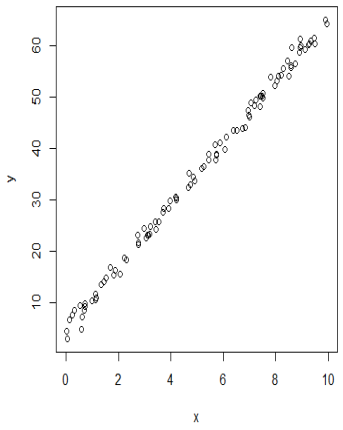
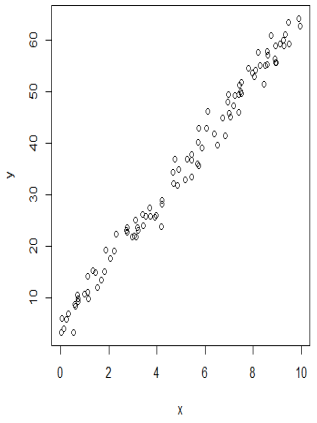
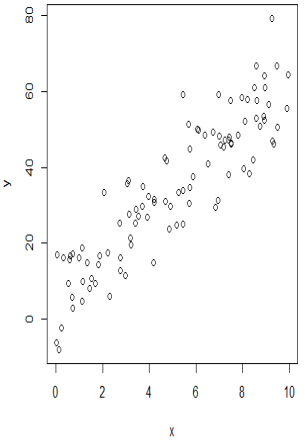
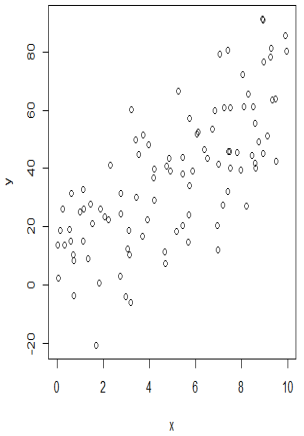
Part 1

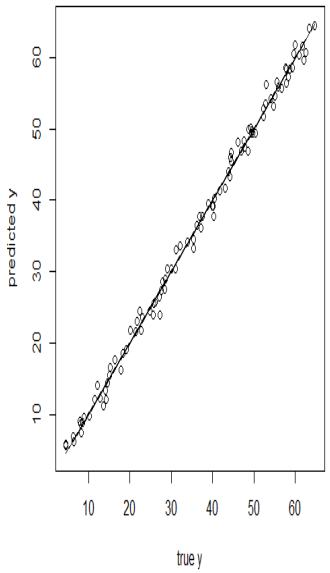
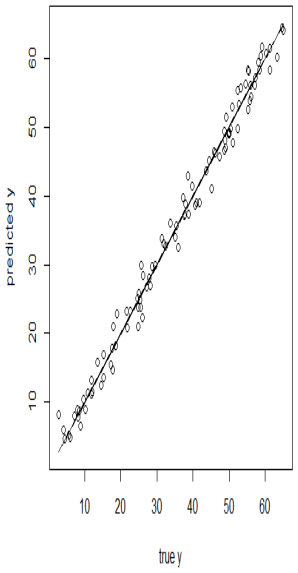
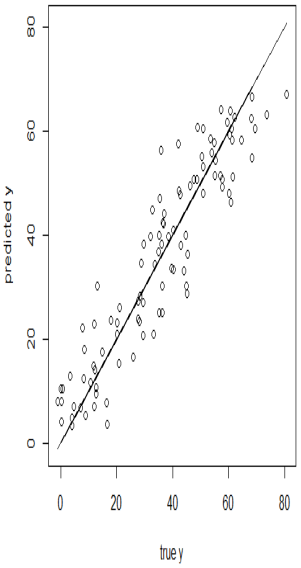
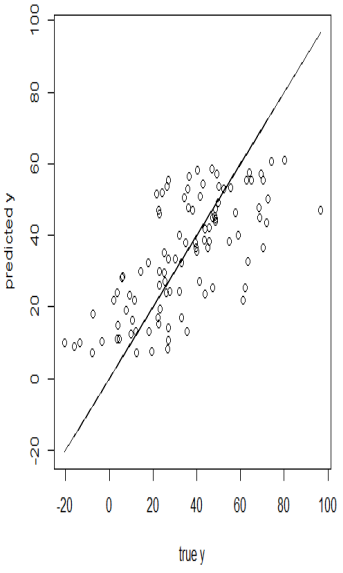
1) the changing of standard deviation makes the data more scattered and that happened because standard deviation measures the spread of data distribution. It measures the typical distance between each data point and the mean. and rnorm(100) gives you a random sample of 100 values from distribution mean = 0 and sd = 1 and each time changing standard deviation will give you a random sample approximately differ from mean by the value of standard deviation. and each time we increase standard deviation, it will add more value(noise) in Y function which causes scattered data.Increasing standard deviation causes increasing the distance between each data point and the mean increases and vice versa.

std =1	std = 2	std = 8	std =15
			

2) OLS trys to calculate coefficients in way to minimize the residual value of each actual y and the predicted y
and the both coefficients b0,b1 are positively correlated to y in all cases of std
In case std = 1 the b0 and b1 is very significant parameters as p value is so small
in case std = 2 the b0,b1 is still very significant parameter
in case std = 8 b0 is not very signift but b1 is very significant
in case std = 15 b0 is not very signift but b1 is very significant
increasing standard deviation effect significantly of parameter b0

std =1 Coefficients: Pr(>|t|) (Intercept) 5.04144 <2e-16 *** x 6.01786 <2e-16 ***	std = 2 Coefficients: Pr(>|t|) (Intercept) 4.97805 <2e-16 *** x 6.05416 <2e-16 ***
std = 8 Coefficients: Pr(>|t|) (Intercept) 2.3163 0.184 x 6.2568 <2e-16 ***	std =15 Coefficients: Pr(>|t|) (Intercept) 5.6653 0.0785 . x 6.2583 <2e-16 ***

3) The smaller is standard deviation , the R square is optimized on the other hand the std when it gets larger, adds more randomness and noises that cause no specific pattern in the prediction of y so it causes more residual values and less R squared value.

std= 1, R-sqr of 0.995581	std = 2 ,R-sqr of 0.9881054	std = 8 , R-sqr of 0.7889698	std =15, R-sqr of 0.5325709
			

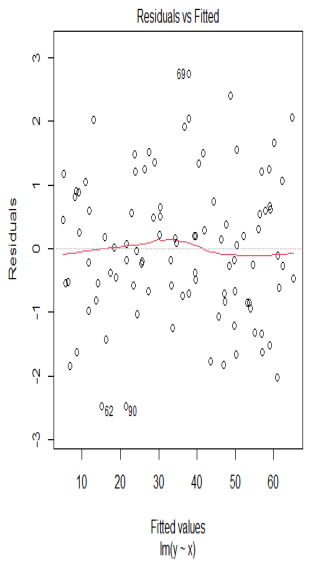
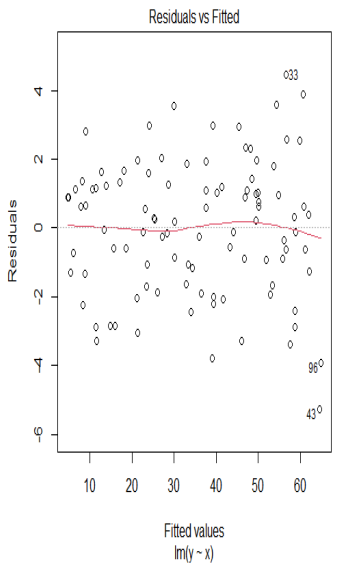
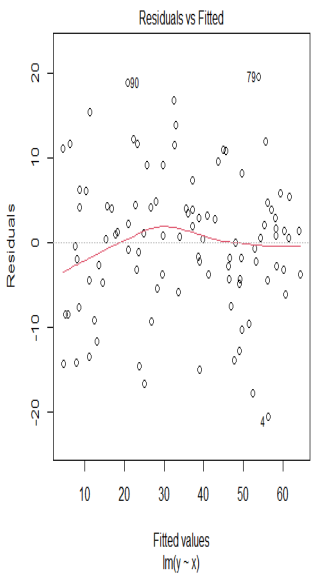
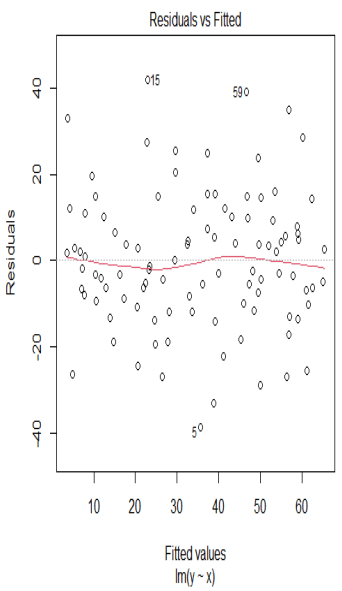
"The residual plot with no pattern is good because it suggests that our use of a linear model is appropriate."

4) as shown in the figure each fitted value have a residual value in y-axis as $Residual = y - \hat{y}$ so in case of std= 1 the range of residuals are 3 to -3 from 0 which is the best case the residuals equal zero ,the red line is close to dash line which indicate that ~ The mean of the errors will tend to zero also the residuals has no pattern and has Residual standard error: **0.9372 so it is good residual plot**

   in case of std= 2 the range of residuals start to be from 4 to -4 which is mean the worst error from predict (fitted)value and the actual y is 4 , the red line is close to dash line also the residuals has no pattern and has Residual standard error: **1.683 so it is good residual plot**

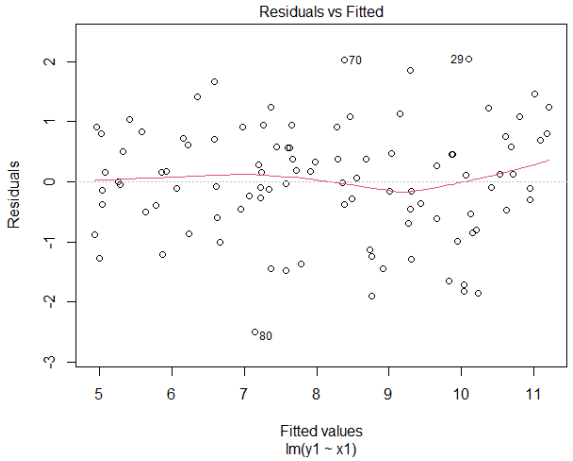
   in case of std=8 the range is 20 to -20, the red line is a little far from dash line which indicate that also the residuals has no pattern and has Residual standard error: **8.543 so it is good residual plot, not the best case**

   in case of std=15 the range is 40 to -40 , the red line is a little far from dash line also the residuals has no pattern and has Residual standard error: **14.7 so it is good residual plot ,not the best case**

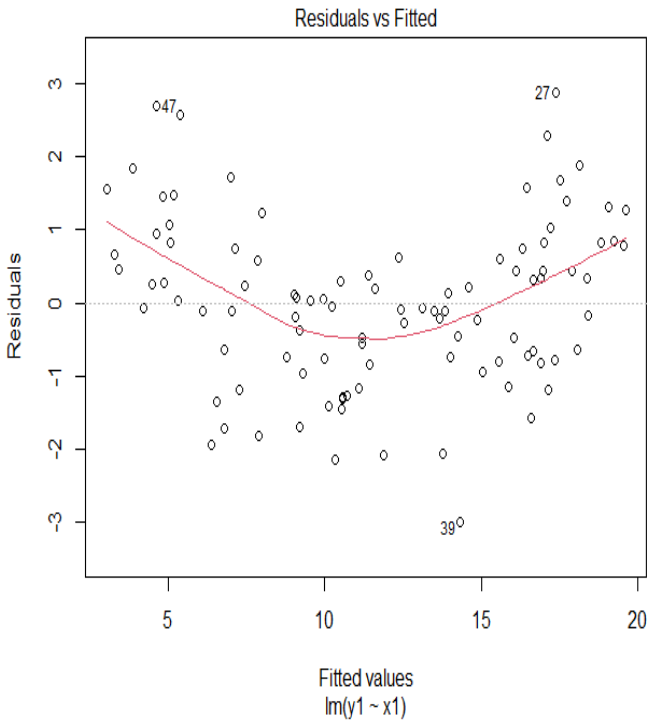
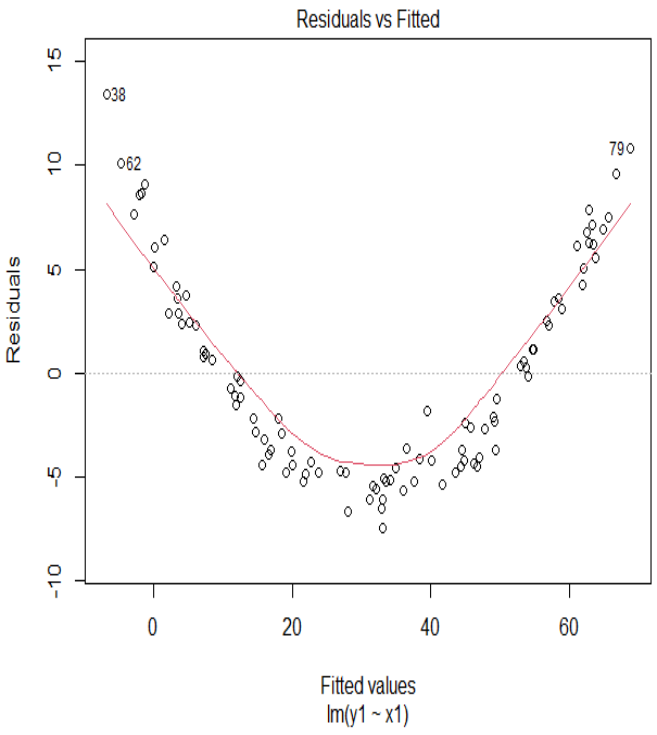
std =1	std = 2	std = 8	std =15
			

Part 2

1) yes, it is good residual plot because there is no pattern

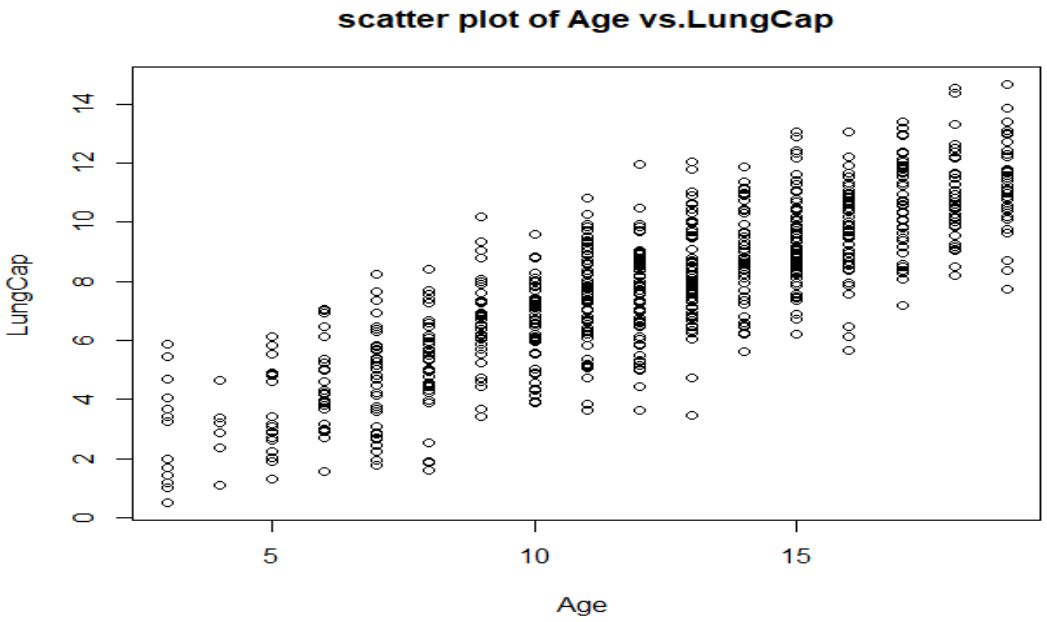


2) When we increase the nonlinear term instead of 0.1 to 10 the residual has a slight pattern but when we increase it to 70 it starts to have a very obvious pattern [u shape] so it is a bad residual plot.

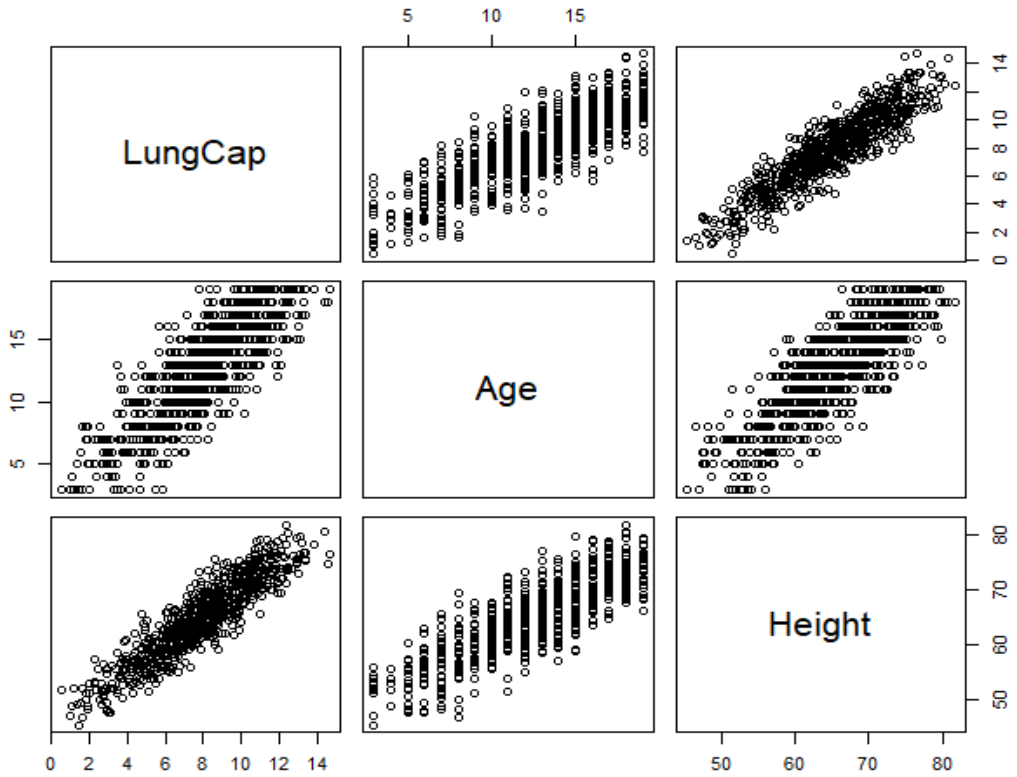
y = 5 + 6*x1 + 10*x1*x1 + rnorm(100)	y = 5 + 6*x1 + 70*x1*x1 + rnorm(100)
	

Part 3

1) LungCap ,Age Height ,Smoke ,Gender ,Caesarean



2) the pairs of lung cap, Age, Height



3) correlation between lung cap, Age, Height

	LungCap	Age	Height
LungCap	1.0000000	0.8196749	0.9121873
Age	0.8196749	1.0000000	0.8357368
Height	**0.9121873**	0.8357368	1.0000000

4) Height

5) Yes,It might be due to the increased surface area of the lungs in relation with increasing height. taller people tend to have larger chests and hence larger total lung capacities and as a short person, I have many breathing problems so I can relate :)

6) summary

```
Call:
lm(formula = y1 ~ x1 + x2 + X3 + x4 + x5)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2094 -0.6829 -0.0005  0.7190  3.1018

Coefficients: (2 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.84527    0.50269 -21.575  < 2e-16 ***
x1            0.15129    0.01826   8.287 5.66e-16 ***
x2            0.26287    0.01025  25.646  < 2e-16 ***
X3                 NA         NA      NA       NA
x4                 NA         NA      NA       NA
x5           -0.40507    0.08106  -4.997 7.32e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.039 on 721 degrees of freedom
Multiple R-squared:  0.8482,  Adjusted R-squared:  0.8476
F-statistic:  1343 on 3 and 721 DF,  p-value: < 2.2e-16
```

7) R-squared: 0.8476 it indicates that ~ 85% of data fit the regression model.

8) it shows that only height , age , gender only effect model in that case we should remove X3, X4 which is smoke and caesarean

	Estimate	Pr(>|t|)
(Intercept)	-10.84527	< 2e-16 ***
x1	0.15129	5.66e-16 ***
x2	0.26287	< 2e-16 ***
x3	NA	NA
x4	NA	NA
x5	-0.40507	7.32e-07 ***

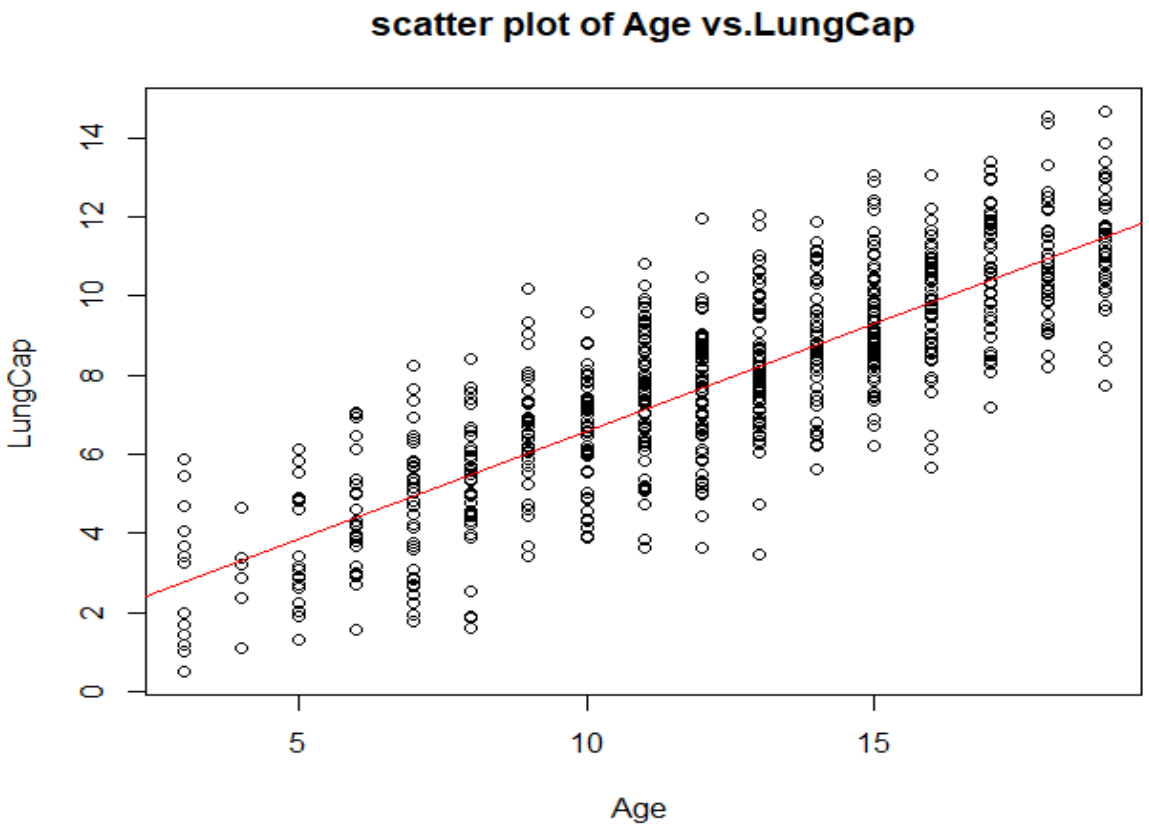
9) we first convert categorical to numerical 0,1
then built model as lm(Y1~X1+X2+X3+X4+X5) the variable "X1,X2,X3,X4,X5" is taken as (independent variable), and the variable "Y1" is (dependent variable).

So, for lm(V~U), abline(lm) will plot the fitted V-values
(y-axis) against the U-values (x-axis) but in our case the Y1-values depend on more than one variable not only age but [height , age and gender] so the line will not be displayed on the plot.as the plot considered only age

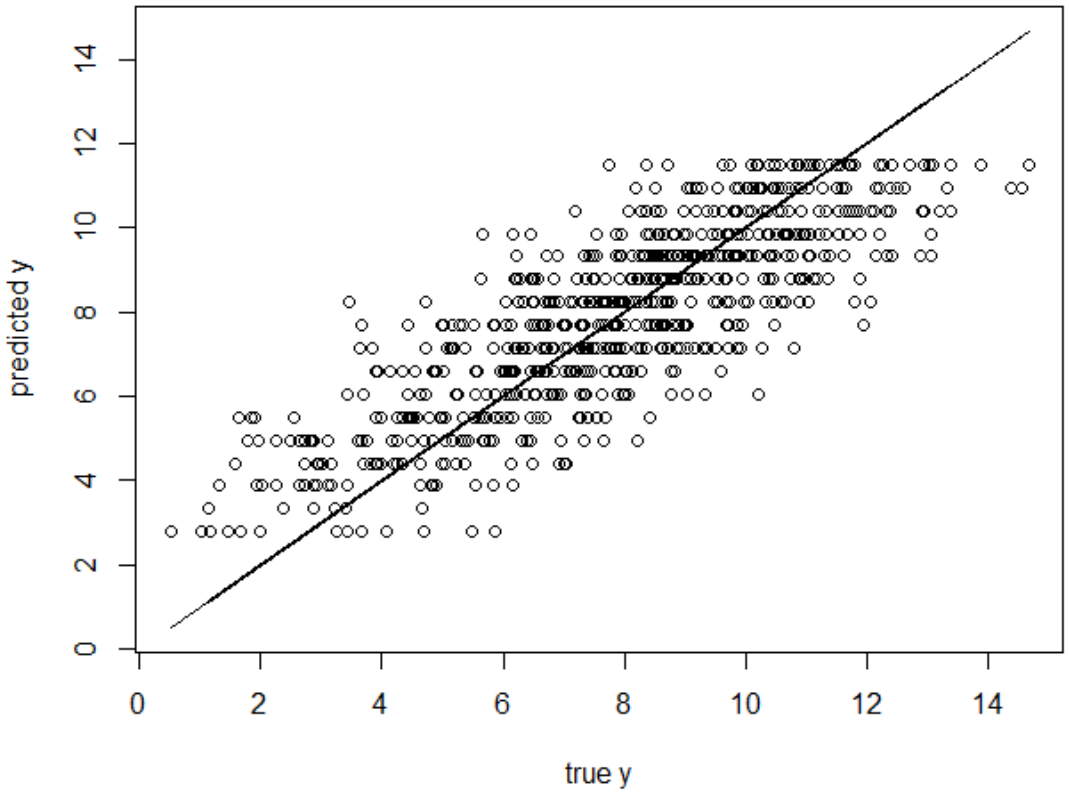
10) it shows that only variable Age which make sense and effect model

	Estimate	Pr(>|t|)
(Intercept)	1.14686	7.06e-10 ***
x1	0.54485	< 2e-16 ***
x3	NA	NA
x4	NA	NA

11) Now the line displayed on the plot as it is the only variable LungCap in this model depends on it.



12) we draw predicted values of LungCap vs. Actual Value Of LungCap.



$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

13) is 1.523824 so **MSE= 2.322038**