Lab 4: Linear Regression

Problem statement:

Develop linear regression model (through least square method) on given data set (drug2.csv) as:

- Create a model (A) (Simple linear regression) to predict response with respect to dose.
- Interpret the model summary
- Draw residuals of model to see the normal distribution.
- Improve model (B) by adding more feature (sex) and again investigate residuals graph.
- Validate your model by performing tests (fitted value vs residuals, fitted values vs actual
- Further improve model (C) through moderation i.e. interaction variable and validate model through aforesaid procedure.
- Calculate the RMSE for all models (A, B, C) and represent as a bar chart/histogram.
- Calculate the standard deviation of residuals of all models (A, B, C) and represent as a bar chart/ histogram.

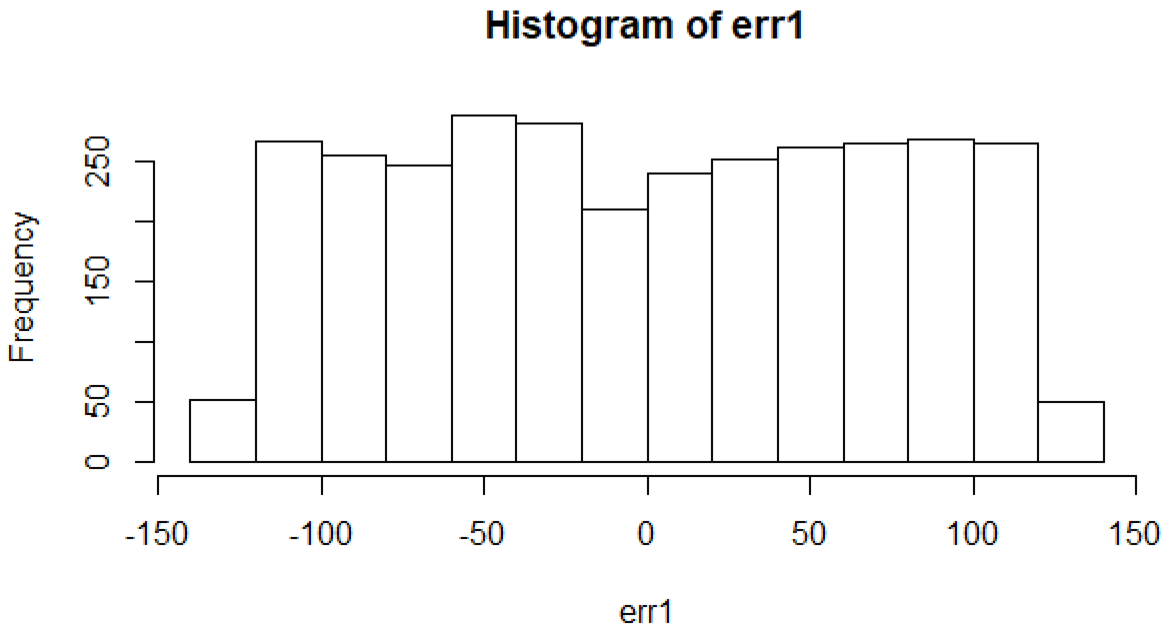
Source Code:

```
#Author: Ashish Upadhyay
#Branch: Computer Science and Engineering
#Semester: 6th
#Dr. SP Mukherjee International Institute of Information Technology, Naya Raipur
#Subject: Machine Learning Lab 4
#Task: Linear Regression Implementation

setwd("C:/Users/Ashish Upadhyay/Documents/Semester6/MachineLearning/Lab")
getwd()
drug = read.csv("drug2.csv")
head(drug)
attach(drug)

#Model A
model1 = lm(response~dose)
summary(model1)
err1 = residuals(model1)
hist(err1)
plot(model1$fitted.values,err1)

#Model B
model2 = lm(response~dose+sex)
summary(model2)
err2 = residuals(model2)
hist(err2)
plot(model2$fitted.values,err2)

#Model C - moderation
product = drug$sex * drug$dose
model3 = lm(drug$response~drug$dose+product+drug$sex)
summary(model3)
err3 = residuals(model3)
hist(err3)
plot(model3$fitted.values,err3)
plot(model3$fitted.values,drug$response)

#RMSE and Strandard Deviation
pred=predict(model3, drug)
actual= drug$response
diff= actual-pred
head(diff)
rmse= sqrt(sum(diff**2)/nrow(drug))
rmse
err4=residuals(model3)
rmse2= sqrt(sum(err4**2)/nrow(drug))
rmse2
```

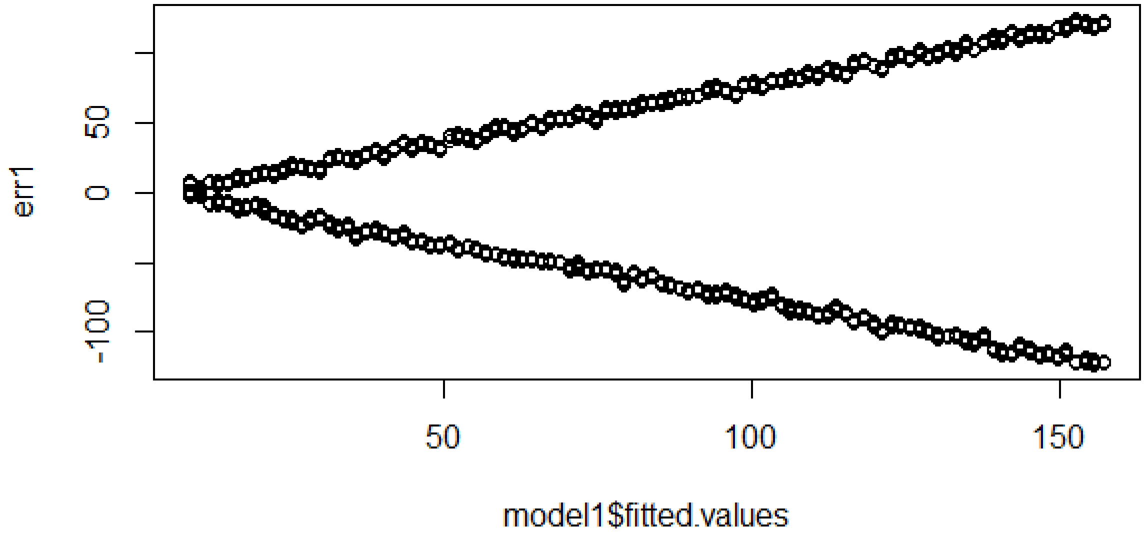
Output:

```
> #Author: Ashish Upadhyay
> #Branch: Computer Science and Engineering
> #Semester: 6th
> #Dr. SP Mukherjee International Institute of Information Technology, Naya Raipur
> #Subject: Machine Learning Lab 4
> #Task: Linear Regression Implementation
>
>
> setwd("C:/Users/Ashish Upadhyay/Documents/Semester6/MachineLearning/Lab")
> getwd()
[1] "C:/Users/Ashish Upadhyay/Documents/Semester6/MachineLearning/Lab"
> drug = read.csv("drug2.csv")
> head(drug)
  sex dose response
1   1  0.1    13.75
2   1  0.2    12.90
3   1  0.3    19.26
4   1  0.4    20.34
5   1  0.5    19.97
6   1  0.6    26.80
> attach(drug)
>
> #Model A
> model1 = lm(response~dose)
> summary(model1)

Call:
lm(formula = response ~ dose)

Residuals:
     Min       1Q   Median       3Q      Max 
-123.514  -62.764    0.401   63.669  124.707 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.2534     2.5778   2.814  0.00493 ** 
dose         15.0020     0.4432  33.852  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

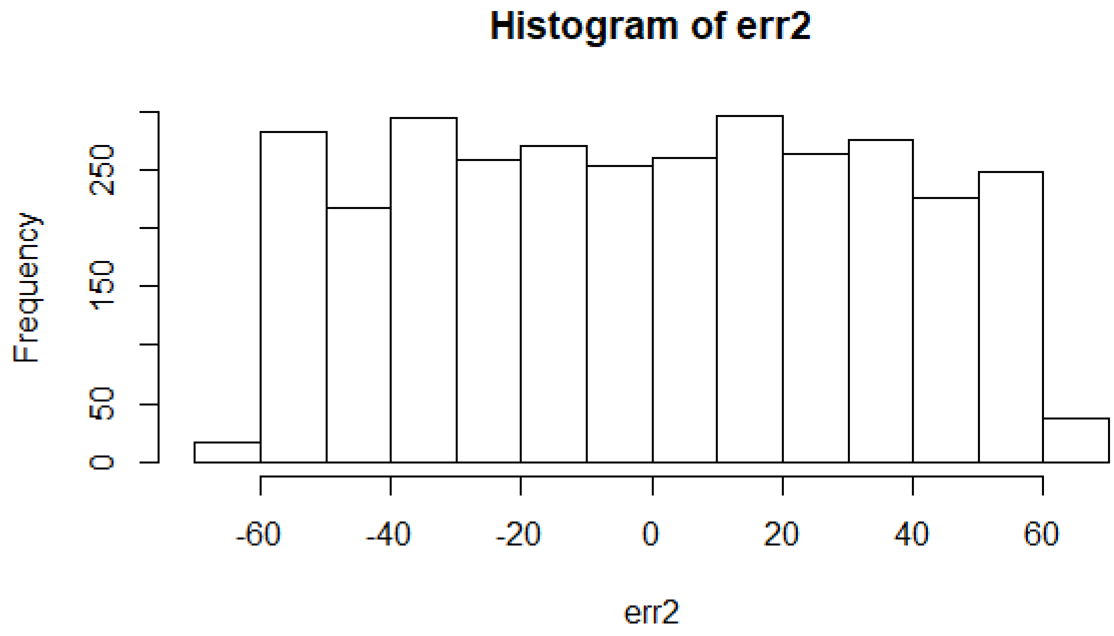
```
Residual standard error: 72.36 on 3198 degrees of freedom
Multiple R-squared:  0.2638,	Adjusted R-squared:  0.2636 
F-statistic:  1146 on 1 and 3198 DF,  p-value: < 2.2e-16

> err1 = residuals(model1)
> hist(err1)
```



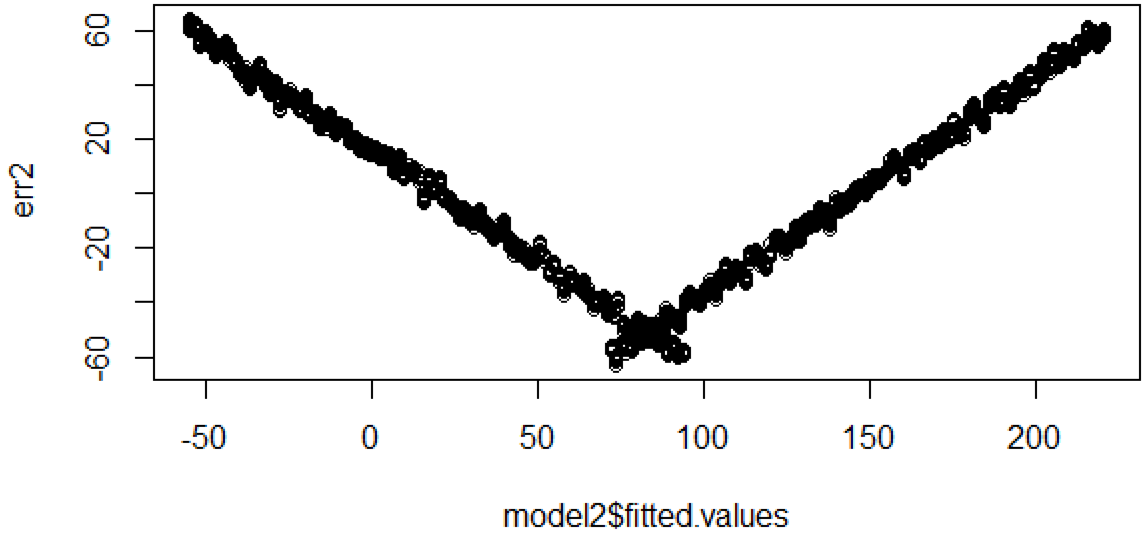
```
> plot(model1$fitted.values,err1)
```



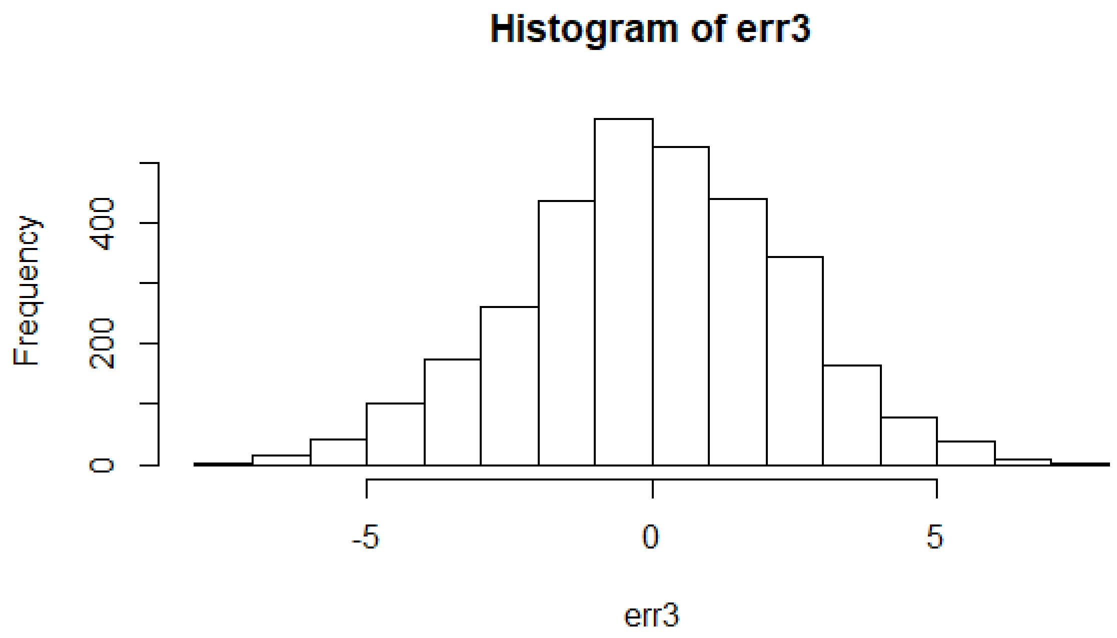
```
> #Model B
> model2 = lm(response~dose+sex)
> summary(model2)

Call:
lm(formula = response ~ dose + sex)

Residuals:
    Min      1Q  Median      3Q     Max 
-62.986 -30.350   0.306  29.360  64.009 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -56.1189     1.3881  -40.43   <2e-16 ***
dose         15.0020     0.2138   70.18   <2e-16 ***
sex         126.7445     1.2341  102.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.91 on 3197 degrees of freedom
Multiple R-squared:  0.8288,	Adjusted R-squared:  0.8286 
F-statistic:  7736 on 2 and 3197 DF,  p-value: < 2.2e-16

> err2 = residuals(model2)
> hist(err2)
```



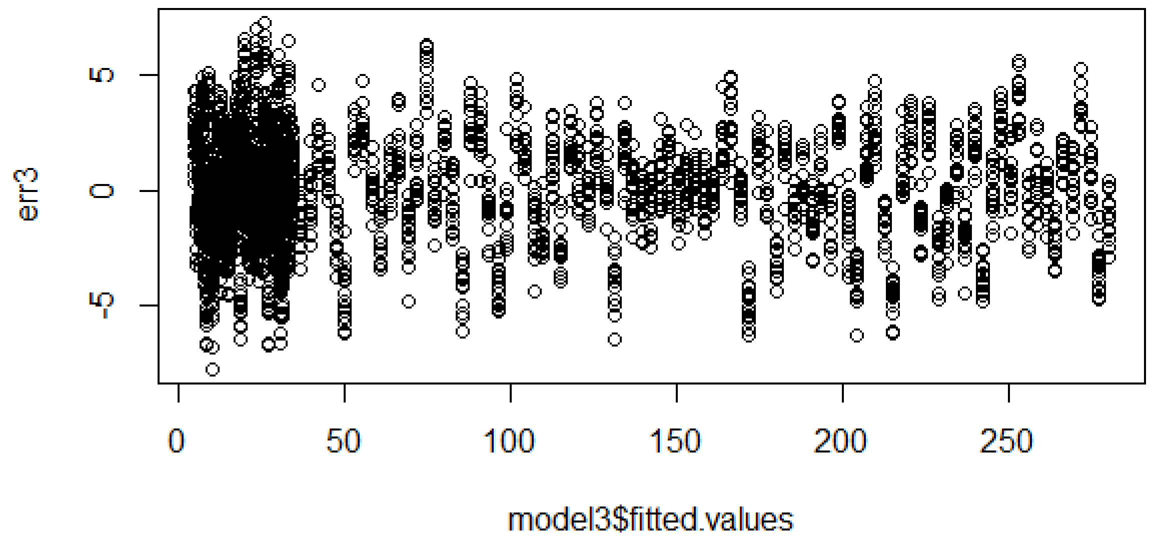
```
> plot(model2$fitted.values,err2)
```



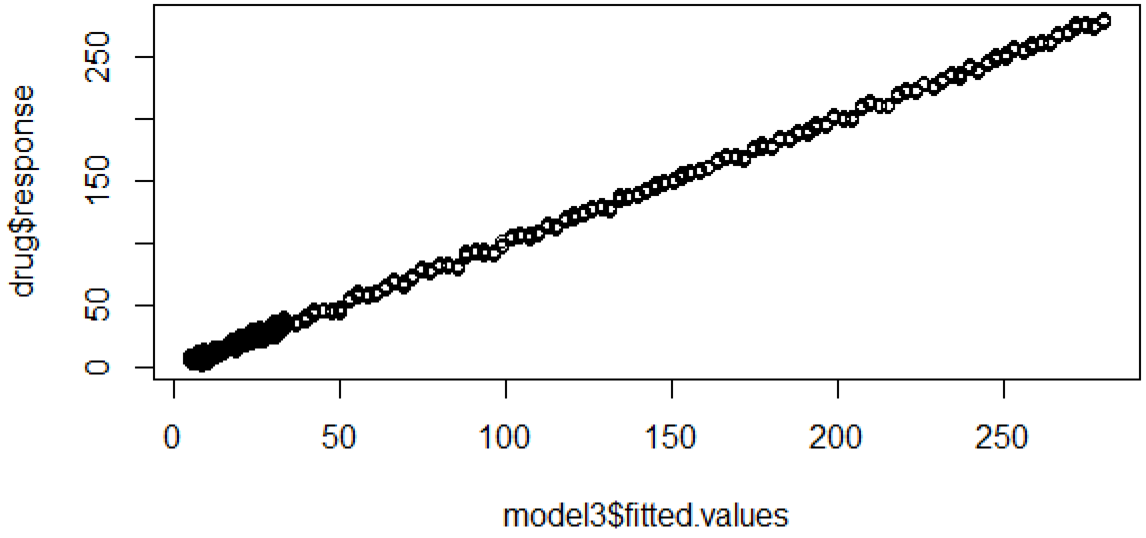
```
> #Model C - moderation
> product = drug$sex * drug$dose
> model3 = lm(drug$response~drug$dose+product+drug$sex)
> summary(model3)

Call:
lm(formula = drug$response ~ drug$dose + product + drug$sex)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.6950 -1.4668 -0.0004  1.5996  7.2181 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.78574    0.11658   41.05   <2e-16 ***
drug$dose    2.94171    0.02004  146.77   <2e-16 ***
product     24.12064    0.02834  850.98   <2e-16 ***
drug$sex     4.93530    0.16487   29.93   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.314 on 3196 degrees of freedom
Multiple R-squared:  0.9992,	Adjusted R-squared:  0.9992 
F-statistic: 1.415e+06 on 3 and 3196 DF,  p-value: < 2.2e-16

> err3 = residuals(model3)
> hist(err3)
```



```
> plot(model3$fitted.values,err3)
```



```
> plot(model3$fitted.values,drug$response)
```



```
> #RMSE and Strandard Deviation
> pred=predict(model3, drug)
> actual= drug$response
> diff= actual-pred
> head(diff)
         1          2          3          4          5          6 
 1.3227276 -2.2335081  1.4202563 -0.2059794 -3.2822150  0.8415493 
> rmse= sqrt(sum(diff**2)/nrow(drug))
> rmse
[1] 2.312749
> err4=residuals(model3)
> rmse2= sqrt(sum(err4**2)/nrow(drug))
> rmse2
[1] 2.312749
```

Lab 5: Polynomial Regression

Problem statement:

Develop linear regression model (through least square method) where "dmf" as dependent variable with respect to various combination of input variable (flor) on given data set (dmf.csv) as:

- Model A (Dependent variable- flor)
- Model B (Dependent variable- flor, square(flor))
- Model C (Dependent variable- flor, square(flor), 1/sqrt(flor))
- Calculate the RMSE for all models (A, B, C) and represent as a bar chart/histogram.
- Calculate the standard deviation of residuals of all models (A, B, C) and represent as a bar chart/ histogram.
- Validate all three models (A, B, C) through various tests and rank models according to efficiency.
- Perform feature engineering through both forward and backward selection methods.
- Declare most perfect model with justification.

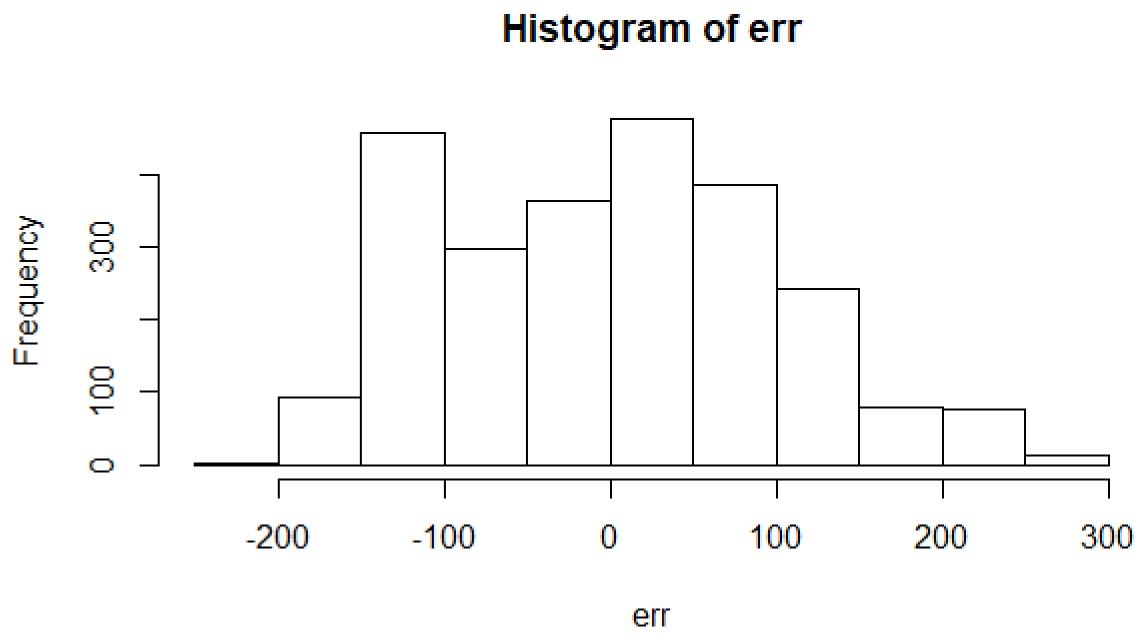
Source Code:

```
#Author: Ashish Upadhyay
#Branch: Computer Science and Engineering
#Semester: 6th
#Dr. SP Mukherjee International Institute of Information Technology, Naya Raipur
#Subject: Machine Learning Lab 5
#Task: Polynomial Regression Implementation

setwd("C:/Users/Ashish Upadhyay/Documents/Semester6/MachineLearning/Lab")
getwd()

dmf = read.csv("dmf.csv")
attach(dmf)

# Basic model: Only flor
model= lm(dmf$dmf~dmf$flor)
summary (model)
err= residuals (model)
hist(err)
plot(model$fitted.values,model$residuals)
plot(model$fitted.values, dmf$dmf)
flor2= dmf$flor^2

# Advance model: flor + flor^2
model2=lm(dmf$dmf~ dmf$flor+flor2)
summary(model2)
err2= residuals(model2)
hist(err2)
plot(model2$fitted.values,model2$residuals)
plot(model2$fitted.values,dmf$dmf)

# More advance model: flor + flor^2 + sqrt(flor)
model3 = lm(dmf$dmf~ dmf$flor+flor2+1/sqrt(flor))
summary(model3)
err3= residuals(model3)
```

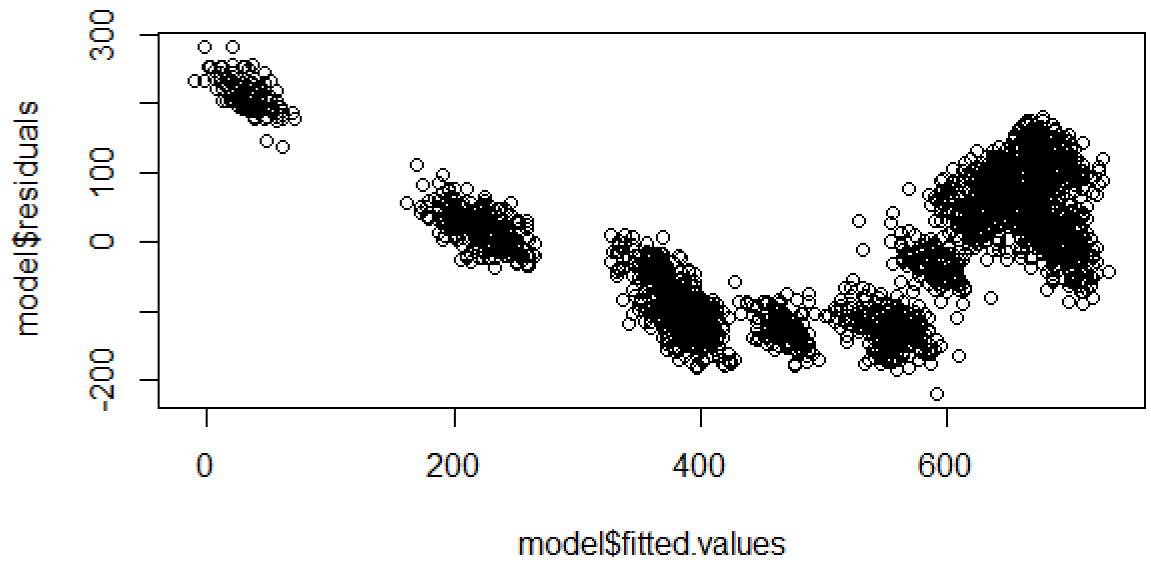
```
hist(err3)
plot(model3$fitted.values,model3$residuals)
plot(model3$fitted.values,dmf$dmf)
```

Output:

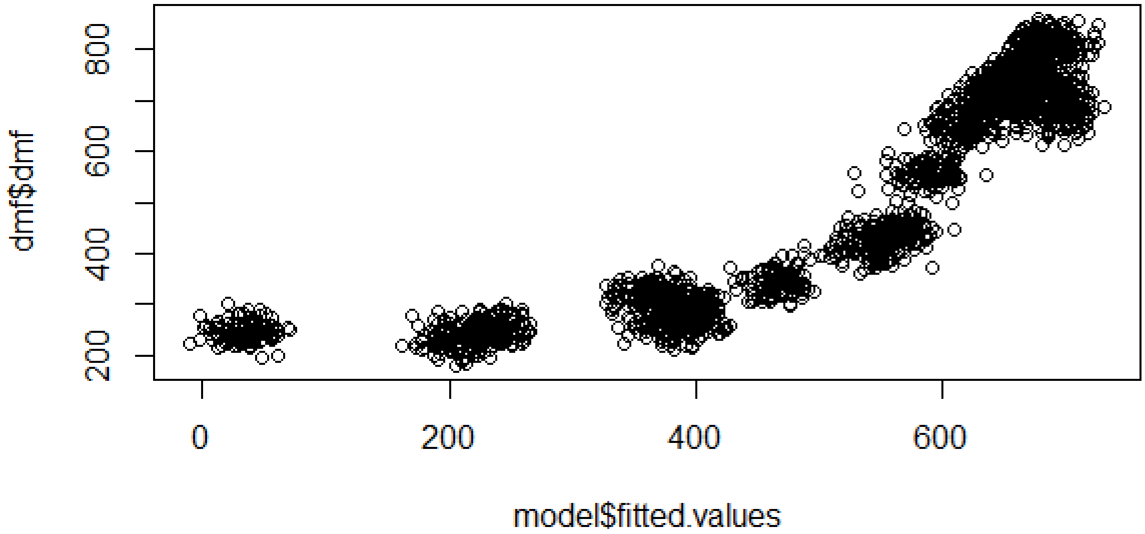
```
> #Author: Ashish Upadhyay
> #Branch: Computer Science and Engineering
> #Semester: 6th
> #Dr. SP Mukherjee International Institute of Information Technology, Naya Raipur
> #Subject: Machine Learning Lab 4
> #Task: Ploynomial Regression Implementation
>
> setwd("C:/Users/Ashish Upadhyay/Documents/Semester6/MachineLearning/Lab")
> getwd()
[1] "C:/Users/Ashish Upadhyay/Documents/Semester6/MachineLearning/Lab"
>
> dmf = read.csv("dmf.csv")
> attach(dmf)
>
> # Basic model: Only flor
> model=lm(dmf$dmf~dmf$flor)
> summary (model)

Call:
lm(formula = dmf$dmf ~ dmf$flor)

Residuals:
    Min      1Q  Median      3Q     Max 
-217.943 -91.930   3.935  70.097 281.904 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  730.929      3.158   231.5   <2e-16 ***
dmf$flor    -252.701      2.700   -93.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 99.79 on 2479 degrees of freedom
Multiple R-squared:  0.7794,	Adjusted R-squared:  0.7793 
F-statistic:  8760 on 1 and 2479 DF,  p-value: < 2.2e-16

> err= residuals (model)
> hist(err)
```



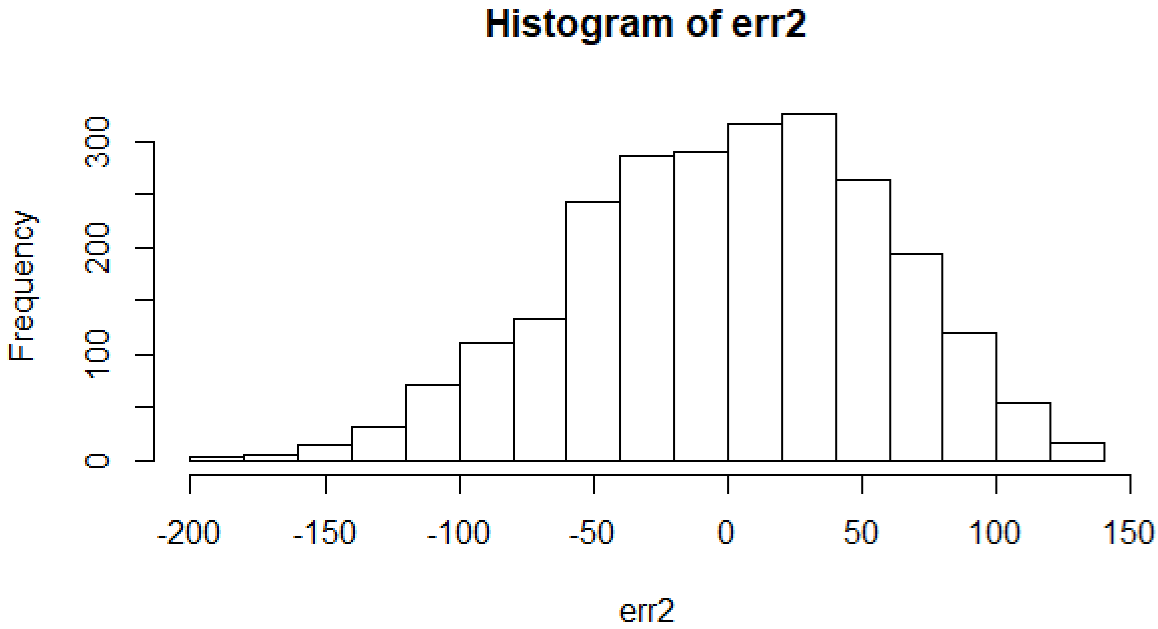
```
> plot(model$fitted.values,model$residuals)
```



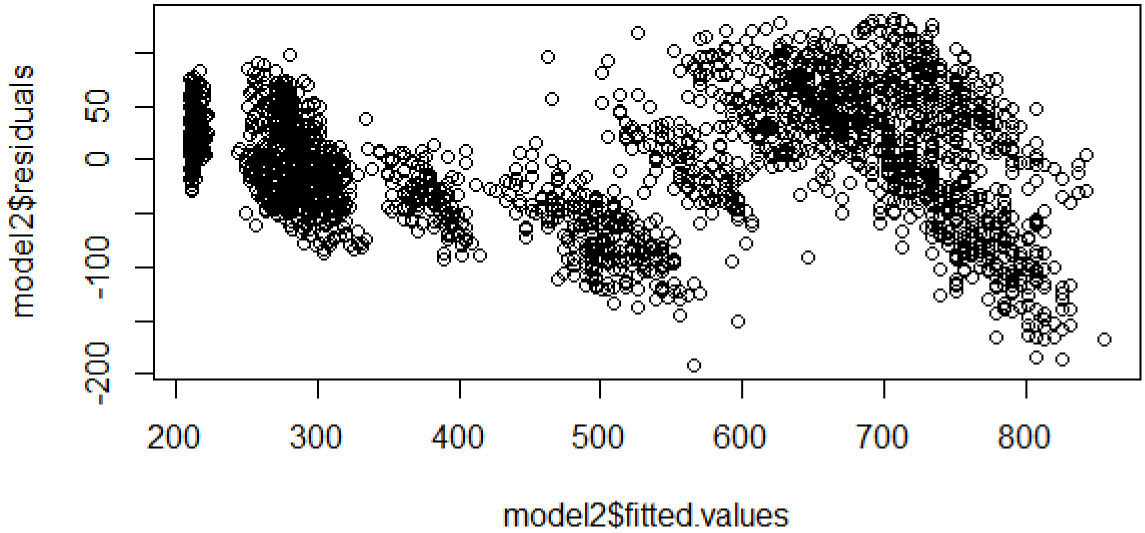
```
> plot(model$fitted.values, dmf$dmf)
```



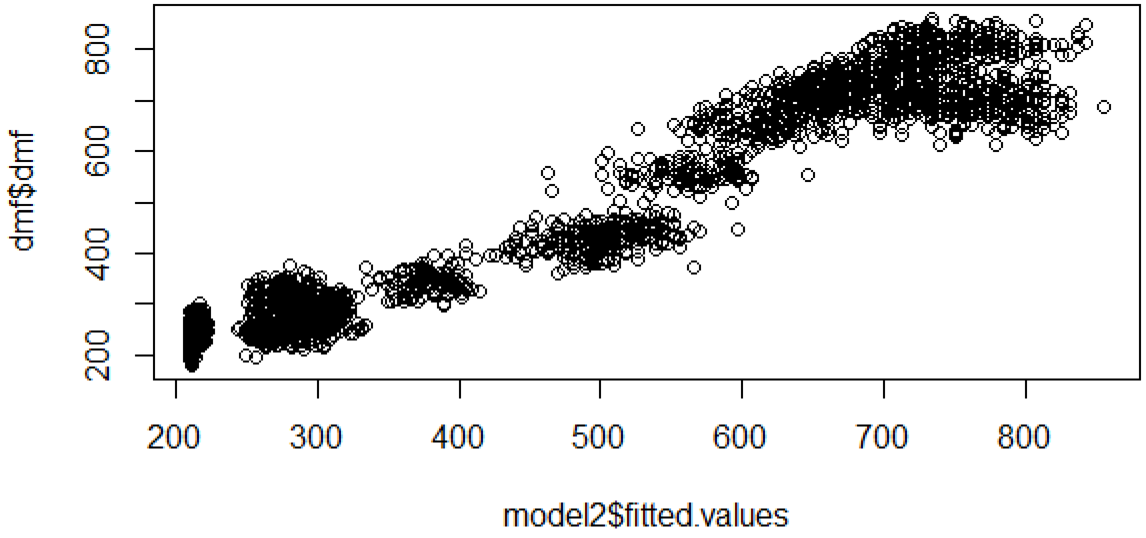
```
> # Advance model: flor + flor^2
> flor2= dmf$flor^2
> model2=lm(dmf$dmf~ dmf$flor+flor2)
> summary(model2)

Call:
lm(formula = dmf$dmf ~ dmf$flor + flor2)

Residuals:
    Min      1Q  Median      3Q     Max 
-191.388  -39.457   2.797  42.347 131.543 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  855.126      2.533  337.55   <2e-16 ***
dmf$flor    -604.861      5.231 -115.64   <2e-16 ***
flor2        141.939      2.013   70.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57.56 on 2478 degrees of freedom
Multiple R-squared:  0.9267,	Adjusted R-squared:  0.9266 
F-statistic: 1.565e+04 on 2 and 2478 DF,  p-value: < 2.2e-16

> err2= residuals(model2)
> hist(err2)
```



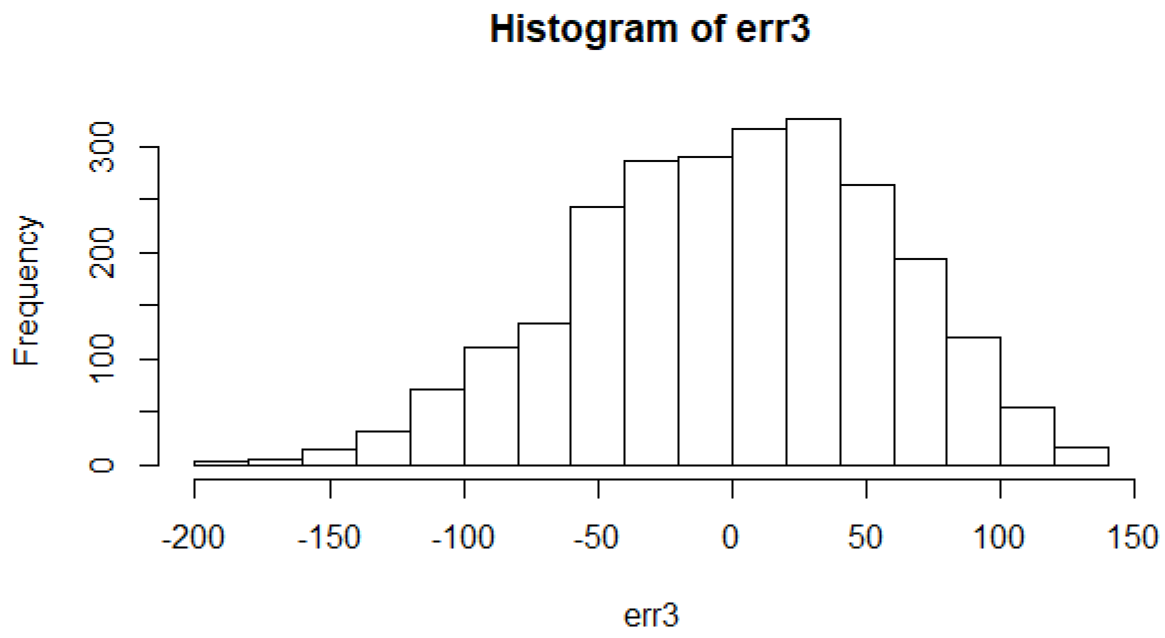
```
> plot(model2$fitted.values,model2$residuals)
```



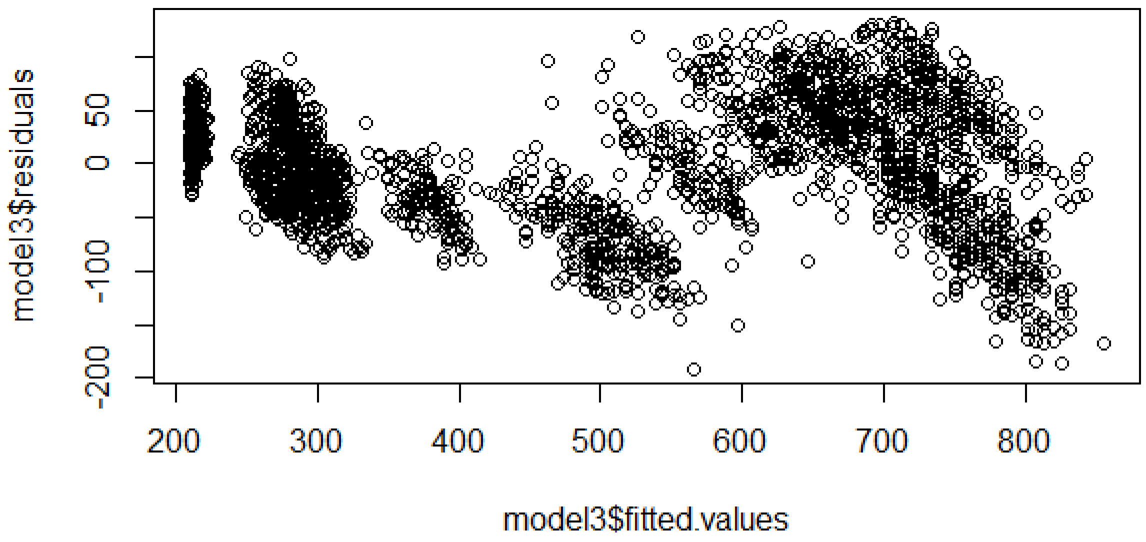
```
> plot(model2$fitted.values,dmf$dmf)
```



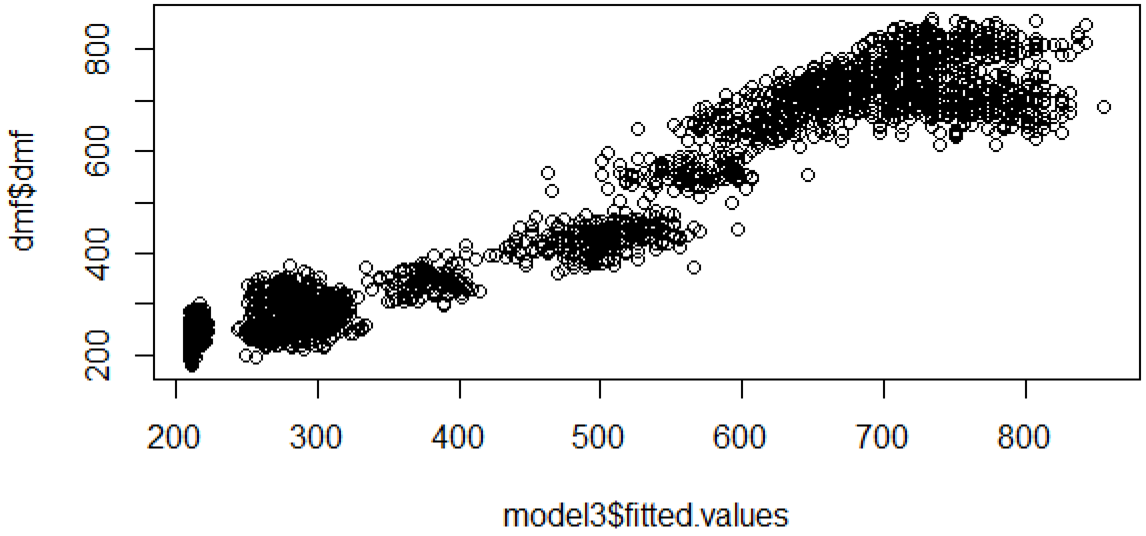
```
>
> # More advance model: flor + flor^2 + sqrt(flor)
> model3 = lm(dmf$dmf~ dmf$flor+flor2+1/sqrt(flor))
> summary(model3)

Call:
lm(formula = dmf$dmf ~ dmf$flor + flor2 + 1/sqrt(flor))

Residuals:
    Min      1Q  Median      3Q     Max
-191.388  -39.457   2.797  42.347  131.543

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept)  855.126     2.533  337.55  <2e-16 ***
dmf$flor   -604.861     5.231 -115.64  <2e-16 ***
flor2       141.939     2.013   70.53  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57.56 on 2478 degrees of freedom
Multiple R-squared:  0.9267,     Adjusted R-squared:  0.9266
F-statistic: 1.565e+04 on 2 and 2478 DF,  p-value: < 2.2e-16

> err3= residuals(model3)
> hist(err3)
```



```
> plot(model3$fitted.values,model3$residuals)
```



```
> plot(model3$fitted.values,dmf$dmf)
```


dmf$dmf
200
400
600
800
200
300
400
500
600
700
800
model3$fitted.values