Assignment 9

Mingrui Duan

Question 1:

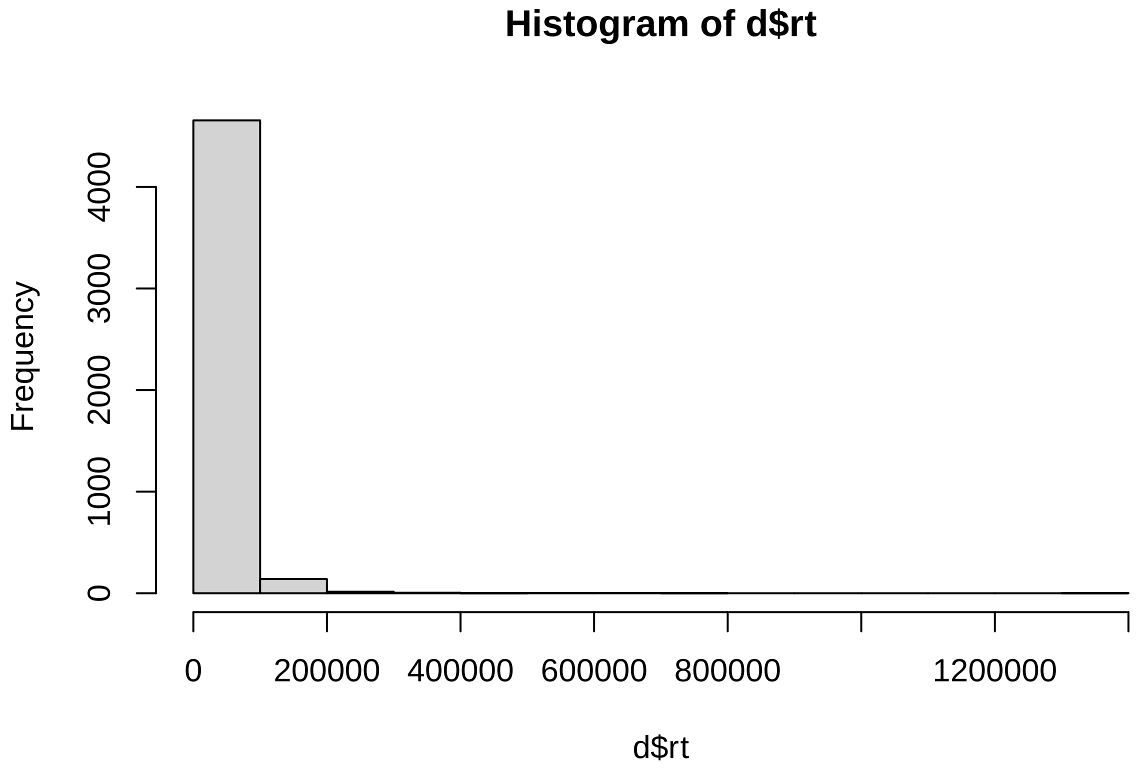
(a) If items get harder, what should we predict about RT and accuracy as item number increases?

As the items get harder, we should predict that the RT will get longer and the accuracy will get lower.

(b) Load the data, check RT for outliers

```
# Load the data
d <- read.csv("data.csv")

# Plot the histogram of the "RT" Column
hist(d$rt)
```



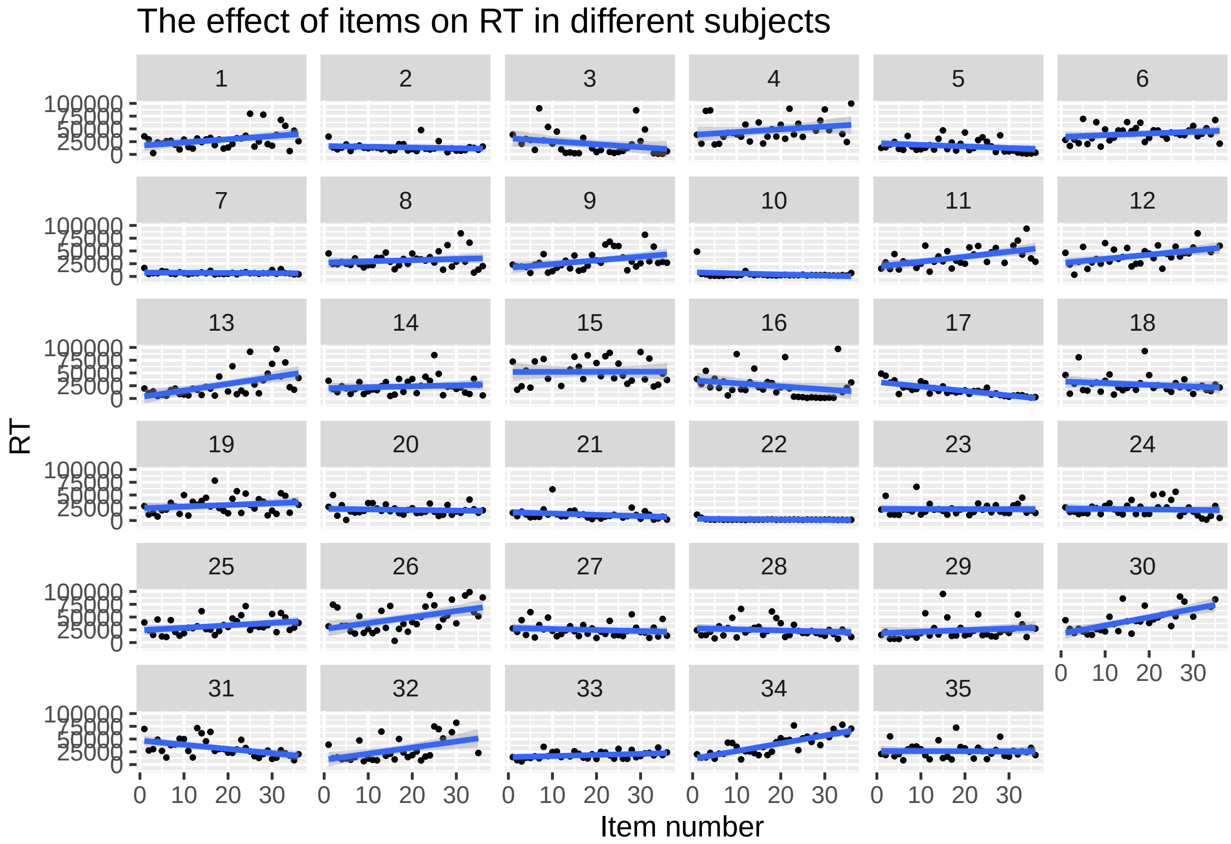
(c) do something to handle them if you find any

```
# From the histogram we can observe that for RT > 100000, these are the outliers
# So we just remove these outliers from the orginal dataset
d <- subset(d, d$rt < 100000)
```

Question 2: Plot the effect of item on RT in a publication-quality plot. Use color to show which were answered correctly, facet by subject, and include a linear regression line for each subject.

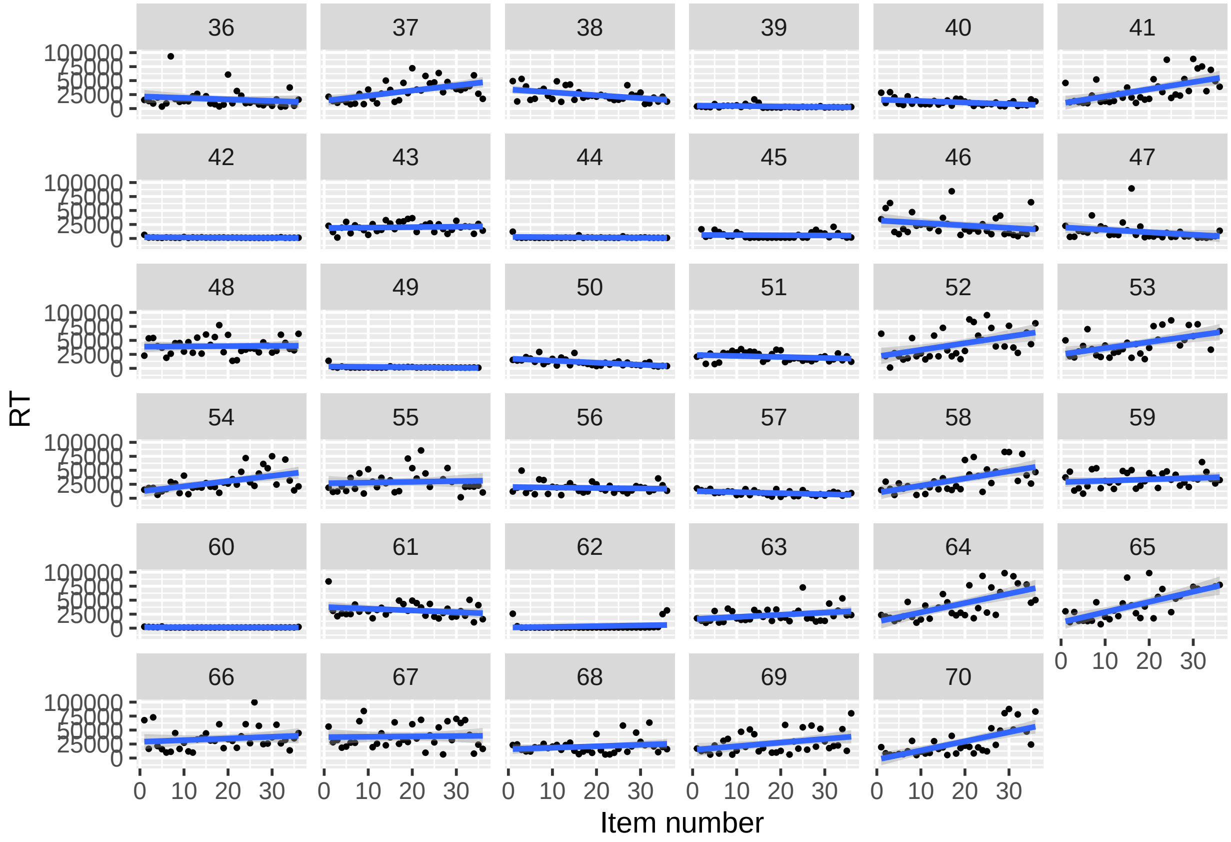
```
library("ggplot2")
# Due to there are too many subjects, so I split these subjects into four charts.
d.1 <- subset(d, d$subject < 36)
ggplot(data = d.1, aes(x = item, y = rt)) +
  geom_point(size = 0.5) +
  stat_smooth(method = "lm") +
  facet_wrap(~ subject) +
  labs(x = "Item number",
       y = "RT") + ggtitle("The effect of items on RT in different subjects")
```

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



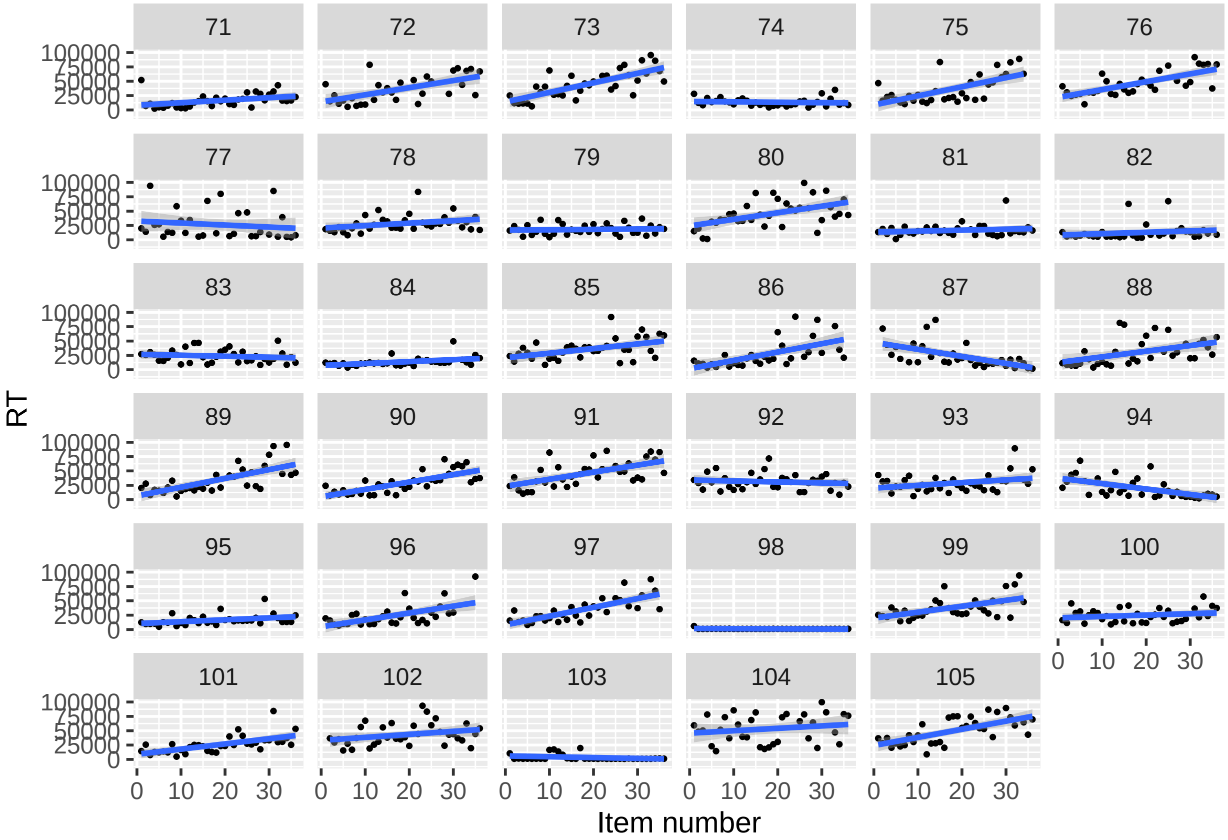
```
d.2 <- subset(d, d$subject > 35 & d$subject < 71)
ggplot(data = d.2, aes(x = item, y = rt)) +
  geom_point(size = 0.5) +
  stat_smooth(method = "lm") +
  facet_wrap(~ subject) +
  labs(x = "Item number",
       y = "RT")
```

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



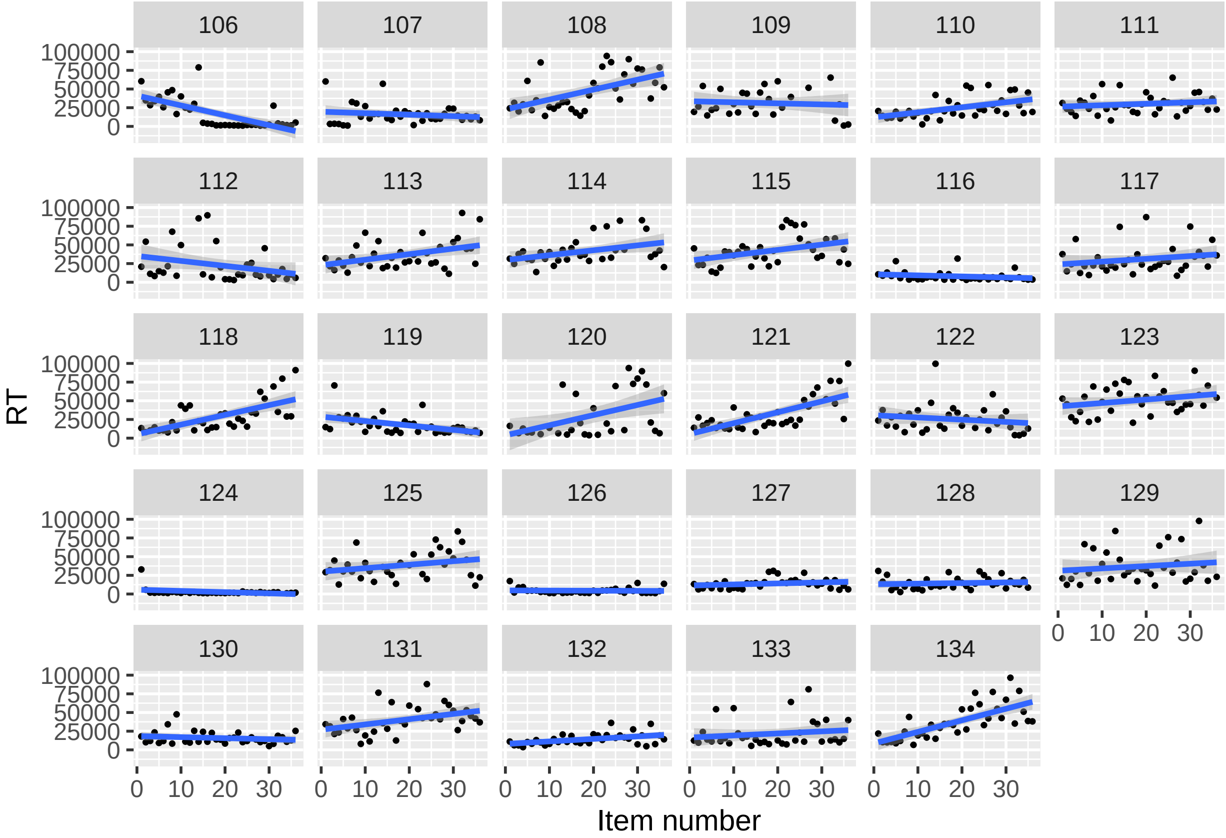
```
d.3 <- subset(d, d$subject > 70 & d$subject < 106)
ggplot(data = d.3, aes(x = item, y = rt)) +
  geom_point(size = 0.5) +
  stat_smooth(method = "lm") +
  facet_wrap(~ subject) +
  labs(x = "Item number",
       y = "RT")
```

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



```
d.4 <- subset(d, d$subject > 105)
ggplot(data = d.4, aes(x = item, y = rt)) +
  geom_point(size = 0.5) +
  stat_smooth(method = "lm") +
  facet_wrap(~ subject) +
  labs(x = "Item number",
       y = "RT")
```

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


106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
100000
75000
50000
25000
0
RT
0
10
20
30
Item number

Question 3: What do you see in the data there? Are there any subjects who should be removed and why (not just response outliers as in Q1, but entire subjects)? Can you spot any? Come up with a criterion for subjects who aren’t trying at all, and remove them (remove them with code, NOT with a list by hand or by editing the spreadsheet). List the subjects you removed.

From the charts we got, we can tell that some subjects had a flat line after conducting linear regression on items~RT, means that there’s no effect of item on RT. I think for those who has flat lines in outcome should be removed. This is because these samples from an online test which may have high noise, so they may not take the test seriously, so even items’ getting more difficult, they won’t take more RT trying to do best in the test.

From these charts, I can tell that subject 2, 7, 10, etc. They are subjects should be removed. For those RT is close to 0(average RT is less than 100000), they should be removed from the further analysis.

```
# Make a list storing the subjects that should be removed
rem_list <- NULL

for (i in 1:134) {
  sub <- subset(d, d$subject == i)
  ave <- mean(sub$rt)
  if (ave < 10000) {
    rem_list <- c(rem_list, i)
  }
}
# Print the list contain those subjects that should be removed
print(rem_list)
```

```
##  [1]   7  10  22  39  42  44  45  49  57  60  62  98 103 116 124 126
```

```
# Remove these subjects
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
```

```
d <- d %>% dplyr::filter(!subject %in% rem_list)
```

Question 4:

(a) Run a model predicting RT from item and correct, and include random slopes and intercepts by subject. Print a summary and write 2-3 sentences like you might find in a real paper to explain whether or not it is true that subjects overall spend more time on later questions.

```r
library(lme4)
```

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

```r
# Run the model
mod <- lmer(rt ~ item + correct + (item + correct | subject),
            data = d,
            control=lmerControl(optimizer="Nelder_Mead"))
summary(mod)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rt ~ item + correct + (item + correct | subject)
##    Data: d
## Control: lmerControl(optimizer = "Nelder_Mead")
##
## REML criterion at convergence: 91060.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7911 -0.6085 -0.1873  0.3940  4.9580
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  subject  (Intercept)  94697479  9731.3
##           item           411010   641.1  -0.47
##           correct       9449036  3073.9  -0.34 -0.29
##  Residual             253971457 15936.5
## Number of obs: 4082, groups:  subject, 118
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 23477.71    1155.09  20.325
## item          377.57      65.01   5.808
## correct     -2081.40     691.07  -3.012
##
## Correlation of Fixed Effects:
##         (Intr) item
## item    -0.562
## correct -0.519  0.066
```

From the fixed effects, the coefficient of item is positive, means that it is true that subjects overall spend more time on later questions.

(b) Compare the coefficients from lmer to a simple lm without subject effects and explain any differences or similarities you see.

```r
# Run a regular linear model
linear.mod <- lm(rt ~ item + correct + 1, data = d)
summary(linear.mod)
```

```
##
## Call:
## lm(formula = rt ~ item + correct + 1, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36543 -14468  -4646   9430  73173
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19647.89     854.14  23.003  < 2e-16 ***
## item          438.20      33.32  13.152  < 2e-16 ***
## correct      2653.09     693.41   3.826 0.000132 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 20040 on 4079 degrees of freedom
## Multiple R-squared:  0.04146,    Adjusted R-squared:  0.04099
## F-statistic: 88.22 on 2 and 4079 DF,  p-value: < 2.2e-16
```

- Differences: The coefficients of “correct” for lm is positive while for the lmer is negative.
- Similarities: Both 2 models’ coefficient of “item” is positive, means that subjects overall spend more time on later questions.

Question 5:

```
tab <- ranef(mod)$subject
```

(a) What is the mean of the item slope adjustments in your regression and why does it have this value?

```
print(mean(tab$item))
```

```
## [1] -7.195624e-12
```

The mean of the item slope adjustments is: -3.697904e-12, beacuse the value of this column is the adjustment to the item slope, means they're the difference between the mean slope and the slope of certain subject. So the mean of all items' slope must be 0, no adjustment to the mean slope.

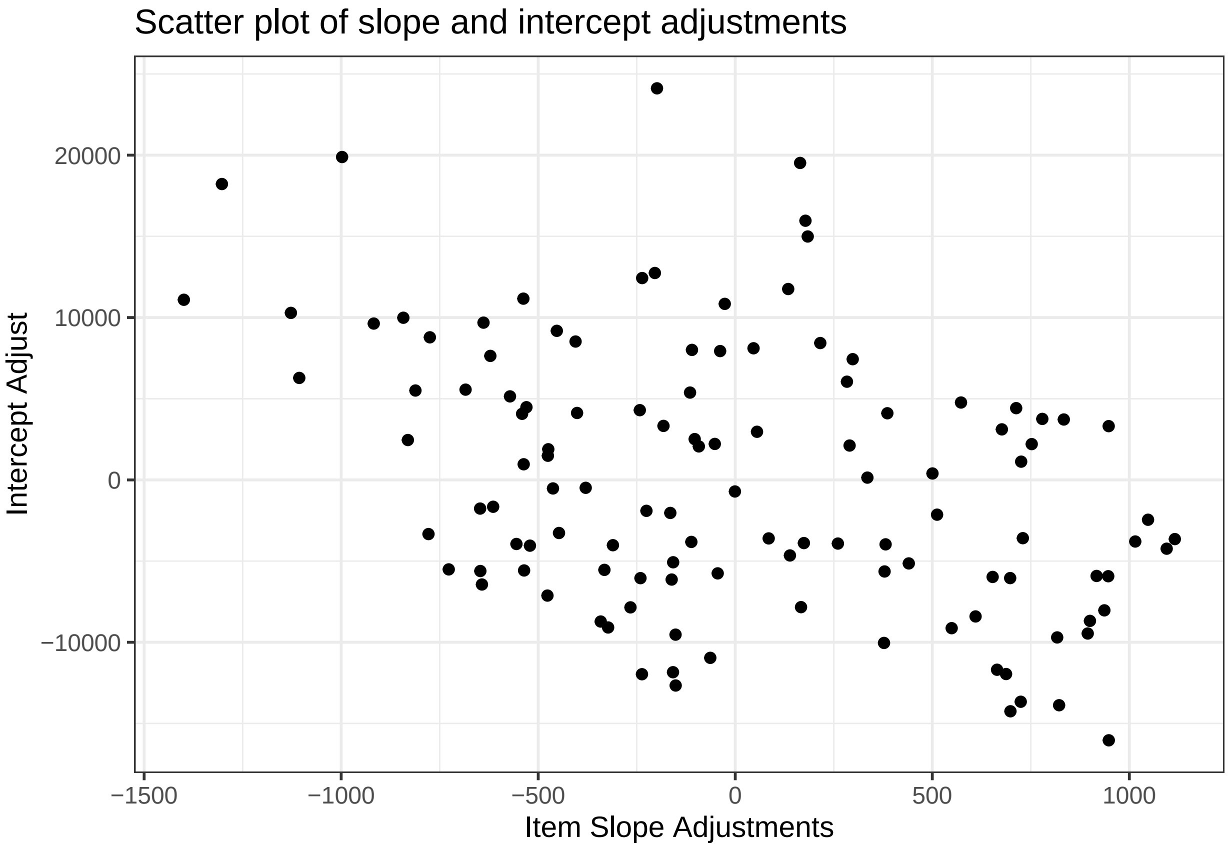
(b) How many subjects here have fit slopes which are positive vs. negative? (Remember ranef doesn't give the slope, it gives the adjustment to the slope)

```
# Find those subjects who have negative slope on item
# If the adjustments lower than the slope of item in the fixed effect
# Then the slope will be negative
neg.slope.item <- subset(tab, tab$item < -320)
print(length(neg.slope.item))
```

```
## [1] 3
```

(c) Make a scatter plot of the slope and intercept adjustments for each subject.

```
ggplot(data = tab, aes(x = item, y = `(Intercept)`)) +
        geom_point() +
        labs(title = "Scatter plot of slope and intercept adjustments",
             x = "Item Slope Adjustments",
             y = "Intercept Adjust") +
        theme_bw()
```



(d) Are they correlated or uncorrelated? Explain why, intuitively, and describe what the correlation means.

I think they are correlated. Because from the scatter plot, we can see that as the slope adjustments goes high, the intercept adjustment goes low. This correlation means that when conducting linear regression, the intercept must be adjusted to fit the change of slope to decrease the error of model fitting(For example, decrease the MSE when fitting).