Tutorial 4 sample answer

Question 1

<table>
<tr><td>a)</td><td>Command/code:

```
TestScore <- read.csv(file.choose())
class_A <- subset(TestScore, Class=="A")
class_B <- subset(TestScore, Class=="B")

t.test(x=class_A$Score, mu=70, alternative="greater")
```

Output:

```
   One Sample t-test

data:  class_A$Score
t = 2.4528, df = 4, p-value = 0.03512
alternative hypothesis: true mean is greater than 70
95 percent confidence interval:
 71.96258      Inf
sample estimates:
mean of x
       85
```

Comment:
In this hypothesis test, the following hypotheses are used,

$$H_0: \mu_A \le 70 \text{ vs } H_1: \mu_A > 70$$

where μ_A is the mean score for class A. The R output shows that the test statistic is 2.4528 and *p*-value for the test is 0.03512. Since the *p*-value is less than α (0.05), we have enough evidence to reject the null hypothesis, and support the alternative hypothesis. We conclude that the mean score for class A is greater than 70.</td></tr>
<tr><td>b)</td><td>Command/code:

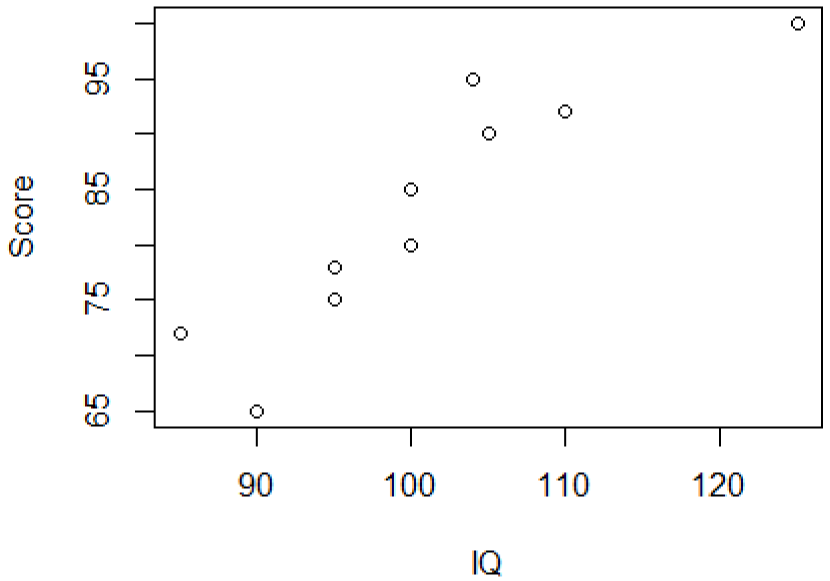
```
var.test(x=class_A$Score, y=class_B$Score)
```

Output:

```
   F test to compare two variances

data:  class_A$Score and class_B$Score
F = 2.2722, num df = 4, denom df = 4,
p-value = 0.4462
alternative hypothesis: true ratio of variances is
not equal to 1
95 percent confidence interval:
  0.2365733 21.8231724
sample estimates:
ratio of variances
          2.272175
```
</td></tr>
</table>

	Comment: The hypotheses are: $$H_0: \sigma_A^2 = \sigma_B^2 \text{ vs } H_1: \sigma_A^2 \neq \sigma_B^2$$ where σ_A^2 is the variance score for class A, and σ_B^2 is the variance score for class B. From the R output, the *p*-value is 0.4462, which is greater than the significance level $\alpha = 0.05$. Therefore there is not enough evidence to reject the null hypothesis. We conclude that the two variances are equal.
c)	Command/code: `t.test(x=class_A$GPA, y=class_B$GPA, var.equal=TRUE, alternative="two.sided")` Output: `Two Sample t-test` `data:  class_A$GPA and class_B$GPA` `t = 1.3112, df = 8, p-value = 0.2262` `alternative hypothesis: true difference in means is not equal to 0` `95 percent confidence interval:` `-0.4096706  1.4896706` `sample estimates:` `mean of x mean of y` `2.70      2.16` Comment: The hypotheses are: $$H_0: \mu_A = \mu_B \text{ vs } H_1: \mu_A \neq \mu_B$$ where μ_A is the mean GPA for class A, and μ_B is the mean GPA for class B. From the R output, the *p*-value is 0.2262, which is greater than the significance level α. Therefore there is not enough evidence to reject the null hypothesis. We conclude that the mean GPAs of the two classes are equal.
d)	Command/code: `summary(aov(GPA ~ Class, data=TestScore))` Output: `            Df Sum Sq Mean Sq F value Pr(>F)` `Class        1  0.729   0.729   1.719  0.226` `Residuals    8  3.392   0.424` Comment: The hypotheses are: $$H_0: \mu_A = \mu_B \text{ vs } H_1: \mu_A \neq \mu_B$$ where μ_A is the mean GPA for class A, and μ_B is the mean GPA for class B. From the R output, the *p*-value is 0.2262, which is greater than the significance level α. Therefore there is not enough evidence to reject the null hypothesis. We conclude that the mean GPAs of the two classes are equal.

	The test using ANOVA gives exactly the same *p*-value as the t-test in (d). ANOVA for two groups are exactly equal as t-test with equal variance. However, ANOVA allows for comparison of mean for more than two groups.
e)	Command/code: `plot(Score ~ IQ, data=TestScore)` `model <- lm(Score ~ IQ, data=TestScore)` `summary(model)` Output:  `Call:` `lm(formula = Score ~ IQ, data = TestScore)` `Residuals:` `    Min      1Q  Median      3Q     Max` `-8.5190 -2.8200  0.3789  2.8412  9.0467` `Coefficients:` `            Estimate Std. Error t value Pr(>|t|)` `(Intercept)  -6.4157    15.6607  -0.410 0.692800` `IQ            0.8882     0.1544   5.754 0.000427 ***` `---` `Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1` `Residual standard error: 5.195 on 8 degrees of freedom` `Multiple R-squared:  0.8054,    Adjusted R-squared: 0.7811` `F-statistic: 33.11 on 1 and 8 DF,  p-value: 0.000427`

	Comment: The fitted regression line is: $$Score = -6.4157 + 0.8882 \times IQ$$ In this case, the p-value for variable IQ is 0.00427, which is smaller than α. Therefore, it can be concluded that the parameter related to IQ is non-zero, and that Score and IQ have a significant linear relationship. Note: The p-value is for testing $\beta_1 = 0$ vs $\beta_1 \neq 0$ where β_1 in this case is the parameter for IQ, which is estimated as 0.8882.

Question 2

a)

Command/code:

```
cement <- read.csv(file.choose())

model1 <- lm(y ~ x1 + x2 + x3 + x4, data=cement)
summary(model1)
```

Output:

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = cement)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1750 -1.6709  0.2508  1.3783  3.9254

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
x1            1.5511     0.7448   2.083   0.0708 .
x2            0.5102     0.7238   0.705   0.5009
x3            0.1019     0.7547   0.135   0.8959
x4           -0.1441     0.7091  -0.203   0.8441
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual standard error: 2.446 on 8 degrees of
freedom
Multiple R-squared:  0.9824,    Adjusted R-squared:
0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

Comment:
Based on the p-values, only the variable x_1 has p-value less than $\alpha = 0.1$. Therefore, only x_1 seems to be significant to the model.

b)

Command/code:

```
model2 <- lm(y ~ x1 + x2, data=cement)
summary(model2)
```

Output:

```
Call:
lm(formula = y ~ x1 + x2, data = cement)

Residuals:
   Min     1Q Median     3Q    Max 
-2.893 -1.574 -1.302  1.363  4.048 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 52.57735    2.28617   23.00 5.46e-10 ***
x1           1.46831    0.12130   12.11 2.69e-07 ***
x2           0.66225    0.04585   14.44 5.03e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 
0.1 ' ' 1

Residual standard error: 2.406 on 10 degrees of 
freedom
Multiple R-squared:  0.9787,    Adjusted R-squared: 
0.9744 
F-statistic: 229.5 on 2 and 10 DF,  p-value: 4.407e-
09
```

Comment:
After removing the variable x_3 and x_4, the p-values for x_1 and x_2 are now very small. Therefore, both of these variables (x_1 and x_2) are significant to the model.

c)

The fitted regression line is:

$$\hat{y} = 52.58 + 1.468x_1 + 0.6623x_2$$

d)

In this case, $x_1 = 8$, $x_2 = 50$, and

$$\hat{y} = 52.58 + 1.468(8) + 0.6623(50) = 97.44$$

The estimated heat evolved is 97.44 calories.

Question 3

<table>
<tr><td>a)</td><td>

Command:

```
property_sales <- read.csv(file.choose())
cheval_sale <- subset(property_sales,
               Neighbourhood=="Cheval")$Sales
hydepark_sale <- subset(property_sales,
                 Neighbourhood=="HydePark")$Sales
var.test(cheval_sale, hydepark_sale)
```

Output:

```
> var.test(cheval_sale, hydepark_sale)

   F test to compare two variances

data:  cheval_sale and hydepark_sale
F = 0.24531, num df = 43, denom df = 33,
p-value = 2.217e-05
alternative hypothesis: true ratio of variances is
not equal to 1
95 percent confidence interval:
 0.1260855 0.4640299
sample estimates:
ratio of variances
         0.2453138
```

Comment:
The hypotheses tested are:

$$H_0: \sigma^2_{cheval} = \sigma^2_{hydepark} \text{ vs } H_1: \sigma^2_{cheval} \neq \sigma^2_{hydepark}$$

where σ^2_{cheval} is the variance for sales in Cheval and $\sigma^2_{hydepark}$ is the variance for sales in Hyde Park.

Since the p-value is very small (2.2×10^{-5}), we have strong evidence to reject the null hypothesis. We conclude that the two variances are not equal

</td></tr>
<tr><td>b)</td><td>

Command:

```
t.test(cheval_sale, hydepark_sale,
       alternative="less", var.equal=FALSE)
```

Output:

```
> t.test(cheval_sale, hydepark_sale,
alternative="less", var.equal=FALSE)

   Welch Two Sample t-test

data:  cheval_sale and hydepark_sale
t = -2.4857, df = 45.444, p-value = 0.008335
alternative hypothesis: true difference in means is
less than 0
95 percent confidence interval:
```

</td></tr>
</table>

	<pre> -Inf -68.12958 sample estimates: mean of x mean of y 455.3955 665.3382</pre> Comment: The hypotheses tested are: $$H_0: \mu_{cheval} \geq \mu_{hydepark} \text{ vs } H_1: \mu_{cheval} < \mu_{hydepark}$$ where μ_{cheval} is the mean for sales in Cheval and $\mu_{hydepark}$ is the mean for sales in Hyde Park. Since the p-value is smaller than $\alpha = 0.05$, we reject the null hypothesis. Therefore, we have evidence to support that the mean for sales in Cheval is smaller than the mean for sales in Hyde Park. Note that the "var.equal=FALSE" in the argument is not needed, as it is the default value for the t-test function.
c)	Command: <pre>summary(aov(Land.value~Neighbourhood, data=property_sales))</pre> Output: <pre>> summary(aov(Land.value~Neighbourhood, data=property_sales)) Df Sum Sq Mean Sq F value Pr(>F) Neighbourhood 3 2463630 821210 81.02 <2e-16 *** Residuals 172 1743289 10135 --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</pre> Comment: The null hypothesis, H_0 states that the mean land values are equal for all four neighbourhoods, whereas the alternative hypothesis states that at least two means are different. The p-value from the ANOVA is very small. Therefore we reject the null hypothesis and conclude that at least a pair of neighbourhood have different mean land values.

d)	Command: `summary(aov(Improvement.value~Neighbourhood,` `        data=property_sales))` Output: `> summary(aov(Improvement.value~Neighbourhood,` `data=property_sales))` `               Df  Sum Sq Mean Sq F value Pr(>F)` `Neighbourhood   3  134935   44978   1.174  0.321` `Residuals     172 6591254   38321` Comment: The null hypothesis, H_0 states that the mean improvement values are equal for all four neighbourhoods, whereas the alternative hypothesis states that at least two means are different. The p-value from the ANOVA is large, larger than $\alpha = 0.05$. Therefore we do not have enough evidence to reject the null hypothesis and conclude that the mean improvement values for all four neighbourhoods are equal.
e)	Command: `fit <- lm(Sales~Land.value+Improvement.value,` `data=property_sales)` `summary(fit)` Output: `> summary(fit)` `Call:` `lm(formula = Sales ~ Land.value + Improvement.value,` `data = property_sales)` `Residuals:` `    Min      1Q  Median      3Q     Max` `-382.46  -49.30    0.79   40.42  540.15` `Coefficients:` `                   Estimate Std. Error t value` `(Intercept)       -16.17600   14.99014  -1.079` `Land.value          1.39329    0.05868  23.746` `Improvement.value   1.33023    0.04640  28.666` `                  Pr(>｜t｜)` `(Intercept)          0.282` `Land.value          <2e-16 ***` `Improvement.value   <2e-16 ***` `---` `Signif. codes:` `0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1`

	`` Residual standard error: 112.9 on 173 degrees of freedom Multiple R-squared: 0.9242, Adjusted R-squared: 0.9233 F-statistic: 1055 on 2 and 173 DF, p-value: < 2.2e-16 ``` Comment: The fitted regression line is: $Sales = -16.176 + 1.393 \times Land.value + 1.330 \times Improvement.value$ The land value is significant in the regression model since the p-value for the parameter related to the land value is very small. If the land value is \$100 000 and the improvement value is \$200 000, the estimated sales price is $Sales = -16.176 + 1.393 \times 100 + 1.330 \times 200 = 389.124 = \$389\ 124$