MODULE 6: R PRACTICE

BY: SARVESH THORVE

FOR: PROF AMIN

Dummy Variables:

Dummy variables are numeric variable used in the regression analysis, they are used to represent subsets or subgroups of the sample in our study. They are often used to distinguish different groups of treatment.

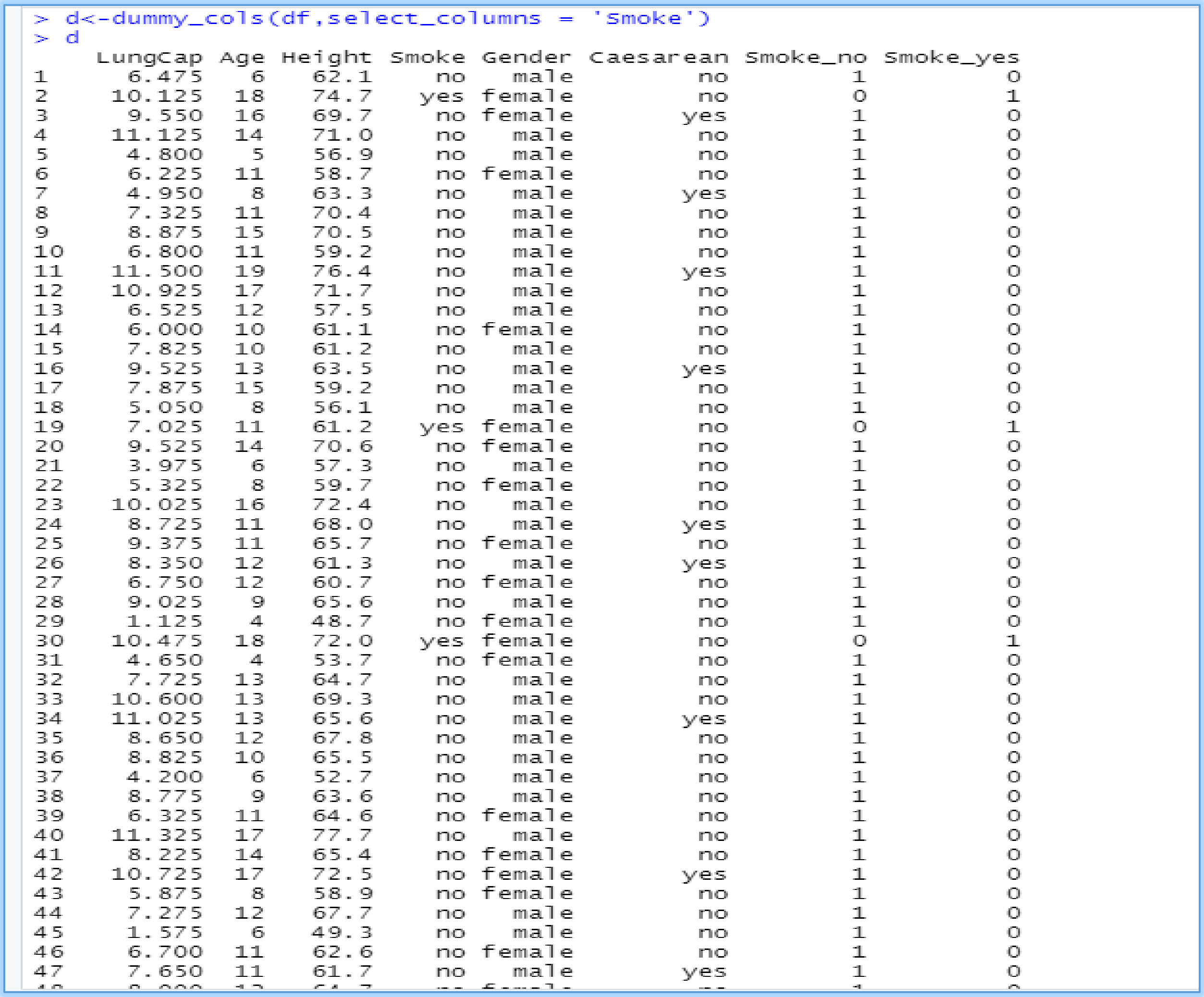
Example: We use 0,1 , a subject is given the value of 0 when they are in the control group and of they are given the value of 1 then they are into the treated group.

I have taken a Dataset of smokers and non-smokers, how it affects their lung capacity and height. It consists of 8 variables and 725 observations.

Impact of the categorical variable on the regression:

> The categorical variable requires to be identified while performing regression because unlike continuous variables, categorical variables cannot by entered in to the regression equation. They need to be recorded into a series of variable first so later they can be entered into the regression equation.

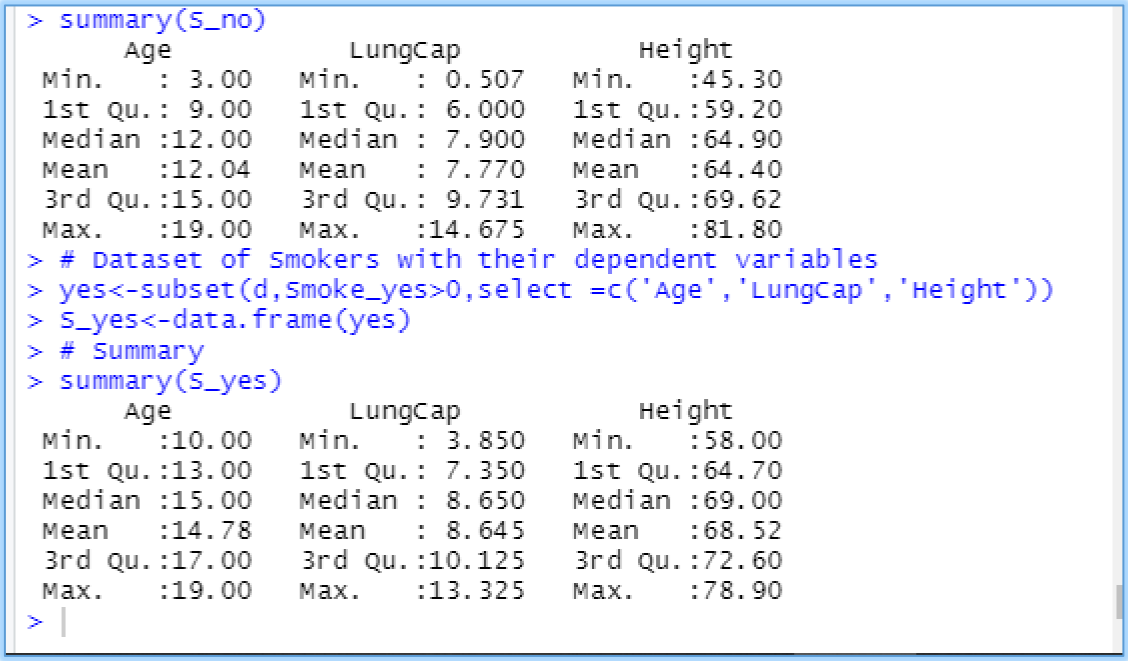
We have created a dummy variable using the Smoke variables, Smokers and Non Smokers. As we can see below we have converted smoke variables in two dummy variables Smoke_no and Smoke_yes which represent people who smoke and people who do not smoke. It is represented with the help of binary numbers 1 and 0.



```
> d<-dummy_cols(df,select_columns = 'Smoke')
> d
    LungCap Age Height Smoke Gender Caesarean Smoke_no Smoke_yes
1     6.475   6   62.1    no   male        no        1         0
2    10.125  18   74.7   yes female        no        0         1
3     9.550  16   69.7    no female       yes        1         0
4    11.125  14   71.0    no   male        no        1         0
5     4.800   5   56.9    no   male        no        1         0
6     6.225  11   58.7    no female        no        1         0
7     4.950   8   63.3    no   male       yes        1         0
8     7.325  11   70.4    no   male        no        1         0
9     8.875  15   70.5    no   male        no        1         0
10    6.800  11   59.2    no   male        no        1         0
11   11.500  19   76.4    no   male       yes        1         0
12   10.925  17   71.7    no   male        no        1         0
13    6.525  12   57.5    no   male        no        1         0
14    6.000  10   61.1    no female        no        1         0
15    7.825  10   61.2    no   male        no        1         0
16    9.525  13   63.5    no   male       yes        1         0
17    7.875  15   59.2    no   male        no        1         0
18    5.050   8   56.1    no   male        no        1         0
19    7.025  11   61.2   yes female        no        0         1
20    9.525  14   70.6    no female        no        1         0
21    3.975   6   57.3    no   male        no        1         0
22    5.325   8   59.7    no female        no        1         0
23   10.025  16   72.4    no   male        no        1         0
24    8.725  11   68.0    no   male       yes        1         0
25    9.375  11   65.7    no female        no        1         0
26    8.350  12   61.3    no   male       yes        1         0
27    6.750  12   60.7    no female        no        1         0
28    9.025   9   65.6    no   male        no        1         0
29    1.125   4   48.7    no female        no        1         0
30   10.475  18   72.0   yes female        no        0         1
31    4.650   4   53.7    no female        no        1         0
32    7.725  13   64.7    no   male        no        1         0
33   10.600  13   69.3    no   male        no        1         0
34   11.025  13   65.6    no   male       yes        1         0
35    8.650  12   67.8    no   male        no        1         0
36    8.825  10   65.5    no   male        no        1         0
37    4.200   6   52.7    no   male        no        1         0
38    8.775   9   63.6    no   male        no        1         0
39    6.325  11   64.6    no female        no        1         0
40   11.325  17   77.7    no   male        no        1         0
41    8.225  14   65.4    no female        no        1         0
42   10.725  17   72.5    no female       yes        1         0
43    5.875   8   58.9    no female        no        1         0
44    7.275  12   67.7    no   male        no        1         0
45    1.575   6   49.3    no   male        no        1         0
46    6.700  11   62.6    no female        no        1         0
47    7.650  11   61.7    no   male       yes        1         0
```

Now, I have made two different dataset for each dummy variable with their dependent and independent variables in them. S_no for people who do not smoke and S_yes for people who do smoke.

Here is the summary of the two dataset, it is shown below:

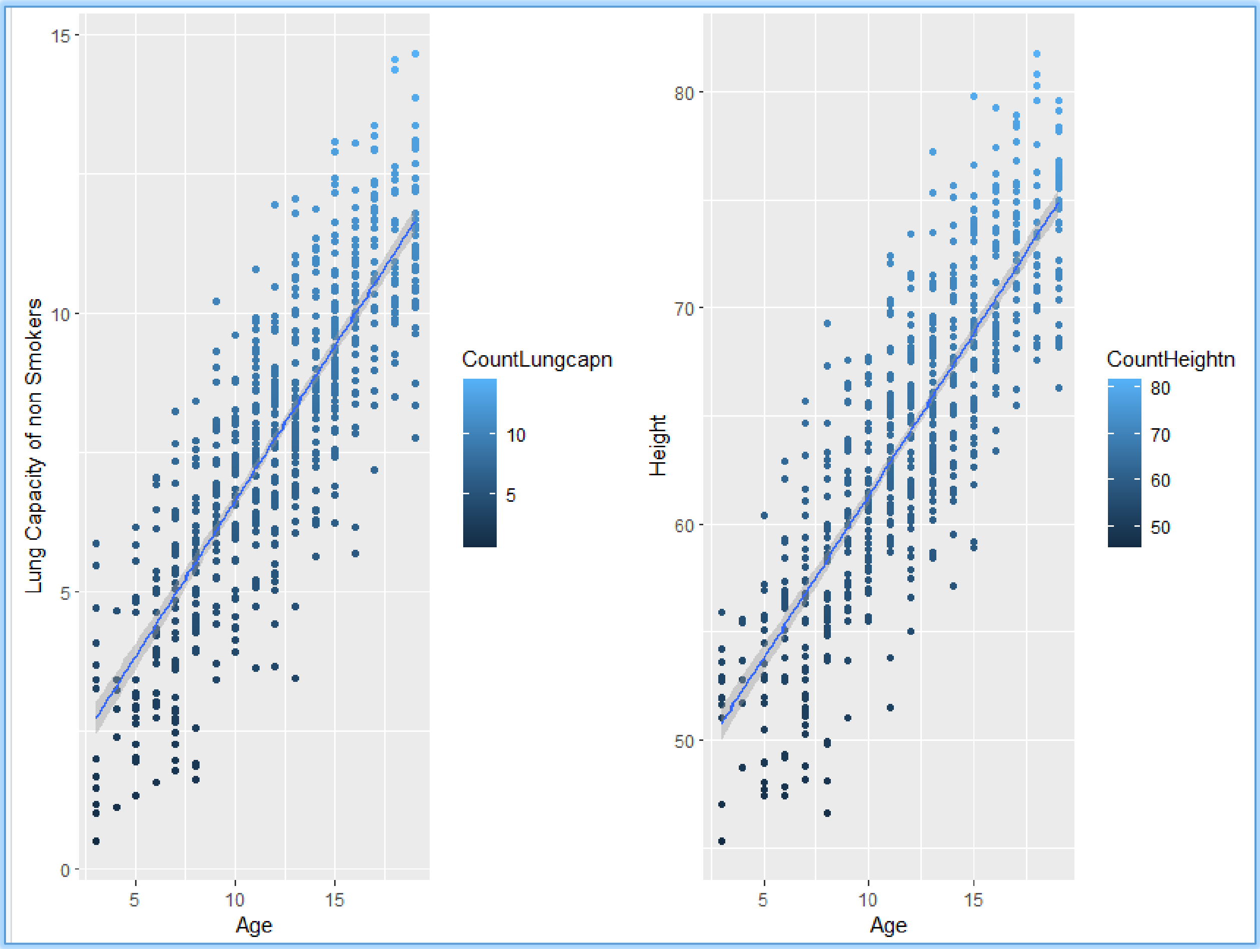


```
> summary(S_no)
      Age          LungCap           Height     
 Min.   : 3.00   Min.   : 0.507   Min.   :45.30  
 1st Qu.: 9.00   1st Qu.: 6.000   1st Qu.:59.20  
 Median :12.00   Median : 7.900   Median :64.90  
 Mean   :12.04   Mean   : 7.770   Mean   :64.40  
 3rd Qu.:15.00   3rd Qu.: 9.731   3rd Qu.:69.62  
 Max.   :19.00   Max.   :14.675   Max.   :81.80  
> # Dataset of Smokers with their dependent variables
> yes<-subset(d,Smoke_yes>0,select =c('Age','LungCap','Height'))
> S_yes<-data.frame(yes)
> # Summary
> summary(S_yes)
      Age          LungCap           Height     
 Min.   :10.00   Min.   : 3.850   Min.   :58.00  
 1st Qu.:13.00   1st Qu.: 7.350   1st Qu.:64.70  
 Median :15.00   Median : 8.650   Median :69.00  
 Mean   :14.78   Mean   : 8.645   Mean   :68.52  
 3rd Qu.:17.00   3rd Qu.:10.125   3rd Qu.:72.60  
 Max.   :19.00   Max.   :13.325   Max.   :78.90  
> 
```

As we can see from the summary, we have created three subsets form our dummy variables. Here Age being the Independent variable and the Lung capacity and Height being the dependent variables.

We count the Lung capacity and height for the smokers and non-smokers from our dataset for dummy variables.

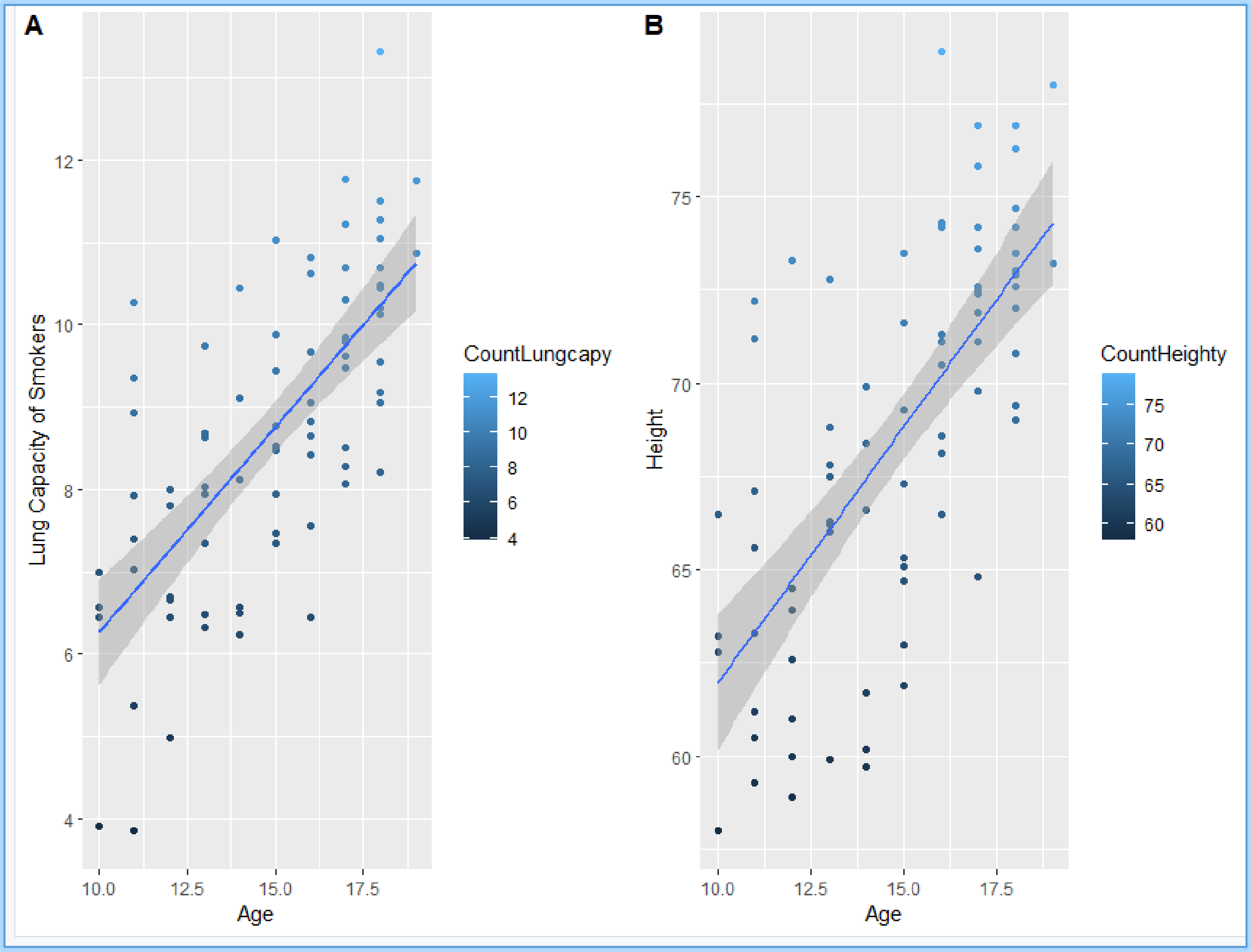
Single regression is performed and separate regression lines are created for each subset data.



Here the two scatterplot show us the regression line for Lung capacity and height for **non-smoker** with the independent variable Age.

In the first plot, we can observe that as the Age increases the Lung capacity for the non-smokers also increases, it is the most between the age 10 and 15. It gets very prominent that the lung capacity is the most dense during this age group and it keeps on increasing.

In the second plot, we can observe that as the Age increases so does our height for the subjects increase. The height of non-smoker has a very gradual scatterplot, it increases with the age and this change is visible from the scatterplot.



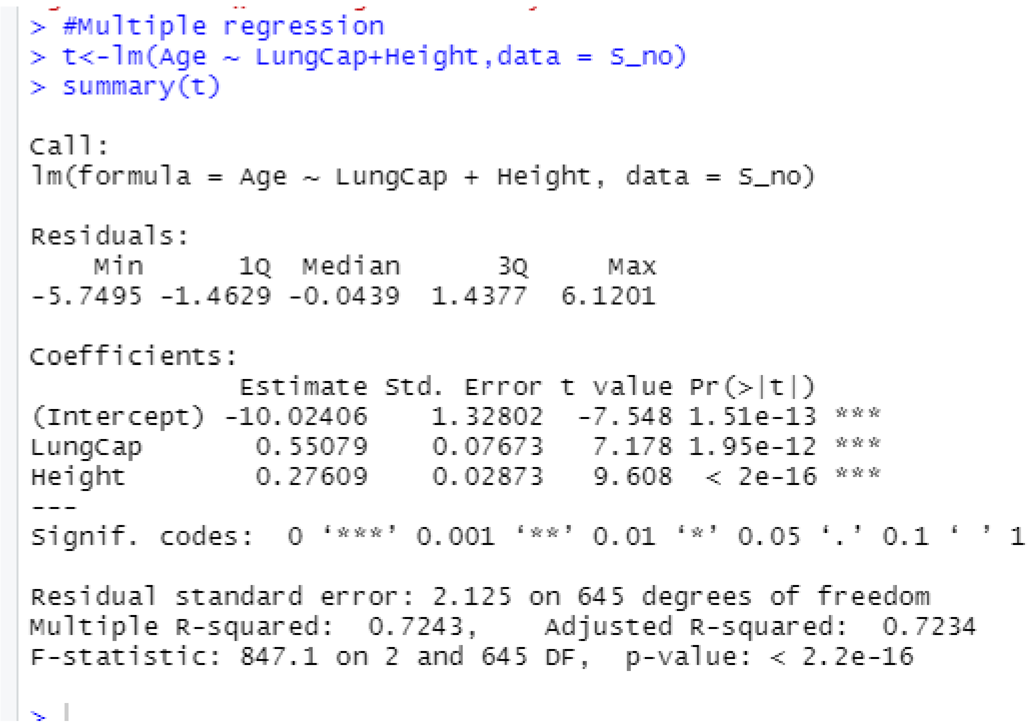
Here the two scatterplot show us the regression line for Lung capacity and height for **Smoker** with the independent variable Age.

In the first plot, we can observe that as the Age increases there is a very moderate increase in the Lung capacity for the non-smokers. As we can see the plots are very scattered and this increases the variance in our plots, telling us that there is no strong relations between two.

In the second plot, we can observe that as the Age increases so does our height for the subjects increase but in a moderate manner. The height of non-smoker has a very scattered, there is a lot of variance that can be seen around the regression line.

We create a multiple regression model for our Lung dataset for subjects who are **Non-smokers.**

The summary for the non-smoker dataset is given below:



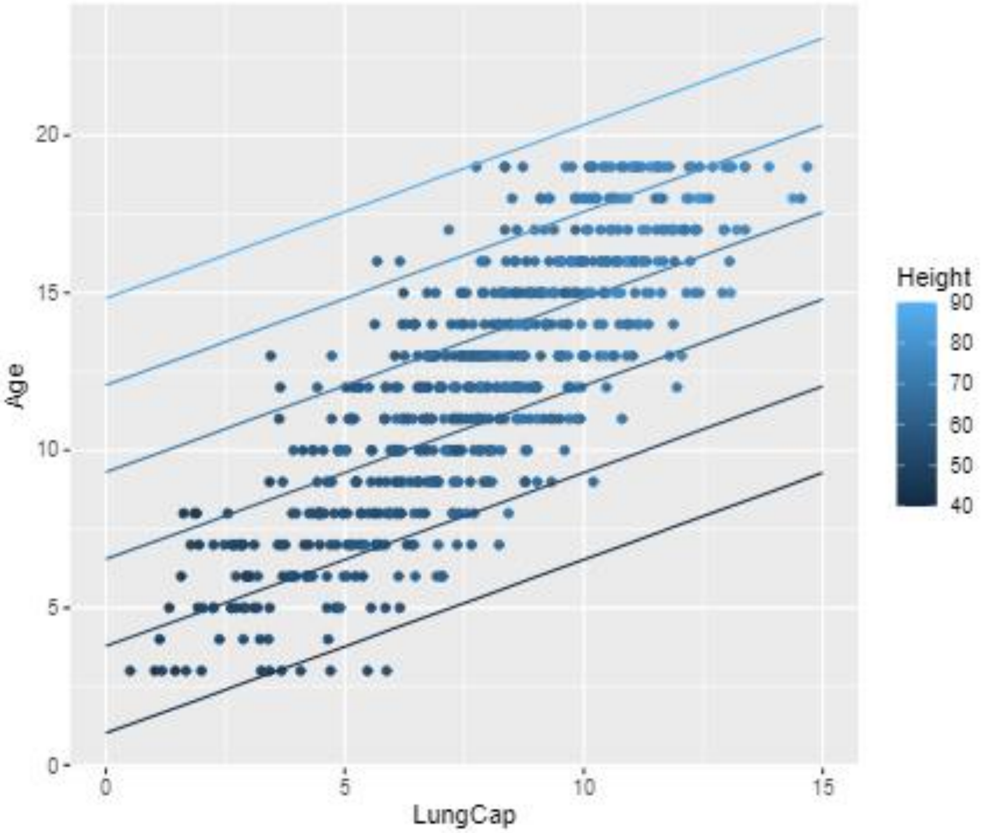
```
> #Multiple regression
> t<-lm(Age ~ LungCap+Height,data = S_no)
> summary(t)

Call:
lm(formula = Age ~ LungCap + Height, data = S_no)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7495 -1.4629 -0.0439  1.4377  6.1201 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -10.02406    1.32802  -7.548 1.51e-13 ***
LungCap       0.55079    0.07673   7.178 1.95e-12 ***
Height        0.27609    0.02873   9.608  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.125 on 645 degrees of freedom
Multiple R-squared:  0.7243,	Adjusted R-squared:  0.7234 
F-statistic: 847.1 on 2 and 645 DF,  p-value: < 2.2e-16

> 
```

We can see that we get a summary of a residuals or errors, estimate of the intercept, standard errors, test statistic and p value.

Residual standard error value tells us about the measure of the variance of observations around the regression line. We get the r squared and the adjusted r square value.

The R square gives us a measure of what percent of the variance in the response variable can be explained by the regression.

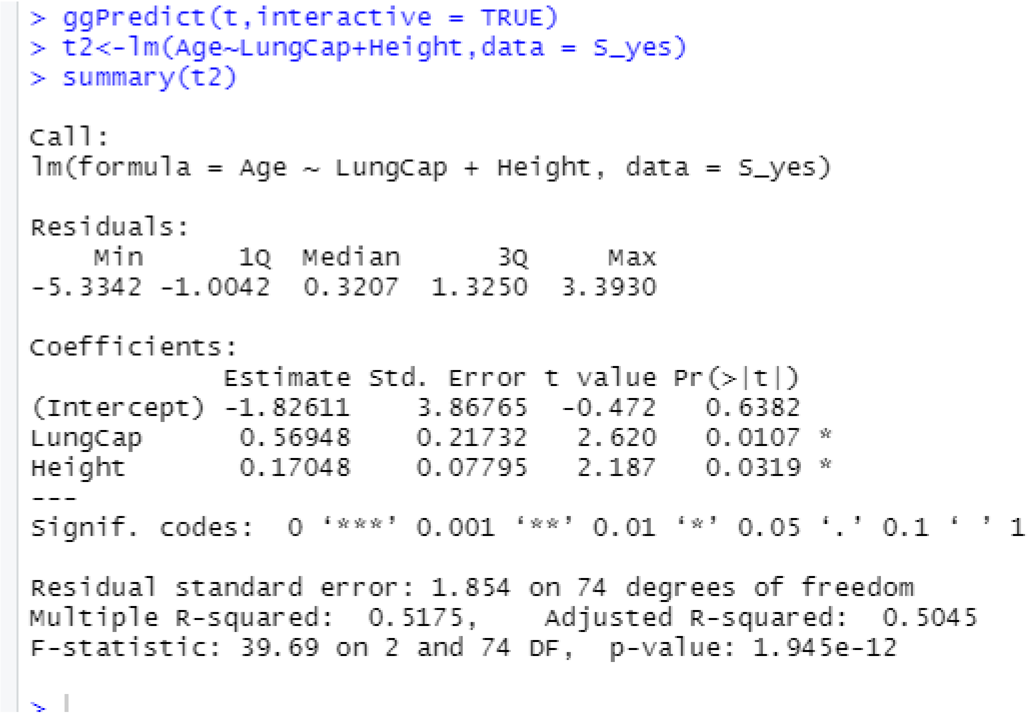
We created a scatterplot with multiple regression for the same model of Non-smokers.



The Lung capacity for non-smokers increases with increase in Age and the height remains stagnant at a particular age (18 as shown) as seen from the scatter plots.

We create a multiple regression model for our Lung dataset for subjects who are **Smokers.**

The summary for the Smoker dataset is given below:



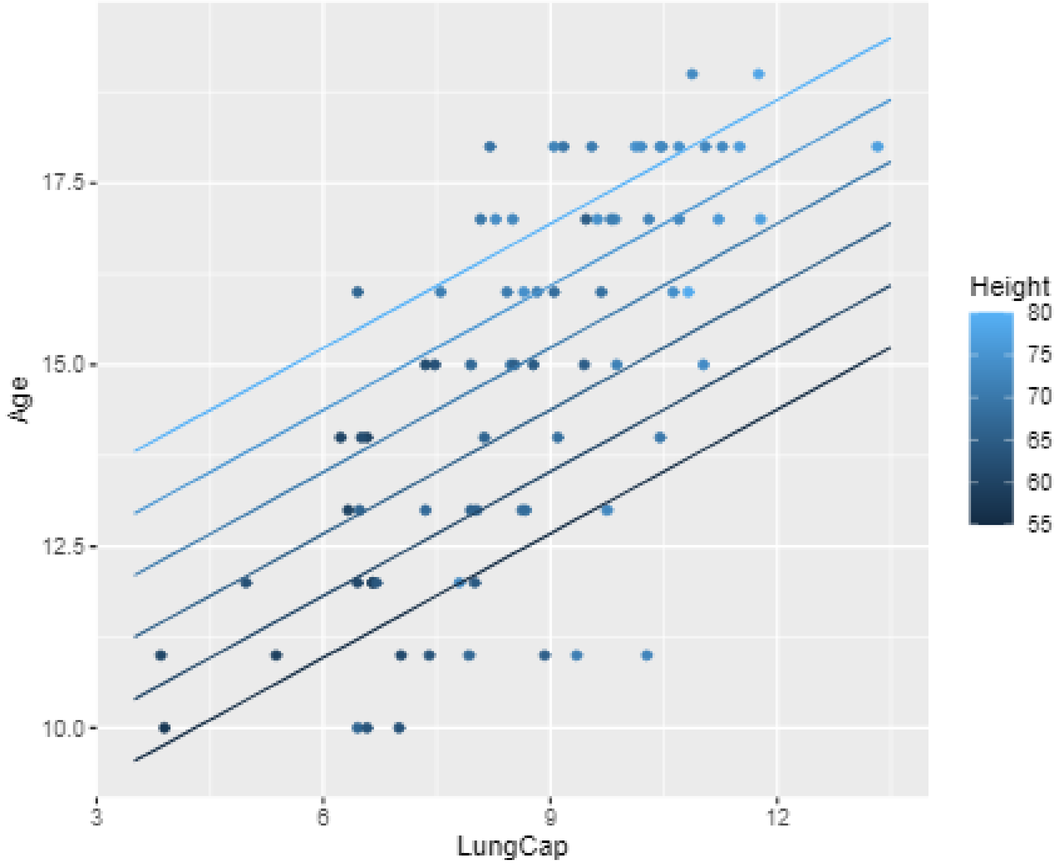
```
> ggPredict(t,interactive = TRUE)
> t2<-lm(Age~LungCap+Height,data = S_yes)
> summary(t2)

Call:
lm(formula = Age ~ LungCap + Height, data = S_yes)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.3342 -1.0042  0.3207  1.3250  3.3930 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -1.82611    3.86765  -0.472   0.6382  
LungCap      0.56948    0.21732   2.620   0.0107 *
Height       0.17048    0.07795   2.187   0.0319 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.854 on 74 degrees of freedom
Multiple R-squared:  0.5175,	Adjusted R-squared:  0.5045 
F-statistic: 39.69 on 2 and 74 DF,  p-value: 1.945e-12

> 
```

We can see that we get a summary of a residuals or errors, estimate of the intercept, standard errors, test statistic and p value.

Residual standard error value tells us about the measure of the variance of observations around the regression line. We get the r squared and the adjusted r square value.

The R square gives us a measure of what percent of the variance in the response variable can be explained by the regression.

We created a scatterplot with multiple regression for the same model of Smokers.



We can see that with the increase in age there is a very slight increase in the lung capacity for smokers and the height stops at a particular age. There is a lot of variance in the scatter plot for Lung capacity as it does not have a strong relation with age and it is very scattered.

References:

- Dummy variables in Regression. (n.d.). Retrieved from Stat Trek: https://stattrek.com/multiple-regression/dummy-variables.aspx

- gallo, A. (2015, November 4). A Refresher on Regression Analysis. Retrieved from Harvard Business Review: https://hbr.org/2015/11/a-refresher-on-regression-analysis

- Regression Analysis. (n.d.). Retrieved from Statistical tools for high-throughput: http://www.sthda.com/english/articles/40-regression-analysis/163-regression-with-categorical-variables-dummy-coding-essentials-in-r/

- Working with Dummy variable. (n.d.). Retrieved from Princeton University Library: https://dss.princeton.edu/online_help/analysis/dummy_variables.htm