Task 1

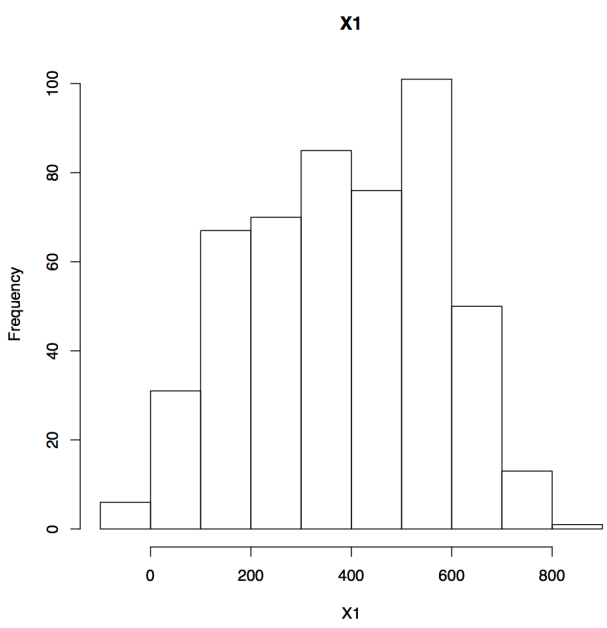
Calculated means and variances are shown in below table.

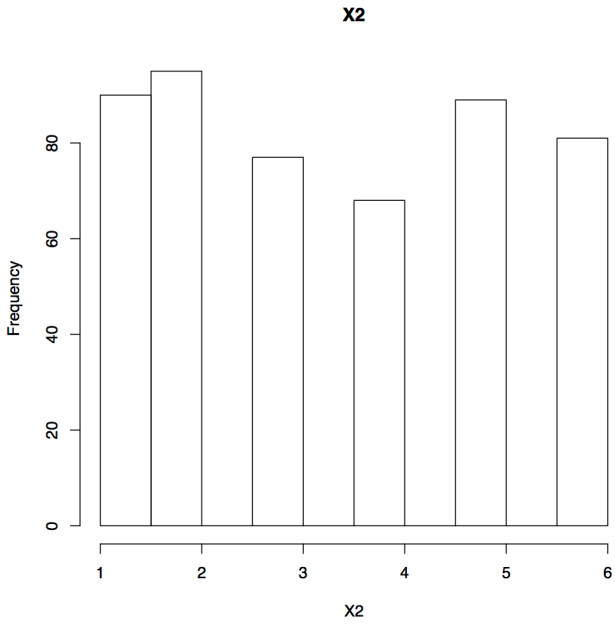
Mean					
X_1	X_2	X_3	X_4	X_5	Y
383.3480653	3.428	0.01886047	15.08967333	0.29943235	883.9504519

Variance					
X_1	X_2	X_3	X_4	X_5	Y
35375.04	3.04	3.28E-04	405.75	0.56	155848.20

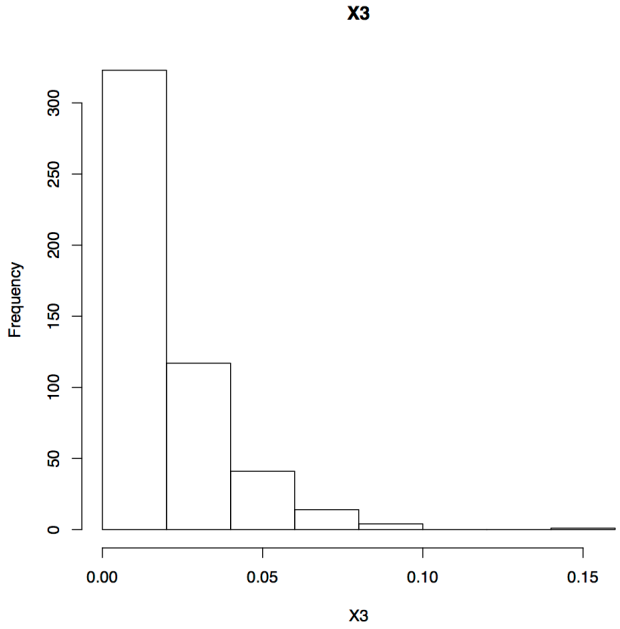
Here X_3 has very small values and hence it's variance is small.

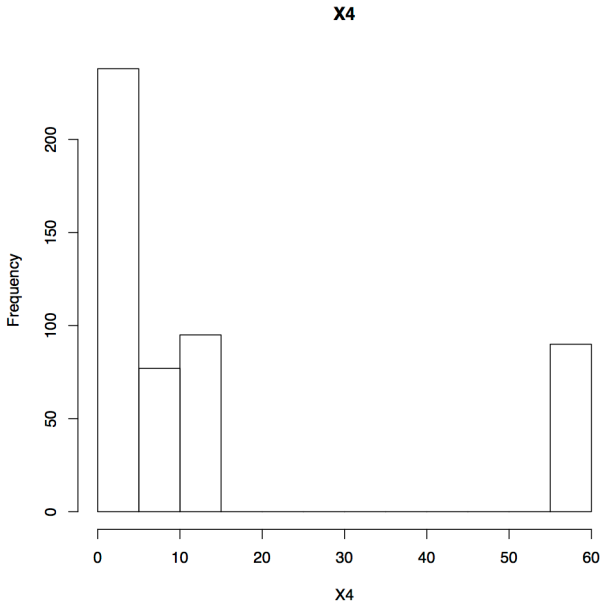
Histograms :



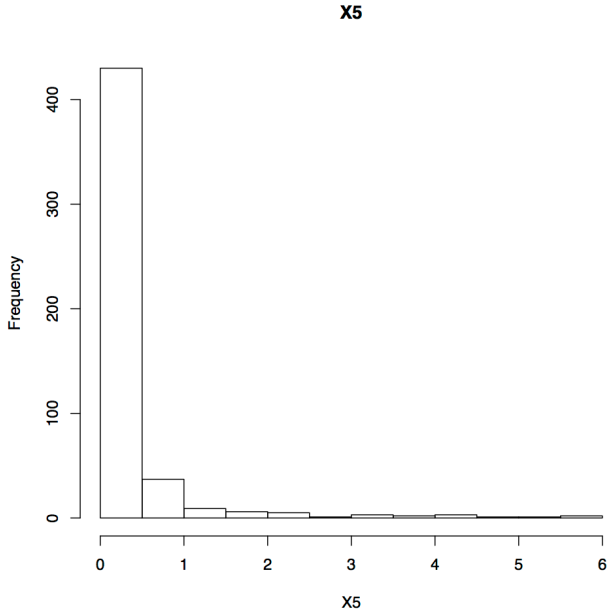


Values of X_1 are distributes on very large interval whereas variable X_2 takes only six values which are 1 to 6.





We can see that X_3 takes very small values. Most of the values are less than 0.05 which explains it's small mean and variance.



Values of X_5 are of similar nature to values of X_3. Most of them are less than 0.5.

CORRELATION MATRIX :

From above correlation matrix we can say that variables ***X_1 and Y*** are positively correlated because their correlation is ***0.9.*** Which means that if X_1 increases Y also increases. So we can say that X_1 contributes more in predicting values of Y. Whereas X_3 and Y have correlation of -0.03 which means that X_3 does not contribute or contributes very little amount in predicting values of Y.

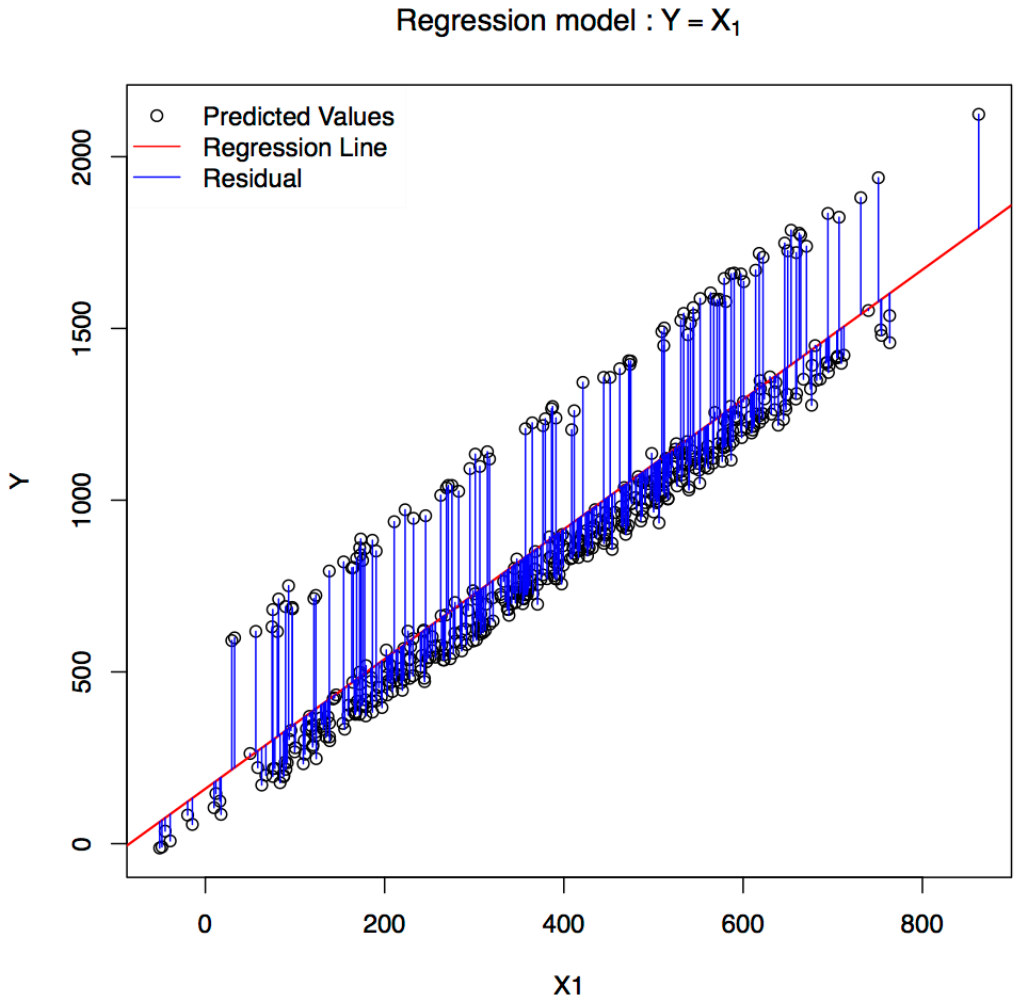
Also we can say that predictor variables do not have much of correlation among themselves.

Task 2

Results for regression model : $Y = a0 + a1X + \varepsilon$, where $X = X1$
Here is the output from ***R*** :

```
Call:
lm(formula = Y ~ X1)

Residuals:
   Min     1Q  Median     3Q    Max
-181.83 -102.08  -73.35   -8.70  415.74

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  159.781     17.503   9.129   <2e-16 ***
X1             1.889      0.041  46.078   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 172.2 on 498 degrees of freedom
Multiple R-squared:   0.81,    Adjusted R-squared:  0.8096
F-statistic:  2123 on 1 and 498 DF,  p-value: < 2.2e-16

[1] "Variance of Predicted Value(Y-cap)"
[1] 126238.2
```

Values of coefficients are made bold above. We can see that coefficient of X_1 has *p-value* almost 0. Which means that we reject H_0 and we can say that coefficient is significant also intercept is meaningful. To summarize,
$a_0 = 159.781$, $a_1 = 1.889$ and $s^2 = 126238.2$
Here is the plot with observed values, fitted values and regression lines. Blue lines show residuals.

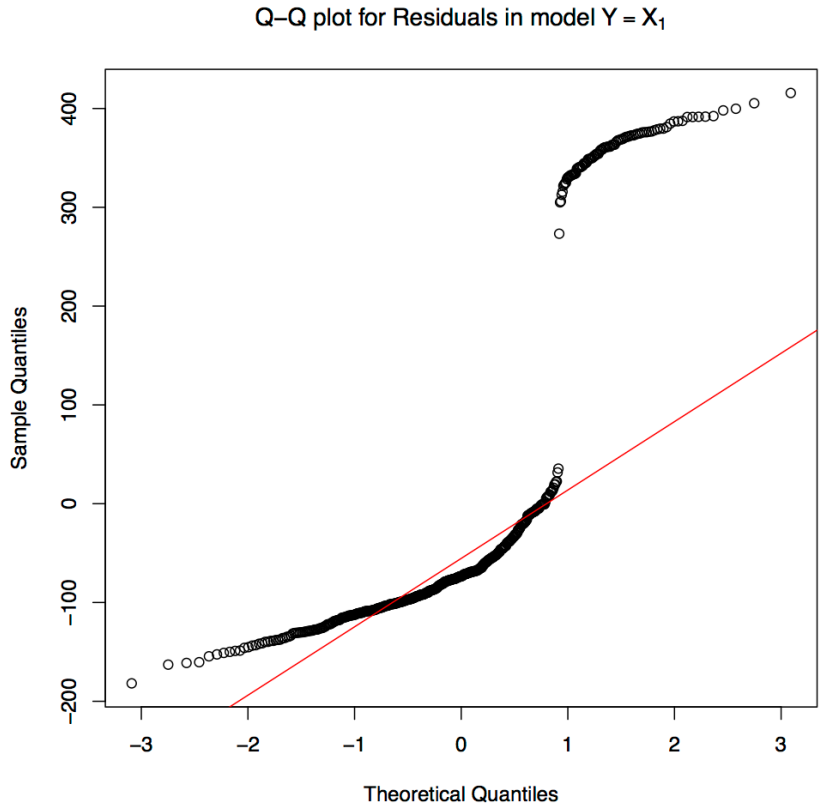


Residual Analysis :
We check for normality of residuals and we do this by looking at Q-Q Plot and also by doing Pearson's Chi-Square test.
For Chi-Square test we use function pearson.test() from the package 'nortest'.
Output of R :

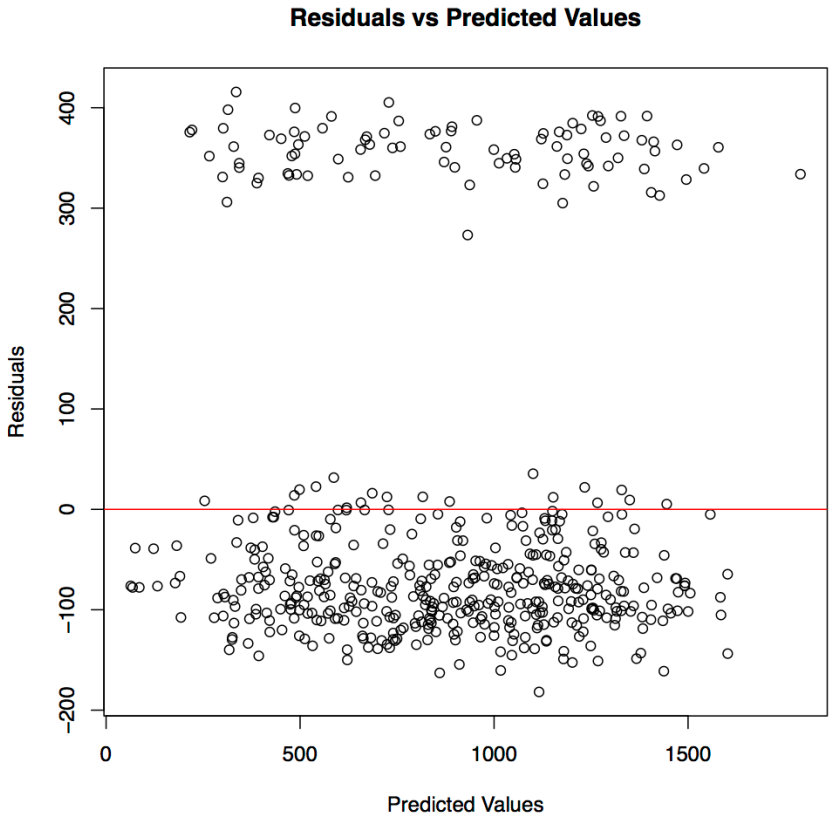
Pearson chi-square normality test
data: x1resid
P = 1080.4, p-value < 2.2e-16

Here *x1resid* is residuals data vector. Here Null Hypothesis H_0 : Distribution is Normal. But here *p-value* suggests that we should reject the Null Hypothesis of distribution being normal. We can see this result also in Q-Q Plot.

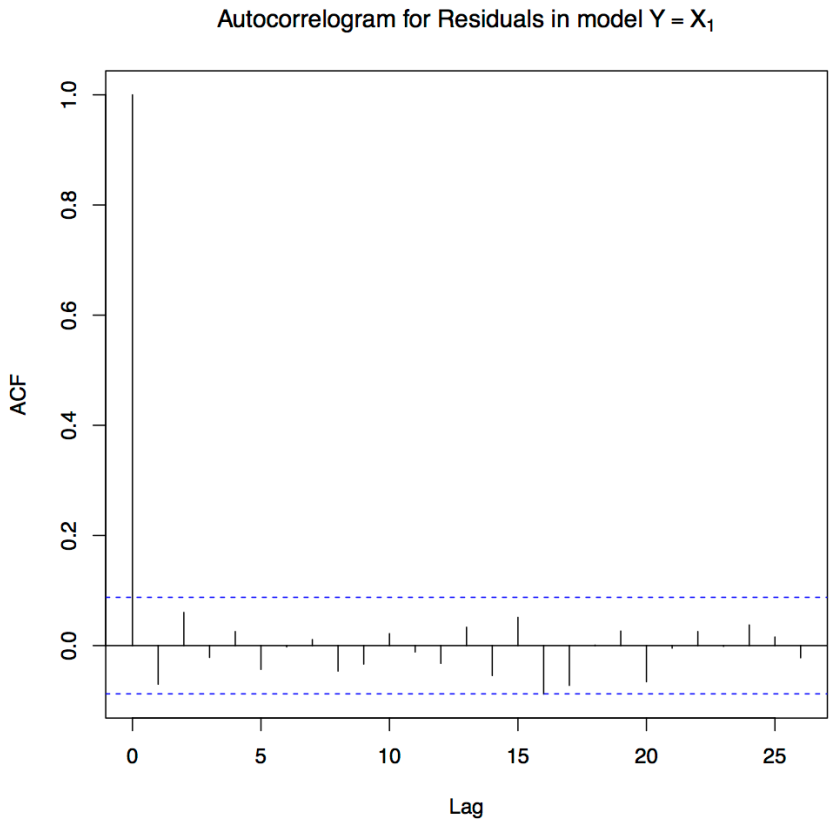
QQ Plot :



So we can say that residuals are not normally distributed.
For the correlation trends in residuals we plot scatter plot.



There don't seem any trends but this plot suggests that our *model is not a good fit.* To verify correlation we can also plot the autocorrelogram.

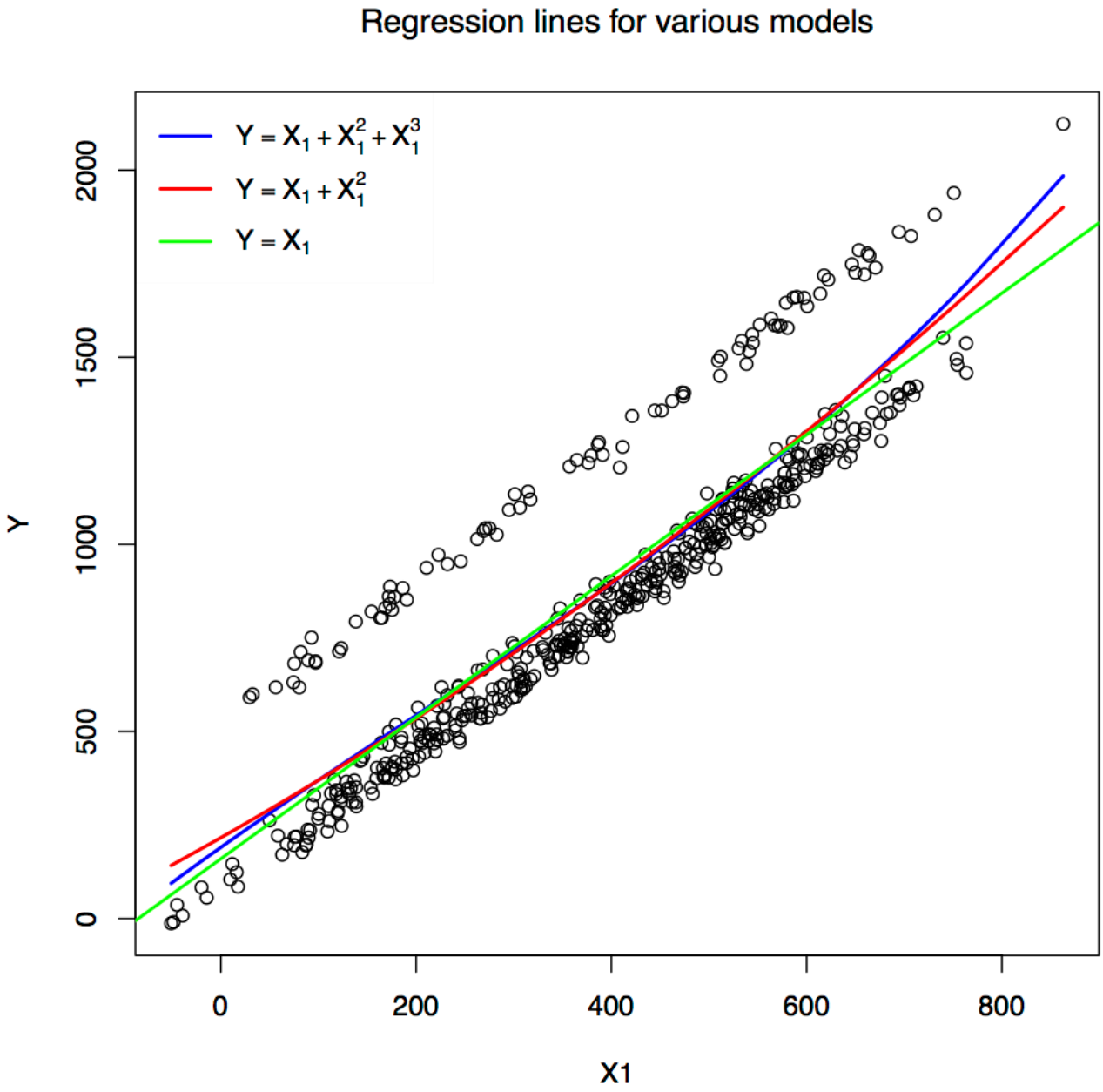


Higher Order polynomials :

Here we apply higher order polynomial to the degree 2 and 3.
Output of R :

```
lm(formula = Y ~ X1 + I(X1^2))

Residuals:
    Min      1Q  Median      3Q     Max
-207.45 -102.52  -69.71   -4.88  421.80

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.161e+02  2.756e+01   7.841 2.74e-14 ***
X1          1.484e+00  1.592e-01   9.320  < 2e-16 ***
I(X1^2)     5.438e-04  2.064e-04   2.635  0.00869 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 171.2 on 497 degrees of freedom
Multiple R-squared:  0.8126,  Adjusted R-squared:  0.8119
F-statistic:  1078 on 2 and 497 DF,  p-value: < 2.2e-16

[1] "Variance of Predicted Value(Y-cap)"
[1] 126646.1

Call:
lm(formula = Y ~ X1 + I(X1^2) + I(X1^3))

Residuals:
```

```
   Min      1Q  Median      3Q     Max
-238.76 -102.94  -71.22   -6.04  415.86

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.902e+02  3.494e+01   5.446 8.14e-08 ***
X1           1.850e+00  3.438e-01   5.381 1.15e-07 ***
I(X1^2)     -6.382e-04  1.005e-03  -0.635    0.526
I(X1^3)      1.048e-06  8.716e-07   1.202    0.230
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 171.2 on 496 degrees of freedom
Multiple R-squared:  0.8132,  Adjusted R-squared:  0.812
F-statistic: 719.6 on 3 and 496 DF,  p-value: < 2.2e-16

[1] "Variance of Predicted Value(Y-cap)"
[1] 126730.9
```



From above results we can say that higher order polynomial fits as good as simple linear polynomial. But there is not any drastic improvement. We can see that in simple linear polynomial ***Residual standard error*** was ***172.2*** and in polynomial cases it is ***171.2.*** But at the same time value of ***F-statistic*** has decreased. From the plot we can observe that there is not any significant *curve* or any kind of curvature and that is the reason that polynomial models do not show drastic improvement. To summarize the results,

Model	Residual Std. Error	R-Squared	Adjusted R-Squared	F-Stat	Variance of predicted Y
$Y = X_1$	172.2	0.81	0.8096	2123	126238.2
$Y = X_1 + X_1^2$	171.2	0.8126	0.8119	1078	126646.1
$Y = X_1 + X_1^2 + X_1^3$	171.2	0.8132	0.812	719.6	126730.9

Task 3

We first carry out multiple regression of the form : Y = X1 + X2 + X3 + X4 + X5 and results are shown as the output of R :

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5)

Residuals:
   Min     1Q  Median     3Q    Max
-82.753 -17.135   1.289  17.695  72.192

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.870218   5.670489  -2.093 0.036829 *
X1            1.881369   0.006486 290.057  < 2e-16 ***
X2           10.296014   1.088122   9.462  < 2e-16 ***
X3          -11.126282  64.470362  -0.173 0.863052
X4            9.122148   0.094143  96.897  < 2e-16 ***
X5            6.233690   1.628435   3.828 0.000146 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.04 on 494 degrees of freedom
Multiple R-squared:  0.9957,  Adjusted R-squared:  0.9956
F-statistic: 2.284e+04 on 5 and 494 DF,  p-value: < 2.2e-16

[1] "Variance of Predicted Value(Y-cap)"
[1] 155176.8
```

We can see that *p-value* for the coefficient of X3 is 0.86 which tells us that we accept the Null Hypothesis that this coefficient is 0 i.e. X3 does not contribute to the prediction of Y. Which we saw from correlation matrix too.
So we again fit the model without variable X_3 and results are shown below,

---

```
Call:
lm(formula = Y ~ X1 + X2 + X4 + X5)

Residuals:
   Min     1Q  Median     3Q    Max
-82.561 -17.167   1.427  17.794  72.325

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.064209   5.552510  -2.173 0.030273 *
X1            1.881409   0.006476 290.535  < 2e-16 ***
X2           10.289022   1.086302   9.472  < 2e-16 ***
X4            9.121705   0.094016  97.023  < 2e-16 ***
X5            6.231686   1.626797   3.831 0.000144 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.02 on 495 degrees of freedom
Multiple R-squared:  0.9957,  Adjusted R-squared:  0.9957
```

F-statistic: 2.86e+04 on 4 and 495 DF, p-value: < 2.2e-16

[1] "Variance of Predicted Value(Y-cap)"
[1] 155176.8

We can see that in our second model ***Residual standard error is 26.02*** which is better than the model with the variable X_3. There is not much difference in other statistics such as R-squared and F-statistics.

We do residual analysis for both the models and results are shown below.

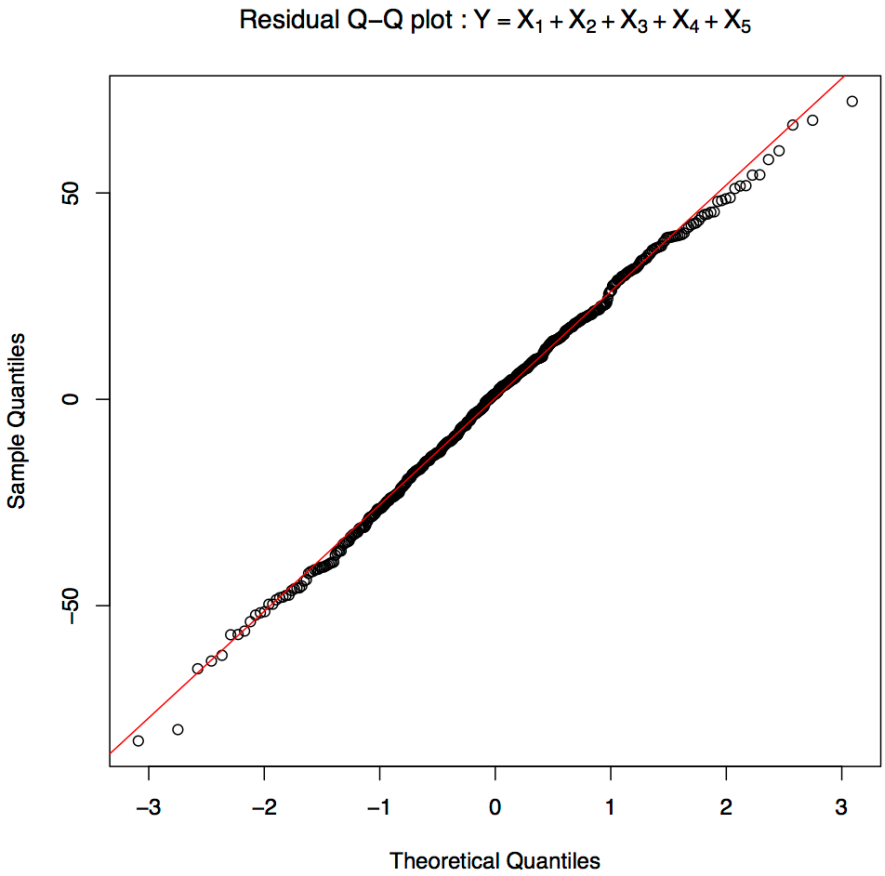
Pearson chi-square normality test

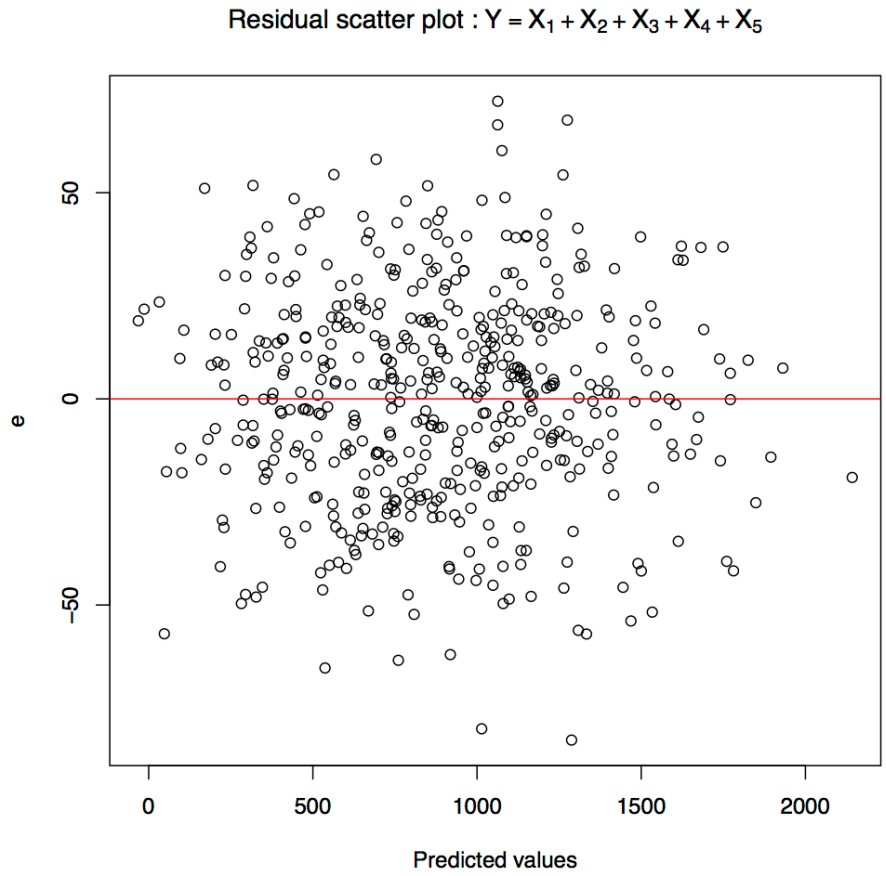
data: mlr_resid (model 1)
P = 21.8, p-value = 0.4719

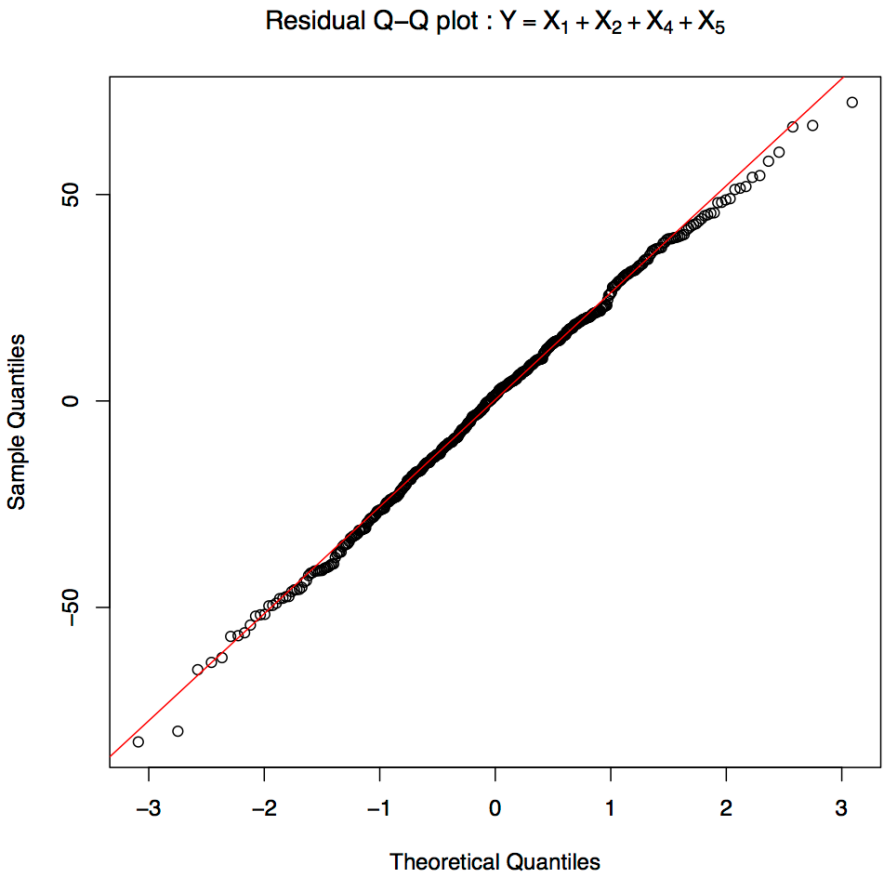
Pearson chi-square normality test

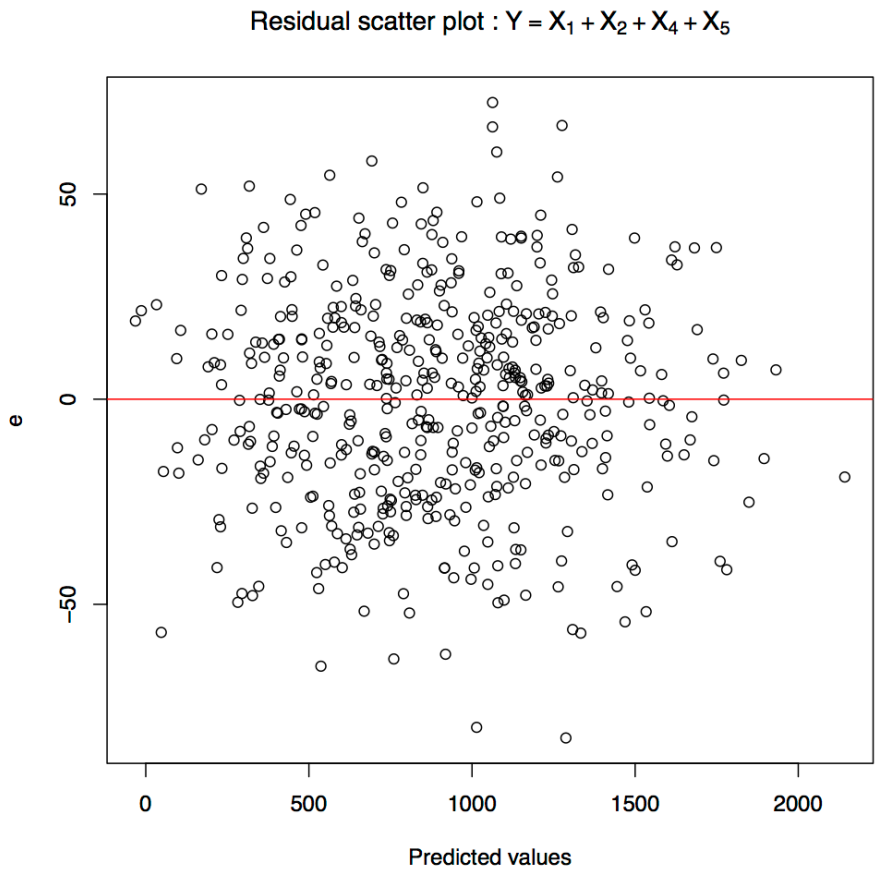
data: i_mlr_resid (model 2 without x3)
P = 27, p-value = 0.2112

Here p-value suggests to accept the null hypothesis which is distribution is normal.




Residual scatter plot : Y = X1 + X2 + X3 + X4 + X5
e
50
0
−50
0
500
1000
1500
2000
Predicted values


Residual Q−Q plot : Y = X1 + X2 + X4 + X5
Sample Quantiles
50
0
−50
−3
−2
−1
0
1
2
3
Theoretical Quantiles



From above graphs we can say that residuals follow normal distribution in both the models and residuals does not have any trend and they show randomness.

To summarize all the results,

Model	Residual Std. Error	R-Squared	Adjusted R-Squared	F-Stat	Variance of predicted Y
$Y = X_1$	172.2	0.81	0.8096	2123	126238.2
$Y = X_1 + X_1^2$	171.2	0.8126	0.8119	1078	126646.1
$Y = X_1 + X_1^2 + X_1^3$	171.2	0.8132	0.812	719.6	126730.9
$Y = X_1 + X_2 + X_3 + X_4 + X_5$	26.04	0.9957	0.9956	2.28E+04	155176.8
$Y = X_1 + X_2 + X_4 + X_5$	26.02	0.9957	0.9957	2.86E+04	155176.8
$Y = X_1 + X_4 + X_5$	28.25	0.9949	0.9949	3.23E+04	155055.1

From the above results we can see that multivariate model has very low residual standard error compared to simple linear and polynomial model.