Assignment 3

Jiarong Ye

September 21, 2018

Q1

1. Use R to randomly assign 10 experimental units to each of three treatments (1, 2, and 3). Then simulate responses for the 30 experimental units satisfying the one-way ANOVA model:

$$Y_{it} = \mu + \tau_i + \epsilon_{it}, i = 1, 2, ..., v, t = 1, 2, , ..., r_i$$

$$\epsilon_{it} \sim N(0, \sigma^2)$$

with $\mu = 4.7, \sigma^2 = 4$, and treatment effects $\tau_1 = -3, \tau_2 = 5, \tau_3 = -2$. Your solution should include your R code and a plot of the simulated values.

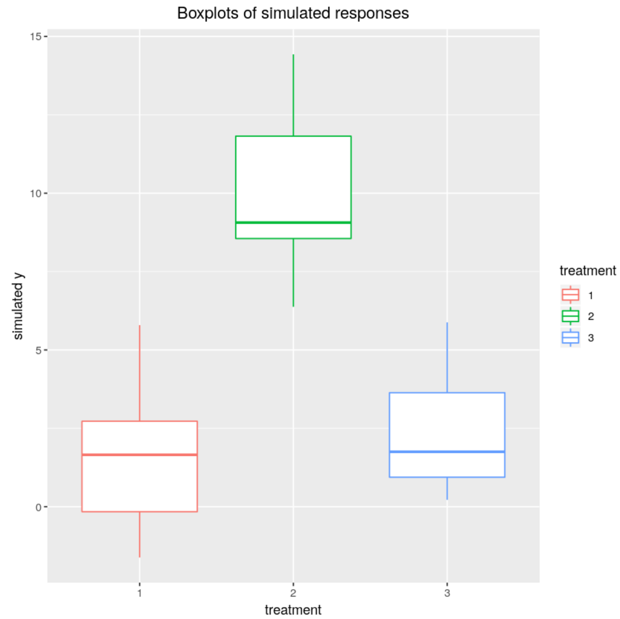
In [40]:
```
treatments.not.random=c(rep("1",10), rep("2",10), rep("3", 10))
treatment=sample(treatments.not.random)
experiment_unit = 1:length(treatment)
table=data.frame(experiment_unit, treatment,row.names=NULL)
table
```

experiment_unit	treatment
1	1
2	3
3	2
4	2
5	2
6	3
7	3
8	1
9	1
10	1
11	2
12	1
13	3
14	3
15	2
16	2
17	3
18	2
19	3
20	2
21	1
22	3
23	3
24	1
25	2
26	1
27	3
28	1
29	1
30	2

In [41]:
```
mu=4.7
tau_1=-3
tau_2=5
tau_3=-2
var=4

means_q1=rep(NA,length(treatment))
means_q1[treatment==1] = mu+tau_1
means_q1[treatment==2] = mu+tau_2
means_q1[treatment==3] = mu+tau_3

y_simulate_q1 = means_q1+rnorm(n = length(means_q1), mean = 0, sd = sqrt(var))
sim_data = data.frame(experiment_unit, treatment, y_simulate_q1)
sim_data
```

experiment_unit	treatment	y_simulate_q1
1	1	-1.6184361
2	3	5.8802735
3	2	6.3770675
4	2	14.3735955
5	2	8.4595723
6	3	3.8001343
7	3	1.0086134
8	1	0.3531091
9	1	1.0026467
10	1	2.3112544
11	2	10.4067991
12	1	3.3594491
13	3	4.4810127
14	3	0.9192064
15	2	12.2877769
16	2	6.5982482
17	3	0.2219408
18	2	14.4265322
19	3	2.1988443
20	2	8.8380465
21	1	5.7932315
22	3	1.3083453
23	3	3.1476776
24	1	-0.3273547
25	2	8.9700277
26	1	-0.9235407
27	3	0.8706975
28	1	2.6829802
29	1	2.7401941
30	2	9.1540193

In [42]:
```
library(ggplot2)
p1<-ggplot(sim_data, aes(x=treatment, y=y_simulate_q1, color=treatment)) +
  geom_boxplot() +
  ylab('simulated y') +
  ggtitle('Boxplots of simulated responses') +
  theme(plot.title = element_text(hjust = 0.5))
p1
```



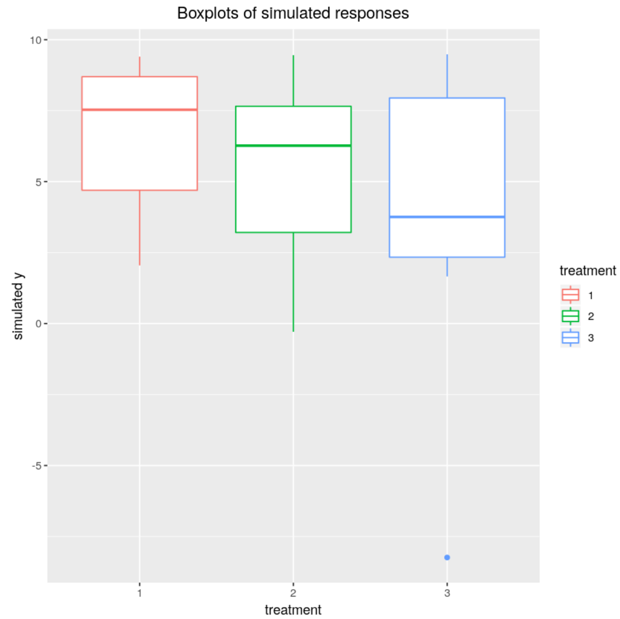
Q2

Consider the situation in Problem 1. The experimenter wants to consider a reduced model where $\tau 1 = \tau 2 = \tau 3 = 0$. Simulate responses for the 30 experimental units satisfying this reduced model. Compare boxplots of simulated responses under this reduced model with boxplots of simulated responses under the full model described in Problem 1 (where there are differences in the treatment effects).

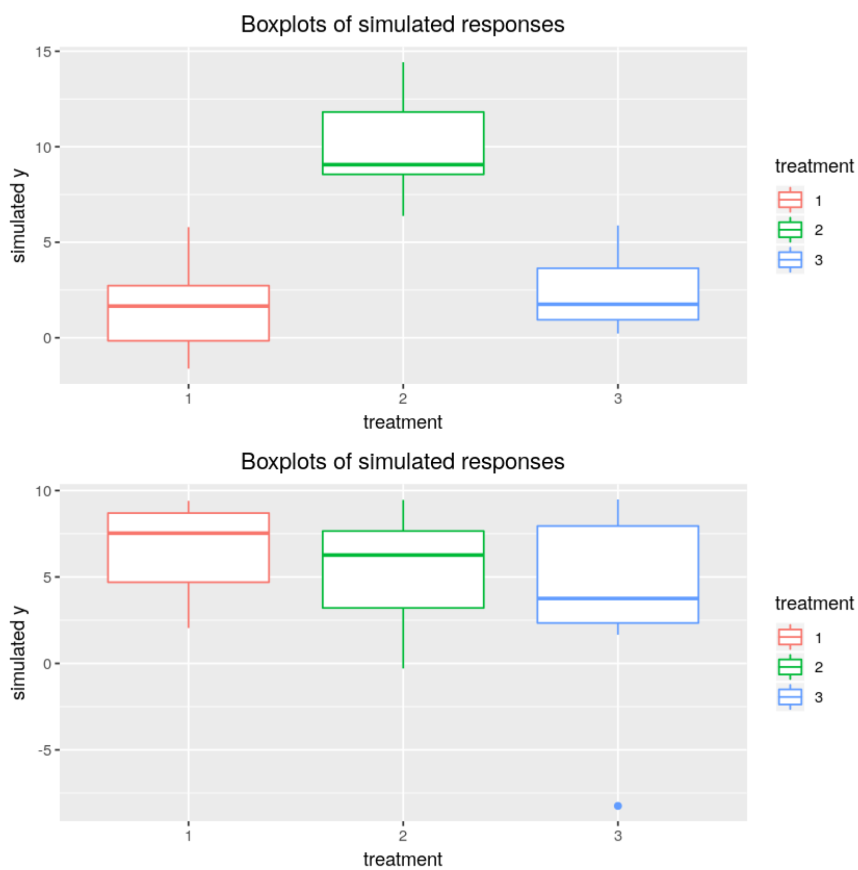
In [48]:
```
tau_1=0
tau_2=0
tau_3=0

means_q2=rep(NA,length(treatment))
means_q2[treatment==1] = mu+tau_1
means_q2[treatment==2] = mu+tau_2
means_q2[treatment==3] = mu+tau_3

y_simulate_q2 = means_q2+rnorm(n = length(means_q2), mean = 0, sd = sqrt(var))
sim_data = data.frame(experiment_unit, treatment, y_simulate_q2)
sim_data
```

experiment_unit	treatment	y_simulate_q2
1	1	4.5871138
2	3	8.9409421
3	2	5.4448399
4	2	-0.2890012
5	2	9.4563168
6	3	5.6918686
7	3	-8.2416622
8	1	7.3036280
9	1	8.7815548
10	1	5.0297383
11	2	-0.2905075
12	1	8.4399276
13	3	3.1789920
14	3	2.0923404
15	2	5.4524241
16	2	9.3299666
17	3	1.6604748
18	2	7.1431085
19	3	4.3337423
20	2	7.0824088
21	1	3.3082532
22	3	3.0837190
23	3	9.4843263
24	1	7.7620453
25	2	2.4644266
26	1	2.0500897
27	3	8.6995005
28	1	9.1221858
29	1	9.4040856
30	2	7.8276827

In [49]:
```
p2<-ggplot(sim_data, aes(x=treatment, y=y_simulate_q2, color=treatment)) +
  geom_boxplot() +
  ylab('simulated y') +
  ggtitle('Boxplots of simulated responses') +
  theme(plot.title = element_text(hjust = 0.5))
p2
```




```
In [50]: library(gridExtra)
         grid.arrange(p1, p2, nrow = 2)
```



As we can see from the two boxplots above, it's apparent that the means of three treatments of the reduced model are more similar to each other, since:

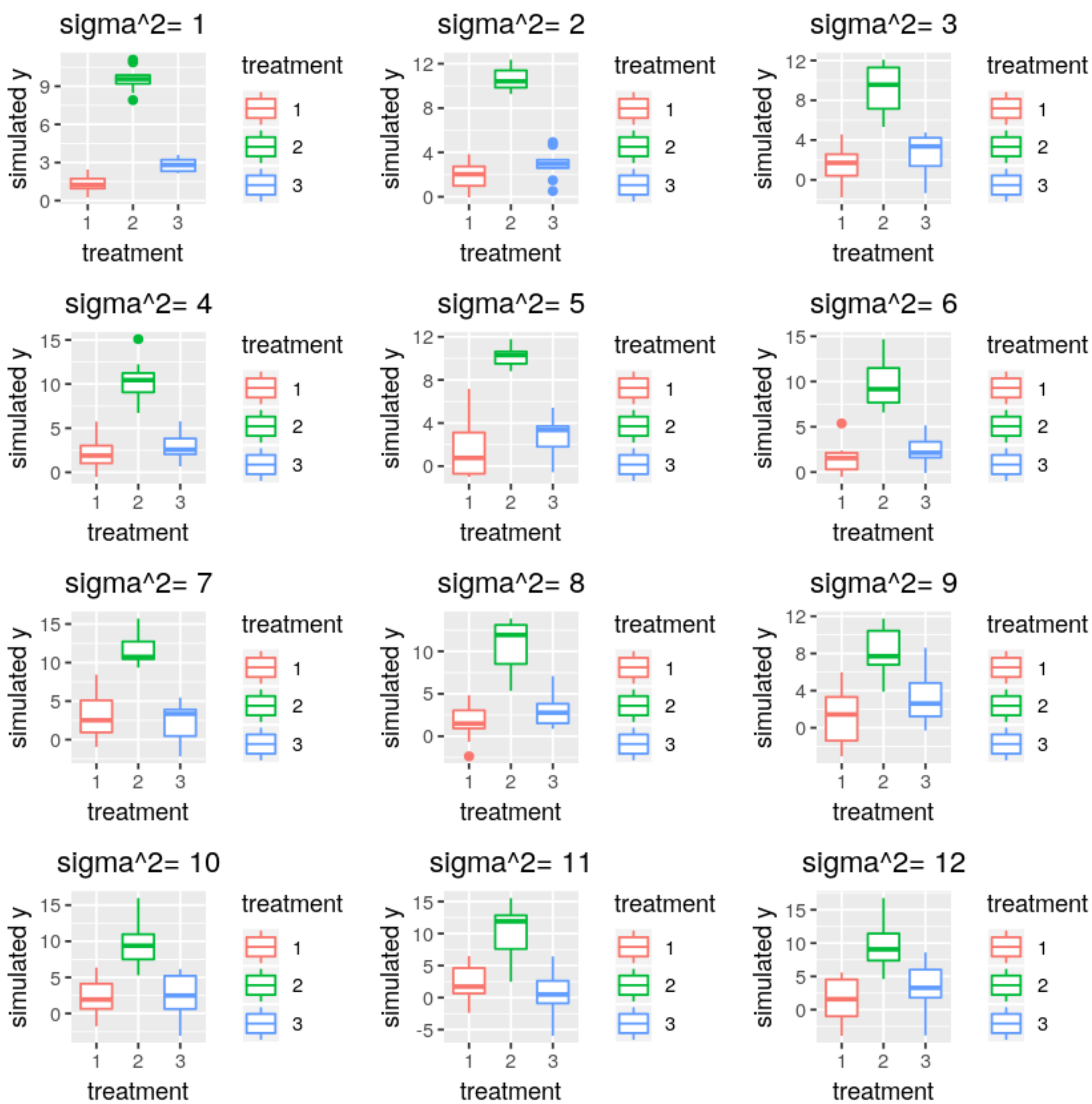
$$Y_{1.} = \mu + \tau_1 + \epsilon = \mu + \epsilon_{1.}$$
$$Y_{2.} = \mu + \tau_2 + \epsilon = \mu + \epsilon_{2.}$$
$$Y_{3.} = \mu + \tau_3 + \epsilon = \mu + \epsilon_{3.}$$

Other than the normally distributed noise, the mean μ of all three treatments are basically the same.

Q3

3.Now explore what happens to data simulated from the model in Problem 1 when the error variance increases. Try multiple values for σ^2 and find a value of σ^2 for which you cannot see any noticable difference in the boxplots of response values from the three treatments.

```
In [47]: vars = 1:12
         boxplots <- list()
         for (var in vars){
             y_simulate = means_q1+rnorm(n = length(means_q1), mean = 0, sd = sqrt(var))
             sim_data = data.frame(experiment_unit, treatment, y_simulate)
             p<-ggplot(sim_data, aes(x=treatment, y=y_simulate, color=treatment)) +
                       geom_boxplot() +
                       ylab('simulated y') +
                       ggtitle(paste('sigma^2=', var)) +
                       theme(plot.title = element_text(hjust = 0.5))
             boxplots[[var]] <- p
             }
         grid.arrange(grobs = boxplots, nrow=4)
```



from the boxplots above we could see that when $\sigma^2 \geq 10$, the boxplots of response values from the three treatments seems similar compared to each other.

Q4

Under the model in Problem 1, what is the distribution of Y_{23} , the response from the 3rd experimental unit to receive treatment 2?

$$Y_{23} = \mu + \tau_2 + \epsilon_{23} = 4.7 + 5 + \epsilon_{23} = 9.7 + \epsilon_{23}$$

$$\epsilon_{23} \sim N(0, 4)$$

$$\therefore Y_{23} \sim N(9.7, 4)$$

Q5

Under the model in Problem 1, what is the distribution of

$$\bar{Y}_2 = \frac{1}{r_2} \sum_{t=1}^{r_2} Y_{2t}$$

Since

$$Y_{2t} = \mu + \tau_2 + \epsilon_{2t} = 4.7 + 5 + \epsilon_{2t} = 9.7 + \epsilon_{2t}$$

$$\epsilon_{2t} \sim N(0, 4)$$

$$\therefore Y_{2t} \sim N(9.7, 4)$$

$$\bar{Y}_2 = \frac{1}{10} \sum_{t=1}^{10} Y_{2t} \sim N(9.7 \cdot 10 \cdot \frac{1}{10}, 4 \cdot 10 \cdot (\frac{1}{10})^2)$$

$$\therefore \bar{Y}_2 \sim N(9.7, 0.4)$$

Q6

Under the model in Problem 1, what is the distribution of the difference between an experimental unit receiving treatment 1 and an experimental unit receiving treatment 2

Since

$$Y_{1t} = \mu + \tau_1 + \epsilon_{1t} = 4.7 - 3 + \epsilon_{1t} = 1.7 + \epsilon_{1t}$$

$$\epsilon_{1t} \sim N(0, 4)$$

$$\therefore Y_{1t} \sim N(1.7, 4)$$

$$Y_{2t} = \mu + \tau_2 + \epsilon_{2t} = 4.7 + 5 + \epsilon_{2t} = 9.7 + \epsilon_{2t}$$

$$\epsilon_{2t} \sim N(0, 4)$$

$$\therefore Y_{2t} \sim N(9.7, 4)$$

Hence

$$E\left[Y_{1t} - Y_{2t}\right] = E\left[Y_{1t}\right] - E\left[Y_{2t}\right] = 1.7 - 9.7 = -8$$

$$Var\left(Y_{1t} - Y_{2t}\right) = Var\left(Y_{1t}\right) + Var\left(Y_{2t}\right) = 4 + 4 = 8$$

$$\therefore Y_{1t} - Y_{2t} \sim N(-8, 8)$$