Question 32 - Method Comparison

An ion-selective electrode (ISE) determination of sulphide from sulphate-reducing bacteria was compared with a gravimetric determination. Each pair of determinations were taken from the same sample.

The results obtained by both methods are expressed in milligrams of sulphide, and are tabulated below.

ISE method	108	12	152	3	106	11	128	12	160	128
gravimetry	105	16	113	1	108	11	141	161	182	118

Two simple linear models are fitted to the data. Model C uses the gravimetric determination as an independent variable used to predict the ISE determination. Conversely, Model D uses the ISE determination as an independent variable used to predict the gravimetric determination. The relevant R output is presented on the following page.

- **Model C**

```
Call:
lm(formula = ISE ~ grav)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.1125    28.8487   0.524    0.615
grav          0.6997     0.2543   2.751    0.025 *
....
```

- **Model D**

```
Call:
lm(formula = grav ~ ISE)
..
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.6215    25.8542   1.494    0.174
ISE           0.6949     0.2526   2.751    0.025 *
....
```

i. (3 marks) Is a simple linear regression model an suitable approach for this type of analysis? Explain why or why not? What alternative type of regression analysis might you recommend?

ii. (2 marks) Provide a brief description of the Bland-Altman plot. Discuss any shortcomings with this approach to method comparison.

Question 33 - Inference Procedures

- The nicotine content in blood can be determined by gas chromatography down to concentrations of 1 ng/ml. The concentration of nicotine was determined in each of two samples of known concentrations 10 ng/ml and 50 ng/ml.

```
Data: Sample (Lo): m = 10 ng/ml, n=14.

8.40, 9.59, 9.38, 9.10, 10.78, 11.41, 9.94,
10.08, 12.11, 9.10, 9.59, 10.36, 10.41, 10.52.

Data: Sample (Hi): m = 50 ng/ml, n=10.

47.5, 48.4, 48.8, 48.4, 46.8,
46.2, 48.6, 50.6, 45.5, 46.1.
```

A research team evaluated both samples to determine whether or not the samples were similar in terms of measures of centrality and dispersion, before the trial commenced.

The following blocks of `R` code (i.e blocks 1 to 6) are based on the data for this assessment.

(a) (10 Marks) Each of the six blocks of code describes a statistical inference procedure. Provide a brief description for each procedure.

(b) (10 Marks) Write a short report on your conclusion for this assessment, clearly indicating which blocks of `R` code you felt were most relevant, and explain why.

Block 1

```
F test to compare two variances

data:  Lo and Hi
F = 0.3945, num df = 13, denom df = 9, p-value = 0.1246
alternative hypothesis:
true ratio of variances is not equal to 1
95 percent confidence interval:
0.1029905 1.3066461
sample estimates:
ratio of variances
0.3945149
```

Block 2

```
> shapiro.test(Lo)

Shapiro-Wilk normality test

data:  Lo
W = 0.9779, p-value = 0.9609
> shapiro.test(Hi)

Shapiro-Wilk normality test

data:  Hi
W = 0.9496, p-value = 0.6634
```

Block 3

```
> t.test(Lo,Hi)

Welch Two Sample t-test

data:  Lo and Hi
t = -67.374, df = 14.016, p-value < 2.2e-16
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
-38.83294 -36.43706
sample estimates:
mean of x mean of y
10.055    47.690
```

Block 4

```
> t.test(Lo,Hi,var.equal=TRUE)

Two Sample t-test

data:  Lo and Hi
t = -72.6977, df = 22, p-value < 2.2e-16
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
-38.70863 -36.56137
sample estimates:
mean of x mean of y
10.055    47.690
```

Block 5

```
> ks.test(Lo,Hi)

        Two-sample Kolmogorov-Smirnov test

data:  Lo and Hi
D = 1, p-value = 1.02e-06
alternative hypothesis: two-sided
```

Block 6

```
wilcox.test(Lo,Hi)

        Wilcoxon rank sum test



data:  Lo and Hi
W = 0, p-value = 1.02e-06
alternative hypothesis:
true location shift is not equal to 0
```

Question 34 - Experimental Design

(Remark : This question will not feature in the 2015 Winter Exam)

(i) Give the principal features of a balanced completely randomised design, and explain the role of replication in such a design.

(ii) State the statistical model for this design, define the terms in the model and state the standard assumptions made about the error term.

(iii) Two basic principles of experimental design are **randomisation** and **replication**. Explain why these are important and how they help to validate an analysis of experimental results.

(iv) Briefly explain the principles of randomisation and replication, in the context of a completely randomised experimental design. Write down the model equation for a completely randomised design having equal numbers of replicates in all treatment groups, defining all the symbols that you use.

Question 35

Short Experimental Design Theory Question

- Short Description on Box-Behnken Design
- Short Description on Central Composite Design
- Rationale for Designs like these

Question 36- Experimental Design Part 2

In an investigation into the extraction of nitrate-nitrogen from air dried soil, three quantitative variables were investigated at two levels. These were the amount of oxidised activated charcoal (A) added to the extracting solution to remove organic interferences, the strength of CaSO4 extracting solution (C), and the time the soil was shaken with the solution (T). The aim of the investigation was to optimise the extraction procedure. The levels of the variables are given here:

		-	+
Activated charcoal (g)	A	0.5	1
CaSO4 (%)	C	0.1	0.2
Time (minutes)	T	30	60

The concentrations of nitrate-nitrogen were determined by ultra-violet spectrophotometry and compared with concentrations determined by a standard technique. The results are given below and are the amounts recovered (expressed as the percentage of known nitrate concentration).

A	C	T	Amounts	(2 Replicates)
-1	-1	-1	45.1	44.6
1	-1	-1	44.9	45.3
-1	1	-1	44.8	46.7
1	1	-1	44.7	44.8
-1	-1	1	33	35
1	-1	1	53.8	51.7
-1	1	1	32.6	33.7
1	1	1	54.2	53.2

i. (8 Marks) Calculate the contrasts, the effects and the sum of squares for the effects.

ii. (8 Marks) Using the computed sums of squares values, complete the ANOVA table (see the `R` code below).

iii. (4 Marks) Comment on the tests for significant for the main effects and interactions. State clearly your conclusions.

iv. (4 Marks) Write down a regression equation that can be used predicting amounts based on the results of this experiment.

```
             Df  Sum Sq  Mean Sq  F value     Pr(>F)
A             1    ...      ...      ...    0.000979 ***
C             1    ...      ...      ...    0.934131
T             1    ...      ...      ...    0.395554
A:C           1    ...      ...      ...    0.944243
A:T           1    ...      ...      ...    0.017582 *
C:T           1    ...      ...      ...    0.072101
A:C:T         1    ...      ...      ...    0.028522 *
Residuals     8  116.2    14.5
```

Question 37

Six analysts each made seven determinations of the paracetamol content of the same batch of tablets. The results are shown below. There are 42 determinations in total. The mean determination for each analysts is also tabulated.

Analyst	Content						
A	84.32	84.61	84.64	84.62	84.51	84.63	84.51
B	84.24	84.13	84.00	84.02	84.25	84.41	84.30
C	84.29	84.28	84.40	84.63	84.40	84.68	84.36
D	84.14	84.48	84.27	84.22	84.22	84.02	84.33
E	84.50	83.91	84.11	83.99	83.88	84.49	84.06
F	84.70	84.36	84.61	84.15	84.17	84.11	83.81

The following R output has been produced as a result of analysis of these data:

Response: Y	Df	Sum Sq	Mean Sq	F value	$Pr(> F)$
Analyst	?	?	?	?	0.00394 **
Residuals	?	?	0.04065		
Total	?	2.3246			

i. (5 marks) Complete the ANOVA table in your answer sheet, replacing the "?" entries with the correct values.

ii. (2 marks) What hypothesis is being considered by this procedure.

iii. (2 marks) What is the conclusion following from the above analysis? State the null and alternative hypothesis clearly.

Question 38 - Assumptions for ANOVA

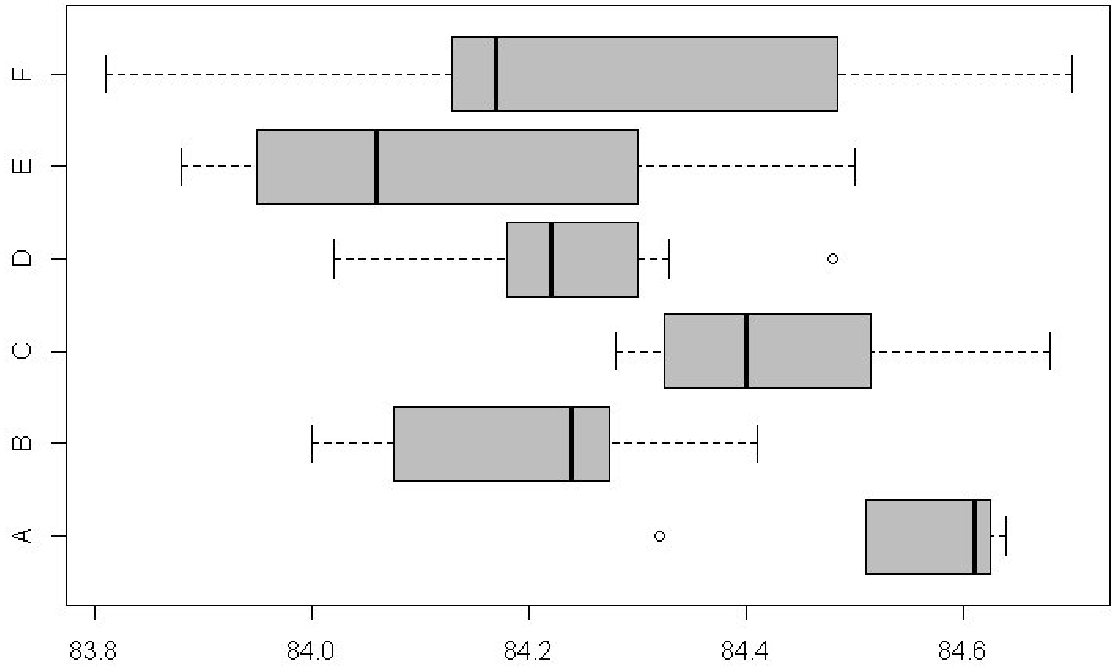
The R code and graphical procedures, below and on the following page, are relevant to checking whether the underlying assumptions are met for the ANOVA model in part (b).

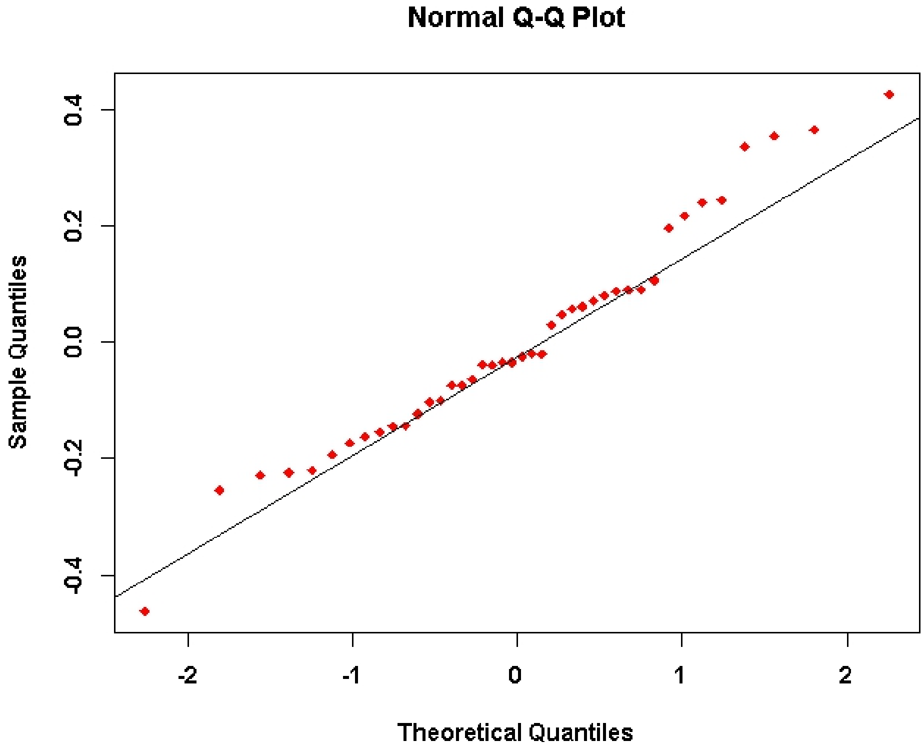
i. (3 marks) What are the assumptions underlying ANOVA?

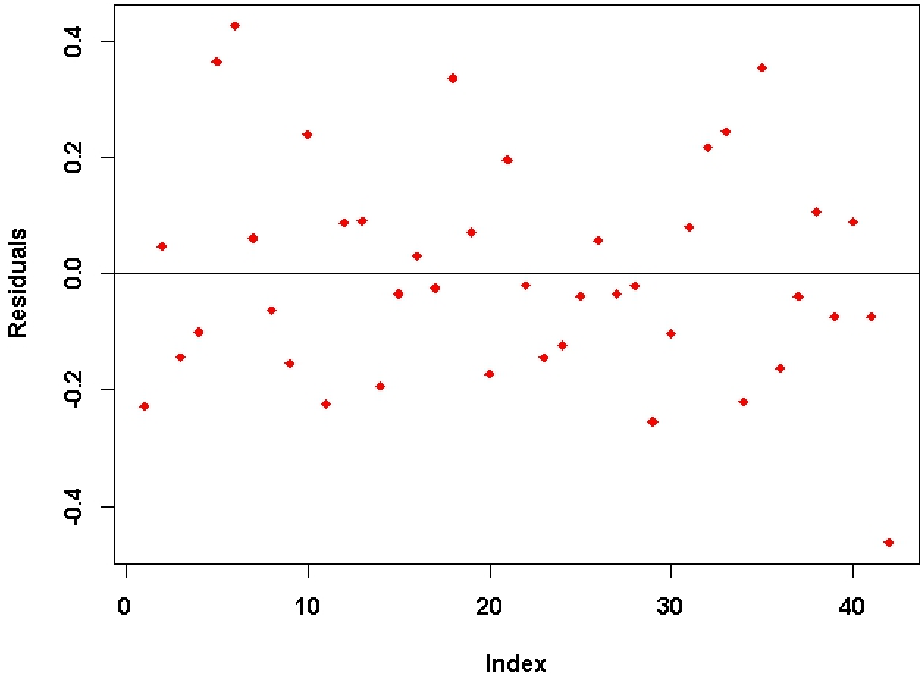
ii. (4 marks) Assess the validity of these assumptions for the ANOVA model in the previous question (Question 37).

```
Shapiro-Wilk normality test

data:  Residuals
W = 0.9719, p-value = 0.3819
```

```
Bartlett test of homogeneity of variances

data:  Experiment
Bartlett's K-squared = 105.9585, df = 1, p-value < 2.2e-16
```




Normal Q-Q Plot
Sample Quantiles
-0.4
-0.2
0.0
0.2
0.4
-2
-1
0
1
2
Theoretical Quantiles


Residuals
-0.4
-0.2
0.0
0.2
0.4
0
10
20
30
40
Index

Question 39 - Control Charts Arithmetic

A normally distributed quality characteristic is monitored through the use of control charts. These charts have the following parameters. All charts are in control.

	LCL	Centre Line	UCL
$\bar{X}$-Chart	614	620	626
R-Chart	0	8.236	18.795

i. (2 marks) What sample size is being used for this analysis?

ii. (2 marks) Estimate the standard deviation of this process.

iii. (2 marks) Compute the control limits for the process standard deviation chart (i.e. the s-chart).

Question 40 - Process Capability Indices

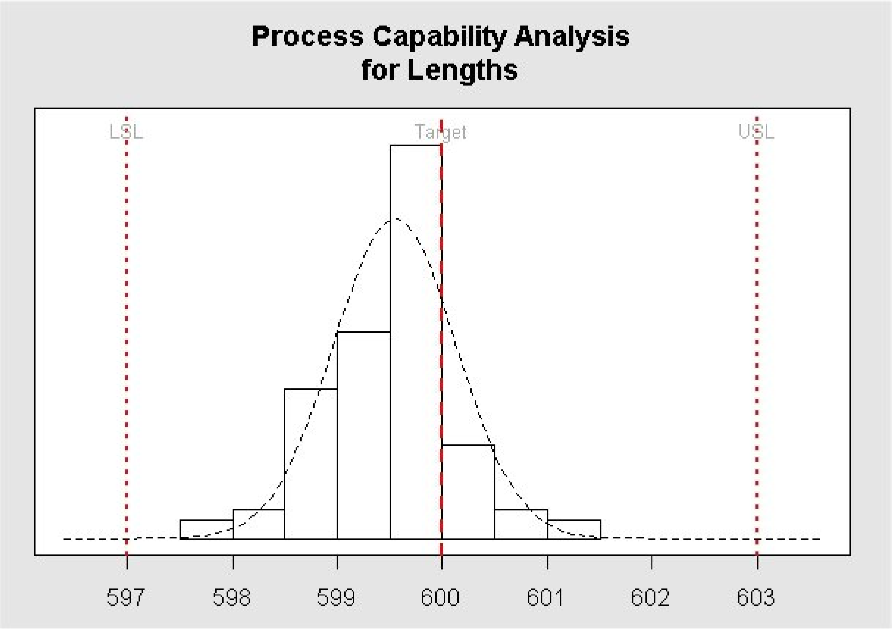
An automobile assembly plant concerned about quality improvement measured sets of five camshafts on twenty occasions throughout the day. The specifications for the process state that the design specification limits at 600±3mm.

i. (2 marks) Determine the *Process Capability Indices* C_p and C_{pk}, commenting on the respective values. You may use the `R` code output on the following page.

ii. (2 marks) The value of C_{pm} is 1.353. Explain why there would be a discrepancy between C_p and C_{pm}.

iii. (2 marks) Comment on the graphical output of the *Process Capability Analysis*, also presented on the next page.

```
Process Capability Analysis

Call:
process.capability(object = obj, spec.limits = c(597, 603))
Number of obs = 100          Target = 600
Center = 599.548         LSL = 597
StdDev = 0.5846948       USL = 603

Capability indices:
Value   2.5%  97.5%
Cp    ...
Cp_l  ...
Cp_u  ...
Cp_k  ...
Cpm   1.353  1.134  1.572
Exp<LSL 0%   Obs<LSL 0%
```



Question 41

Answer the following questions.

i (1 marks) Differentiate common causes of variation in the quality of process output from assignable causes.

ii. (1 marks) What is tampering in the context of statistical process control?

iii (4 marks) Other than applying the *Three Sigma* rule for detecting the presence of an assignable cause, what else do we look for when studying a control chart? Support your answer with sketches.

Question 42- Control Charts Arithmetic

A normally distributed quality characteristic is monitored through the use of control charts. These charts have the following parameters. All charts are in control.

	LCL	Centre Line	UCL
$\bar{X}$-Chart	542	550	558
R-Chart	0	8.236	16.504

i (2 marks) What sample size is being used for this analysis?

ii. (2 marks) Estimate the standard deviation of this process.

iii. (2 marks) Compute the control limits for the process standard deviation chart (i.e. the s-chart).

Question 43 - Process Capability Indices

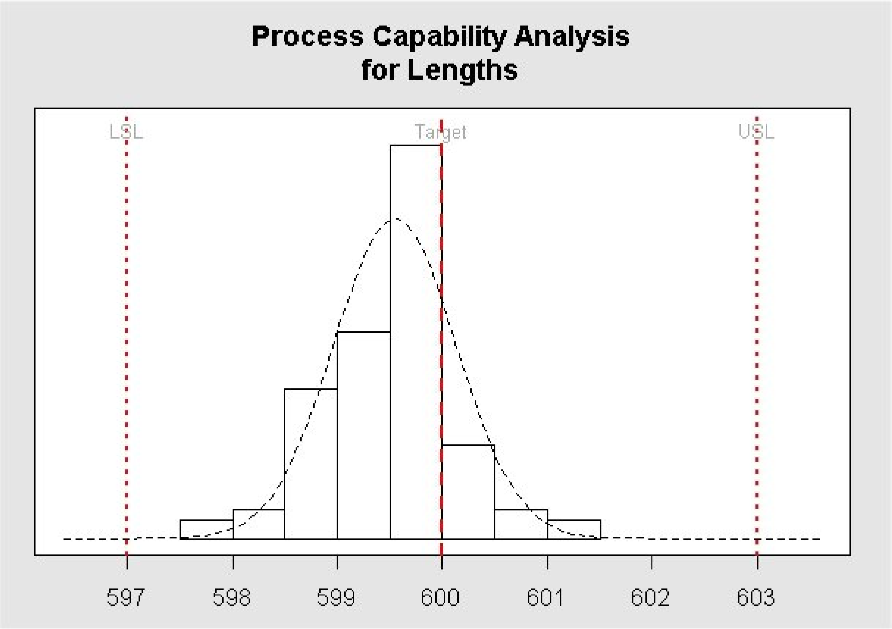
An automobile assembly plant concerned about quality improvement measured sets of five camshafts on twenty occasions throughout the day. The specifications for the process state that the design specification limits at 600±3mm.

i. (4 marks) Determine the *Process Capability Indices* C_p and C_{pk}, commenting on the respective values. You may use the `R` code output on the following page.

ii. (2 marks) The value of C_{pm} is 1.353. Explain why there would be a discrepancy between C_p and C_{pm}.

iii. (2 marks) Comment on the graphical output of the *Process Capability Analysis*, also presented on the next page.

```
Process Capability Analysis

Call:
process.capability(object = obj, spec.limits = c(597, 603))
Number of obs = 100          Target = 600
Center = 599.548         LSL = 597
StdDev = 0.5846948       USL = 603

Capability indices:
Value   2.5%  97.5%
Cp    ...
Cp_l  ...
Cp_u  ...
Cp_k  ...
Cpm   1.353  1.134  1.572
Exp<LSL 0%   Obs<LSL 0%
```



Question 44 - Factorial Design

An experiment is run on an operating chemical process in which the aim is to reduce the amount of impurity produced. Three continuous variables are thought to affect impurity, these are concentration of NaOH, agitation speed and temperature. As an initial investigation two settings are selected for each variable these are

Factor:	low level	highlevel
Concentration of NaOH	40%	45%
Agitation speed (rpm)	15	25
Temperature (°F)	170	200

Readings were recorded of the impurity produced from the chemical process for each combination of the levels of these factors, and each combination was tested twice.

Conc NaOH	Agitation	Temperature	Impurity
-	-	-	90,70
+	-	-	100,120
-	+	-	90,110
+	+	-	120,150
-	-	+	110,100
+	-	+	100,130
-	+	+	100,80
+	+	+	160,140

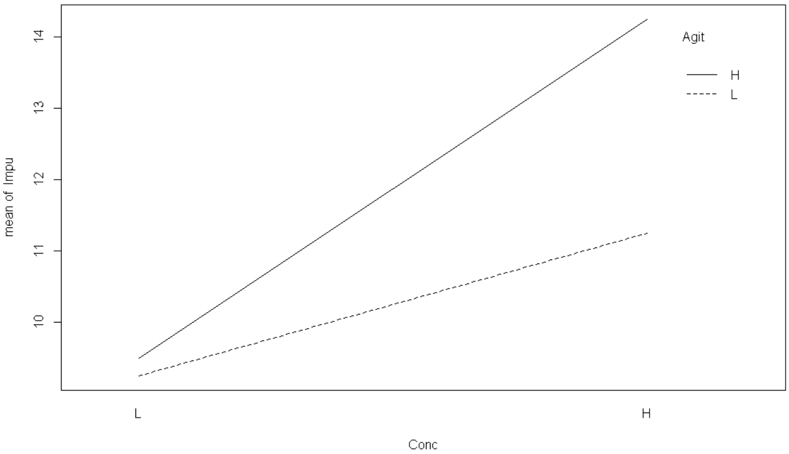
i. (8 Marks) Calculate the contrasts, the effects and the sum of squares for the effects.

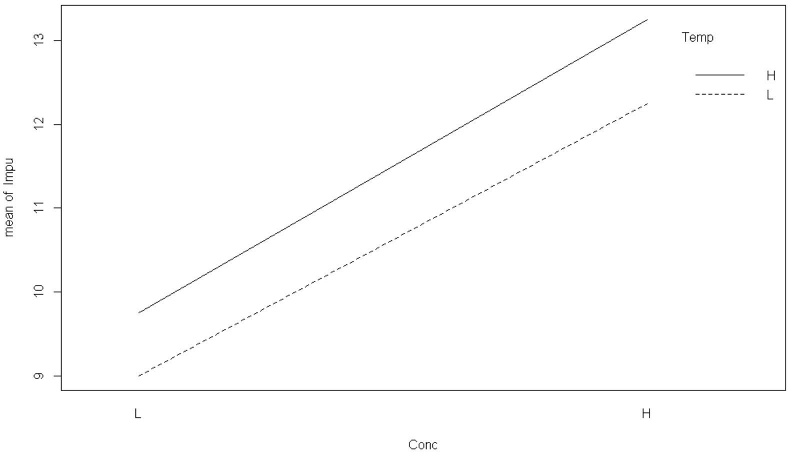
ii. (8 Marks) Using the computed sums of squares values, complete the ANOVA table (see the `R` code below).

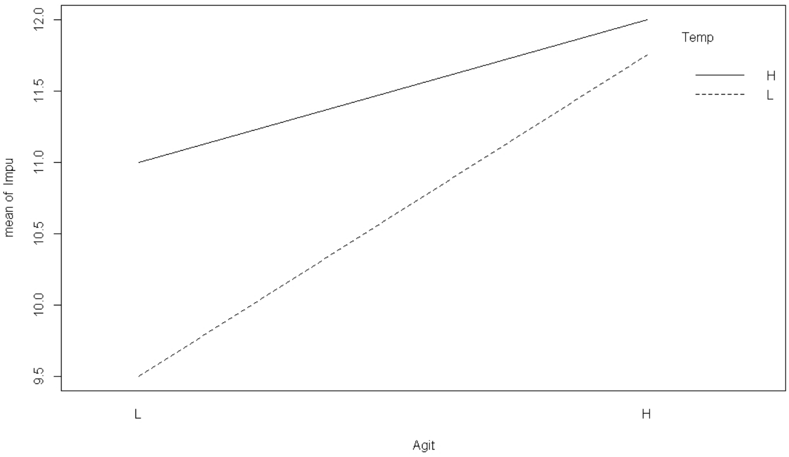
iii. (4 Marks) Comment on the tests for significant for the main effects and interactions. State clearly your conclusions.

iv. (4 Marks) Write down a regression equation that can be used predicting impurity based on the results of this experiment.

```
Df Sum Sq Mean Sq F value  Pr(>F)
Conc             1    ...      ...   ...    0.00253 **
Agit             1    ...      ...   ...    0.07093 .
Temp             1    ...      ...   ...    0.29485
Conc:Agit        1    ...      ...   ...    0.48239
Conc:Temp        1    ...      ...   ...    0.87675
Agit:Temp        1    ...      ...   ...    0.44646
Conc:Agit:Temp   1    ...      ...   ...    0.18751
Residuals        8   1950     244
```


mean of Impu
14
13
12
11
10
Agit
H
L
L
H
Conc


mean of Impu
13
12
11
10
9
Temp
H
L
L
H
Conc


mean of Impu
12.0
11.5
11.0
10.5
10.0
9.5
Temp
H
L
L
H
Agit

Question 45 - Statistical Process Control

Answer the following questions.

i. (1 marks) What is the purpose of maintaining control charts?

ii. (1 marks) What is the *Three Sigma* rule in the context of statistical process control?

iii. (4 marks) Other than applying the *Three Sigma* rule for detecting the presence of an assignable cause, what else do we look for when studying a control chart? Limit your answer to three examples. Support your answer with sketches.

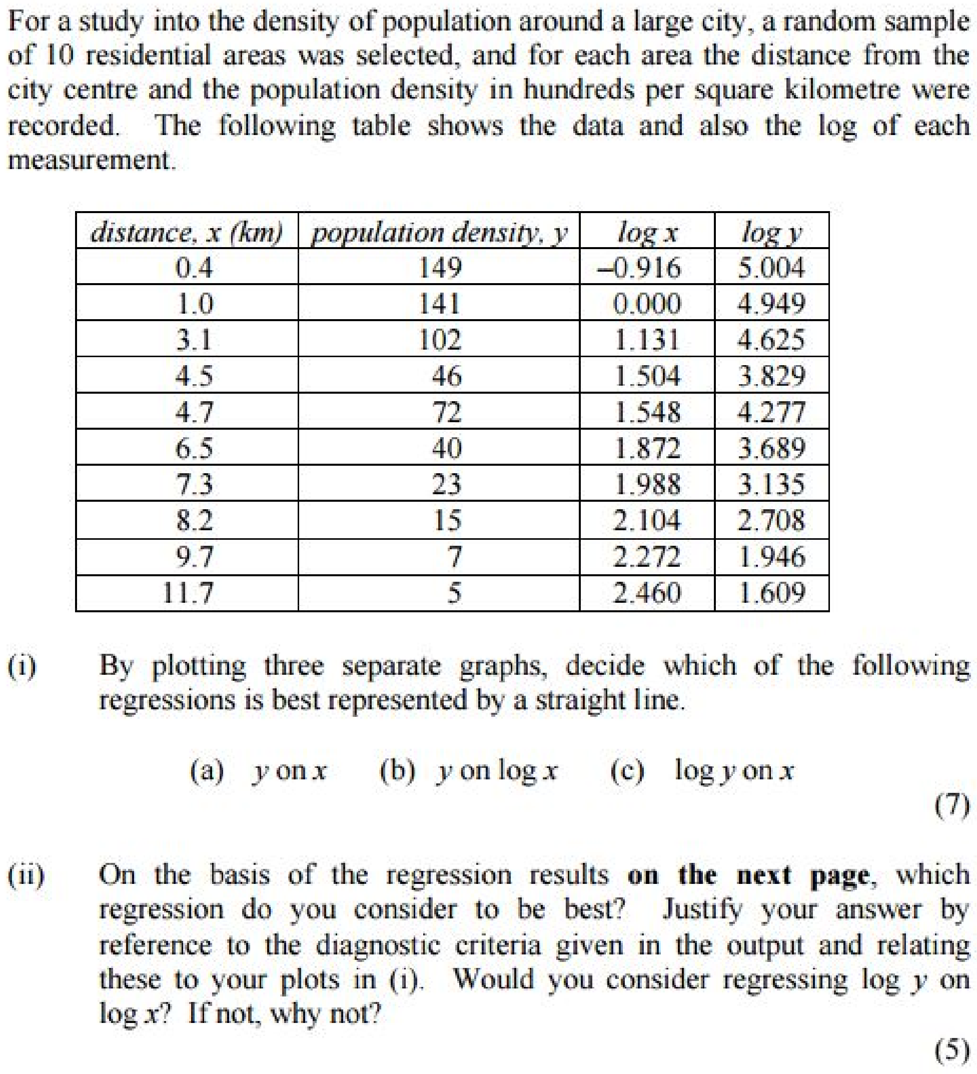
iv. (2 Marks) What is a CUSUM chart? What type of departures from the production target value is this type of chart useful for detecting?

Question 46 - Residual Diagnostics

Expect a question on Hypothesis Tests from car R package, and Model diagnostic plots.

- `ncvTest()` - Non Constant Error Variance
- `outlierTest()` - Outliers
- `durbinWatsonTest()` - Autocorrelation
- Cook's Distances
- Diagnostic Plot 1 (Fitted Vs Residual)
- Diagnostic Plot 2 (Residual Normality)

Question 47 - Regression Analysis



For a study into the density of population around a large city, a random sample of 10 residential areas was selected, and for each area the distance from the city centre and the population density in hundreds per square kilometre were recorded. The following table shows the data and also the log of each measurement.

distance, x (km)	*population density, y*	*log x*	*log y*
0.4	149	–0.916	5.004
1.0	141	0.000	4.949
3.1	102	1.131	4.625
4.5	46	1.504	3.829
4.7	72	1.548	4.277
6.5	40	1.872	3.689
7.3	23	1.988	3.135
8.2	15	2.104	2.708
9.7	7	2.272	1.946
11.7	5	2.460	1.609

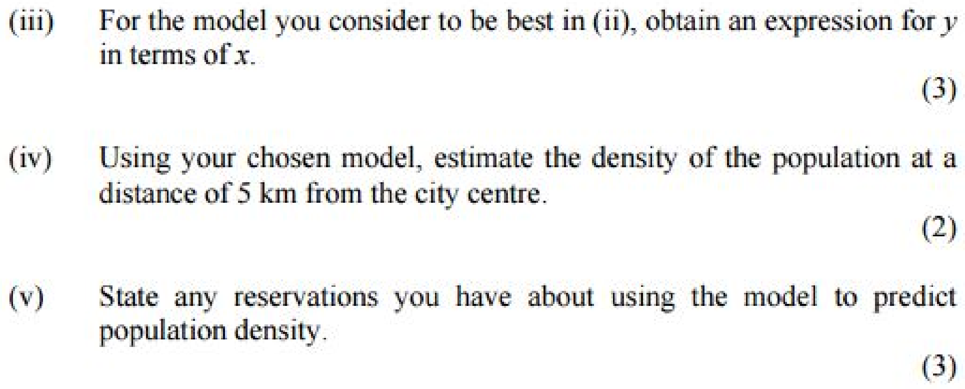
(i) By plotting three separate graphs, decide which of the following regressions is best represented by a straight line.

(a) *y* on *x* (b) *y* on log *x* (c) log *y* on *x*

(7)

(ii) On the basis of the regression results **on the next page**, which regression do you consider to be best? Justify your answer by reference to the diagnostic criteria given in the output and relating these to your plots in (i). Would you consider regressing log *y* on log *x*? If not, why not?

(5)



(iii) For the model you consider to be best in (ii), obtain an expression for *y* in terms of *x*.

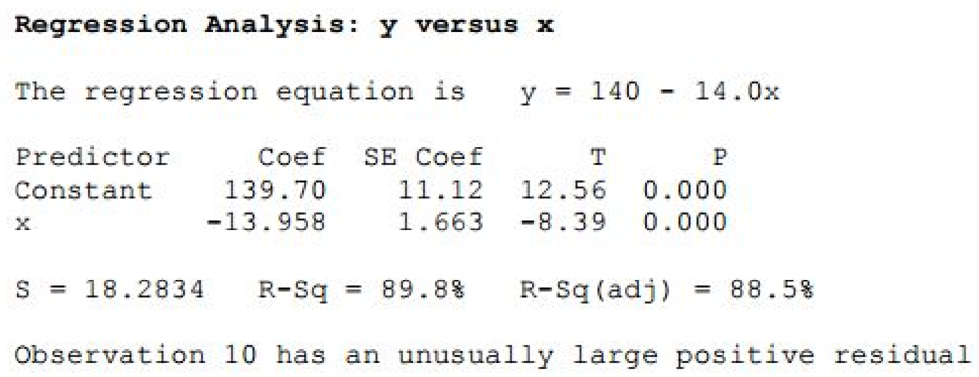
(3)

(iv) Using your chosen model, estimate the density of the population at a distance of 5 km from the city centre.

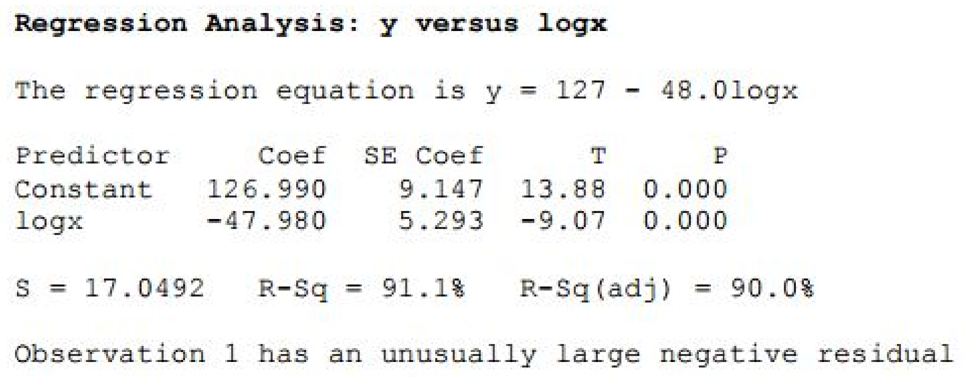
(2)

(v) State any reservations you have about using the model to predict population density.

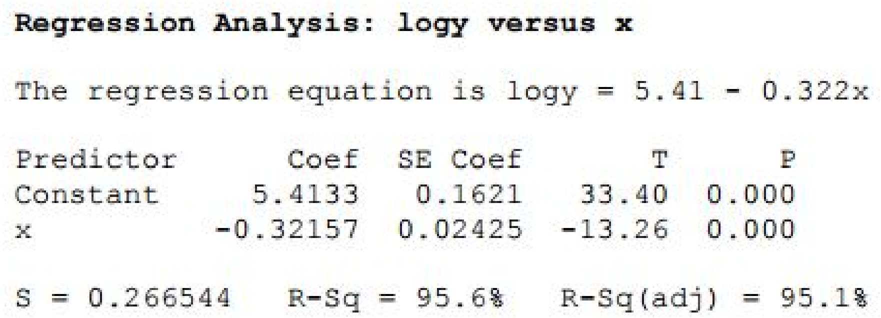
(3)



```
Regression Analysis: y versus x

The regression equation is   y = 140 - 14.0x

Predictor     Coef  SE Coef      T      P
Constant    139.70    11.12  12.56  0.000
x          -13.958    1.663  -8.39  0.000

S = 18.2834   R-Sq = 89.8%   R-Sq(adj) = 88.5%

Observation 10 has an unusually large positive residual
```



```
Regression Analysis: y versus logx

The regression equation is y = 127 - 48.0logx

Predictor     Coef  SE Coef      T      P
Constant   126.990    9.147  13.88  0.000
logx       -47.980    5.293  -9.07  0.000

S = 17.0492   R-Sq = 91.1%   R-Sq(adj) = 90.0%

Observation 1 has an unusually large negative residual
```



```
Regression Analysis: logy versus x

The regression equation is logy = 5.41 - 0.322x

Predictor      Coef   SE Coef       T      P
Constant     5.4133    0.1621   33.40  0.000
x          -0.32157   0.02425  -13.26  0.000

S = 0.266544   R-Sq = 95.6%   R-Sq(adj) = 95.1%
```