STAT 8030 HW 5

Question 1:

1. Summary after fitting the model can be seen below:

```
> #1.1
> lm_LD <- lm(y~x1+x2+x3, data = Exercise1)
> lm_LD

Call:
lm(formula = y ~ x1 + x2 + x3, data = Exercise1)

Coefficients:
(Intercept)           x1           x2           x3
    0.26592     -0.02125      0.01430      4.17811

> summary(lm_LD)

Call:
lm(formula = y ~ x1 + x2 + x3, data = Exercise1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.100557 -0.063233  0.007131  0.045971  0.134691

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.265922   0.194585   1.367   0.1919
x1          -0.021246   0.007974  -2.664   0.0177 *
x2           0.014298   0.017217   0.830   0.4193
x3           4.178111   1.522625   2.744   0.0151 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07729 on 15 degrees of freedom
Multiple R-squared:  0.3639,    Adjusted R-squared:  0.2367
F-statistic:  2.86 on 3 and 15 DF,  p-value: 0.07197

> |
```

From the summary above we can see that the R-square value of this model with all predictors is a low 0.2367 which suggests that only 23% of the datapoints are a good fit for this model. We can also see that the f-statistic suggests that not all predictors are a good fit. Considering the p-values of the predictors, we can also say that x1 and x3 are better predictors.

2. Taking above findings into consideration, lets fit a model with x1 and x3 predictor variables and check if it would have the same degree of fit as of the original model. The summary for this reduced model is shown below:

```
> #1.2
> reduced <- lm(y~x1+x3, data = Exercise1)
> reduced

Call:
lm(formula = y ~ x1 + x3, data = Exercise1)

Coefficients:
(Intercept)           x1           x3
    0.28552     -0.02044      4.12533

> summary(reduced)

Call:
lm(formula = y ~ x1 + x3, data = Exercise1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.12333 -0.07416  0.01238  0.04884  0.12668

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.285517   0.191267   1.493   0.1550
x1          -0.020444   0.007838  -2.608   0.0190 *
x3           4.125330   1.506472   2.738   0.0146 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07654 on 16 degrees of freedom
Multiple R-squared:  0.3347,    Adjusted R-squared:  0.2515
F-statistic: 4.024 on 2 and 16 DF,  p-value: 0.0384

> |
```

When we look at the R-squared value above we observe that it is not very different with the R-squared value observed in the original model. Hence, even the reduced model has the same degree of fit. In order to further prove it, let’s do the hypothesis:

Assumption - H0: Original model and the reduced model are similar

```
> anova(reduced, lm_LD)
Analysis of Variance Table

Model 1: y ~ x1 + x3
Model 2: y ~ x1 + x2 + x3
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1     16 0.093729
2     15 0.089609  1   0.00412 0.6897 0.4193
> |
```

Looking at the p-value, which is 0.4193 after doing the anova above, we can see that it is greater than the alpha value 0.05. Hence, we fail to reject the null hypothesis. Therefore, we can conclude that with this reduced model it possible to obtain essentially the same degree of fit as that of the original model.

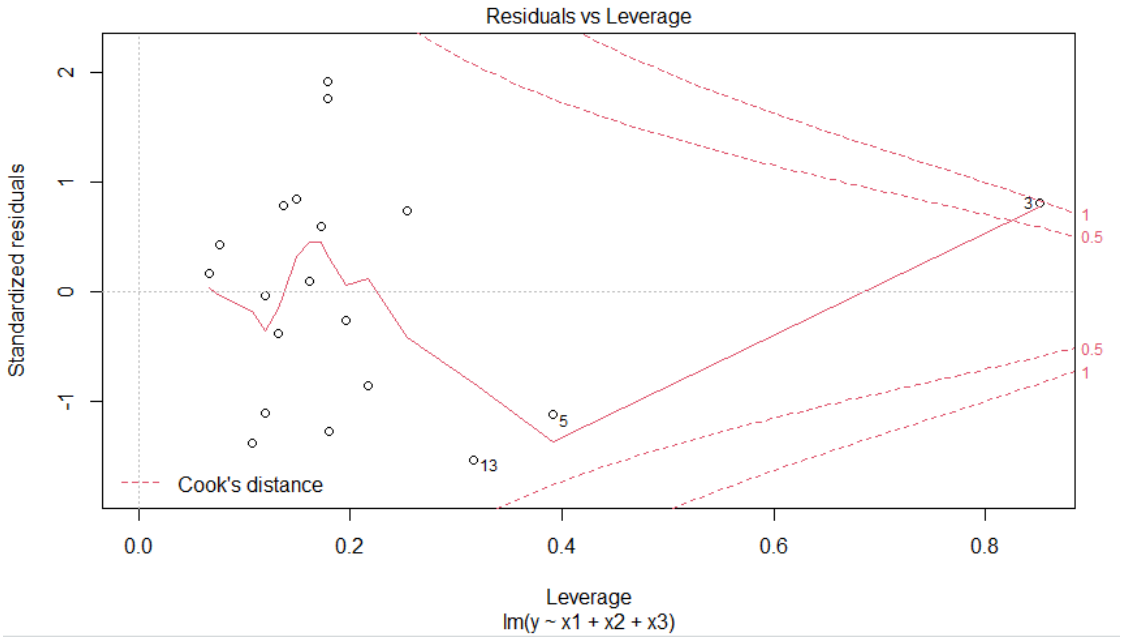
3. Leverage values for all the observations are shown below:

```
> #1.3
> L <- as.data.frame(hatvalues(lm_LD))
> L
   hatvalues(lm_LD)
1        0.17798270
2        0.17934099
3        0.85091457
4        0.10761585
5        0.39153825
6        0.16115958
7        0.13688107
8        0.25367448
9        0.06701578
10       0.11968672
11       0.11950583
12       0.17239599
13       0.31618336
14       0.13140699
15       0.07617481
16       0.21661460
17       0.19522441
18       0.14872221
19       0.17796183
> L[order(-L['hatvalues(lm_LD)']), ]
 [1] 0.85091457 0.39153825 0.31618336 0.25367448 0.21661460 0.19522441 0.17934099 0.17798270 0.17796183 0.17239599
[11] 0.16115958 0.14872221 0.13688107 0.13140699 0.11968672 0.11950583 0.10761585 0.07617481 0.06701578
```

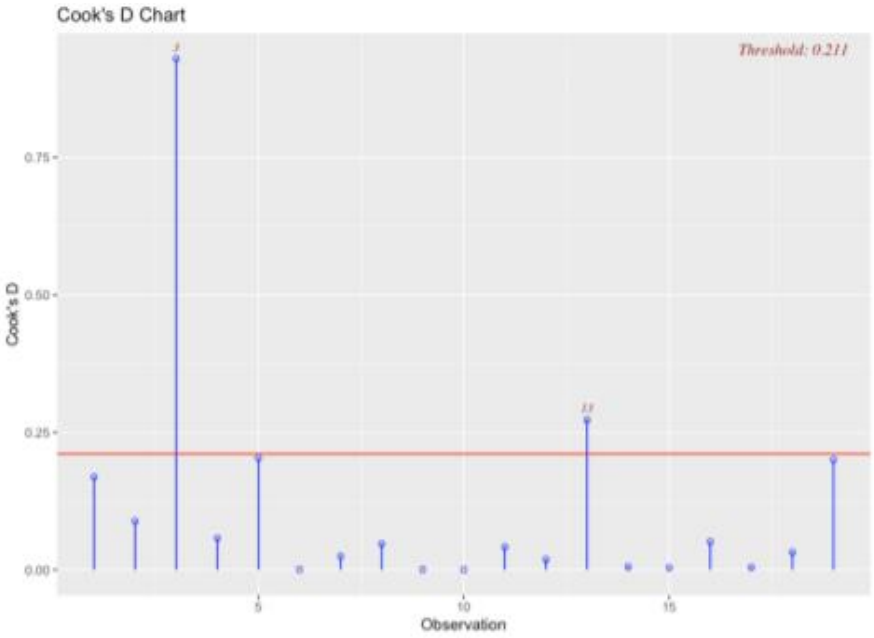
Cooks distance for all observations are shown below:

```
> #Cooks distance
> CD <- cooks.distance(lm_LD)
> CD
           1            2            3            4            5            6            7            8            9
1.688268e-01 8.854024e-02 9.296160e-01 5.718456e-02 2.029162e-01 4.874208e-04 2.461564e-02 4.685795e-02 4.883028e-04
          10           11           12           13           14           15           16           17           18
5.229549e-05 4.143644e-02 1.889847e-02 2.726019e-01 5.370022e-03 3.733265e-03 5.099189e-02 4.249284e-03 3.162543e-02
          19
1.999403e-01
>
```

Plot for Leverage values vs Residuals is shown below:



Plot for Cooks distance vs Residuals is shown below:



Looking at the leverage values we see that the highest leverage value is 0.8509 and the threshold being 4/n which is 0.21. We also notice from the last plot that the datapoints 3 and 13 have exceeded the threshold. Thus, these points might be the influential datapoints.

Question 2:

1. We will be using One factor random effect model. Let's assume Xij to be the percentage of ingredients. i being the batch number and j being the container number.
   So, Xij = μ+Aj+ eij

   Here, i =1,2..10 batches and j=1,2..5 containers and μ is the mean percentage and Aj is the random effect of the ith randomly selected batches and eij are the normal random variables with variance sigma square.

2. Shown below is the summary for the random effects model fit using restricted maximum likelihood:

```
> P <- read.csv(file = "C:/Users/sshri/Desktop/Fall 2021/Stat 8030/paint_batches.csv",header=TRUE)
> library(lme4)
> PR <- lmer(Percentage~(1|Batch),data=P)
> summary(PR)
Linear mixed model fit by REML ['lmerMod']
Formula: Percentage ~ (1 | Batch)
   Data: P

REML criterion at convergence: 219.4

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.81576 -0.66464 -0.05532  0.62646  1.97050

Random effects:
 Groups   Name        Variance Std.Dev.
 Batch    (Intercept) 0.2359   0.4857
 Residual             4.5585   2.1351
Number of obs: 50, groups:  Batch, 10

Fixed effects:
            Estimate Std. Error t value
(Intercept)   5.2420     0.3388   15.47
> |
```

Estimating the variance as shown below:

```
> #Estimating variance
> S <- anova(P.lm)$`Mean Sq`[2]
> SB <- ( anova(P.lm)$`Mean Sq`[1] - anova(P.lm)$`Mean Sq`[2] )/a
> SB
[1] 0.2359322
> |
```

From the summary above, we can say that the estimate variance of random effect is 0.2359

3. Hypothesis testing:
   Null hypothesis – Ho: variance = 0
   Alternate hypothesis – Ha: variance > 0

```
> P$Batch <- as.factor(P$Batch)
> P.lm <-lm(Percentage~Batch, data = P)
>
> anova(P.lm)
Analysis of Variance Table

Response: Percentage
          Df  Sum Sq Mean Sq F value Pr(>F)
Batch      9  51.643  5.7381  1.2588 0.2889
Residuals 40 182.338  4.5584
> |
```

We know that f-ratio = mean square between/mean square error
Which means f-ratio = 5.7381/4.5584 which is = 1.2588

Now, p-value from the above image is = 0.2889

We see that the p-value is greater than the alpha value 0.05. Hence, we fail to reject the null hypothesis. Hence, there is no evidence that there is a significant batch effect.

Question 3:

1. This Study consists of 2 factors which are locations and chemicals. From the mentioned point about the type of environmental conditions in which the chemical is placed might affect the effectiveness of the treatment to kill fire ants. Looking at this, the factor more appropriate to be included as random effect is location. Also, because the researcher is randomly selecting 5 locations from a large set of locations.

2. Mixed effect model is shown below:

   Zijk = μ + ti + Aj + eijk

   Here, i = 1,2,3,4 and j = 1,2,3,4,5 locations, the ti is the fixed effects referring to the chemicals. μ is the overall mean and Aj is the random effect and eijk are the normal random variables with the variance of sigma square and mean 0.

3. Hypothesis testing to show if there is a significant effect of the factor chemical is shown below:

   Null Hypothesis: Ho: t1= t2= t3= t4 = 0
   Alternate Hypothesis: Ha: At least one of ti is non-zero

```
> #Question 3
>
> FA <- read.csv(file = "C:/Users/sshri/Desktop/Fall 2021/Stat 8030/fire_ants.csv",header=TRUE)
> FA$Locations <- as.factor(FA$Locations)
> FA$Chemicals <- as.factor(FA$Chemicals)
>
> FAC <- lmer(NumberKilled~Chemicals+(1|Locations),data=FA,REML=FALSE)
> FAR <- lmer(NumberKilled~(1|Locations),data=FA,REML=FALSE)
boundary (singular) fit: see ?isSingular
> anova(FAR, FAC)
Data: FA
Models:
FAR: NumberKilled ~ (1 | Locations)
FAC: NumberKilled ~ Chemicals + (1 | Locations)
    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
FAR    3 185.27 190.34 -89.637   179.27
FAC    6 109.59 119.72 -48.795    97.59 81.684  3  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

   From above, we can see that the test statistic value i.e. Chisq value is 81.684 and the p-value is <2.2e-16 which is less than the alpha value 0.05. Hence, we reject the null hypothesis. Thus, there is evidence that there is significant effect of the factor chemicals.

4. Hypothesis testing to check if there is a significant effect of the factor location is shown below:

   Null Hypothesis: Ho: variance =0
   Alternate Hypothesis: Ha: variance > 0

```
> #3.4
> FAL <- lmer(NumberKilled~Chemicals+(1|Locations),data=FA,REML=FALSE)
> FALR <- lm(NumberKilled~Chemicals, data=FA)
> anova(FAL, FALR)
Data: FA
Models:
FALR: NumberKilled ~ Chemicals
FAL: NumberKilled ~ Chemicals + (1 | Locations)
     npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
FALR    5 107.64 116.08 -48.819   97.638
FAL     6 109.59 119.72 -48.795   97.590 0.0476  1     0.8274
> |
```

From above we can see that the test statistic value i.e. Chisq value is 0.0476 and the p-value is 0.8274 which is greater than the alpha value 0.05. Hence, we fail to reject the null hypothesis. Thus, there is no evidence that proves that there is significant effect of the factor locations.