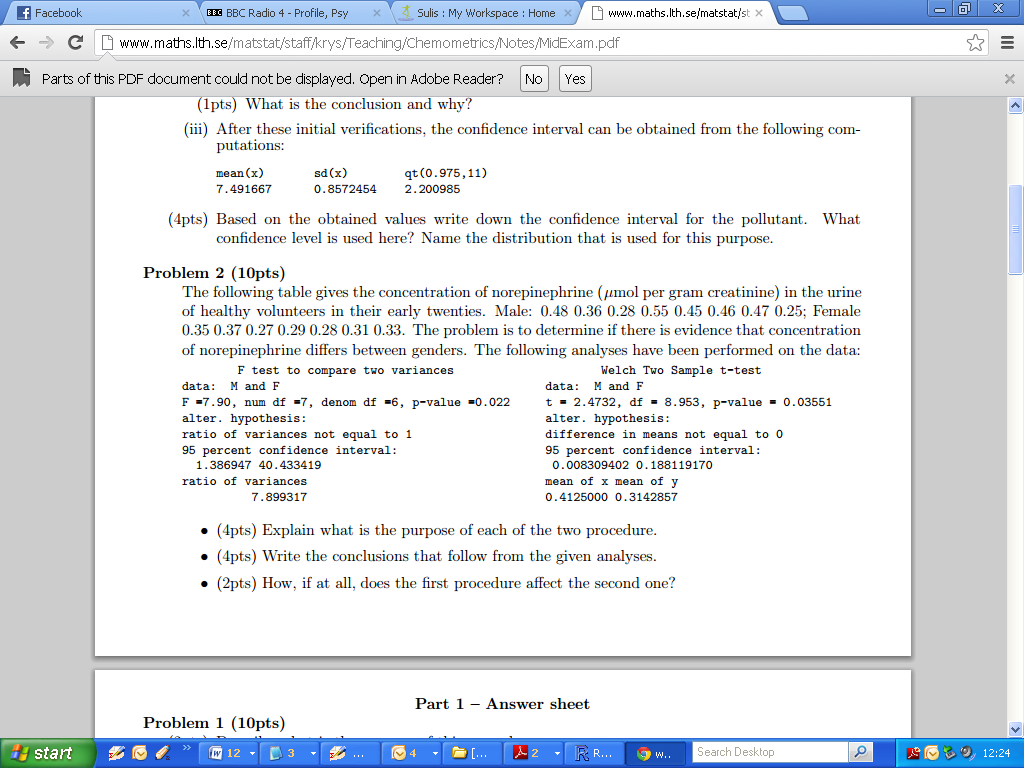
**Question 1**



**Question 2**

Explain the following terms

* Overfitting
* Multicollinearity

Explain why the adjusted R-square value may be much smaller that the multiple R-square value for a given fitted model.

**Question 3**

Explain the following terms

* Influence
* Leverage
* Cook’s Distance

Write a brief explanation of how robust regression differs from linear models computed using the Ordinary Least Squares method, making reference to one particular weighting method.

**Question 4**

As part of a class exercise, a ﬁrst-grade teacher recorded the heights in inches of his 73 students, hoping to determine whether the class was consistent with government standards on height. The government states that the true average height of all ﬁrst-graders is 46 inches. The following R output was produced.

|  |
| --- |
| > t.test(X, mu=50)  One Sample t-test  data: X  t = -8.4411, df = 72, p-value = 2.301e-12  alternative hypothesis: true mean is not equal to 50  95 percent confidence interval:  40.84772 44.34469  sample estimates:  mean of x  42.59621 |

(i) Are the students in this class consistent with government standards? Justify your answer. [3 marks]

(ii) Interpret the 95% conﬁdence interval produced for the above example. [3 marks]

**Question 5**

With reference to only three of the following six ***R*** outputs, write a short report comparing the data sets “***Exp”*** and “***Cont”***, in terms of centrality, distribution and dispersion.

**Code Segment A**

|  |
| --- |
| > shapiro.test(Exp)  Shapiro-Wilk normality test  data: Exp  W = 0.8776, p-value = 0.006174  > shapiro.test(Cont)  Shapiro-Wilk normality test  data: Cont  W = 0.893, p-value = 0.009294 |

**Code Segment B**

|  |
| --- |
| > var.test(Exp,Cont)  F test to compare two variances  data: Exp and Cont  F = 0.6646, num df = 24, denom df = 26, p-value = 0.3178  alternative hypothesis: true ratio of variances is not equal to 1  95 percent confidence interval:  0.2997211 1.4929571  sample estimates:  ratio of variances  0.6646071 |

**Code Segment C**

|  |
| --- |
| > wilcox.test(Exp,Cont)  Wilcoxon rank sum test  data: Exp and Cont  W = 347, p-value = 0.8702  alternative hypothesis: true location shift is not equal to 0 |

**Code Segment D**

|  |
| --- |
| > t.test(Exp, Cont)  Welch Two Sample t-test  data: Exp and Cont  t = -0.2894, df = 49.238, p-value = 0.7735  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -5.172546 3.870205  sample estimates:  mean of x mean of y  19.37792 20.02909 |

**Code Segment E**

|  |
| --- |
| > t.test(Exp, Cont, var.equal = TRUE)  Two Sample t-test  data: Exp and Cont  t = -0.2871, df = 50, p-value = 0.7752  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -5.206675 3.904333  sample estimates:  mean of x mean of y  19.37792 20.02909 |

**Code Segment F**

|  |
| --- |
| > ks.test(Exp,Cont)  Two-sample Kolmogorov-Smirnov test  data: Exp and Cont  D = 0.1585, p-value = 0.8387  alternative hypothesis: two-sided |

**Question 6**

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 (A): mean = 10 ng/ml, n=12.

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

Data: Sample (B): mean = 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=c(8.40, 9.59, 9.38, 9.10, 10.78, 11.41, 9.94, 10.08, 12.11, 9.10, 9.59, 10.36)  > B=c(47.5, 48.4, 48.8, 48.4, 46.8, 46.2, 48.6, 50.6, 45.5, 46.1)  >  > var.test(A,B)  F test to compare two variances  data: A and B  F = 0.4514, num df = 11, denom df = 9, p-value = 0.2141  alternative hypothesis: true ratio of variances is not equal to 1  95 percent confidence interval:  0.115379 1.619474  sample estimates:  ratio of variances  0.4513712  > t.test(A,B,var.equal=TRUE)  Two Sample t-test  data: A and B  t = -67.5402, df = 20, p-value < 2.2e-16  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -38.86779 -36.53887  sample estimates:  mean of x mean of y  9.986667 47.690000 |

* Do the data sets have variances that differ significantly? Justify your answer. Write down the null and alternative hypotheses being considered.
* Write down the estimate for the difference of means.
* Construct a 95% confidence interval for the difference of means.