Part 1 – Question sheet

Problem 1 (20pts) The reproducibility of a method for the determination of a pollutant in water was investigated by taking twelve samples from a single batch of water and determining the concentration of pollutant in each. The following results were obtained:

```
x=c(5.98, 8.80, 6.89, 8.49, 8.48, 7.47, 7.97, 5.94, 7.32, 6.64, 6.94, 3.51)
```

It is expected that from this sample a 95% confidence interval for the concentration of pollutant will be obtained.

(i) There is a concern that the data may contain an outlier. Thus the following procedure has been performed on the data:

```
grubbs.test(x, two.sided=T)
        Grubbs test for one outlier
data:  x
G = 2.4180, U = 0.4202, p-value = 0.04811
alternative hypothesis: lowest value 3.51 is an outlier
```

(2pts) Describe what is the purpose of this procedure.

(2pts) Write the conclusion that follows from it.

(2pts) What should be done to the data in a consequence?

(ii) After the test for outlier, another preliminary procedure has been performed in the following

```
n=length(x)
y=x[1:(n-1)]
y
[1]  5.98 8.80 6.89 8.49 8.48 7.47 7.97 5.94 7.32 6.64 6.94

ks.test(y,pnorm,mean(y),sd(y))
One-sample Kolmogorov-Smirnov test
data:  y
D = 0.1442, p-value = 0.9518
```

(2pts) Explain what is the purpose of this procedure.

(2pts) What is the conclusion and why?

(iii) After these initial verifications, the confidence interval can be obtained from the following computations

```
mean(y)
[1] 7.356364
sd(y)
[1] 0.991315
qt(0.975,10)
[1] 2.228139
```

(4pts) Based on the obtained values write down the confidence interval for the pollutant;

(4pts) If the initial analysis was not performed the following values would be used for the computations

```
mean(x)
[1] 7.035833
sd(x)
[1]  1.458165
qt(0.975,11)
[1] 2.200985
```

Evaluate the confidence interval based on these values;

(2pts) Explain which of these two intervals should be recommended and point reasons for the choice.

Part 1 – Answer sheet

Problem 1 (20pts)

(2pts) Describe what is the purpose of this procedure

This is a statistical test called Grubbs test that allows for identification of a value in the data that appears unusually large or unusually small. Such a data point is called an outlier.

(2pts) Write a conclusion that follows from it.

Since the p-value reported by the program is smaller than the standard 5% significance level, we conclude that there an outlier is detected in the data. The outlier indicated by the program is the lowest value in the data which is 3.51.

(2pts) What should be done to the data in a consequence?

Since the data point 3.51 is an outlier, i.e. a value which can not be explained by the normal variability in the sample, it is recommended to remove this point from the data and perform further analysis on the remaining datapoints.

(2pts) Explain what is the purpose of this procedure.

The performed analysis is called Kolmogorov-Smirnov test. The purpose of it is to detect if the distribution of the data deviates significantly from the normal distribution.

(2pts) What is the conclusion and why?

We observe that the p-value reported by the test is more than 95% and thus it gives us no reason for claiming any deviation from normality.

(4pts) Based on the obtained values write down the confidence interval for the pollutant;

The confidence interval is given by

$$[\bar{y} - ts/\sqrt{n}, \bar{y} + ts/\sqrt{n}] \approx [7.36 - 2.23 * 0.99/\sqrt{11}, 7.36 + 2.23 * 0.99/\sqrt{11}] \approx [6.69, 8.03]$$

(4pts) Evaluate the confidence interval based on all initial values;

The confidence interval is given by

$$[\bar{x} - ts/\sqrt{n}, \bar{x} + ts/\sqrt{n}] \approx [7.04 - 2.23 * 1.46/\sqrt{12}, 7.04 + 2.23 * 1.46/\sqrt{12}] \approx [6.10, 7.98]$$

(2pts) Explain which of these two intervals should be recommended and point reasons for the choice.

It is recommended to use the first interval. First, it is based on the data with the unusual small value removed so that its value is not influencing computed interval limits. We also see the first computed interval is shorter than the second one, which means that it is more accurate with determining a possible location of the true pollutant concentration.

Problem 2 (20pts) A company developed a new flame atomic-absorption spectroscopic method of determining antimony in the atmosphere. It has commissioned Laboratory A and B to compare it with the recommended calorimeter method by giving each of them funding for carrying out twenty measurements.

Laboratory A has taken the following approach. They collected a sample of size twenty from an urban atmosphere and randomly assignment one of the two measurements methods to each specimen so that the number of measurements for each of the methods was the same and so equal to ten. The following results were obtained:

```
NewMethod 14.28 22.46 18.40 16.44 13.22 17.24 18.08 18.54 18.90 15.60
RecMethod 20.50 17.54 19.02 14.40 19.36 12.04 18.66 17.84 17.80 18.98
```

Laboratory B has taken a different approach. They collected a sample of size ten from an urban atmosphere and the measurements have been made twice on each specimen, one by the new method and one by the recommendent method. The following results were obtained:

```
NewMethod  20.58 19.70 18.78 16.68 19.66 14.88 18.26 19.94 16.56 16.92
RecMethod  17.90 18.66 16.66 13.32 18.16 14.30 17.08 14.64 17.48 18.46
```

The company recieved the following two reports from the laboratories.

```
***Report I:
**Part 1: Grubbs test for one outlier
data:  NewMethod
G = 1.9608, U = 0.5253, p-value = 0.2756
alternative hypothesis: highest value 22.46
is an outlier
data:  RecMethod
G = 2.1994, U = 0.4028, p-value = 0.08771
alternative hypothesis: lowest value 12.04
is an outlier
**Part 2: Normality assumption
One-sample Kolmogorov-Smirnov test
data:  NewMethod
D = 0.173, p-value = 0.8778
alternative hypothesis: two-sided
data:  RecMethod
D = 0.2884, p-value = 0.313
alternative hypothesis: two-sided
**Part 3: Equality of variances
F test to compare two variances
data:  NewMethod and RecMethod
F = 1.0715, num df = 9, denom df = 9, p-value = 0.9197
alternative hypothesis: true ratio of variances
is not equal to 1
**Part 4: Testing means
Two Sample t-test
data: NewMethod and RecMethod
t = -0.2584, df = 18, p-value = 0.799
alternative hypothesis: true difference in means
is not equal to 0
```

```
***Report II:
Diff=NewMethod-RecMethod
**Part 1: Grubbs test for one outlier
data:  Diff
G = 1.8869, U = 0.5604, p-value = 0.3666
alternative hypothesis: highest value 5.3
is an outlier
**Part 2: Normality assumption
One-sample Kolmogorov-Smirnov test
data:  Diff
D = 0.1172, p-value = 0.9961
alternative hypothesis: two-sided
**Part 3: Testing for difference
One Sample t-test
data:  Diff
t = 2.4216, df = 9, p-value = 0.03851
alternative hypothesis: true mean is not
equal to 0
```

Please, answer the questions in the following page.

Part 2 – Answer sheet

Problem 1 (20pts)

(4pts) Due to some misplacement of the paperwork, it is not known which record is from which Laboratory. Can you identify reports?

(1pts) Report I is received from Laboratory: **[✓ A]/B** and Report II from Laboratory: **A/[✓ B]** (circle your answer)

(3pts) Give reasons for your choice:

The approach taken by Laboratory B ties the measurements obtained by one method with the ones obtained by the other one. Such paired results of measurements require the paired sample test that is performed on the differences between measurements obtained on the same specimen. It is clear that Report II is presenting such analysis as we see that all analysis is performed on `Diff` *which is the difference of the original data.*

(6pts) Write the conclusions that follows from each report.

(3pts) Report 1:*First, two initial analyses were performed on measurements to check if there are outliers in the data and if the data follow normal distribution. The test for outliers (Grubbs test) returned p-values of approximately 0.28 and 0.09 for new and recommended methods, respectively. Thus no significant outlier has been detected. The Kolmogorov-Smirnov test for normality of the data produced 0.88 and 0.31, respectively. We conclude no deviation from normality assumption. The next test is checking if variability in the two data sets is the same. This is the test for equality of variance. The reported* p-value *is very high (92%) and conclusion is that the variability is the same. Finally, the two-sample test for the means is performed yielding the* p-value *of* 79%. *We conclude that no significant difference between the two methods has been detected.*

(3pts) Report 2:*First, the differences of the data produced by the two methods have been computed. The test for outliers (Grubbs test) on differences returned* p-value *of approximately 0.37 – no significant outlier has been detected. The Kolmogorov-Smirnov test for normality of the data produced* p-value *of nearly one – no deviation from the normality assumption. Finally, the test for the mean of differences being equal to zero is performed yielding the* p-value *of approximately* 3.9%. *We conclude that there is a significant difference between the two methods.*

(5pts) Which of the approaches you like better and explain why?

The approach taken by Laboratory B is considered more favorable since by taking the difference of measurements by the two methods on a single specimen the variability of specimen is eliminated. Smaller variability often translates to more accurate statistical analysis. It also made the analysis simpler.

(3pts) Which of the approaches produced more valuable information? Explain why.

The report by Laboratory A contains no rejections of the null hypotheses which is thus non-conclusive – no significant results can be reported. On the other hand, the report by Laboratory B contains a conclusive statement on 5% significance level that the methods of measurements differ which is valuable information for the company.

(2pts) Do you think that the problem with one of the analyses is due to the errors in analyzing data **Yes/[✓ No]** (circle your answer), do you think that the problem was with the design of collecting data **[✓ Yes]/No** (circle your answer).

Part 3 – Question sheet

Problem 1 (20pts)

The following results were obtained when each of a series of standard silver solutions was analysed by flame atomic-absorption spectrometry. The analysis of these by means of R is also presented below.

Concentration, ng/ml	Absorbance
10	0.251
15	0.390
20	0.498
25	0.625
30	0.763
0	0.003
5	0.127

```
Conc=c(10,15,20,25,30,0,5)
Abs=c(0.251,0.390,0.498,0.625,0.763,0.003,0.127)
Call:
lm(formula = Abs ~ Conc)
Residuals:
         1          2          3          4          5          6          7
-0.0027500  0.0104286 -0.0073929 -0.0062143  0.0059643  0.0008929 -0.0009286

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0021071  0.0047874    0.44    0.678
Conc        0.0251643  0.0002656   94.76 2.48e-09 ***
---
Residual standard error: 0.007026 on 5 degrees of freedom
Multiple R-squared: 0.9994,Adjusted R-squared: 0.9993
F-statistic:  8980 on 1 and 5 DF,  p-value: 2.481e-09
```

Based on this information do the following

(3pts) Determine the slope and intercept of the calibration plot and make a sketch of the linear fit to the date. Include data points on the graph as well.

(2pts) Determine the 95% confidence limits for the slope and intercept. (The corresponding quantiles of Student t-distribution are given by $qt(0.025, 5) = -2.57$ and $qt(0.975, 5) = 2.57$.)

(4pts) Based on the above calibration fit find the silver concentration for a sample giving an absorbance of 0.42 in a single determination. Estimate the 95% confidence limits for the silver concentration. The following formula for the standard deviation of the concentration determination should be used for the purpose

$$s_{x_0} = \frac{s_{y/x}}{b}\sqrt{\frac{1}{m}+\frac{1}{n}+\frac{(y_0-\bar{y})^2}{b^2\sum_{i=1}^{n}(x_i-\bar{x})^2}}.$$

The following computations in R allows for effective computation of the above standard deviation

```
> mean(Abs)
[1] 0.3795714
> mean(Conc)
[1] 15
> sum((Conc-mean(Conc))^2)
[1] 700
```

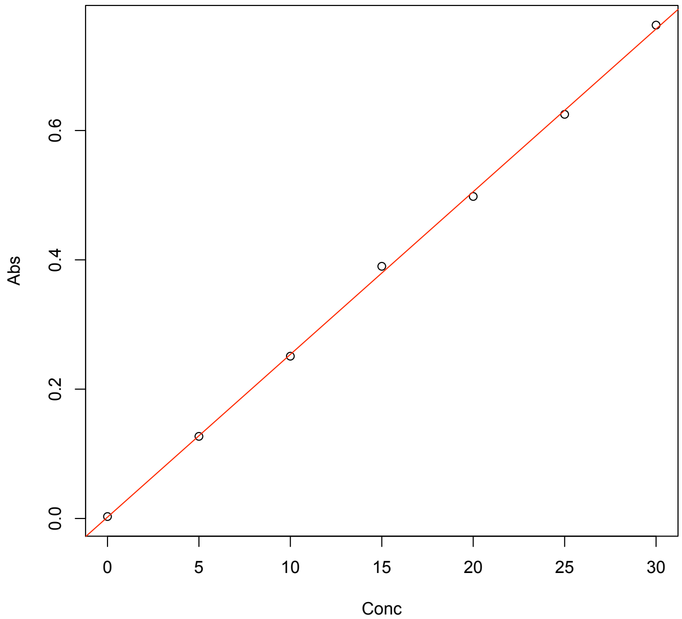
(6pts) Find the silver concentration for a sample giving absorbance values of 0.30, 0.31, 0.29 in three separate analyses of the same sample. Estimate the confidence limits for the concentration in this case. How the accuracy improved comparing to the one in the previous question? (In your computations you should use the information from the previous question.)

(5pts) Estimate the limit of detection of the silver analysis.

Part 3 – Answer sheet

Problem 1 (20pts)

(3pts) The slope of the fit is $\boxed{\text{b=0.025}}$ and the intercept is $\boxed{\text{a=0.0021}}$. The following is the graph of data and the fit:



(2pts) The 95% confidence limits for the slope are $\boxed{\text{[0.0243,0.0257]}}$ and the corresponding ones for the intercept $\boxed{\text{[-0.0102,0.0144]}}$.

(4pts) The predicted silver concentration for a sample giving an absorbance of 0.42 is $\boxed{16.61}$.

The 95% confidence limits for the so-determined silver concentration is $\boxed{\text{[15.84,17.38]}}$.

(6pts) The predicted silver concentration for a sample giving absorbance values of 0.30, 0.31, 0.29 is $\boxed{11.83}$ and the 95% confidence limits for this prediction are $\boxed{\text{[11.34,12.34]}}$.

By examining the obtained confidence intervals in this and the previous question we see that $\boxed{\text{by taking three observations the length of confidence interval reduced from about 1.5 to 1 units.}}$

(5pts) The response to the blank signal which is equal to $\boxed{y_B = 0.0021}$.

The standard error of the response is given by $\boxed{s_B = s_{y/x} = 0.007026}$.

The limits of detection in the terms of absorption is $\boxed{y_B + 3s_B = 0.0021 + 3 * 0.007026 \approx 0.0232}$

The smallest concentration that is reliably detected is $\boxed{x_0 = (0.0232 - a)/b = 0.83820}$.

Part 4 – Question sheet

Problem 1 (10pts)

The fluorescence of each of a series of acidic solutions of quinine with concentrations 0,10,20,30,40,50 was determined five times. The mean values and standard deviations of these determinations have been obtained as follows:

```
Means:  4.0  21.2  44.6  61.8  78.0 105.2
StDev:  0.71 0.84 0.89 1.64 2.24 3.03
```

The following two analyses have been performed on the data

```
lm(formula = Means ~ Conc)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9238     2.1648   1.351    0.248
Conc          1.9817     0.0715  27.715 1.01e-05 ***
---
Residual standard error: 2.991 on 4 degrees of freedom
Multiple R-squared: 0.9948,Adjusted R-squared: 0.9935
F-statistic: 768.1 on 1 and 4 DF,  p-value: 1.008e-05
```

```
weights=SdInt^(-2)/mean(SdInt^(-2))
lm(formula = Means ~ Conc, weights = weights)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.48066    1.15736   3.007   0.0397 *
Conc         1.96315    0.06765  29.018  8.4e-06 ***
---
Residual standard error: 2.034 on 4 degrees of freedom
Multiple R-squared: 0.9953,Adjusted R-squared: 0.9941
F-statistic:   842 on 1 and 4 DF,  p-value: 8.396e-06
```

Answer the following questions.

(3pts) What kind of analyses have been performed above? Write down the fits found by each of the two analyses.

(4pts) Describe differences between the two methods. When one is preferable over the other?

(3pts) Find the concentrations that follow from each of these two fits for the observed intensity of 45.

Problem 2 (10pts)

In an experiment to determine hydrolysable tannins in plants by absorption spectroscopy the following results were obtained:

```
Absorbance (Abs)            0.084 0.183 0.326 0.464 0.643
Concentration (Conc), mg/ml 0.123 0.288 0.562 0.921 1.420
```

The following two analyses have been performed on the data

```
Conc2=Conc^2

lm(formula = Abs ~ Conc + Conc2)
Residuals:
        1         2         3         4         5
-0.004572  0.003127  0.008089 -0.009119  0.002474

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01651    0.01186   1.392  0.29841
Conc         0.59973    0.03953  15.172  0.00432 **
Conc2       -0.11288    0.02483  -4.546  0.04514 *
---
Residual standard error: 0.009628 on 2 degrees of freedom
Multiple R-squared: 0.9991,Adjusted R-squared: 0.9981
F-statistic:  1065 on 2 and 2 DF,  p-value: 0.0009384
```

```
lm(formula = Abs ~ Conc)

Residuals:
        1         2         3         4         5
-0.026572  0.002299  0.028842  0.014259 -0.018828

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05829    0.02060    2.83 0.066209 .
Conc         0.42502    0.02544   16.71 0.000467 ***
---
Residual standard error: 0.02646 on 3 degrees of freedom
Multiple R-squared: 0.9894,Adjusted R-squared: 0.9858
F-statistic: 279.1 on 1 and 3 DF,  p-value: 0.0004669
```

Answer the following questions.

(3pts) What kind of analyses have been performed above? Write down the fits found by each of them.

(4pts) Describe differences between the two fits. Examine the residuals and R^2 coefficients. Explain their role in assessing quality of the fits. Do you see any evidence that one fit is preferable over the other? Explain why.

(3pts) Find the concentrations that follows from each of these two fits for the observed intensity of 0.5.

Part 4 – Answer sheet

Problem 1

(3pts) The analysis presented on the left hand side corresponds to: $\boxed{\text{unweighted linear regression}}$ and the resulting fit to the data is $\boxed{y = 2.92 + 1.98x}$.

The analysis presented on the right hand side corresponds to: $\boxed{\text{weighted linear regression}}$ and the resulting fit to the data is $\boxed{\text{y=3.48+1.96x}}$.

(4pts) The two methods differ in the following aspects:

The unweighted linear regression does not require values of standard deviations of the response variables and is accurate only if the variability of the response is more or less constant across the measured range of values (homoscedasticity). The weighted linear regression allows to account for different variability around the straigth line at different locations (heteroscedasticity) but requires additional information on standard deviations of the response at each point in the data. These standard deviations are used to compute wehights.

The analysis presented on the right hand side is preferable over the other one if

the standard deviation values differ for various points in the data.

(3pts) The concentration corresponding to the intensity 45.0 is equal to:

for the left hand side fit: $\boxed{21.25}$

for the right hand side fit: $\boxed{21.18}$

Problem 2

(3pts) The analysis presented on the left hand side corresponds to: $\boxed{\text{quadratic regression}}$ and the resulting fit to the data is $\boxed{y = 0.017 + 0.60x - 0.113x^2}$.

The analysis presented on the right hand side corresponds to: $\boxed{\text{linear regression}}$ and the resulting fit to the data is $\boxed{y = 0.058 + 0.425x}$.

(4pts) The two methods differ in the following aspects:

The quadratic regression allows fitting the data that are not necessarily in linear relation allowing for a curvature of the fit according to the quadratic relation. The linear regression can not model properly curvatures that maybe present in the dependence between variables. The linear regression requires one less parameter to be fit to the data.

After examining the residuals for each fit we may say that:

The residuals in linear fit appears to follow quadratic curvature by being negative at the smallest and largest values being relatively large in absolute values. This is not observed in residuals to the quadratic fit. This suggests that the quadratic fit maybe better.

After examining the R^2 and adjusted R^2 for each fit we may say that:

We observe a significant improvement in the values of both R^2 and adjusted R^2 in the quadratic fit as compared with the ones for the linear fit again suggesting that the quadratic fit is a better one.

(3pts) The concentration corresponding to the intensity 0.5 is:
for the left hand side fit: $\boxed{0.989}$

for the right hand side fit: $\boxed{1.04}$

Part 5 – Question sheet

Problem 1 (20pts)

Four standard solutions were prepared, each containing 16.00% (by weight) of chloride. Three titration methods, each with a different technique of end-point determination, were used to analyse each standard solution. The order of the experiments was randomized. The results for the chloride found (% w/w) are shown below:

	Method		
Solution	A	B	C
1	16.03	16.13	16.09
2	16.05	16.13	16.15
3	16.02	15.94	16.12
4	16.12	15.97	16.10

(3pts) Identify two factors and their levels. Are they controllable? Which of them is of primary interest and which is a nuisance factor?

(4pts) Explain why it is not possible to analyze interactions between factors in this problem. What would have to be done in order to study the interactions? Do you expect interactions to play a role in this example?

(4pts) The following code has been run in *R* and the resulting numerical values are presented as well:

```
> Perc=c(16.03,16.13,16.09,16.05,16.13,16.15,16.02,15.94,16.12,16.12,15.97,16.10)
> Sol=c(rep("1",3),rep("2",3),rep("3",3),rep("4",3))
> Meth=c(rep(c("A","B","C"),4))
> PercM=matrix(Perc,ncol=3,byrow=T)
> r=4
> c=3
> sum((Perc-mean(Perc))^2)
[1] 0.05129167
> MTr=apply(PercM,2,mean)
> r*sum((MTr-mean(Perc))^2)
[1] 0.01201667
> MBl=apply(PercM,1,mean)
> c*sum((MBl-mean(Perc))^2)
[1] 0.01109167
```

From these computations identify all sums of squares and report the degrees of freedom associated to each of them.

(3pts) Using the above computations test whether there are significant differences between the results obtained by different methods. The following value can be used for this purpose: `qf(0.95,2,6)=5.143`

(3pts) Test whether there are significant differences between the concentration of chloride in different solutions. The following value can be used for this purpose: `qf(0.95,3,6)=4.757`

(3pts) The following output table is a result of performing ANOVA on the data using *R*-function `lm`. Some values in this printout have been removed. Using the computations you have for the previous questions, fill out the missing values in the printout

```
Results=lm(Perc~Sol+Meth)
anova(Results)

Analysis of Variance Table

Response: Perc
          Df    Sum Sq   Mean Sq F value Pr(>F)
Sol          0.0110917            0.7871  0.5435
Meth       2 0.0120167 0.0060083          0.3446
Residuals  6           0.0046972
```

Part 5 – Answer sheet

Problem 1

(3pts) Factor 1: titration methods at levels A, B, C. Controllable? **[✓ Yes]**/**No** (circle your answer)

Factor 2: solutions at levels 1, 2, 3, 4. Contrallable? **Yes**/**[✓ No]** (circle your answer)

Of primary interest is Factor: 1 – titration method

and a nuisance factor is Factor: 2 – solutions.

(4pts) Explain why it is not possible to analyze interactions between factors in this problem.

In this example we have $c = 3$ and $r = 4$, so the number degrees of freedom for studying variability in the data is $cr - 1 = 11$. For analysis of variability between levels of Factor 1 $c-1 = 2$ degrees of freedom have to be spent, while for variability between levels of Factor 2 $r - 1 = 3$ degrees of freedom are needed. For analysis of variability of interactions there is required $cr - c - r + 1 = 12 - 3 - 4 + 1 = 6$ degrees of freedom. Thus the total number of degrees that is needed adds to 11 which leaves no degrees of freedom for residuals which are used as the denominator in the test statistics. For this reason there is not enough data (degrees of freedom) to test for interactions.

What would have to be done in order to study the interactions?

Replicates of measurements for each combination of levels of the two factors are needed for the analysis of interactions.

Do you expect interactions to play a role in this example?

Since the solutions are levels of an uncontrollable nuisance factor, the interactions are not of principal interest as they are hard to interpret. However adding them can disclose some otherwise hidden effects.

(4pts) The sums of squares and the corresponding degrees of freedom are:

$SS_{Tot} \approx 0.0513$, $SS_{Tr} \approx 0.0120$, $SS_{Bl} \approx 0.0111$, $SS_{Res} \approx 0.0282$. The corresponding degrees of freedom are $cr - 1 = 11$, $c - 1 = 2$, $r - 1 = 3$, and $11 - 2 - 3 = 6$, respectively.

(3pts) By computing the following test statistic $F_1 = 6 * SS_{Tr}/(2 * SS_{Res}) \approx 3 * 0.0120/0.0282 \approx 1.28$

and comparing it with the quantile 5.143 of F distribution with 2, and 6 degrees of freedom,

we conclude that there are no significant differences between titration methods.

(3pts) By computing the following test statistic $F_2 = 6 * SS_{Bl}/(3 * SS_{Res}) \approx 2 * 0.0111/0.0282 \approx 0.787$

and comparing it with the quantile 4.757 of F -distribution with 3, and 6 degrees of freedom, we conclude that

there are no significant differences between the concentration of chloride in different solutions.

(3pts) The following output table is a result of performing ANOVA on the data using R-function `lm`. Some values in this printout have been removed. Using the computations you have for the previous questions, fill out the missing values in the printout

```
Analysis of Variance Table

Response: Perc
          Df    Sum Sq   Mean Sq F value Pr(>F)
Sol        3 0.0110917 0.0036972  0.7871 0.5435
Meth       2 0.0120167 0.0060083  1.2791 0.3446
Residuals  6 0.0281833 0.0046972
```