was recorded repeatedly (approximately weekly, except in week seven when two recordings were taken) over a 9-week period. The question of most interest is whether the growth profiles of the three groups differ”. (p. 22 [here](https://mooc.helsinki.fi/pluginfile.php/29659/course/section/2208/MABS4IODS-Part6.pdf)).

So, let’s get to work, shall we!

Just to add interest (and some challenges), this week we are swapping data sets. In practice, this means that instead of conducting Chapter 8 of MABS exercises using the BPSR data we are using RATS. Same with Chapter 9, BPSR instead of RATS.

Oh, rats!



Please note that from section 1C onwards I encountered a serious problem with R Markdown. The chapter itself knitted absolutely fine to html, but as I tried to knit my index.Rmd file, the knitting failed. After various attempts at fixing the issue, and spending 2 days with it, I had to admit defeat. So, in order to have a chapter to submit at all this week, I have had to disable various code chunks. Where this is the case, I have included a screen shot of my codes and outputs. My apologies for the resultant somewhat messy report this week!

## Part 1. Implementing the Analyses of Chapter 8 of MABS Using the RATS Data

Turns out I need to include the code from meet\_and\_repeat.R here in order for the Markdown to knit to html…

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

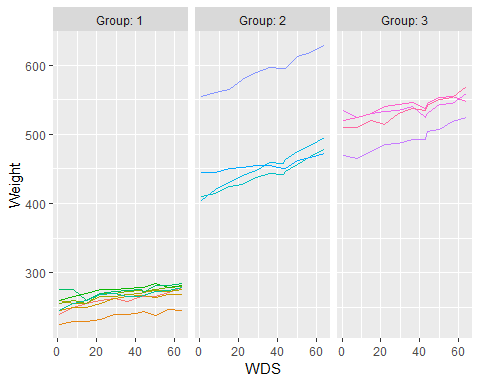
library(tidyr)  
RATS <- read.table("https://raw.githubusercontent.com/KimmoVehkalahti/MABS/master/Examples/data/rats.txt", header = TRUE, sep = '\t')  
  
RATS$ID <- factor(RATS$ID)  
RATS$Group <- factor(RATS$Group)  
RATSL <- RATS %>%  
 gather(key = WD, value = Weight, -ID, -Group) %>%  
 mutate(WDS = as.integer(substr(WD,3,4)))

### 1A. Individuals on the plot

First, let’s look at some graphical representations of the long RATS data.

To begin we shall plot the Weight values for all 16 rats, and we will differentiate between the diet groups into which the rats have been placed. These simple graphs make a number of features of the data readily apparent.

#Access the package ggplot2  
library(ggplot2)  
  
# Draw the plot  
p <- ggplot(RATSL, aes(x = WDS, y = Weight, col = ID))  
p1 <- p + geom\_line()  
p2 <- p1 + scale\_linetype\_manual(values = rep(1:10, times=4))  
p3 <- p2 + facet\_grid(. ~ Group, labeller = label\_both)  
p4 <- p3 + theme(legend.position = "none")  
p5 <- p4 + scale\_y\_continuous(limits = c(min(RATSL$Weight), max(RATSL$Weight)))  
p5



Inspecting the graphs above, we can see how the rats’ weight increases over time in all the different groups. These plots are helpful in visualizing the weight growth profiles of individual rats. As we can see, rats with lower starting body weight have been grouped together, and that group shows the least actual (as opposed to proportional) weight increase.

### 1B. The Golden Standardise

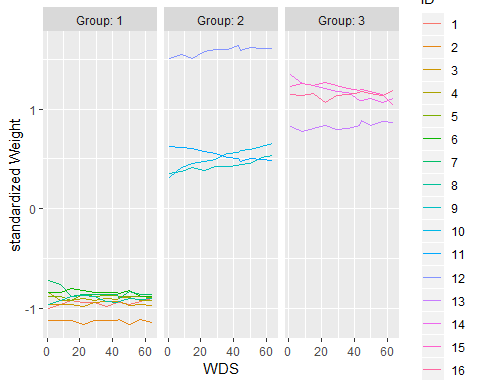
Let’s take a look at how our data behaves once we standardize the Weightvariable. We will perform this by subtracting the relevant mean from the original value and then dividing by the standard deviation. Here’s the math for it:

And here comes the code for the graph:

# Standardise the variable Weight  
RATSL <- RATSL %>%  
 group\_by(WDS) %>%  
 mutate(stdweight = (Weight - mean(Weight))/sd(Weight) ) %>%  
 ungroup()  
  
# Glimpse the data  
glimpse(RATSL)

## Observations: 176  
## Variables: 6  
## $ ID <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1...  
## $ Group <fct> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 1, 1...  
## $ WD <chr> "WD1", "WD1", "WD1", "WD1", "WD1", "WD1", "WD1", "WD...  
## $ Weight <int> 240, 225, 245, 260, 255, 260, 275, 245, 410, 405, 44...  
## $ WDS <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 8, 8...  
## $ stdweight <dbl> -1.0011429, -1.1203857, -0.9613953, -0.8421525, -0.8...

# Plot again with the standardised bprs  
ggplot(RATSL, aes(x = WDS, y = stdweight, col = ID)) +  
 geom\_line() +  
 scale\_linetype\_manual(values = rep(1:10, times=4)) +  
 facet\_grid(. ~ Group, labeller = label\_both) +  
 scale\_y\_continuous(name = "standardized Weight")



Interesting results! In Group 1, then standardized weights of the rats were much closer to one another at the end of the experiment - see how the lines get closer to each other towards the end?

In Group 2 we can see all sorts going on. The rats with the lowest standardized weight had a definite increase, while the largest rat had a moderate increase and the medium-sized one’s standardized weight declined.

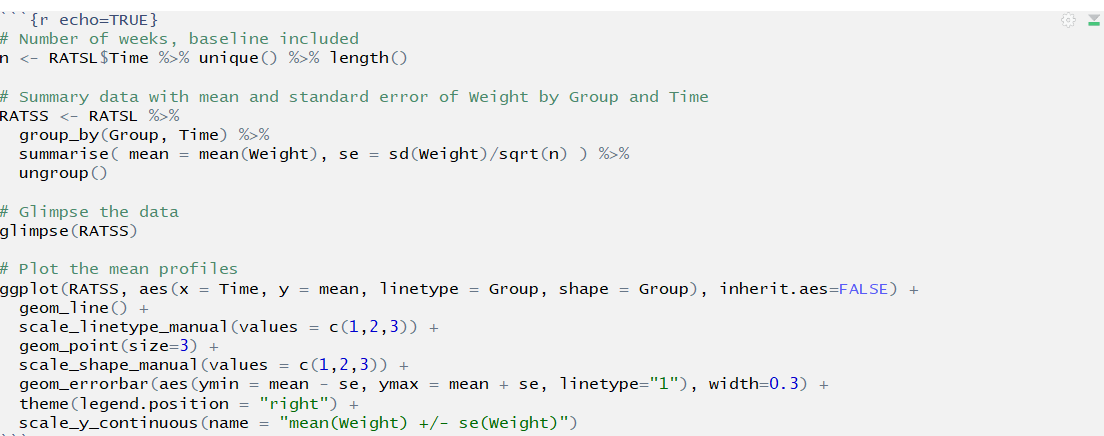
In Group 3 the weights were closer to one another at the end of the experiment, similarly to Group 1.

### 1C. Good things come in Summary graphs

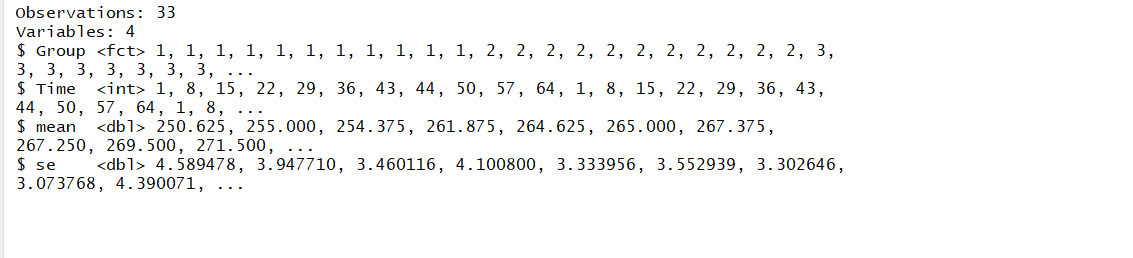
With a fairly large numbers of observations, graphical displays of individual profiles are often of little use and researchers commonly produce graphs showing average (mean) profiles for each group along with some indication of the variation of the observations at each time point, in this case the standard error of mean:

Although in our RATS dataset the number of individuals is much smaller (16) than in the BPRS set (40), we will do this nonetheless.

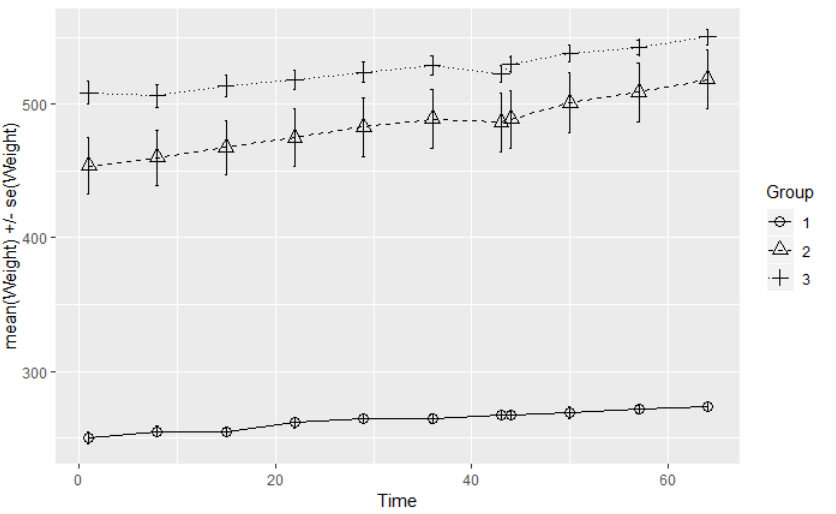
**Below are the first of the screen prints that became necessary**



**code chunk**



**console output**



**graph**

Note that unlike in the BPRS data, there is no overlap in the mean profiles of the three groups. This suggests there is a difference between the three groups in respect to the mean Weight values. Little rats remain little rats!

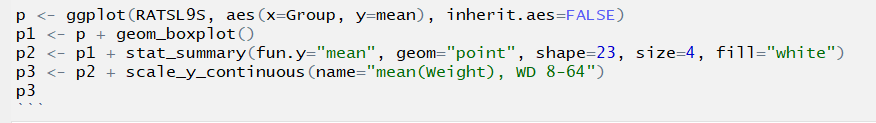
### 1D. Find the outlaw… Outlier!

It is time to find the outlier rat - if there is one! We will look into the post-diet values of RATS. The mean of the 9 weeks of data will be our summary measure. We will first calculate this measure and look at boxplots of the measure for each diet group.

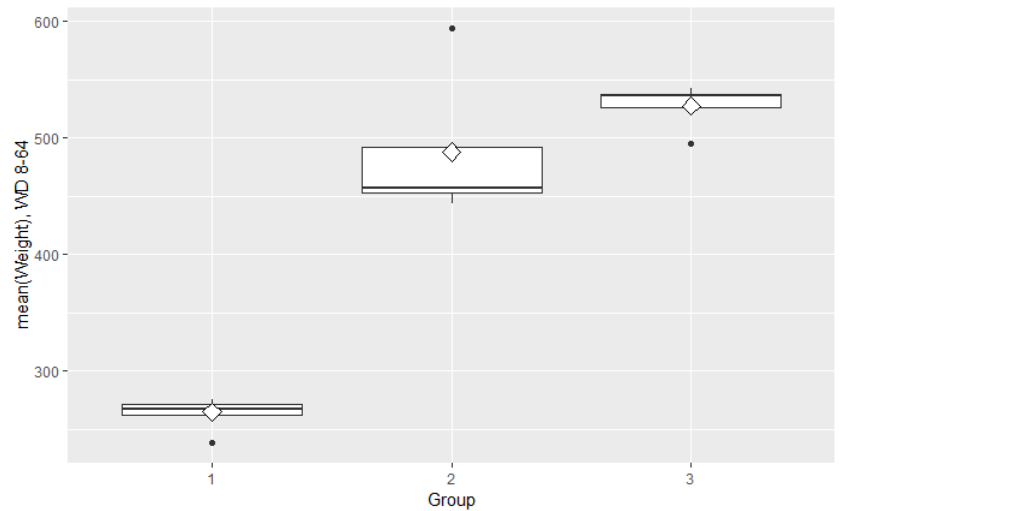
**In the below code chunk, everything works fine up until the ggplot part, suggesting that that is where the problem lies… However, for the life of me I can NOT figure out why it fails when knitting the index, and not before… So, after the output you will again find my screen prints.**

RATSL9S <- RATSL %>%  
 filter(WDS > 1) %>%  
 group\_by(Group, ID) %>%  
 summarise(mean=mean(Weight)) %>%  
 ungroup()  
  
glimpse(RATSL9S)

## Observations: 16  
## Variables: 3  
## $ Group <fct> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3  
## $ ID <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16  
## $ mean <dbl> 263.2, 238.9, 261.7, 267.2, 270.9, 276.2, 274.6, 267.5, ...



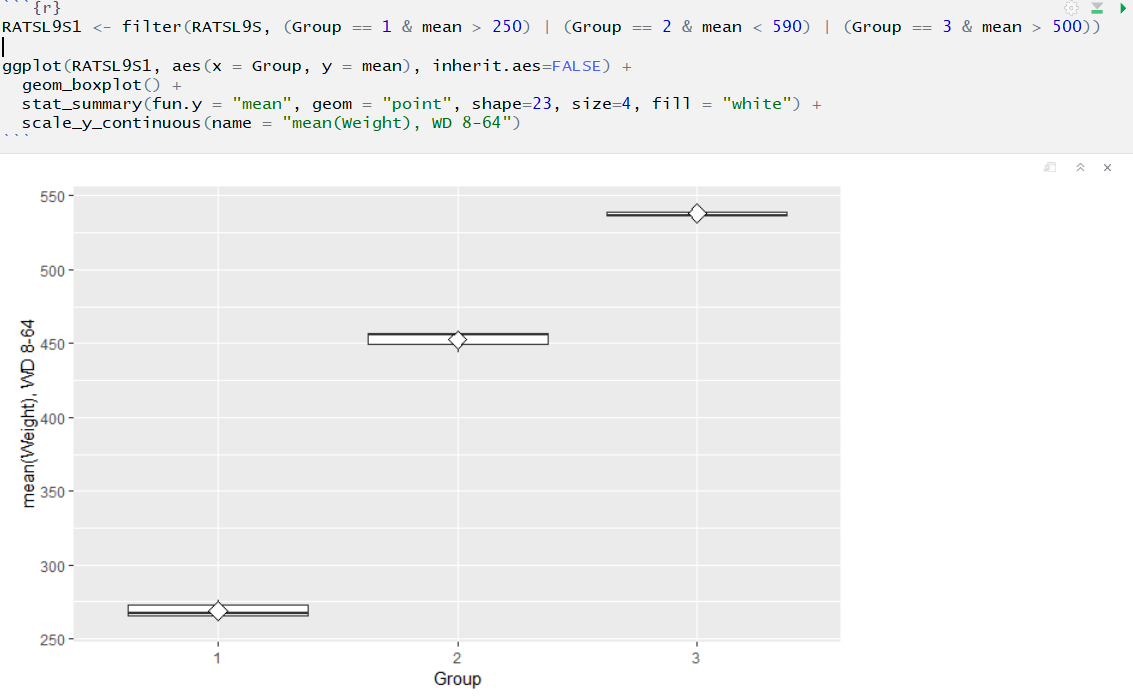
**ggplot code**



**graph**

The group that shows us the most interesting results here is Group 2. The mean is on the very edge of the box, and the distribution is very skewed. The second group also shows us that one outlier (outlaw rat!), near the 600 gram mark. Both Group 1 and Group 3 also show a single outlier. Let’s remove those pesky rodents…

**And again, an unfortunate screenprint coming your way…**



**code AND graph**

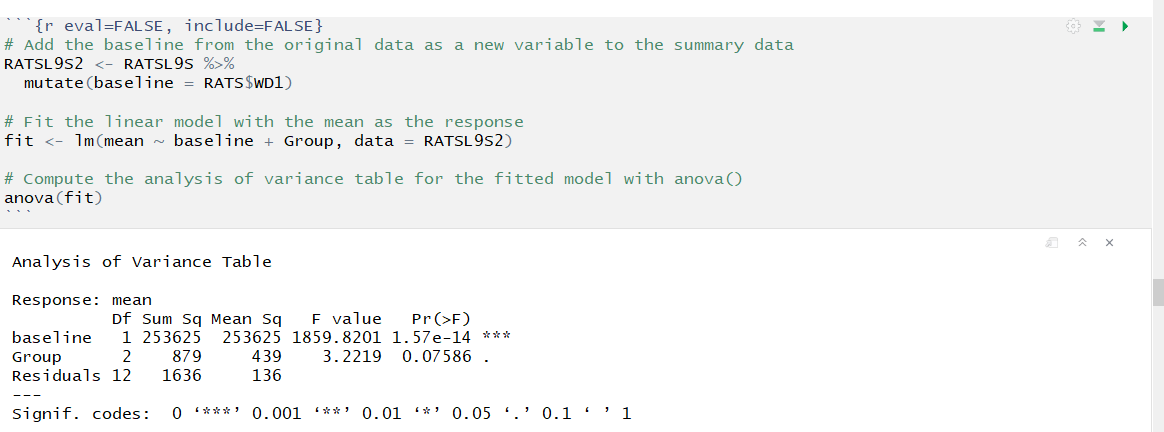
Excluding the outliers changes the box plots quite significantly, especially for Group 2. The boxes being so short means that the standardized weights of the rats in each group are very close to one another.

### 1E. T for test and A for Anova

The next step that we will carry out is a more formal test to assess the differences between the diet groups. This is where a t-test would come in, if we were working on the BPSR data. The t-test is appropriate when there are two group means. For a comparison of more than two group means (as in RATS) the one-way analysis of variance (ANOVA) is the appropriate method instead of the t-test.

The baseline measurements of the outcome are often correlated with the summary measure in longitudinal study. We will include the baseline in the analysis, that is the starting Weight of the rats on WD1.

**As I had to disable the code chunk above, the code that would normally appear below, cannot recognize RATSL9S, so I’ve had to disable the one below also… More screenprints!**



**code AND output**

When we look at the baseline, the significance is great.But here we can also see that the variable Group achieves the 95% confidence interval, meaning that the diet group that the rats belong in is significant.

Did I already say “Ah, rats!”? Well, we can say that again! The Index file knitting issue remains a mystery. Let’s hope we have more luck with the second part…

## Part 2. Implementing the Analyses of Chapter 9 of MABS Using the BPRS Data

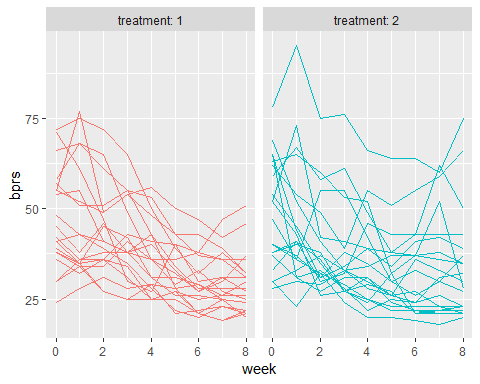
We will start with the code from meet\_and\_repeat.R again…

BPRS <- read.table("https://raw.githubusercontent.com/KimmoVehkalahti/MABS/master/Examples/data/BPRS.txt", sep =" ", header = T)  
BPRS$treatment <- factor(BPRS$treatment)  
BPRS$subject <- factor(BPRS$subject)  
BPRSL <- BPRS %>% gather(key = weeks, value = bprs, -treatment, -subject)  
BPRSL <- BPRSL %>% mutate(week = as.integer(substr(weeks,5,5)))

### 2A. Plot first, ask questions later

Let’s begin by plotting the data and identifying the observations in each treatment group.

# Plot the BPRSL data  
ggplot(BPRSL, aes(x = week, y = bprs, group = subject)) +  
 geom\_line(aes(col = treatment)) +  
 scale\_linetype\_manual(values = rep(1:10, times=4)) +  
 facet\_grid(. ~ treatment, labeller = label\_both) +  
 theme(legend.position = "none") +   
 scale\_y\_continuous(limits = c(min(BPRSL$bprs), max(BPRSL$bprs)))



Here we can see all the individuals in each treatment group, and how their bprs score behaves throughout the weeks. In the DataCamp exercise (for RATS data), the different groups were all presented in the same graph. However, that does not work with our BPRS data as there are so many subjects. Hence the two separate graphs for the two treatment groups. We can see that, for the most part, the bprs scores are declining throughout the weeks, indicating that the treatments are at least somewhat effective. Another observation of interest here is that especially within the treatment 1 group, the scores tend to converge towards the end of the treatment weeks. As is the case so often, though, there are exceptions to the rule: we can observe two subjects in each group whose bprs scores start increasing before the treatment weeks are over.

### 2B. Holding on to independence: The Linear model

To begin this analysis, we will ignore some home truths! We will pretend that all of the data observations are independent of one another, and ignore the fact that 18 bprs scores come from the same subject. We have a data set consisting of 360 observations that can easily be analyzed using linear regression.

In this section, we will fit a linear regression model. The response variable is bprs and the explanatory variables are week and treatment.

# create a regression model BPRS\_reg  
BPRS\_reg <- lm(bprs ~ week + treatment, data = BPRSL)  
  
# print out a summary of the model  
summary(BPRS\_reg)

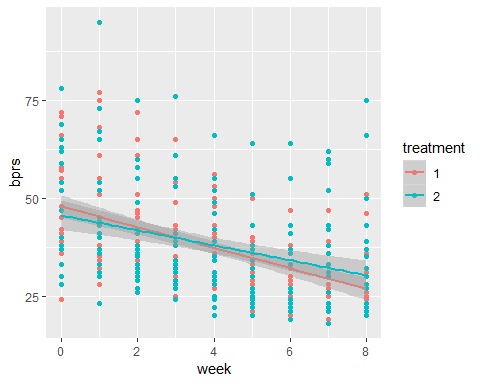
##   
## Call:  
## lm(formula = bprs ~ week + treatment, data = BPRSL)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.454 -8.965 -3.196 7.002 50.244   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 46.4539 1.3670 33.982 <2e-16 \*\*\*  
## week -2.2704 0.2524 -8.995 <2e-16 \*\*\*  
## treatment2 0.5722 1.3034 0.439 0.661   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.37 on 357 degrees of freedom  
## Multiple R-squared: 0.1851, Adjusted R-squared: 0.1806   
## F-statistic: 40.55 on 2 and 357 DF, p-value: < 2.2e-16

The summary table above shows the results from fitting a linear regression model with the bprs score as the response variable and treatment and week as the explanatory variables. This model ignores the repeated-measures structure of the data.

In this summary, a worthwhile observation is the significance of the regression week. Also, conditional on week, treatment group 2 differs from treatment group 1.

So, what does a graphical representation of the above data look like?

p1 <- ggplot(BPRS\_reg, aes(x = week, y = bprs, col = treatment))   
p2 <- p1 + geom\_point()   
p3 <- p2 + geom\_smooth(method = "lm")  
p3



According to this graph, the bprs scores of treatment group 1 start higher and decline faster throughout the weeks than the scores of the subjects in group 2. This indicates that the treatment that the subjects receive in group 1 is more effective.

### 2C. The Random Intercept Model

The model that we created in the previous step assumed that the weekly bprs scores are independent of one another. However, this is highly unlikely. Now we will move onto looking at some more appropriate models and graphical representations of our data.

We will fit a random intercept model of the bprs score data using week and treatment as the explanatory variables. This enables the linear regression fit for each subject to differ in *intercept* from the other subjects.

The package we will use is lme4. It offers some handy tools for linear and generalized linear mixed-effects models. We will be using the now-familiar ~ operator and also the vertical bar for distinguishing random-effects terms.

# access library lme4  
library(lme4)  
  
# Create a random intercept model  
BPRS\_ref <- lmer(bprs ~ week + treatment + (1 | subject), data = BPRSL, REML = FALSE)  
  
# Print the summary of the model  
summary(BPRS\_ref)

## Linear mixed model fit by maximum likelihood ['lmerMod']  
## Formula: bprs ~ week + treatment + (1 | subject)  
## Data: BPRSL  
##   
## AIC BIC logLik deviance df.resid   
## 2748.7 2768.1 -1369.4 2738.7 355   
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -3.0481 -0.6749 -0.1361 0.4813 3.4855   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## subject (Intercept) 47.41 6.885   
## Residual 104.21 10.208   
## Number of obs: 360, groups: subject, 20  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 46.4539 1.9090 24.334  
## week -2.2704 0.2084 -10.896  
## treatment2 0.5722 1.0761 0.532  
##   
## Correlation of Fixed Effects:  
## (Intr) week   
## week -0.437   
## treatment2 -0.282 0.000

The standard deviation of the subject is 6.885.

### 2D. Slippery slopes: Random Intercept and Random Slope Model

Now we will fit a random intercept and random slope model to the bprs scores data. This will allow the linear regression fits for each individual to differ in not only intercept but also in slope. This will make it possible to account for the differences in the subjects’ bprs scores and also the effects of time.

# create a random intercept and random slope model  
BPRS\_ref1 <- lmer(bprs ~ week + treatment + (week | subject), data = BPRSL, REML = FALSE)  
  
# print a summary of the model  
summary(BPRS\_ref1)

## Linear mixed model fit by maximum likelihood ['lmerMod']  
## Formula: bprs ~ week + treatment + (week | subject)  
## Data: BPRSL  
##   
## AIC BIC logLik deviance df.resid   
## 2745.4 2772.6 -1365.7 2731.4 353   
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -2.8919 -0.6194 -0.0691 0.5531 3.7976   
##   
## Random effects:  
## Groups Name Variance Std.Dev. Corr   
## subject (Intercept) 64.8202 8.0511   
## week 0.9608 0.9802 -0.51  
## Residual 97.4307 9.8707   
## Number of obs: 360, groups: subject, 20  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 46.4539 2.1052 22.066  
## week -2.2704 0.2977 -7.626  
## treatment2 0.5722 1.0405 0.550  
##   
## Correlation of Fixed Effects:  
## (Intr) week   
## week -0.582   
## treatment2 -0.247 0.000

# perform an ANOVA test on the two models  
anova(BPRS\_ref1, BPRS\_ref)

## Data: BPRSL  
## Models:  
## BPRS\_ref: bprs ~ week + treatment + (1 | subject)  
## BPRS\_ref1: bprs ~ week + treatment + (week | subject)  
## Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)   
## BPRS\_ref 5 2748.7 2768.1 -1369.4 2738.7   
## BPRS\_ref1 7 2745.4 2772.6 -1365.7 2731.4 7.2721 2 0.02636 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The chi-squared statistic of the likelihood ratio test between BPRS\_ref and BPRS\_ref is within the 95% level of significance, indicating a good fit for the model.

### 2E. Time to interact: Random Intercept and Random Slope Model with interaction

Now for the final part: fitting a random intercept and slope model that allows for a treatment group x time interaction.

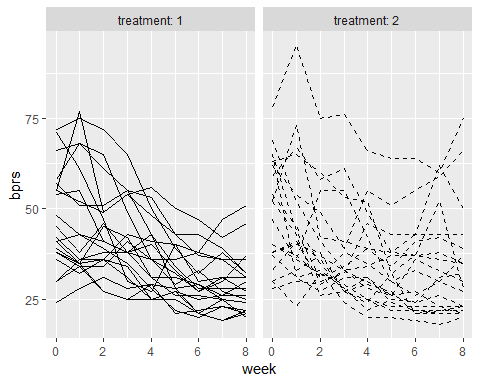
# create a random intercept and random slope model  
BPRS\_ref2 <- lmer(bprs ~ week \* treatment + (week | subject), data = BPRSL, REML = FALSE)  
  
# print a summary of the model  
summary(BPRS\_ref2)

## Linear mixed model fit by maximum likelihood ['lmerMod']  
## Formula: bprs ~ week \* treatment + (week | subject)  
## Data: BPRSL  
##   
## AIC BIC logLik deviance df.resid   
## 2744.3 2775.4 -1364.1 2728.3 352   
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -3.0512 -0.6271 -0.0767 0.5288 3.9260   
##   
## Random effects:  
## Groups Name Variance Std.Dev. Corr   
## subject (Intercept) 65.0016 8.0624   
## week 0.9688 0.9843 -0.51  
## Residual 96.4699 9.8219   
## Number of obs: 360, groups: subject, 20  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 47.8856 2.2522 21.262  
## week -2.6283 0.3589 -7.323  
## treatment2 -2.2911 1.9090 -1.200  
## week:treatment2 0.7158 0.4010 1.785  
##   
## Correlation of Fixed Effects:  
## (Intr) week trtmn2  
## week -0.650   
## treatment2 -0.424 0.469   
## wek:trtmnt2 0.356 -0.559 -0.840

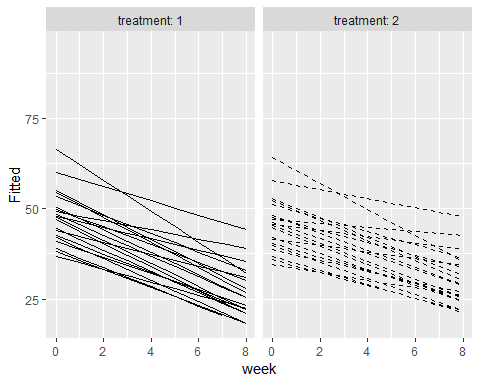
# perform an ANOVA test on the two models  
anova(BPRS\_ref2, BPRS\_ref1)

## Data: BPRSL  
## Models:  
## BPRS\_ref1: bprs ~ week + treatment + (week | subject)  
## BPRS\_ref2: bprs ~ week \* treatment + (week | subject)  
## Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)   
## BPRS\_ref1 7 2745.4 2772.6 -1365.7 2731.4   
## BPRS\_ref2 8 2744.3 2775.4 -1364.1 2728.3 3.1712 1 0.07495 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# draw the plot of BPRSL  
ggplot(BPRSL, aes(x = week, y = bprs, group = subject)) +  
 geom\_line(aes(linetype = treatment)) +  
 scale\_linetype\_manual(values = rep(1:10, times=4)) +  
 facet\_grid(. ~ treatment, labeller = label\_both) +  
 theme(legend.position = "none") +   
 scale\_y\_continuous(limits = c(min(BPRSL$bprs), max(BPRSL$bprs)))



# Create a vector of the fitted values  
Fitted <- fitted(BPRS\_ref2)  
  
# Create a new column fitted to BPRSL  
BPRSL <- BPRSL %>%  
 mutate(Fitted)  
  
# draw the plot of BPRSL  
ggplot(BPRSL, aes(x = week, y = Fitted, group = subject)) +  
 geom\_line(aes(linetype = treatment)) +  
 scale\_linetype\_manual(values = rep(1:10, times=4)) +  
 facet\_grid(. ~ treatment, labeller = label\_both) +  
 theme(legend.position = "none") +   
 scale\_y\_continuous(limits = c(min(BPRSL$bprs), max(BPRSL$bprs)))



Here it is worth noting that the first two graphs are the same as before, and it shows the subjects whose scores started climbing back up before the treatment weeks were over.

The ANOVA test on the two models BPRS\_ref1 and BPRS\_ref tells us that the chi-squared value is 3.1712, and the p-value tells us that the fit against the comparison model is within the 95% significance level. This indicates a good fit.

Next we created a vector of the fitted values of the model with the function fitted() and added this new vector as a column in BPRSL. Then we used these fitted values to draw a new plot of BPRSL.

The second set of graphs are interesting. They show us the *fitted* values of the bprs scores. These graphics underline the fact that the interaction model does not fit the observed data *quite* as well as it did with the RATS data.

Looking at the graphs of the fitted values, it appears that the slope of declining bprs scores is more acute in the first treatment group, indicating that the first form of treatment was more effective. Out of curiosity, I tried to find out what the treatments were that the subjects received in the two groups, but unfortunately this information was not readily available anywhere. While looking, I did manage to find a pdf of the Davis 2002 book, “Statistical Methods for the Analysis of Repeated Measurements”. If you’re interested in learning more about methods for repeated measurements (i.e. longitudinal data), I recommend taking a look at this volume.

Again, thanks for reading to the end. Sadly, this was the last of the weekly exercises. I hope you’ve enjoyed yourself as much as I have!



**Goodbye**