Chapter6

Pekka Tolli

7 December 2018

Week 6 - Longitudinal analysis§

This is the last IODS exercise for now. Time for longitudinal analysis.

Part 1 RATS:

On this first part I'll analyse the longitudinal rats data set by Kimmo Vehkalahti from a nutrition study on three groups of rats. The data each rat's weight being recorded once a week for 9 weeks. Thus the data is longitudinal, i.e. it tracks each rat over a time period. I will test if the nutrition changes have impact on the weight change of a rat.

```
RATSL <- read.csv("C:/Users/pekka/Documents/GitHub/IODS-project/Data/rats.txt", sep  ="\t", h
eader = T)

RATSL$Group <- factor(RATSL$Group)
RATSL$ID <- factor(RATSL$ID)

str(RATSL)
```

```
## 'data.frame':    176 obs. of  5 variables:
##  $ ID    : Factor w/ 16 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Group : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 2 2 ...
##  $ WD    : Factor w/ 11 levels "WD1","WD15","WD22",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Weight: int  240 225 245 260 255 260 275 245 410 405 ...
##  $ Time  : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
dim(RATSL)
```

```
## [1] 176   5
```

```
summary(RATSL)
```

```
##        ID      Group        WD         Weight           Time
##  1      : 11   1:88   WD1    :16   Min.   :225.0   Min.   : 1.00
##  2      : 11   2:44   WD15   :16   1st Qu.:267.0   1st Qu.:15.00
##  3      : 11   3:44   WD22   :16   Median :344.5   Median :36.00
##  4      : 11          WD29   :16   Mean   :384.5   Mean   :33.55
##  5      : 11          WD36   :16   3rd Qu.:511.2   3rd Qu.:50.00
##  6      : 11          WD43   :16   Max.   :628.0   Max.   :64.00
##  (Other):110          (Other):80
```

```
head(RATSL)
```

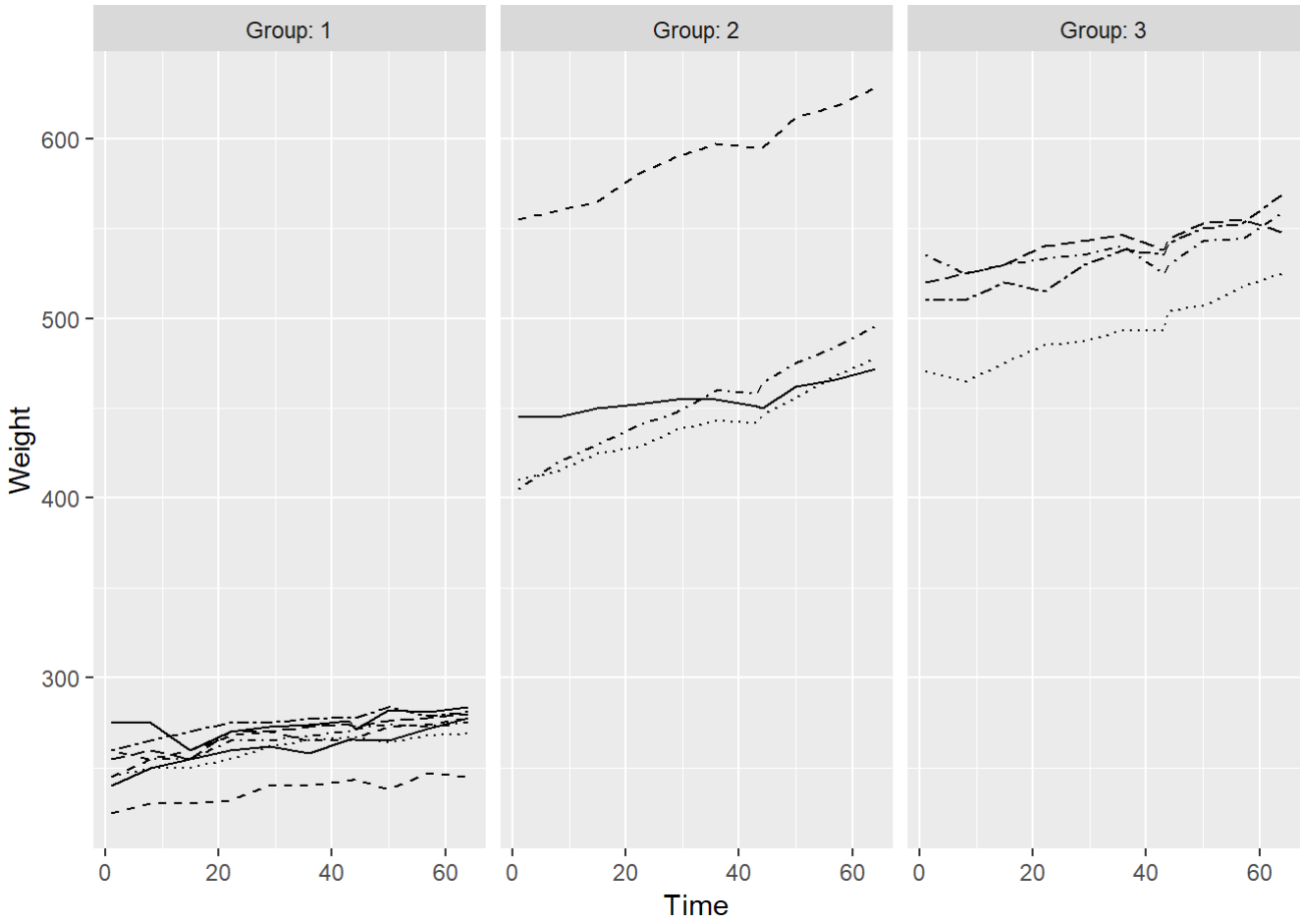
```
##   ID Group  WD Weight Time
## 1  1     1 WD1    240    1
## 2  2     1 WD1    225    1
## 3  3     1 WD1    245    1
## 4  4     1 WD1    260    1
## 5  5     1 WD1    255    1
## 6  6     1 WD1    260    1
```

We have data set with 5 variables and 176 rows containing observations of 16 rats across 9 weeks of time.

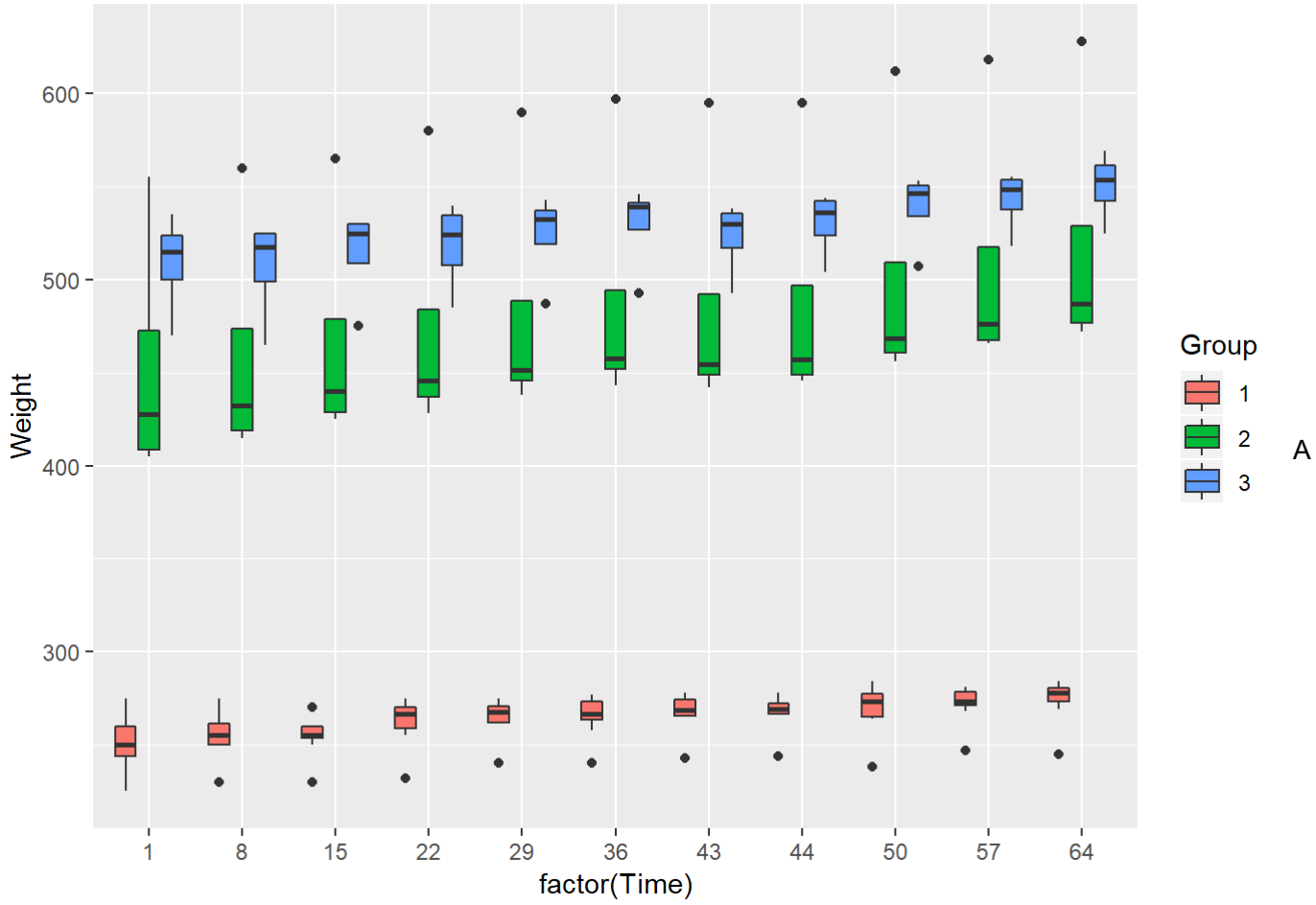
Visualizing RATS data

Let's draw basic line charts of the unstandardized RATS data classified by nutrition group:

```
ggplot(RATSL, aes(x = Time, y = Weight, linetype = ID)) +
  geom_line() +
  scale_linetype_manual(values = rep(1:10, times=4)) +
  facet_grid(. ~ Group, labeller = label_both) +
  theme(legend.position = "none") +
  scale_y_continuous(limits = c(min(RATSL$Weight), max(RATSL$Weight)))
```




```
ggplot(RATSL, aes(y=Weight, x=factor(Time), fill=Group)) +
  geom_boxplot()
```



couple of observations emerge: First, practically all rats have increased their weight during the study. Second, group 1 rats were lighter at the beginning and the end of the study. Third, group 2 has a potential outlier (a big rat).

Standardizing the RATS data

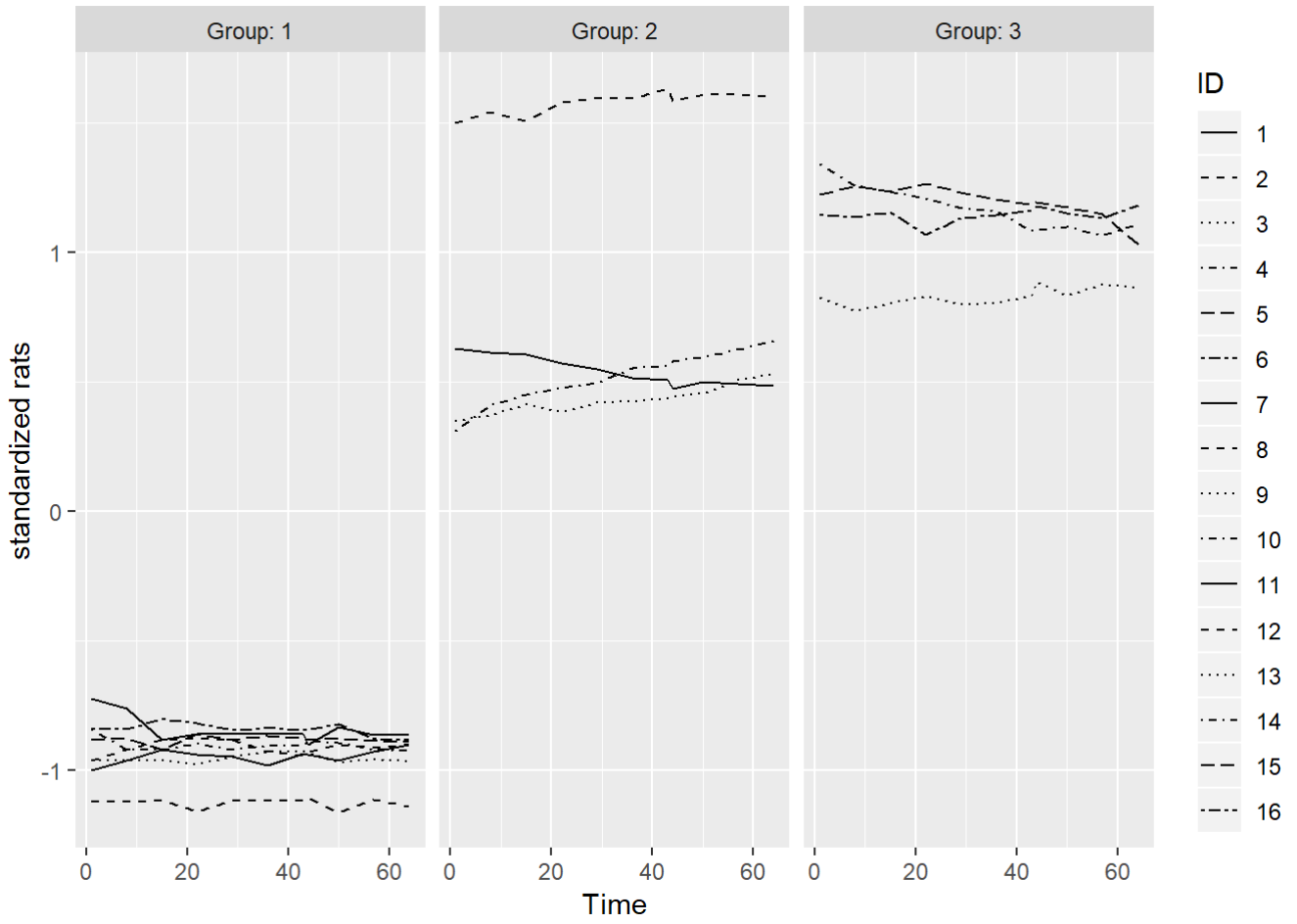
To do proper analysis, let's standardize the data:

```
RATSL <- RATSL %>%
  group_by(Time) %>%
  mutate(standard_RATSL = (Weight - mean(Weight))/sd(Weight) ) %>%
  ungroup()
glimpse(RATSL)
```

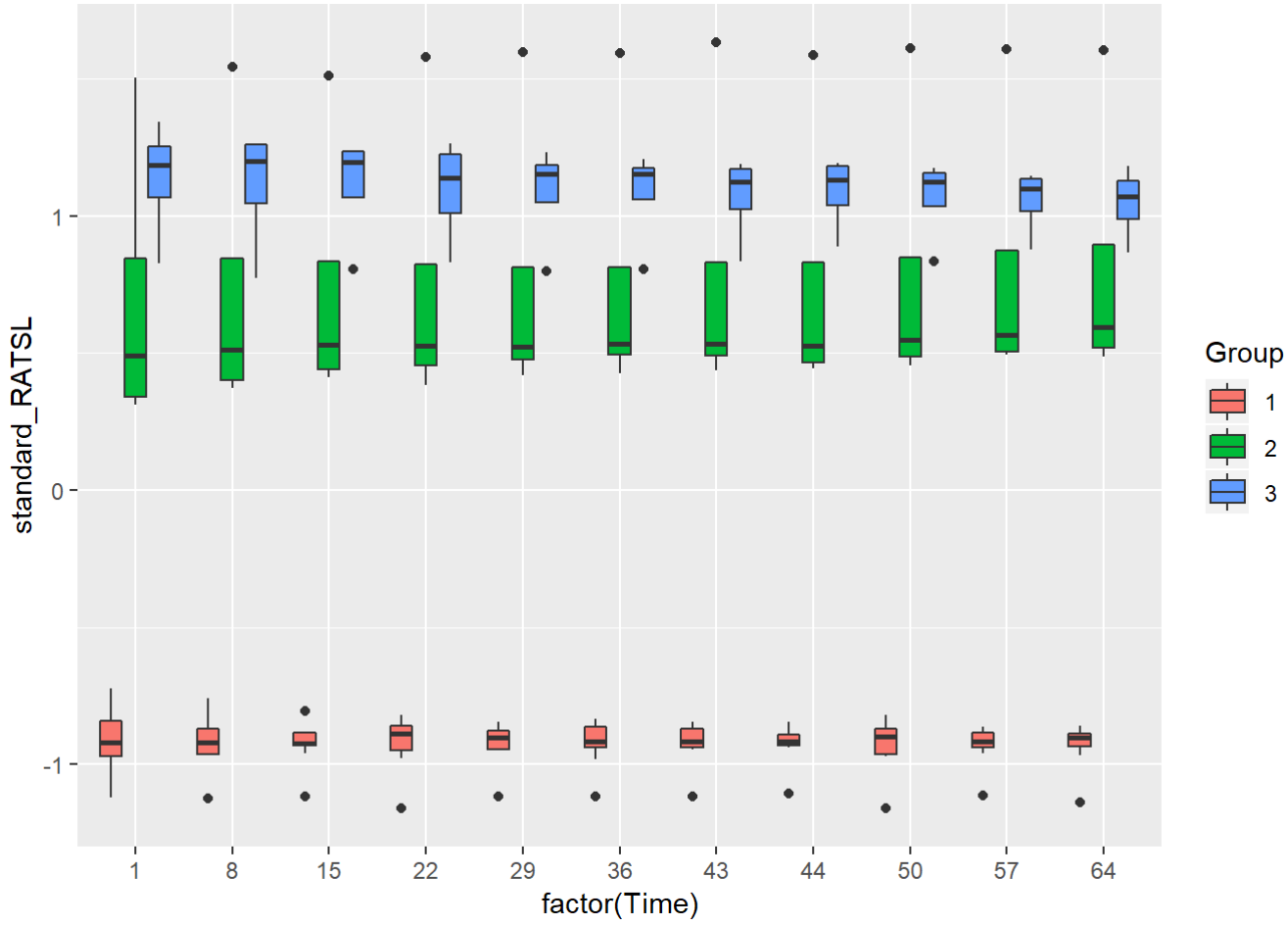
```
## Observations: 176
## Variables: 6
## $ ID             <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, ...
## $ Group          <fct> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3,...
## $ WD             <fct> WD1, WD1, WD1, WD1, WD1, WD1, WD1, WD1, WD1, WD...
## $ Weight         <int> 240, 225, 245, 260, 255, 260, 275, 245, 410, 40...
## $ Time           <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ standard_RATSL <dbl> -1.0011429, -1.1203857, -0.9613953, -0.8421525,...
```

And let's see again how the standardized data looks like:

```
ggplot(RATSL, aes(x = Time, y = standard_RATSL, linetype = ID)) +
  geom_line() +
  scale_linetype_manual(values = rep(1:10, times=4)) +
  facet_grid(. ~ Group, labeller = label_both) +
  scale_y_continuous(name = "standardized rats")
```




```
ggplot(RATSL, aes(y=standard_RATSL, x=factor(Time), fill=Group)) +
  geom_boxplot()
```

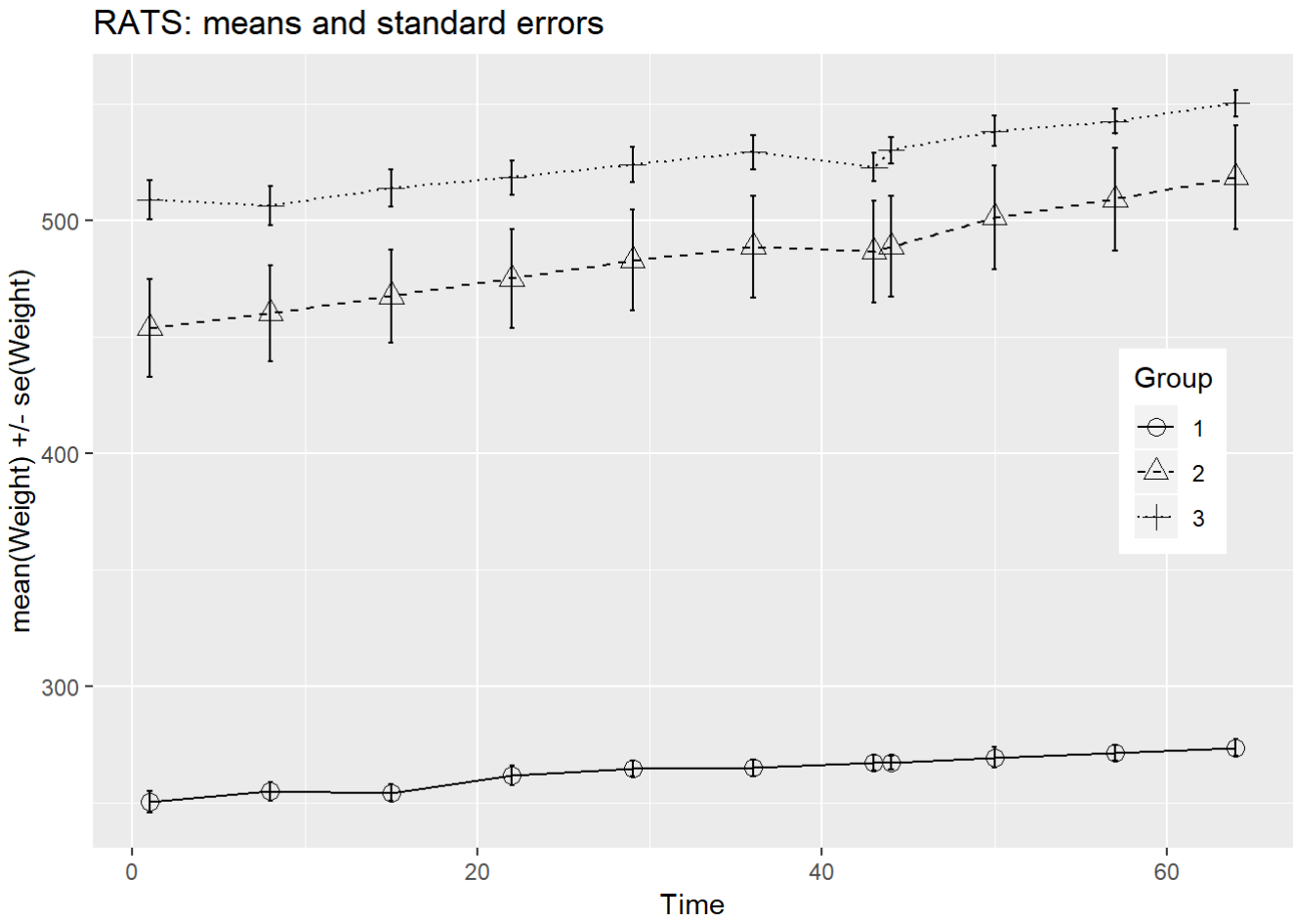


We now can see that Y-axis with weights has now standardized scale. Variability inside the groups was not removed but we made data more suitable for analysis

```
n <- RATSL$Time %>% unique() %>% length()
RATSS <- RATSL %>%
  group_by(Group, Time) %>%
  summarise( mean = mean(Weight), se = sd(Weight)/sqrt(n) ) %>%
  ungroup()
glimpse(RATSS)
```

```
## Observations: 33
## Variables: 4
## $ Group <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2,...
## $ Time  <int> 1, 8, 15, 22, 29, 36, 43, 44, 50, 57, 64, 1, 8, 15, 22, ...
## $ mean  <dbl> 250.625, 255.000, 254.375, 261.875, 264.625, 265.000, 26...
## $ se    <dbl> 4.589478, 3.947710, 3.460116, 4.100800, 3.333956, 3.5529...
```

```
ggplot(RATSS, aes(x = Time, y = mean, linetype = Group, shape = Group)) +
  geom_line() +
  scale_linetype_manual(values = c(1,2,3)) +
  geom_point(size=3) +
  scale_shape_manual(values = c(1,2,3)) +
  geom_errorbar(aes(ymin = mean - se, ymax = mean + se, linetype="1"), width=0.3) +
  theme(legend.position = c(0.9,0.5)) +
  scale_y_continuous(name = "mean(Weight) +/- se(Weight)") +
  ggtitle("RATS: means and standard errors")
```



The plot above represents the mean and standard deviations of weight of each three groups over time. As discussed, group 1 has the lowest mean of weight, while Groups 2 and 3 are closer to each other.

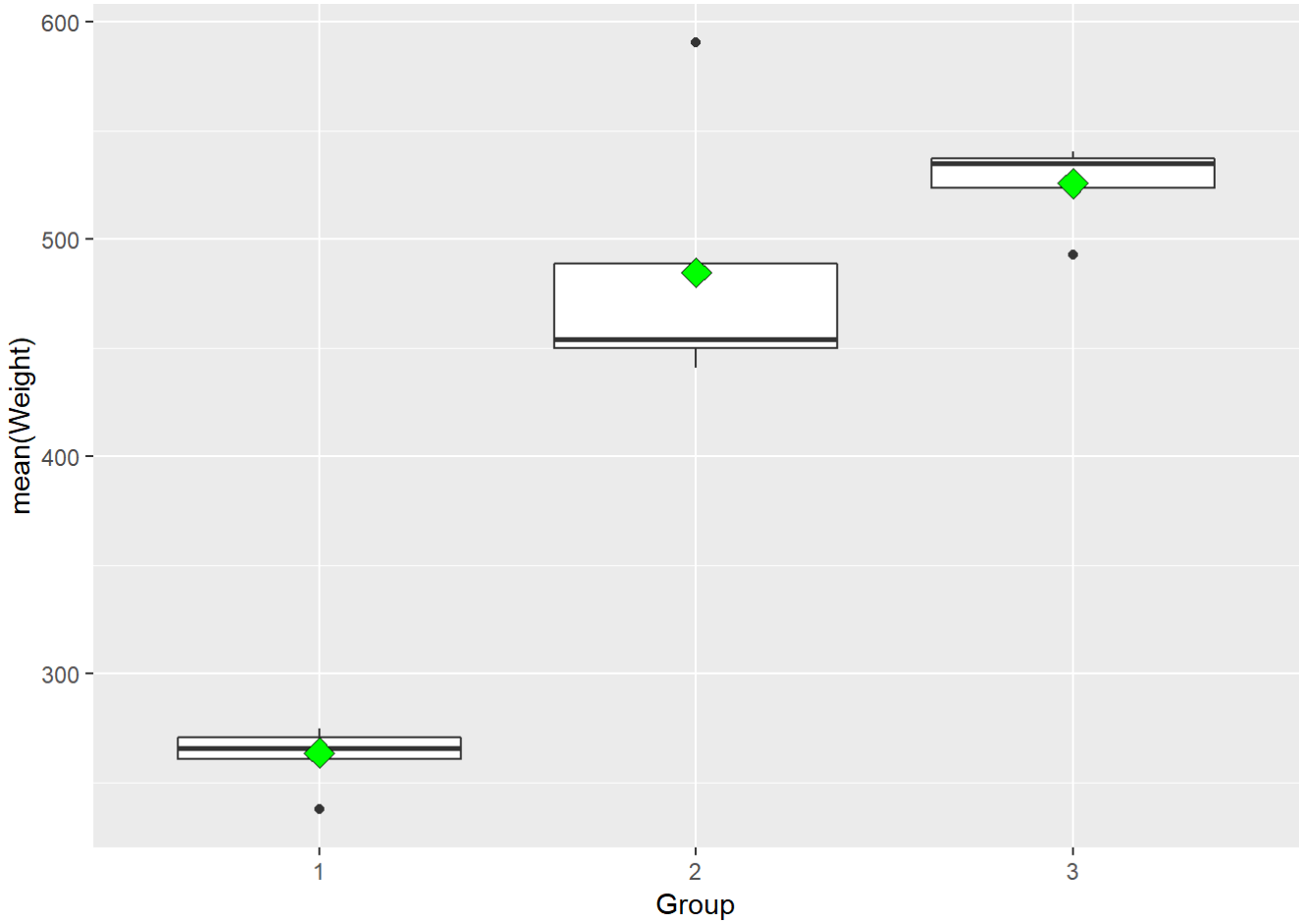
Removing outliers

Lets see if the data has any outliers. If yes, remove them.

```
RATSL8S <- RATSL %>%
  filter(Time > 0) %>%
  group_by(Group, ID) %>%
  summarise( mean=mean(Weight) ) %>%
  ungroup()
glimpse(RATSL8S)
```

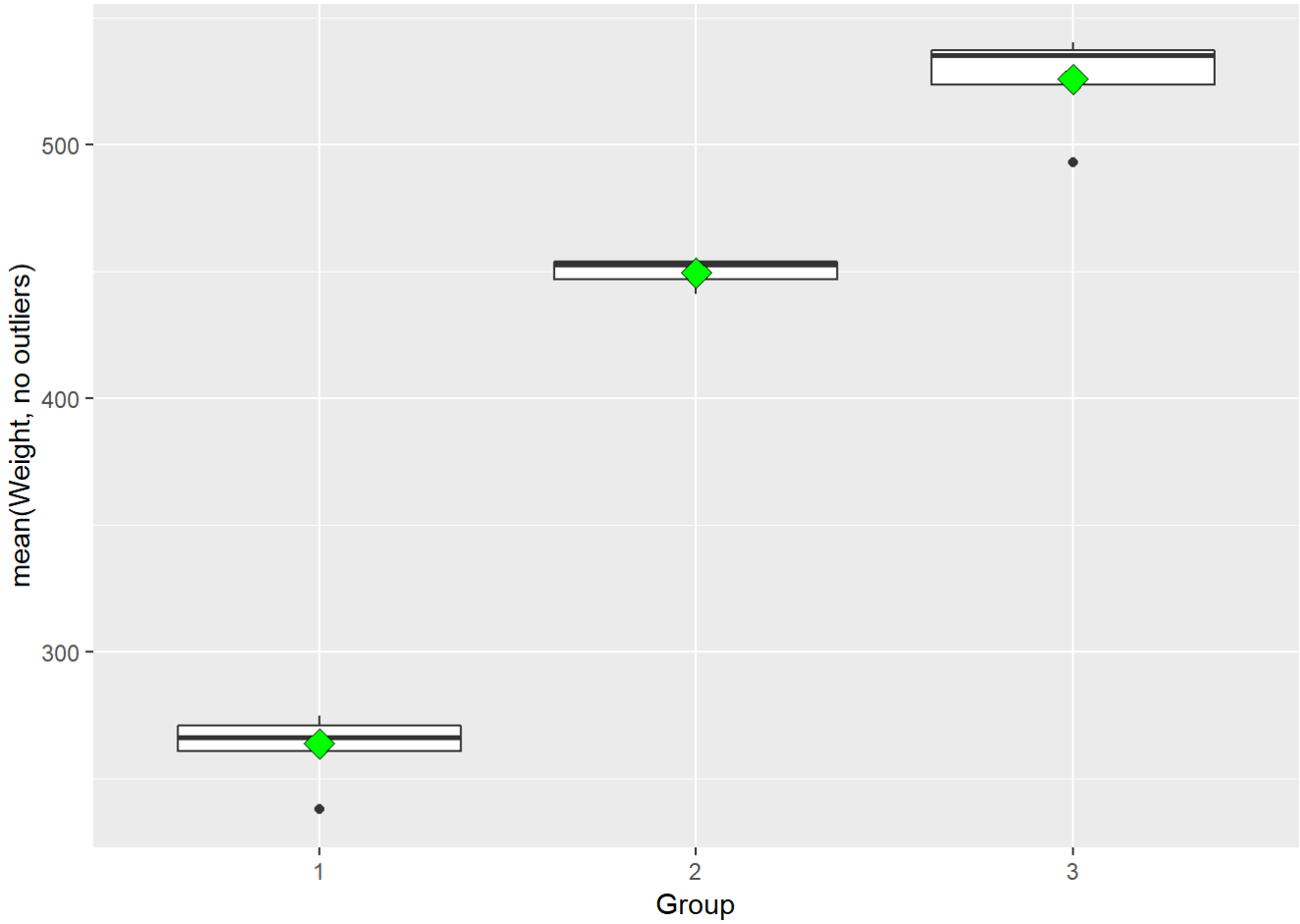
```
## Observations: 16
## Variables: 3
## $ Group <fct> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3
## $ ID    <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
## $ mean  <dbl> 261.0909, 237.6364, 260.1818, 266.5455, 269.4545, 274.72...
```

```
ggplot(RATSL8S, aes(x = Group, y = mean)) +
  geom_boxplot() +
  stat_summary(fun.y = "mean", geom = "point", shape=23, size=4, fill = "green") +
  scale_y_continuous(name = "mean(Weight)")
```



In group 2 there is one observation quite far off from the others as pointed out in the initial observation. Let's remove it

```
RATSL8S1 <- RATSL8S %>%
  filter(mean < 550)
ggplot(RATSL8S1, aes(x = Group, y = mean)) +
  geom_boxplot() +
  stat_summary(fun.y = "mean", geom = "point", shape=23, size=4, fill = "green") +
  scale_y_continuous(name = "mean(Weight, no outliers)")
```



Now the data looks better, no clear outliers anymore.

Time for checking the real differences between the groups. It looks from the plots that groups behaved bit differently but let's see it with ANOVA. First, add the baseline from the original data as a new variable to the summary data. Let's then fit a linear model and run anova on it.

```
RATSL8S1 <- RATSL8S %>%
  mutate(baseline = filter(RATSL, Time==1)$Weight)
RATSL8S1$mean <- as.numeric(RATSL8S1$mean)

fit <- lm(mean ~ baseline + Group, data = RATSL8S1)

anova(fit)
```

```
## Analysis of Variance Table
##
## Response: mean
##           Df Sum Sq Mean Sq   F value    Pr(>F)
## baseline   1 252125  252125 2237.0655 5.217e-15 ***
## Group      2    726     363    3.2219   0.07586 .
## Residuals 12   1352     113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Unsurprisingly the anova clarifies the groups differ by their weight gains.

Part 2: BPRS

Let's then analyse the BPRS data, which is about the psychological treatment study. The long format of BPRS (BPRSL) has 360 rows and 5 columns i.e. variables. The data contains results of 40 male patients who were randomly assigned to two treatment groups and each subject was rated on BPRS (brief psychiatric rating scale) before when the treatment began (week 0) and once a week for eight weeks.

```
BPRSL <- read.csv("C:/Users/pekka/Documents/GitHub/IODS-project/Data/bprs.txt", sep  ="\t", h
eader = T)

BPRSL$treatment <- factor(BPRSL$treatment)
BPRSL$subject <- factor(BPRSL$subject)

str(BPRSL)
```

```
## 'data.frame':    360 obs. of  5 variables:
##  $ treatment: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ subject  : Factor w/ 20 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ weeks    : Factor w/ 9 levels "week0","week1",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ bprs     : int  42 58 54 55 72 48 71 30 41 57 ...
##  $ week     : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
dim(BPRSL)
```

```
## [1] 360   5
```

```
summary(BPRSL)
```

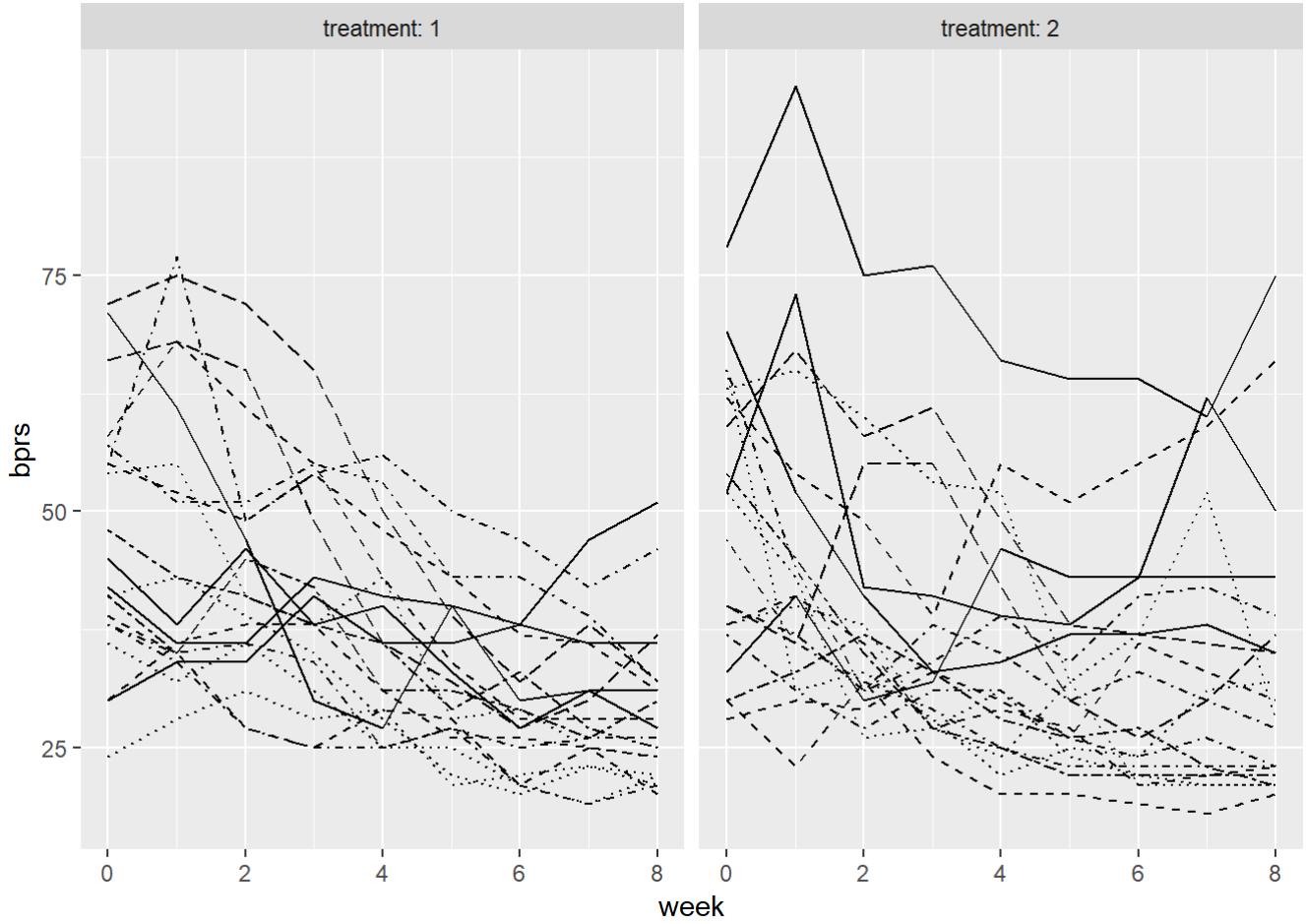
```
##  treatment    subject        weeks          bprs            week
##  1:180     1      : 18   week0  : 40   Min.   :18.00   Min.   :0
##  2:180     2      : 18   week1  : 40   1st Qu.:27.00   1st Qu.:2
##            3      : 18   week2  : 40   Median :35.00   Median :4
##            4      : 18   week3  : 40   Mean   :37.66   Mean   :4
##            5      : 18   week4  : 40   3rd Qu.:43.00   3rd Qu.:6
##            6      : 18   week5  : 40   Max.   :95.00   Max.   :8
##            (Other):252   (Other):120
```

```
head(BPRSL)
```

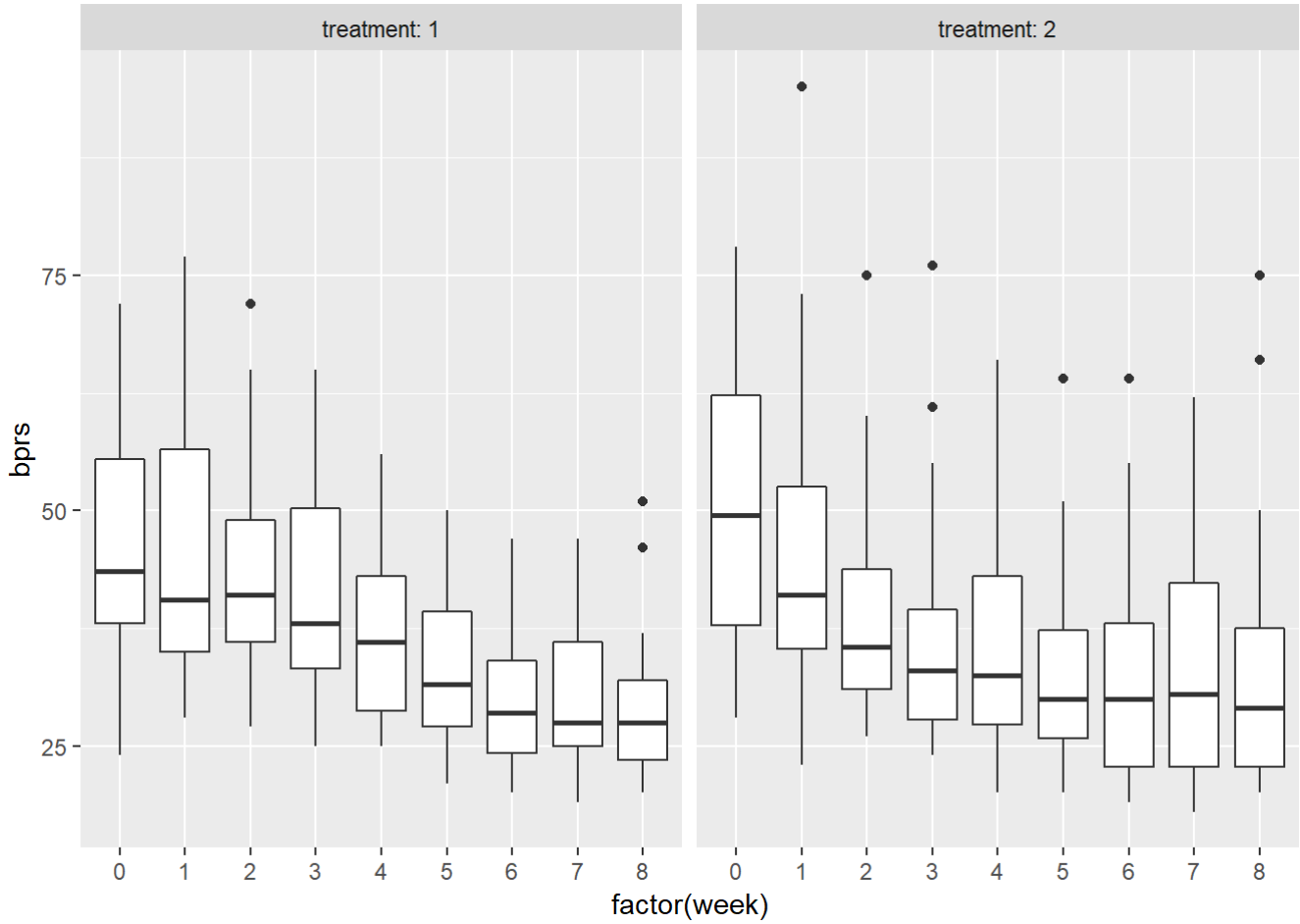
```
##   treatment subject weeks bprs week
## 1         1       1 week0   42    0
## 2         1       2 week0   58    0
## 3         1       3 week0   54    0
## 4         1       4 week0   55    0
## 5         1       5 week0   72    0
## 6         1       6 week0   48    0
```

Let's visualize the data:

```
ggplot(BPRSL, aes(x = week, y = bprs, linetype = subject)) +
  geom_line() +
  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)))
```




```
ggplot(BPRSL, aes(y=bprs, x=factor(week))) +
  geom_boxplot() +
  facet_grid(. ~ treatment, labeller = label_both)
```



Two treament groups appear to be quite close to each other. But it seems there is link between the stating and ending score of a patient.

Linear regression model

Let's then fit linear regression model with treatment and week as explanatory variables and bprs as dependent variable.

```
BPRS_fit <- lm(bprs ~ week + treatment, data = BPRSL)
summary(BPRS_fit)
```

```
##
## 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
```

Hmmm. regular linear regression model is not super good fit for this data as expected. Results imply that the treatment group doesn't have statistically significant impact. The time has which is obvious as the treatment is expected to decrease bprs score. For longitudinal data, a linear model is not great because it does not take into account that there are the same individuals apperaing in the data over time, i.e. it treats observations over time as independent of each other.

Therefore, let's fit a random intercepts model, which is more suitable for the longitudinal data.

Creating random intercept model, LMER

```
library(lme4)
```

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

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
BPRS_ref <- lmer(bprs ~ week + treatment + (1 | subject), data = BPRSL, REML = FALSE)

anova(BPRS_ref)
```

```
## Analysis of Variance Table
##           Df  Sum Sq Mean Sq  F value
## week       1 12371.5 12371.5 118.7136
## treatment  1    29.5    29.5   0.2828
```

```
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 random intercept model allows the linear regression fit for each patient to differ in intercept from patients.

Let’s fit also a random slope model for the data, which assumes heterogeneity in slopes.

```
BPRS_ref1 <- lmer(bprs ~ week + treatment + (week | subject), data = BPRSL, REML = FALSE)
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
```

Then, run ANOVA test on the random intercept and the random slope model.

```
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 anova results imply a reasonably good fit of the models with p-value of 0.026. That means the models are statistically significantly different from each other.

Let’s then check for the week*treatment interaction.

```
BPRS_ref2 <- lmer(bprs ~ week * treatment + (week | subject), data = BPRSL, REML = FALSE)
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
```

And do one more ANOVA to compare this with the above 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
```

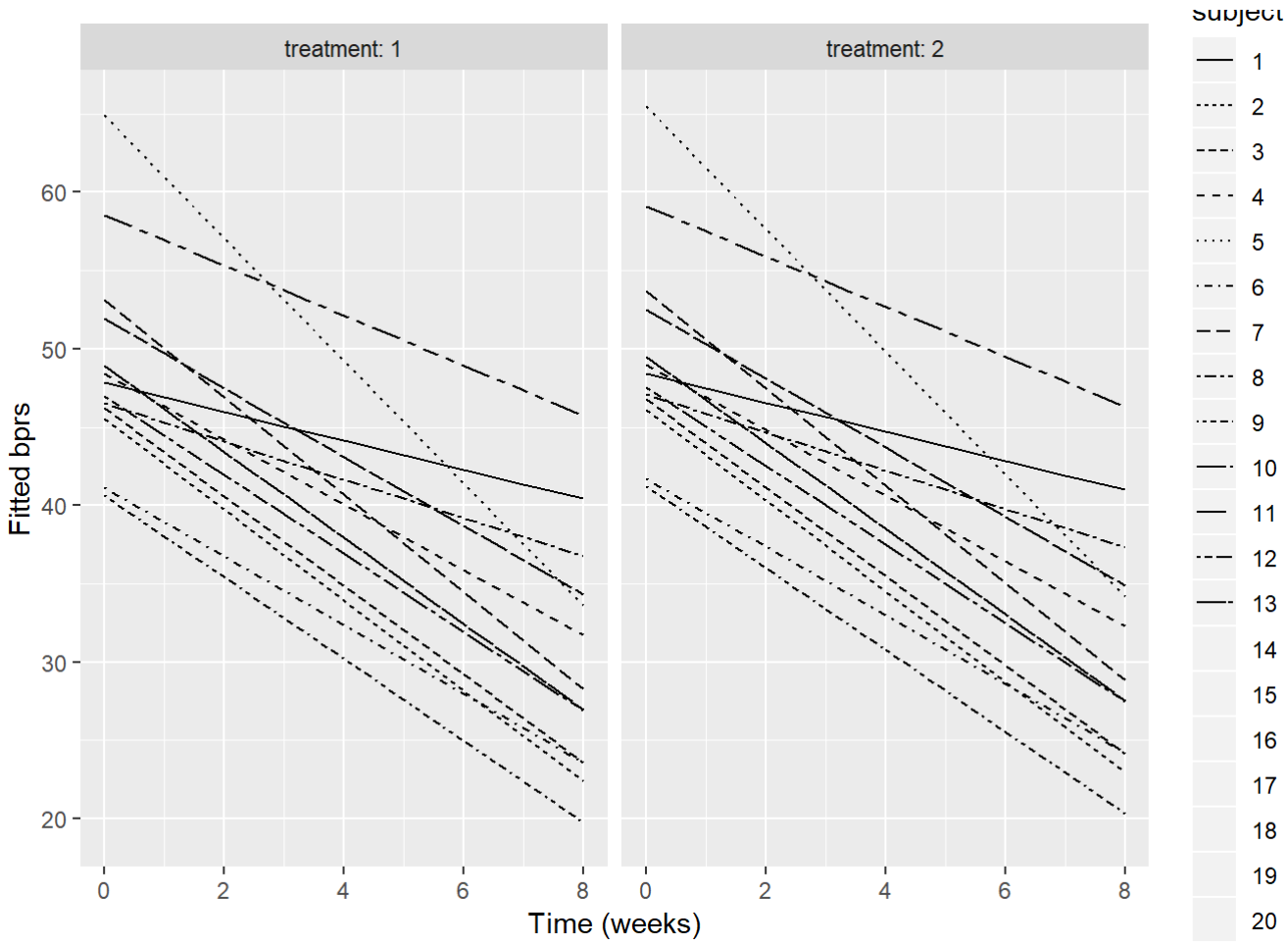
Now the anova results show that p-value is 0.076, i.e. greater than 0.05. Interaction does not seem to improve the previous model tested.

Anyway, let’s see how the fitted model fits with the data:

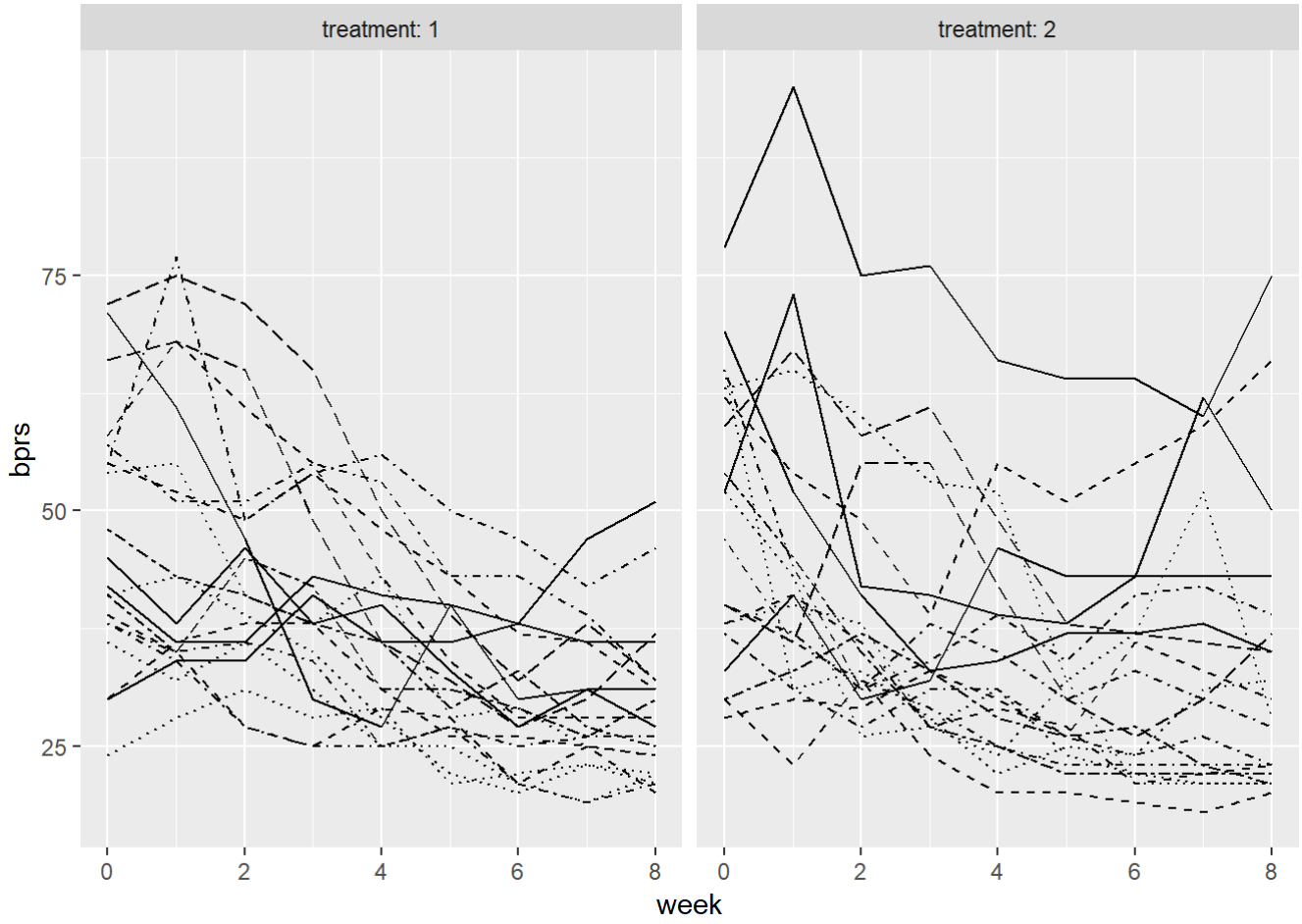
```
Fitted <- fitted(BPRS_ref1)

BPRSL <- BPRSL %>%
  mutate(Fitted)

ggplot(BPRSL, aes(x = week, y = Fitted, group = subject)) +
  geom_line(aes(linetype = subject)) +
   facet_grid(. ~ treatment, labeller = label_both) +
  scale_x_continuous(name = "Time (weeks)", breaks = seq(0, 8, 2)) +
  scale_y_continuous(name = "Fitted bprs") +
  theme(legend.position = "right")
```




```
ggplot(BPRSL, aes(x = week, y = bprs, linetype = subject)) +
  geom_line() +
  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)))
```



Hmmm, not bad. Looks like the model is not that far off from the actual data.