E1

```
SNPs <- c(0.040,0.100,0.400,0.550,0.340,0.620,0.001,0.010,0.800,0.005)
```

```
round(cbind('FDR' = p.adjust(SNPs, method = "fdr"),
            'BON' = p.adjust(SNPs, method = "bonferroni")),4)
```

FDR	BON
0.1000	0.40
0.2000	1.00
0.5714	1.00
0.6875	1.00
0.5667	1.00
0.6889	1.00
0.0100	0.01
0.0333	0.10
0.8000	1.00
0.0250	0.05

By correction using FDR (false discovery rate)we can clearly see that

SNPs 7,8 and 10 are statistically significant

Similarly after using Bonferroni,

SNP 7 has statistically significant effect and SNP 10 is exactly 0.05 which is also significant based on condition.

Q2

Null hypothesis: there is no difference in terms of means in between 3 lung function groups.

Alternative hypothesis: There is a difference in means between at least one set of groups which will be compared

```r
# Create data set from table
lung <- data.frame( group=c( rep('A',5), rep('B',12), rep('C',5) ),
react=c(20.8,4.1,30,24.7,13.8,
7.5,7.5,11.9,4.5,3.1,8,4.7,28.1,10.3,10,5.1,2.2,
9.2,2,2.5,6.1,7.5) )
```

```r
oneway.test(react ~ group, data=lung, var.equal = T)
```

```
        One-way analysis of means

data:  react and group
F = 4.9893, num df = 2, denom df = 19, p-value = 0.01813
```

```r
library(DescTools)
anova <- aov(react ~ group, data=lung)
PostHocTest(anova, method = c('lsd'))
```

```
  Posthoc multiple comparisons of means : Fisher LSD
    95% family-wise confidence level

$group
       diff    lwr.ci    upr.ci   pval
B-A -10.105 -18.01927 -2.190731 0.0151 *
C-A -13.220 -22.62355 -3.816447 0.0084 **
C-B  -3.115 -11.02927  4.799269 0.4203

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

But in our test we got p < 0.05 in 2 sets of group i.e. 0.0151, 0.0084

Hence we reject null hypothesis

So while comparing group B to A, Group C to A are statistically different as per the P value of . 0.0151, 0.0084

2b

```r
PostHocTest(anova, method = c('hsd'))
```

```
  Posthoc multiple comparisons of means : Tukey HSD
    95% family-wise confidence level

$group
       diff    lwr.ci     upr.ci   pval
B-A -10.105 -19.71110 -0.4988964 0.0382 *
C-A -13.220 -24.63375 -1.8062481 0.0217 *
C-B  -3.115 -12.72110  6.4911036 0.6932

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null hypothesis: there is no difference in terms of means in between 3 lung function groups.

Alternative hypothesis: There is a difference in means between at least one set of groups which will be compared

But in our test we got p < 0.05 in 2 sets of group i.e. 0.0151, 0.0084

Hence we reject null hypothesis

So while comparing group B to A, Group C to A are statistically different as per the P value of . 0.0151, 0.0084

2c

```
1 oneway.test(react ~ group, data=lung, var.equal = F)
```

```
        One-way analysis of means (not assuming equal variances)

data:  react and group
F = 3.9682, num df = 2.0000, denom df = 8.9319, p-value = 0.05845
```

Null hypothesis: there is no difference in terms of means in between 3 lung function groups.

Alternative hypothesis: There is a difference in means between at least one set of groups which will be compared

But in our test we got p = 0.05845

Because the overall p value is not less than .05, there is no significant evidence that any of the group means are different than each other (here we fail to reject the null hypothesis from 2B), so further testing is not warranted to compare the means of each pair of groups.

3a

```r
subset <- data.frame(carotenoids$betaplas, carotenoids$calories, carotenoids$smoke)
smoke1 <- mean(subset$carotenoids.betaplas [subset$carotenoids.smoke==1])
sprintf("%s is the mean for betaplas when smoke = 1", smoke1)
smoke2 <- mean(subset$carotenoids.betaplas [subset$carotenoids.smoke==2])
sprintf("%s is the mean for betaplas when smoke = 2", smoke2)
smoke3 <- mean(subset$carotenoids.betaplas [subset$carotenoids.smoke==3])
sprintf("%s is the mean for betaplas when smoke = 3", smoke3)
sd1 <- sd(subset$carotenoids.betaplas [subset$carotenoids.smoke==1])
sprintf("%s is the standard deviation for betaplas when smoke = 3", sd1)
sd2 <- sd(subset$carotenoids.betaplas [subset$carotenoids.smoke==2])
sprintf("%s is the standard deviation for betaplas when smoke = 3", sd2)
sd3 <- sd(subset$carotenoids.betaplas [subset$carotenoids.smoke==3])
sprintf("%s is the standard deviation for betaplas when smoke = 3", sd3)
std.error(subset$carotenoids.betaplas [subset$carotenoids.smoke==1])
std.error(subset$carotenoids.betaplas [subset$carotenoids.smoke==2])
std.error(subset$carotenoids.betaplas [subset$carotenoids.smoke==3])

```

'206.050955414013 is the mean for betaplas when smoke = 1'

'193.469565217391 is the mean for betaplas when smoke = 2'

'121.325581395349 is the mean for betaplas when smoke = 3'

'193.208562618102 is the standard deviation for betaplas when smoke = 3'

'191.639524626138 is the standard deviation for betaplas when smoke = 3'

'78.8116262060946 is the standard deviation for betaplas when smoke = 3'

15.4197219792349

17.8704778162762

12.0186603220171

Cross-checking with professor's suggested method

```r
carotenoids$smoke_factor <- factor(carotenoids$smoke, levels=c(1,2,3),
                                   labels=c('Never','Former','Current'))

# Write functions to calculate N and SE
n.sum <- function(x) sum(!is.na(x))
se_calc <- function(x) sd(x) / sqrt(sum(!is.na(x)))

library(doBy)
library(kableExtra)

sum_tab <- summaryBy(betaplas ~ smoke_factor, data=carotenoids,
                     FUN=list(n.sum, mean, sd, se_calc))
sum_tab # print the object to see the "raw" output
```

smoke_factor	betaplas.n.sum	betaplas.mean	betaplas.sd	betaplas.se_calc
Never	157	206.0510	193.20856	15.41972
Former	115	193.4696	191.63952	17.87048
Current	43	121.3256	78.81163	12.01866

Got same result

3b

```
model <- lm(betaplas ~ smoke_factor + calories, data=carotenoids)
coef(model)
```

(Intercept)	209.595793362558
smoke_factorFormer	-12.2756104198893
smoke_factorCurrent	-84.2811519663666
calories	-0.00206890721479301

```
carotenoids$X1 <- carotenoids$smoke==2
carotenoids$X2 <- carotenoids$smoke==3
model_alt <- lm(betaplas ~ X1 + X2 + calories, data=carotenoids)
coef(model_alt)
```

(Intercept)	209.595793362558
X1TRUE	-12.2756104198893
X2TRUE	-84.2811519663666
calories	-0.00206890721479301

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon,$$

Y = plasma beta carotene value,

X_1= indicator for former smokers, (smoker2)

X_2= indicator for current smokers, (smoker 3)

X_3 = calories

E = error

$$\epsilon \sim N(0, \sigma^2_{Y|X}).$$

Assumptions are: homoscedasticity, linear relationships, normality, independent variable

3c

Null Hypothesis: $\beta1=\beta2=\beta3=0$

Alternative **Hypothesis**: at least one β is not 0.

```r
summary(model)
```

```
Call:
lm(formula = betaplas ~ smoke_factor + calories, data = carotenoids)

Residuals:
    Min      1Q  Median      3Q     Max
-203.36  -98.52  -45.00   43.66 1209.62

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         209.595793  29.801831   7.033 1.29e-11 ***
smoke_factorFormer  -12.275610  22.417618  -0.548  0.58437
smoke_factorCurrent -84.281152  31.448087  -2.680  0.00775 **
calories             -0.002069   0.015195  -0.136  0.89178
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 181.7 on 311 degrees of freedom
Multiple R-squared:  0.02332,   Adjusted R-squared:  0.0139
F-statistic: 2.475 on 3 and 311 DF,  p-value: 0.06153
```

```r
null_model <- lm(betaplas ~ 1, data=carotenoids)
anova(model, null_model, test='F')
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
311	10270401	NA	NA	NA	NA
314	10515638	-3	-245236.9	2.475355	0.06152536

F=2.475 and p=0.06153, we fail to reject Null hypothesis.

We cannot conclude that any of our beta coefficients are different from 0.

This model does not contribute significantly to the prediction of beta-carotene above and beyond the average beta-carotene level.

Due to F-test, we can see that all the slopes of the variables are nearly 0, hence not significant.

Similarly both calories and smoking status does not contribute significantly in the prediction of Y.

3d

Null hypothesis (partial F-test) : $\beta1=\beta2=0$

Alternative hypothesis: At least either $\beta1$ or $\beta2$ are not 0

```r
model2 <- lm(betaplas ~ calories, data=carotenoids)
#anova function for partial F-test
anova(model, model2, test='F')
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
311	10270401	NA	NA	NA	NA
313	10510453	-2	-240051.1	3.634517	0.02752413

Here in case of partial F test, the smoking status groups are not equal to 0.

P value is 0.027524 which is <0.05, so Null hypothesis is Rejected.

We can say that smoking status significantly contributes in the prediction of plasma beta carotene and beyond calories alone.

3e

Null hypothesis: All of the slopes of the variable = 0, so do not significantly contributes in prediction of Y

Alternative Hypothesis: At least one of the slopes is not equal to 0 so significantly contributes in prediction of Y

```
model3 <- lm(betaplas ~ smoke_factor, data=carotenoids)
```

```
summary(model3)
```

```
Call:
lm(formula = betaplas ~ smoke_factor, data = carotenoids)

Residuals:
    Min      1Q  Median      3Q     Max
-206.05  -98.05  -45.47   43.60 1208.95

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          206.05      14.48  14.230  < 2e-16 ***
smoke_factorFormer   -12.58      22.27  -0.565  0.57251
smoke_factorCurrent  -84.73      31.23  -2.713  0.00704 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 181.4 on 312 degrees of freedom
Multiple R-squared:  0.02326,   Adjusted R-squared:  0.017
F-statistic: 3.715 on 2 and 312 DF,  p-value: 0.02543
```

```
anova(model3, null_model, test='F')
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
312	10271014	NA	NA	NA	NA
314	10515638	-2	-244624.7	3.715451	0.02542822

P<0.05, null hypothesis is rejected and we can depict that smoking status has significant contribution in predicting plasma beta carotene levels.

We dropped calories from the model we see that our overall F-test has F=3.715 and p=0.025, suggesting we reject H0 and at least one of the β coefficients is not 0.

But the overall performance in **Part C**, model includes calories as 1 factor. The result was "*None of the variables/Factors are significant on the prediction of Y*"

In **Part D**, a **partial F-test**, we got to know that Smoking Status has significant contribution towards prediction of Y above the Contribution of Calories alone (Smoking status >contribution Calories)

Hence in **Part E**, we fitted a complete new model, by excluding calories and result of F test matched with Partial F test of Previously made model.