STOR 590 HW9 Solution

Taebin Kim

4/15/2020

Page 252 Exercise 3

Part (a)

After preprocessing the dataset, we plot the data using two panels, one for each sex, showing how weight increases with age.

```
library(faraway)
data(nepali)
nepali <- na.omit(subset(nepali, select = -ht))
nepali$sex <- ifelse(nepali$sex == 1, 'M', 'F')
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
ggplot(nepali, aes(age, wt, color = sex)) + geom_point() + facet_wrap(. ~sex)
```

We can observe that weight increases with for both sexes. Also note that the variance of weight increases with age for Female.

Part (b)

We fit a fixed effects model with weight as the response and age, sex, mother's age, literacy and other deaths in the family as predictors.

```
lm1 <- aov(wt ~ age + sex + mage + lit + died, nepali)
summary(lm1)
```

```
##              Df Sum Sq Mean Sq  F value   Pr(>F)    
## age           1   5840    5840 3109.901  < 2e-16 ***
## sex           1     23      23   12.421 0.000447 ***
## mage          1     53      53   28.076 1.48e-07 ***
## lit           1     30      30   15.768 7.75e-05 ***
## died          1      2       2    1.166 0.280435    
## Residuals   871   1636       2                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result shows that `age`, `sex`, `mage`, and `lit` are significant.

Part (c)

We fit a mixed effects model with weight as the response. We include an interaction between age and sex and main effects in the other two predictors, and we use a random intercept term for the child. Since we concluded that `age`, `sex`, `mage`, and `lit` are significant, we only consider these four predictors.

```
library(lme4)
```

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

```
lmer1_1 <- lmer(wt ~ age*sex + lit + mage + (1|id), nepali)
sumary(lmer1_1)
```

```
## Fixed Effects:
##             coef.est coef.se
## (Intercept) 4.38     0.49   
## age         0.13     0.00   
## sexM        0.37     0.26   
## lit         0.76     0.47   
## mage        0.05     0.02   
## age:sexM    0.00     0.00   
## 
## Random Effects:
##  Groups   Name        Std.Dev.
##  id       (Intercept) 1.33    
##  Residual             0.44    
## ---
## number of obs: 877, groups: id, 197
## AIC = 1805.5, DIC = 1730.4
## deviance = 1760.0
```

```r
lmer1_2 <- lmer(wt ~ age*sex*lit + mage + (1|id), nepali)
sumary(lmer1_2)
```

```
## Fixed Effects:
##               coef.est coef.se
## (Intercept)    4.43     0.49
## age            0.13     0.00
## sexM           0.35     0.27
## lit           -0.06     0.81
## mage           0.05     0.02
## age:sexM       0.00     0.00
## age:lit        0.03     0.02
## sexM:lit       0.15     1.57
## age:sexM:lit -0.03     0.04
##
## Random Effects:
##  Groups   Name        Std.Dev.
##  id       (Intercept) 1.32
##  Residual             0.44
## ---
## number of obs: 877, groups: id, 197
## AIC = 1816.2, DIC = 1715.5
## deviance = 1754.9
```

```r
lmer1_3 <- lmer(wt ~ age*sex*mage + lit + (1|id), nepali)
sumary(lmer1_3)
```

```
## Fixed Effects:
##                coef.est coef.se
## (Intercept)     4.08     0.88
## age             0.15     0.02
## sexM            0.64     1.23
## mage            0.06     0.03
## lit             0.75     0.47
## age:sexM       -0.02     0.02
## age:mage        0.00     0.00
## sexM:mage      -0.01     0.04
## age:sexM:mage  0.00     0.00
##
## Random Effects:
##  Groups   Name        Std.Dev.
##  id       (Intercept) 1.34
##  Residual             0.44
## ---
## number of obs: 877, groups: id, 197
## AIC = 1841.7, DIC = 1697.7
## deviance = 1758.7
```

```r
lmer1_4 <- lmer(wt ~ age*sex*lit*mage + (1|id), nepali)
sumary(lmer1_4)
```

```
## Fixed Effects:
```

```
##                   coef.est coef.se
## (Intercept)          4.09     0.88
## age                  0.15     0.02
## sexM                 0.84     1.25
## lit                  7.82     6.14
## mage                 0.06     0.03
## age:sexM            -0.02     0.02
## age:lit             -0.23     0.12
## sexM:lit          -542.79   269.61
## age:mage             0.00     0.00
## sexM:mage           -0.02     0.04
## lit:mage            -0.31     0.24
## age:sexM:lit        11.91     5.94
## age:sexM:mage        0.00     0.00
## age:lit:mage         0.01     0.00
## sexM:lit:mage       28.53    14.22
## age:sexM:lit:mage   -0.62     0.31
##
## Random Effects:
##  Groups   Name        Std.Dev.
##  id       (Intercept) 1.32
##  Residual             0.44
## ---
## number of obs: 877, groups: id, 197
## AIC = 1863.4, DIC = 1659.9
## deviance = 1743.7
```

Note that the AICs indicate that the best model is the first model. We could also use the 'KRmodcomp' function to verify this result. Thus, we obtain more precise numbers with the following code.

```
lmer1 <- lmer1_1
summary(lmer1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: wt ~ age * sex + lit + mage + (1 | id)
##    Data: nepali
##
## REML criterion at convergence: 1789.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.4700 -0.4886  0.0007  0.5198  4.1880
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept) 1.766    1.329
##  Residual             0.191    0.437
## Number of obs: 877, groups:  id, 197
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 4.3812771  0.4859578   9.016
## age         0.1338857  0.0034868  38.398
```

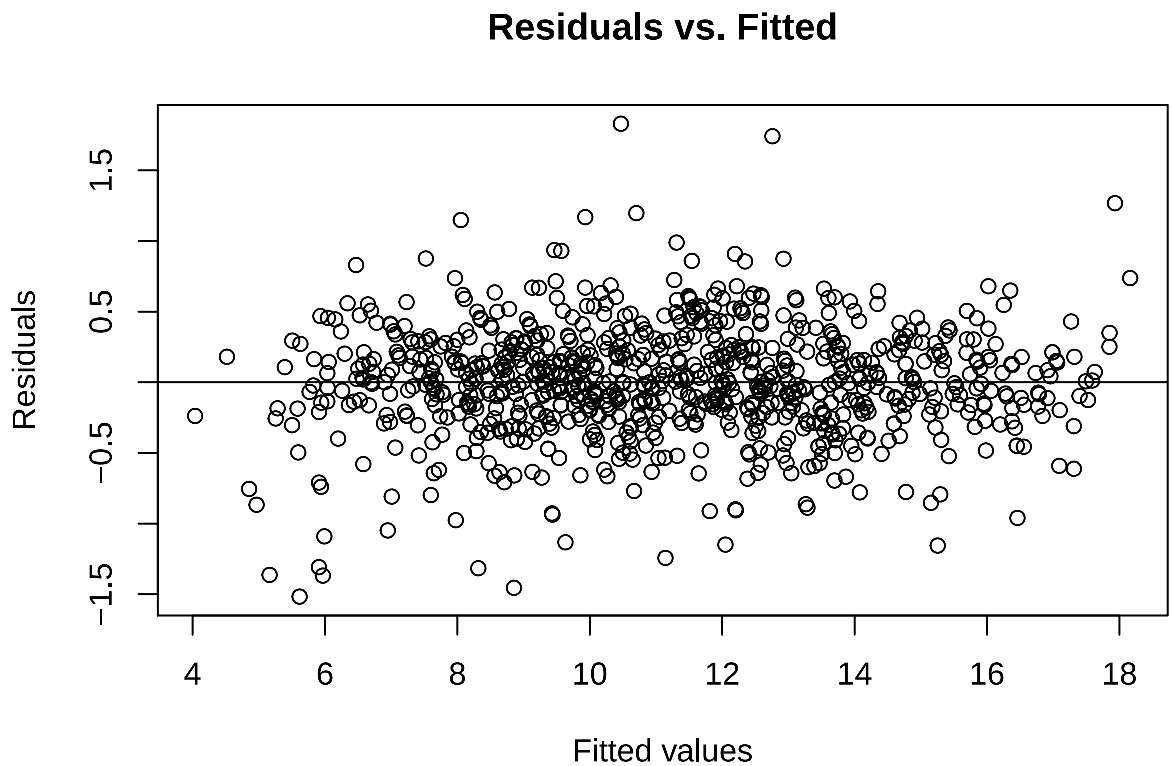
```
## sexM        0.3698066  0.2640962   1.400
## lit         0.7560264  0.4716539   1.603
## mage        0.0508327  0.0157632   3.225
## age:sexM    0.0007549  0.0048474   0.156
##
## Correlation of Fixed Effects:
##          (Intr) age    sexM   lit    mage
## age      -0.182
## sexM     -0.304  0.478
## lit      -0.195 -0.010  0.063
## mage     -0.919 -0.086  0.024  0.145
## age:sexM  0.185 -0.714 -0.682  0.005  0.003
```

The predicted difference in child weight between a 15- and a 25-year-old mother is $10 * 0.0508327 = 0.508327$. Weights of identical twins would expected to be the same since they have all the same values of predictors, and this seems reasonable.

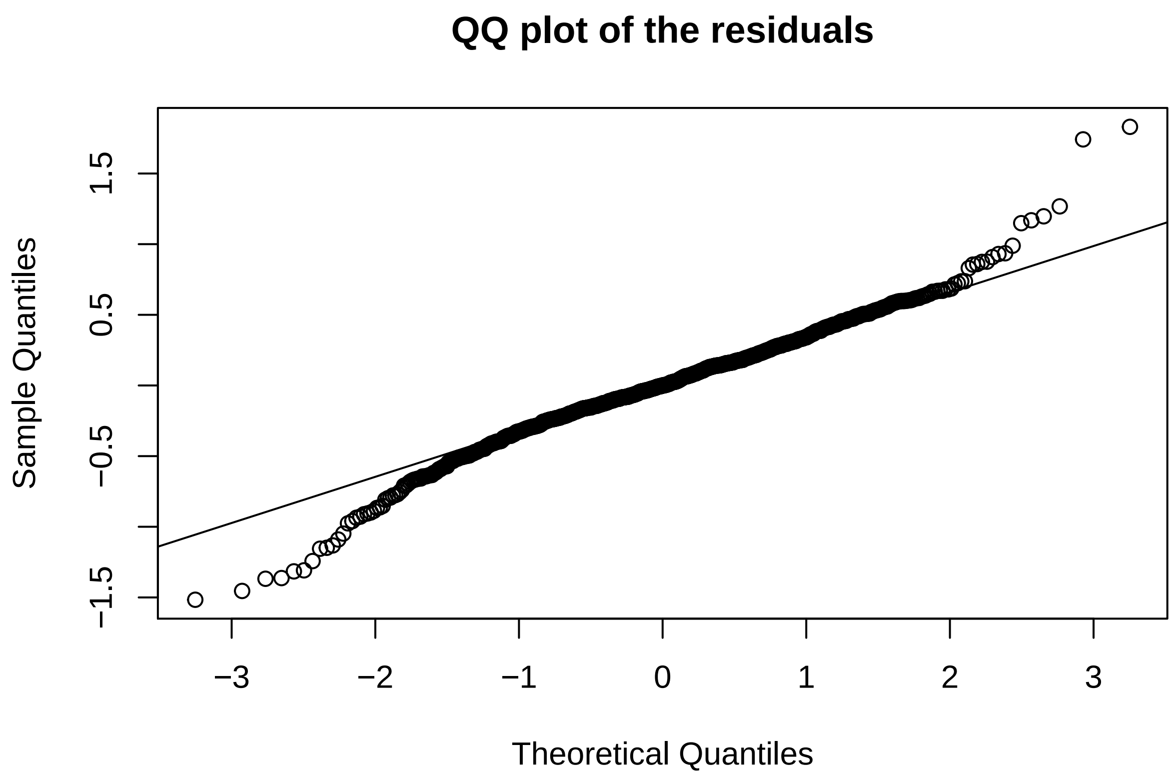
Part (d)

We make the following plots and interpret: (i) Residuals vs. Fitted plot, (ii) QQ plot of the residuals, (iii) QQ plot of the random effects.

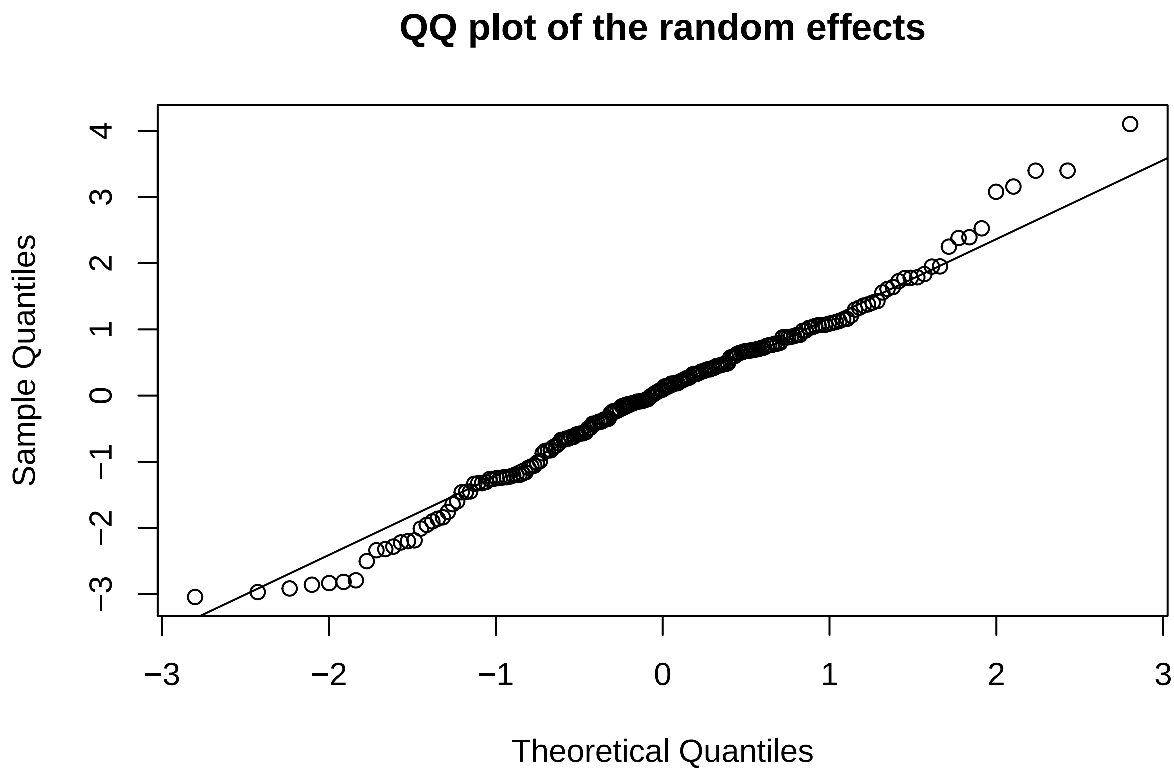
```r
plot(residuals(lmer1) ~ fitted(lmer1), main = "Residuals vs. Fitted", xlab = "Fitted values", ylab = "Re
abline(h=0)
```



```r
qqnorm(residuals(lmer1), main = "QQ plot of the residuals")
qqline(residuals(lmer1))
```




```r
qqnorm(ranef(lmer1)$"id"[[1]], main = "QQ plot of the random effects")
qqline(ranef(lmer1)$"id"[[1]])
```



The plots show that there is no evidence of lack of fit.

Part (e)

We fit a model with age and mother's age as the only fixed effects and compare it to the previous model.

```r
lmer2 <- lmer(wt ~ age + mage + (1|id), nepali)
summary(lmer2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: wt ~ age + mage + (1 | id)
##    Data: nepali
##
## REML criterion at convergence: 1785.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.4706 -0.4862 -0.0019  0.5169  4.2063
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept) 1.8045   1.3433
##  Residual             0.1907   0.4367
## Number of obs: 877, groups:  id, 197
##
```

```
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 4.731261   0.459734  10.291
## age         0.134367   0.002443  54.990
## mage        0.046498   0.015756   2.951
##
## Correlation of Fixed Effects:
##      (Intr) age
## age  -0.082
## mage -0.958 -0.118
```

```
library(pbkrtest)
```

```
## Warning: package 'pbkrtest' was built under R version 3.6.3
```

```
KRmodcomp(lmer1, lmer2)
```

```
## F-test with Kenward-Roger approximation; time: 0.97 sec
## large : wt ~ age * sex + lit + mage + (1 | id)
## small : wt ~ age + mage + (1 | id)
##          stat      ndf      ddf F.scaling p.value
## Ftest   2.0928   3.0000 309.6111    0.9989  0.1011
```

The p-value indicates that the previous model is not significantly better.

Part (f)

We elaborate the previous model to include a random slope in age and use AIC to choose between this model and the previous one.

```
lmer3 <- lmer(wt ~ age + mage + (age|id), nepali)
summary(lmer3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: wt ~ age + mage + (age | id)
##    Data: nepali
##
## REML criterion at convergence: 1706.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.02185 -0.45017  0.02223  0.49853  2.59522
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 1.93545  1.39120
##           age         0.00129  0.03592  -0.57
##  Residual             0.14125  0.37583
## Number of obs: 877, groups:  id, 197
##
## Fixed effects:
```

```
##             Estimate Std. Error t value
## (Intercept) 4.898072   0.429071  11.416
## age         0.136878   0.003384  40.445
## mage        0.044120   0.014635   3.015
##
## Correlation of Fixed Effects:
##      (Intr) age
## age  -0.151
## mage -0.953 -0.068
```

```r
sumary(lmer3)
```

```
## Fixed Effects:
##             coef.est coef.se
## (Intercept) 4.90     0.43
## age         0.14     0.00
## mage        0.04     0.01
##
## Random Effects:
##  Groups   Name        Std.Dev. Corr
##  id       (Intercept) 1.39
##           age         0.04     -0.57
##  Residual             0.38
## ---
## number of obs: 877, groups: id, 197
## AIC = 1720.8, DIC = 1668.5
## deviance = 1687.7
```

Since the AIC is smaller, we choose the new model. For our chosen model, weight of children are expected to increase by 0.136878kg per year. This rate could vary with a standard deviation of 0.03592kg/yr.

Part (g)

We extract information about panchayat, ward, household and birth order from the `id` variable. Then we fit a random intercept mixed effects model which allows for the nested random effects structure of child within household within ward whithin panchayat. We also construct bootstrap confidence intervals.

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
##
## 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
```

```r
nepali <- nepali %>%
  mutate(panchayat = substring(id, 1, 2),
         ward = substring(id, 3, 4),
         household = substring(id, 5, 5),
         child = substring(id, 6))
lmer4 <- lmer(wt ~ age + mage + (1|panchayat) + (1|panchayat:ward) + (1|panchayat:ward:household) + (1|p
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.00221673
## (tol = 0.002, component 1)
```

```r
summary(lmer4)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: wt ~ age + mage + (1 | panchayat) + (1 | panchayat:ward) + (1 |
##     panchayat:ward:household) + (1 | panchayat:ward:household:child)
##    Data: nepali
##
## REML criterion at convergence: 1768.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.4836 -0.4806 -0.0011  0.5339  4.1781
##
## Random effects:
##  Groups                         Name        Variance Std.Dev.
##  panchayat:ward:household:child (Intercept) 1.26068  1.1228
##  panchayat:ward:household       (Intercept) 0.37393  0.6115
##  panchayat:ward                 (Intercept) 0.03572  0.1890
##  panchayat                      (Intercept) 0.24034  0.4902
##  Residual                                   0.19086  0.4369
## Number of obs: 877, groups:
## panchayat:ward:household:child, 197; panchayat:ward:household, 112; panchayat:ward, 19; panchayat, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 5.067435   0.553918   9.148
## age         0.135053   0.002388  56.564
## mage        0.037596   0.016125   2.332
##
## Correlation of Fixed Effects:
##      (Intr) age
## age  -0.064
## mage -0.802 -0.116
## convergence code: 0
## Model failed to converge with max|grad| = 0.00221673 (tol = 0.002, component 1)
```

```r
set.seed(590)
confint(lmer4, method = "boot")
```

```
## Computing bootstrap confidence intervals ...
```

```
##
## 115 message(s): boundary (singular) fit: see ?isSingular
## 264 warning(s): Model failed to converge with max|grad| = 0.0020018 (tol = 0.002, component 1) (and o

##                   2.5 %     97.5 %
## .sig01       0.95291161 1.29196428
## .sig02       0.01370239 0.85064572
## .sig03       0.00000000 0.47649972
## .sig04       0.00000000 1.00583002
## .sigma       0.41180844 0.46159387
## (Intercept) 4.08243793 6.20614697
## age          0.13019860 0.14007733
## mage         0.00651768 0.06779261
```

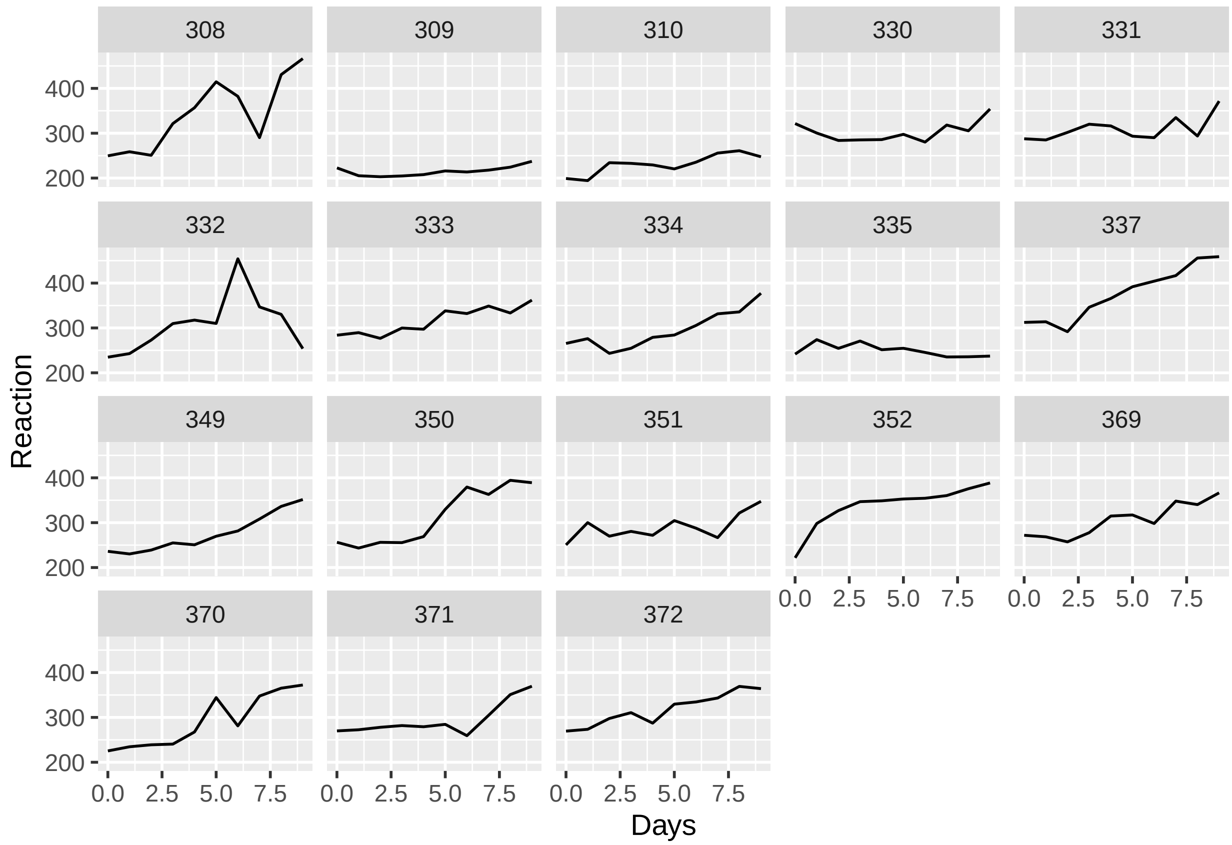
We can see that `panchayat:ward:household:child`and `panchayat:ward:household`are significant, and that the random effects are relatively large compared to the fixed effects.

Page 253 Exercise 5

Part (a)

We plot the data taking care to distinguish the trajectories of the different subjects.

```
data(sleepstudy)
ggplot(sleepstudy, aes(Days, Reaction)) + geom_line() + facet_wrap(. ~ Subject)
```



The plot show that reaction time increases as the number of days of sleep deprivation increases with some exceptions.

Part (b)

We fit a mixed effects model that describes how the reaction time varies linearly with days and allows for random variation in both the slope and intercepts of the subject lines.

```r
lmer1 <- lmer(Reaction ~ Days + (Days|Subject), sleepstudy)
summary(lmer1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
##    Data: sleepstudy
##
## REML criterion at convergence: 1743.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9536 -0.4634  0.0231  0.4633  5.1793
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 611.90   24.737
##           Days         35.08    5.923   0.07
##  Residual             654.94   25.592
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  251.405      6.824  36.843
## Days          10.467      1.546   6.771
##
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.138
```

Since the coefficient estimate of `Days` is much larger than its standard error and random variation, it would be unusual for one to have a reaction time that does not increase over time.

Part (c)

We allow for quadratic effects in the previous model.

```r
lmer2 <- lmer(Reaction ~ Days + I(Days^2) + (Days|Subject), sleepstudy)
summary(lmer2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + I(Days^2) + (Days | Subject)
##    Data: sleepstudy
##
```

```
## REML criterion at convergence: 1742.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.0093 -0.4489  0.0422  0.5036  5.2702
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 613.12   24.761
##           Days         35.11    5.925   0.06
##  Residual             651.97   25.534
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 255.4494     7.5135  33.999
## Days          7.4341     2.8189   2.637
## I(Days^2)     0.3370     0.2619   1.287
##
## Correlation of Fixed Effects:
##           (Intr) Days
## Days      -0.418
## I(Days^2)  0.418 -0.836
```

```
KRmodcomp(lmer1, lmer2)
```

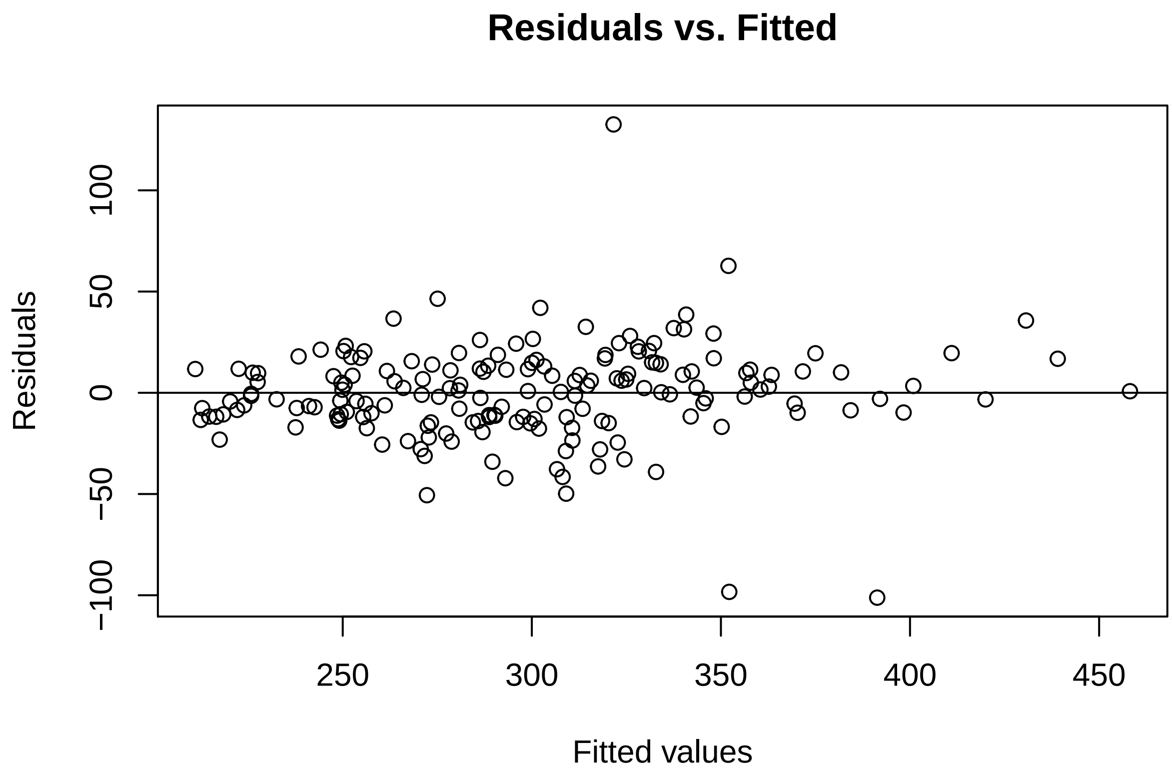
```
## F-test with Kenward-Roger approximation; time: 0.14 sec
## large : Reaction ~ Days + I(Days^2) + (Days | Subject)
## small : Reaction ~ Days + (Days | Subject)
##           stat      ndf      ddf F.scaling p.value
## Ftest   1.6558   1.0000 143.0000         1  0.2003
```

Since the p-value is small, we conclude that the quadratic term is not significant.

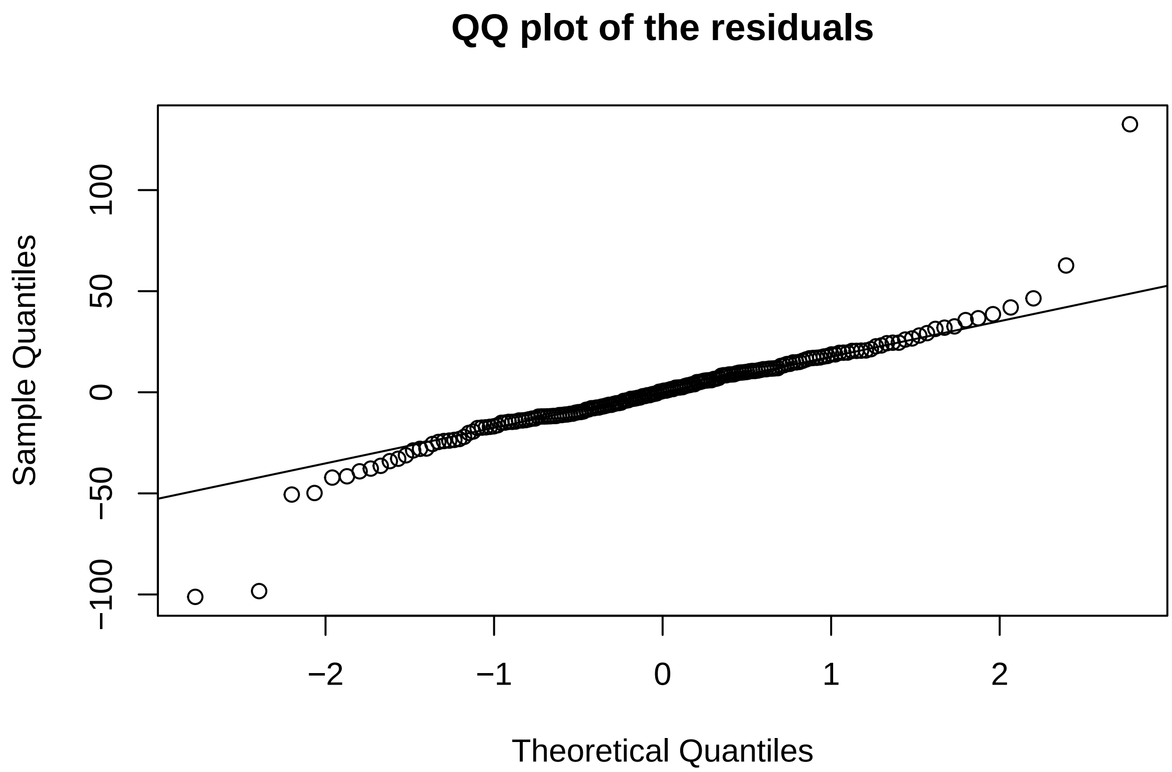
Part (d)

We make the following diagnostic plots and interpret: (i) Residuals vs. Fitted plot, (ii) QQ plot of the residuals, (iii) QQ plot of both random effects, (iv) a scatterplot of the random effects.

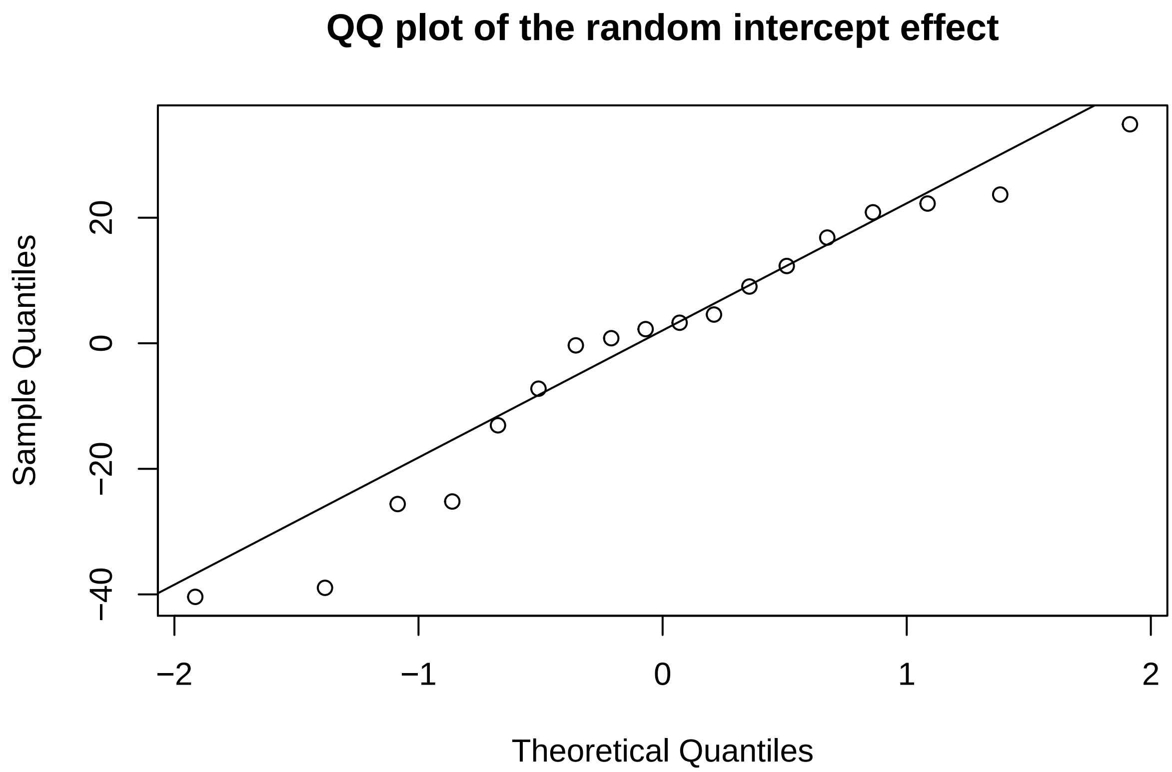
```
plot(residuals(lmer1) ~ fitted(lmer1), main = "Residuals vs. Fitted", xlab = "Fitted values", ylab = "Re
abline(h=0)
```



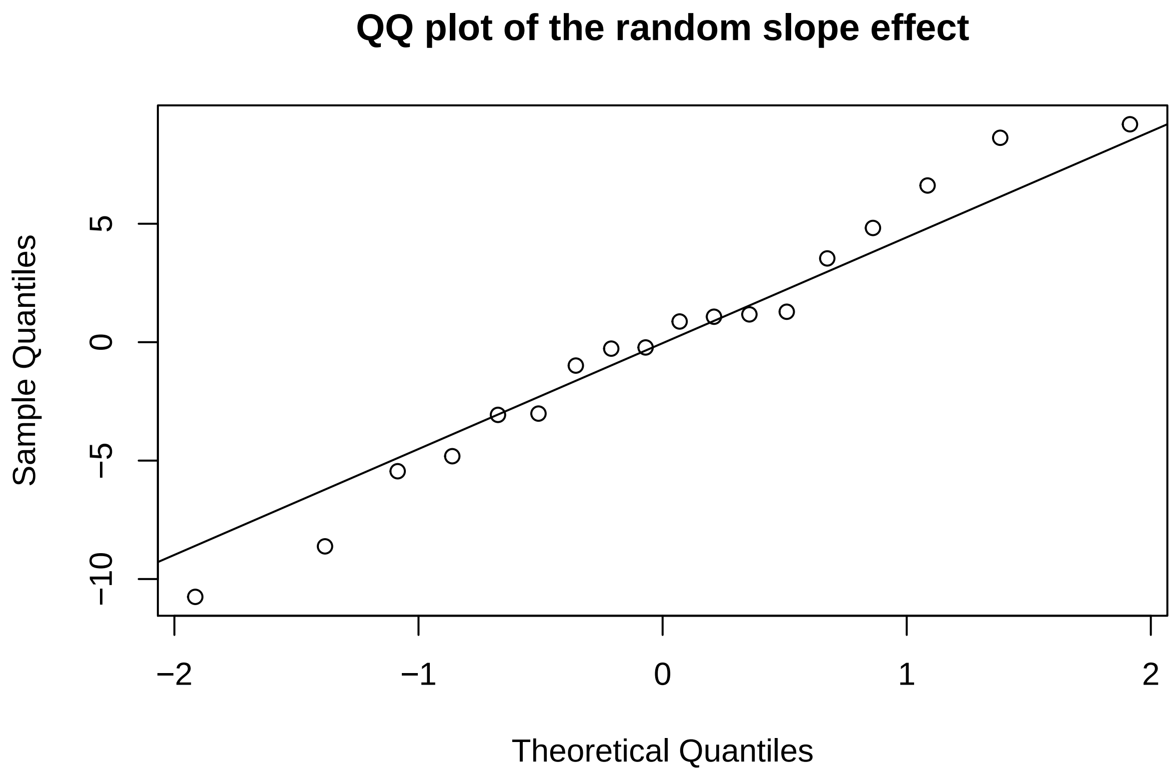

```r
qqnorm(residuals(lmer1), main = "QQ plot of the residuals")
qqline(residuals(lmer1))
```



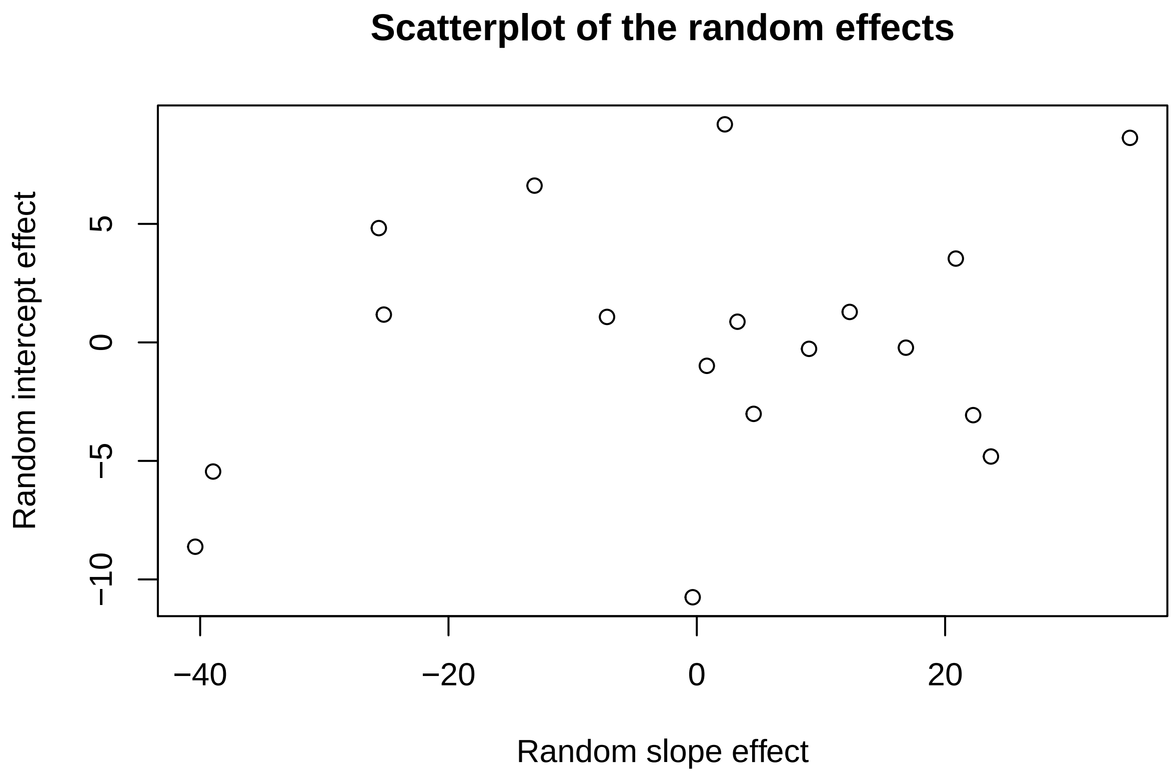

```
qqnorm(ranef(lmer1)$"Subject"[[1]], main = "QQ plot of the random intercept effect")
qqline(ranef(lmer1)$"Subject"[[1]])
```




```
qqnorm(ranef(lmer1)$"Subject"[[2]], main = "QQ plot of the random slope effect")
qqline(ranef(lmer1)$"Subject"[[2]])
```




```
plot(ranef(lmer1)$"Subject"[[2]] ~ ranef(lmer1)$"Subject"[[1]], main = "Scatterplot of the random effec
```



The plots show that there is no evidence of lack of fit. However, we can see that there are three possible outliers.

Part (e)

```
sleepstudy1 <- sleepstudy[-order(abs(residuals(lmer1)), decreasing = TRUE)[1:3],]
lmer2 <- lmer(Reaction ~ Days + (Days|Subject), sleepstudy1)
summary(lmer2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
##    Data: sleepstudy1
##
## REML criterion at convergence: 1638.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.67893 -0.55470 -0.00955  0.56348  2.47740
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 705.27   26.557
##           Days         44.71    6.687   -0.06
##  Residual             374.16   19.343
```

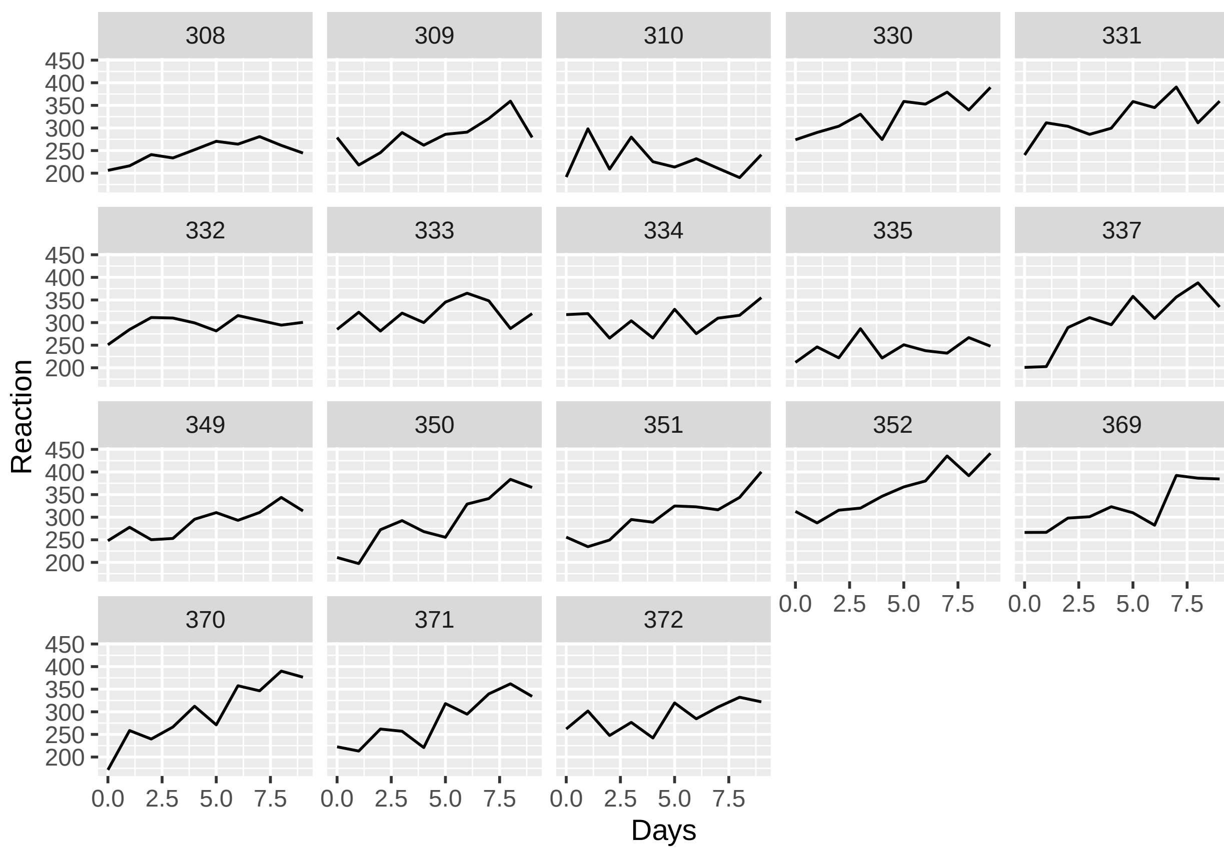
```
## Number of obs: 177, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  250.135      6.812  36.721
## Days          10.880      1.656   6.568
##
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.157
```

The largest change in the model fit is the random slope effect.

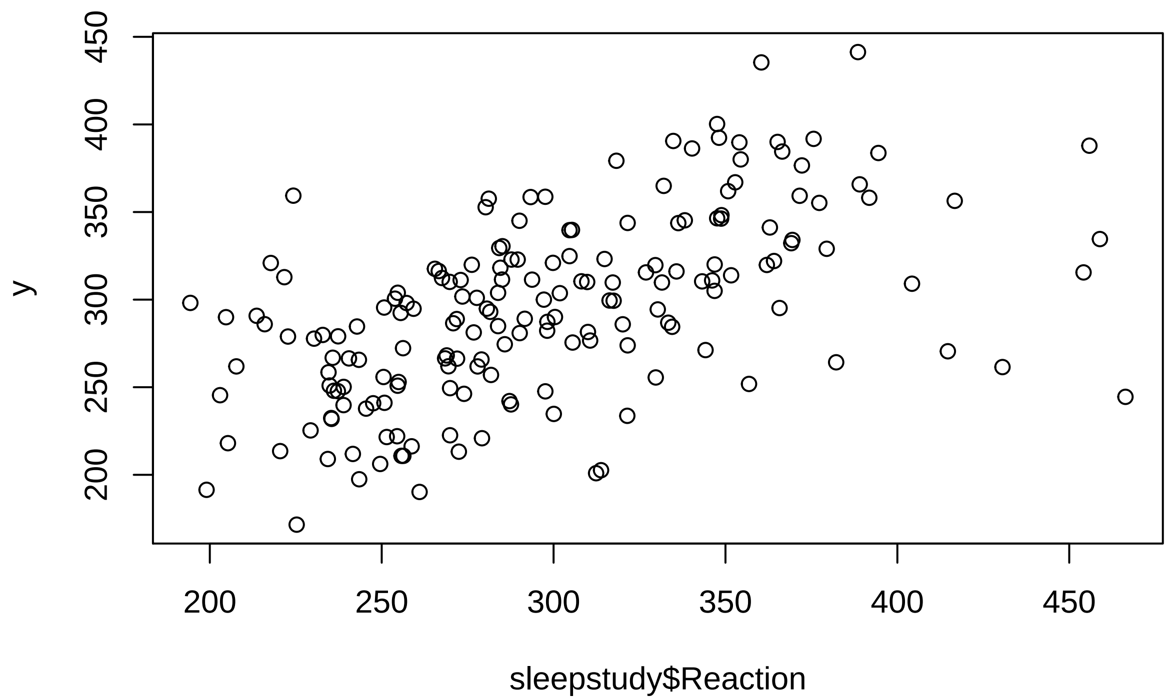
Part (f)

We simulate the response under our first model and plot it.

```r
set.seed(590)
y = unlist(simulate(lmer1))
sleepstudy2 <- sleepstudy
sleepstudy2$Reaction <- y
ggplot(sleepstudy2, aes(Days, Reaction)) + geom_line() + facet_wrap(. ~ Subject)
```



```r
plot(y ~ sleepstudy$Reaction)
```



We can observe that there is a difference between the simulated data and the actual data.