hw_06

justyn koenig

11/9/2021

Question 1

#data set created for plant height (x) and grain yield (y) of rice. There are eight varieties of rice in this data set.

plant height is the predictor variable and grain yield is the response variable in the simple linear regression.

First the simple linear regression model is fit. Part A:

The model summary above allows us to note that the fitted regression equation would be Y = 10.137455 + (-0.037175)(X).

The two numbers in the equation above were found under the estimate column within the coefficients section on the printed model summary. I believe that this equation is showing that the larger the grain yield, the smaller the height of the plant would be. This would make sense due to the plant putting more energy into forming grain, rather than growth in height.

#The slope of the equation is -0.037175. Estimates : B1 = 10.137455 and B0 = -0.037175. # the correlation between x and y is low (-0.868707)

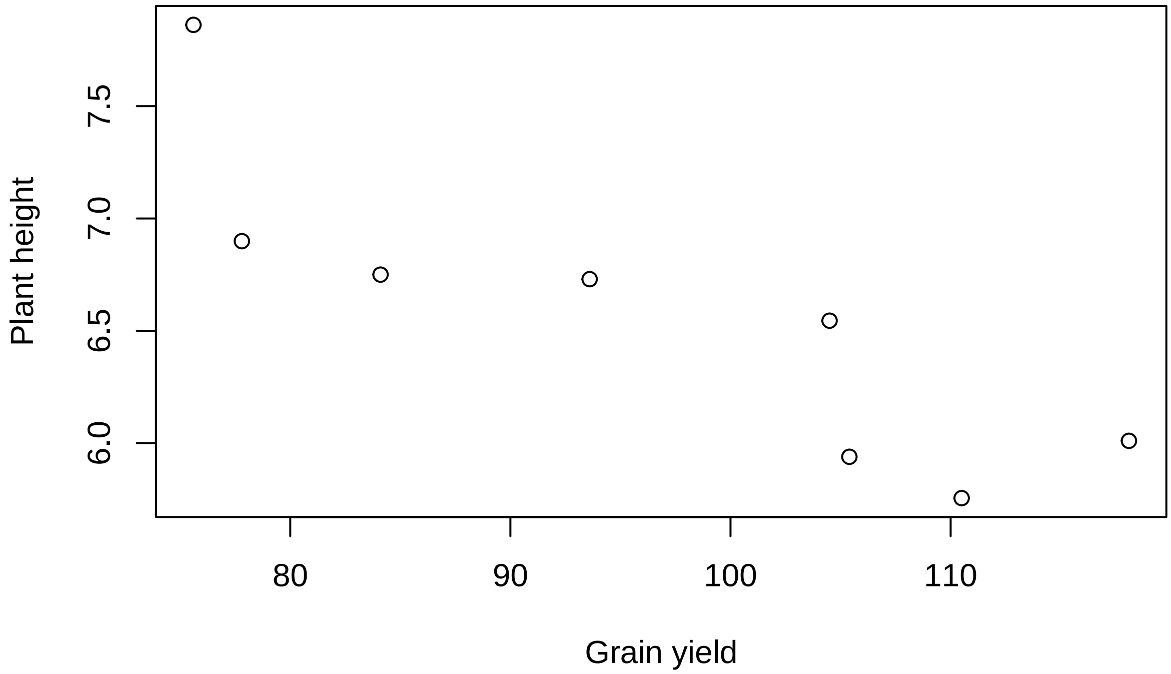
```
Plantdata = data.frame(Plant_height = c(110.5, 105.4, 118.1, 104.5, 93.6, 84.1, 77.8, 75.6), Grain_yiel
Plantdata
```

```
##   Plant_height Grain_yield
## 1        110.5       5.755
## 2        105.4       5.939
## 3        118.1       6.010
## 4        104.5       6.545
## 5         93.6       6.730
## 6         84.1       6.750
## 7         77.8       6.899
## 8         75.6       7.862
```

```
model = lm(Plantdata$Grain_yield~Plantdata$Plant_height)
summary(model)
```

```
## 
## Call:
## lm(formula = Plantdata$Grain_yield ~ Plantdata$Plant_height)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.34626 -0.27605 -0.09448  0.27023  0.53495 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            10.137455   0.842265  12.036    2e-05 ***
## Plantdata$Plant_height -0.037175   0.008653  -4.296  0.00512 ** 
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.3624 on 6 degrees of freedom
## Multiple R-squared:  0.7547, Adjusted R-squared:  0.7138 
## F-statistic: 18.46 on 1 and 6 DF,  p-value: 0.005116
```

```
Plantdata = data.frame(Plant_height = c(110.5, 105.4, 118.1, 104.5, 93.6, 84.1, 77.8, 75.6), Grain_yiel
plot(Plantdata$Grain_yield~Plantdata$Plant_height, xlab = "Grain yield", ylab = "Plant height")
```



```
plot
```

```
## function (x, y, ...)
## UseMethod("plot")
## <bytecode: 0x000000001220f0f0>
## <environment: namespace:base>
```

```
cor(Plantdata$Plant_height, Plantdata$Grain_yield)
```

```
## [1] -0.868707
```

Intercept = 10.13746

```
fit_Plantdata = lm(Plantdata$Grain_yield~Plantdata$Plant_height, data = Plantdata)
summary(fit_Plantdata)
```

```
##
## Call:
## lm(formula = Plantdata$Grain_yield ~ Plantdata$Plant_height,
##     data = Plantdata)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34626 -0.27605 -0.09448  0.27023  0.53495
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.137455   0.842265  12.036    2e-05 ***
## Plantdata$Plant_height -0.037175   0.008653  -4.296  0.00512 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3624 on 6 degrees of freedom
## Multiple R-squared:  0.7547, Adjusted R-squared:  0.7138
## F-statistic: 18.46 on 1 and 6 DF,  p-value: 0.005116
```

```
coef(fit_Plantdata)[1]
```

```
## (Intercept)
##    10.13746
```

Plant height = -0.03717469

```
coef(fit_Plantdata)["Plantdata$Plant_height"]
```

```
## Plantdata$Plant_height
##            -0.03717469
```

B

#The pvalue is less than 0.05 so the null hypothesis is rejected and the two variances are not equal. F value = 18.455.

```
anova(fit_Plantdata)
```

```
## Analysis of Variance Table
##
## Response: Plantdata$Grain_yield
##                        Df  Sum Sq Mean Sq F value   Pr(>F)
## Plantdata$Plant_height  1 2.42357 2.42357  18.455 0.005116 **
## Residuals               6 0.78794 0.13132
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

B

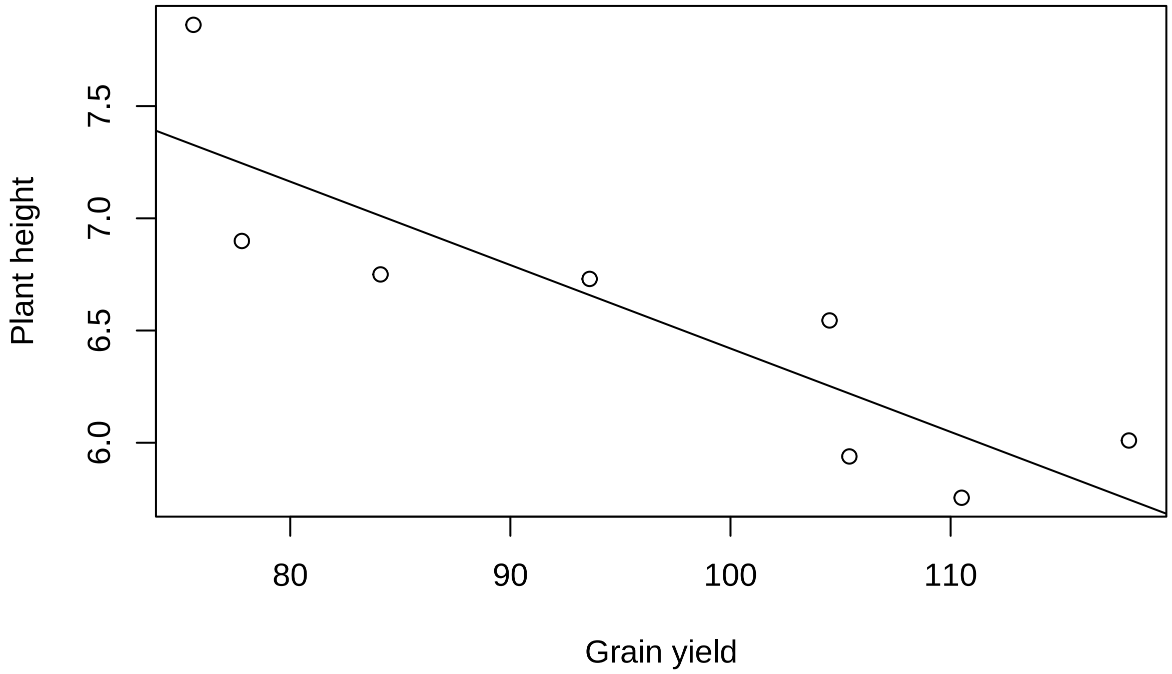
Reject the null because p value is less than 0.05. Get the same p value for T test as F test of 0.005. There is evidence of a strong relationship between plant height and grain yield. F value = 18.455.

```
summary(fit_Plantdata)
```

```
##
## Call:
## lm(formula = Plantdata$Grain_yield ~ Plantdata$Plant_height,
##     data = Plantdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34626 -0.27605 -0.09448  0.27023  0.53495
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.137455   0.842265  12.036    2e-05 ***
## Plantdata$Plant_height -0.037175   0.008653  -4.296  0.00512 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3624 on 6 degrees of freedom
## Multiple R-squared:  0.7547, Adjusted R-squared:  0.7138
## F-statistic: 18.46 on 1 and 6 DF,  p-value: 0.005116
```

c.

```
plot(Plantdata$Grain_yield~Plantdata$Plant_height, xlab = "Grain yield", ylab = "Plant height")
abline(fit_Plantdata)
```



calculate the 95 % CI

t value = -2.446912

#b0 + t_-2.446912 * se_b0 #-0.037175 + -2.446912 * 0.008653 = #-0.0583481295 #-0.037175 + t(8-2, 0.05/2) * 0.008653

= upper level CI

#b0 - t_-2.446912 * se_b0 #-0.037175 - -2.446912 * 0.008653 = -0.0160018705 #-0.037175 - t_(8-2, 0.05/2) * 0.008653

the 95% CI shows that the range of values lies between -0.05834895 and -0.01600043.

```
b0 = -0.037175
se_b0 = 0.008653
qt(0.05/2, 8-2)
```

```
## [1] -2.446912
```

```
Plantdata
```

```
##   Plant_height Grain_yield
## 1        110.5       5.755
## 2        105.4       5.939
## 3        118.1       6.010
## 4        104.5       6.545
## 5         93.6       6.730
## 6         84.1       6.750
## 7         77.8       6.899
## 8         75.6       7.862
```

```
fit_Plantdata = lm(Grain_yield~Plant_height, data = Plantdata)
confint(fit_Plantdata)
```

```
##                    2.5 %      97.5 %
## (Intercept)   8.07650745 12.19840320
## Plant_height -0.05834895 -0.01600043
```

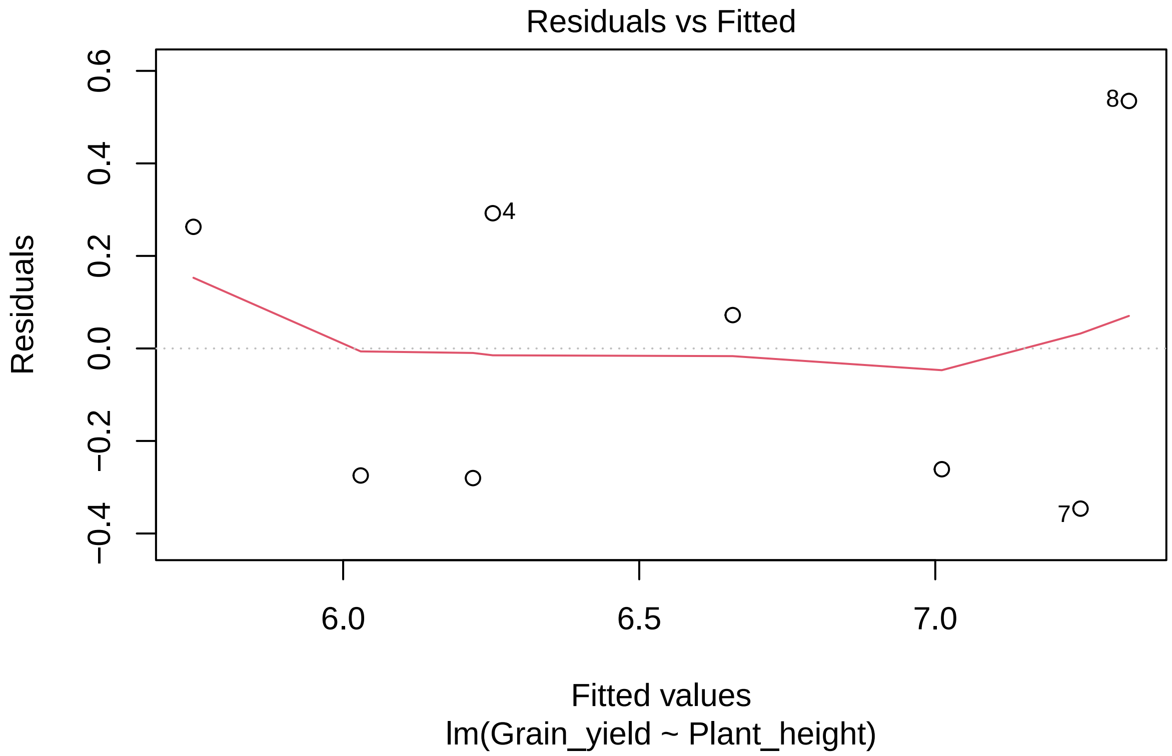
Below is the fitted regression line

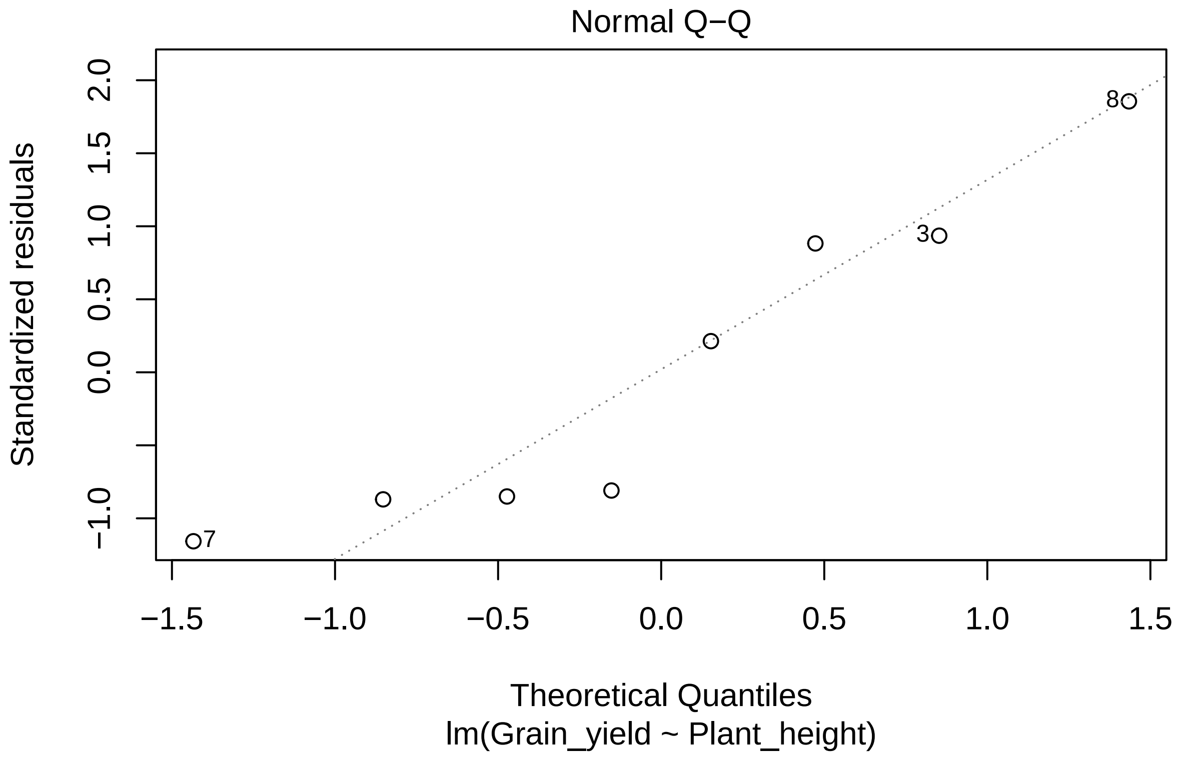
#Y = 10.137455 + (-0.037175)(X)

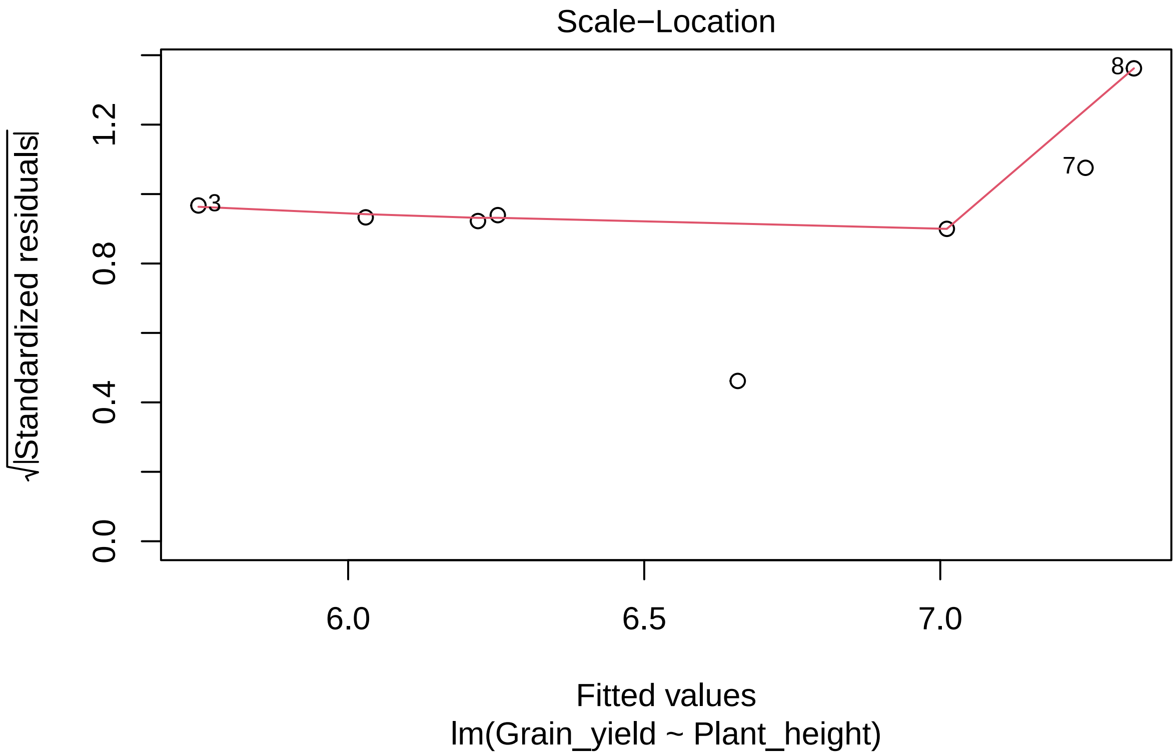
Printed below are the residuals.

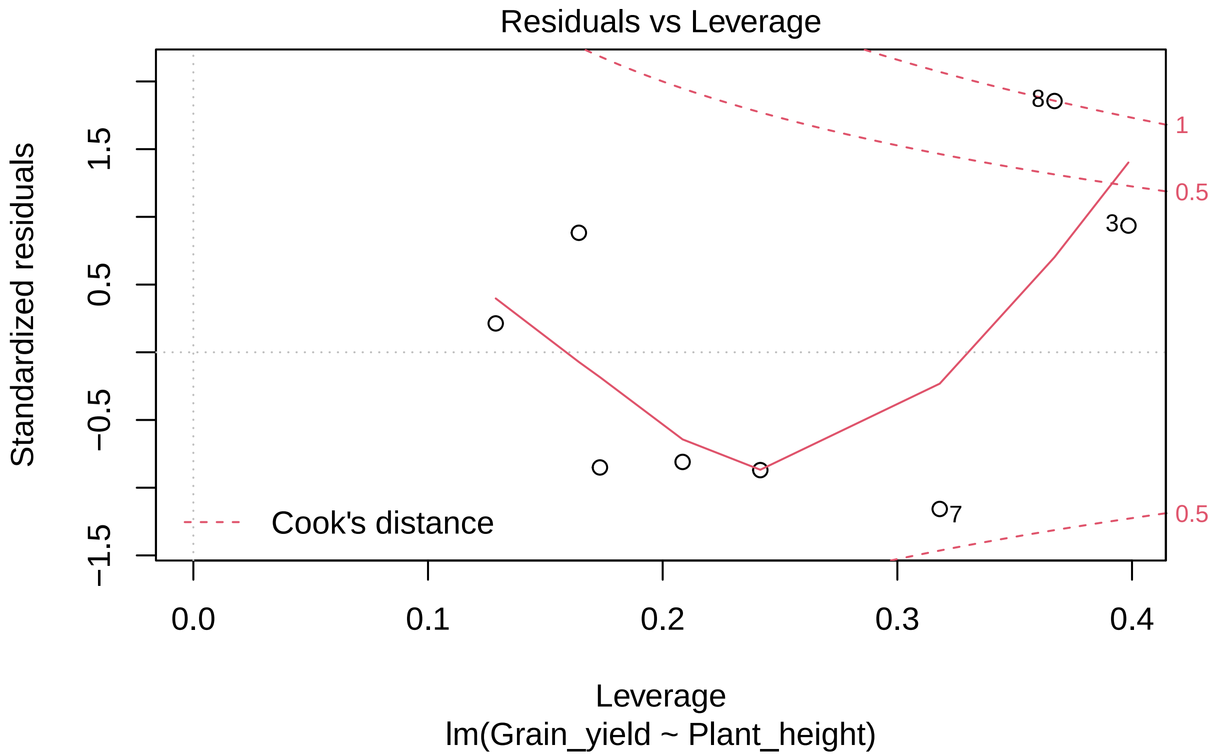
```
# d.

plot(fit_Plantdata)
```


Residuals vs Fitted
Residuals
0.6
0.4
0.2
0.0
−0.2
−0.4
4
8
7
6.0
6.5
7.0
Fitted values
lm(Grain_yield ~ Plant_height)


Normal Q−Q
Standardized residuals
2.0
1.5
1.0
0.5
0.0
−1.0
8
3
7
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles
lm(Grain_yield ~ Plant_height)


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
6.0
6.5
7.0
Fitted values
lm(Grain_yield ~ Plant_height)
3
7
8



```
summary(fit_Plantdata)
```

```
##
## Call:
## lm(formula = Grain_yield ~ Plant_height, data = Plantdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34626 -0.27605 -0.09448  0.27023  0.53495
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.137455   0.842265  12.036    2e-05 ***
## Plant_height -0.037175   0.008653  -4.296  0.00512 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3624 on 6 degrees of freedom
## Multiple R-squared:  0.7547, Adjusted R-squared:  0.7138
## F-statistic: 18.46 on 1 and 6 DF,  p-value: 0.005116
```

```
resid(fit_Plantdata)
```

```
##          1          2          3          4          5          6          7
## -0.2746519 -0.2802428  0.2628757  0.2922999  0.0720958 -0.2610638 -0.3462643
##          8
##  0.5349514
```

e.

The estimate od the error variance (MSE) is 0.1049614

```
mean(summary(fit_Plantdata$residuals^2))
```

```
## [1] 0.1049614
```

f.

#Confidence interval used to estimate the expected yield of a rice variety.

```
predict(fit_Plantdata, newdataplant = Plantdata(x = 100), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 6.029652 5.593799 6.465505
## 2 6.219243 5.850145 6.588341
## 3 5.747124 5.187376 6.306872
## 4 6.252700 5.893295 6.612105
## 5 6.657904 6.339603 6.976206
## 6 7.011064 6.606184 7.415944
## 7 7.245264 6.745186 7.745342
## 8 7.327049 6.789884 7.864214
```

g.

Prediction interval used to predict the yield of a new rice variety.

both g and g have the same fit values, the lower levels of the 95% CI are smaller for g and larger for f , and the upper levels of the 95% CI are larger in g and smaller in f, in comparing the values of f and g. G is wider.

```
predict(fit_Plantdata, newdataplant = Plantdata(x = 100), interval = "prediction")
```

```
## Warning in predict.lm(fit_Plantdata, newdataplant = Plantdata(x = 100), : predictions on current data
##        fit      lwr      upr
## 1 6.029652 5.041600 7.017704
## 2 6.219243 5.258768 7.179718
## 3 5.747124 4.698508 6.795741
## 4 6.252700 5.295908 7.209492
## 5 6.657904 5.715782 7.600027
## 6 7.011064 6.036278 7.985849
## 7 7.245264 6.227248 8.263281
## 8 7.327049 6.290311 8.363787
```

h.

The r squared value (coefficient of determination) is 0.7546518. This means that 75.46% of the variation in the grain yield can be explained by the height of the rice plants. A R squared value of 1 means that the explanatory variables can be used to explain the variance observed in the response variable. A value of 0 means that the explanatory variables cannot explain the variance in the response variable. Bigger the R squared = Better the explanatory variables can be used as predictors of the repsonse variables.

```
summary(fit_Plantdata)$r.squared
```

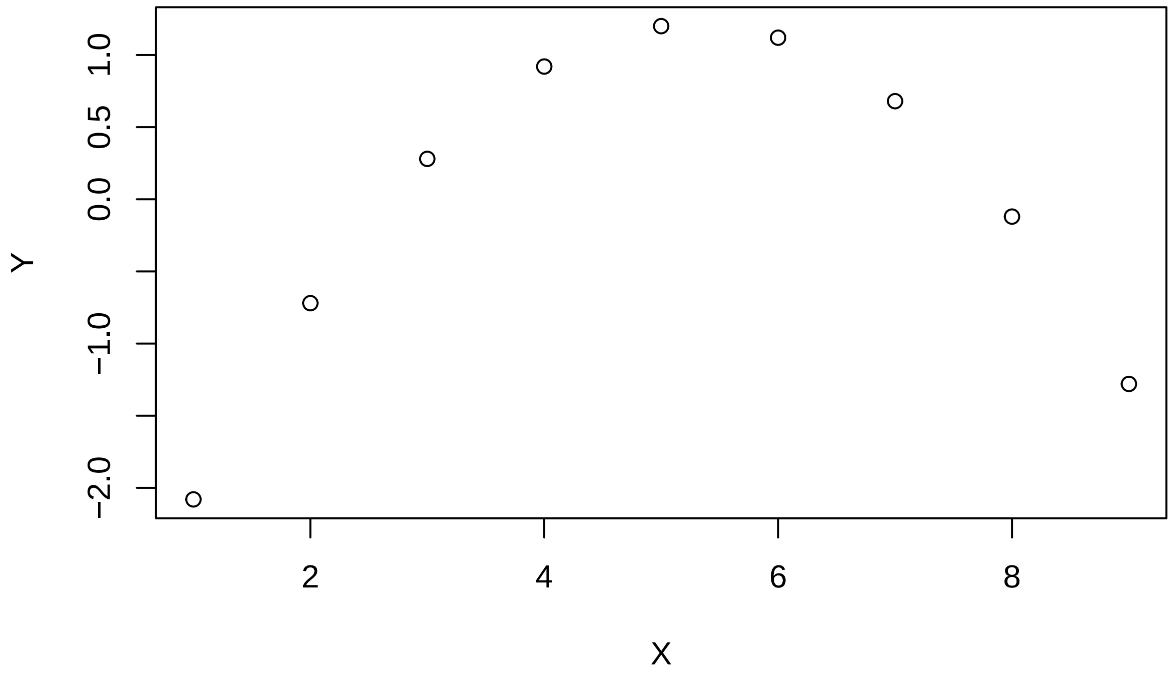
```
## [1] 0.7546518
```

Question 2

```
#
Demodataset = data.frame(x = c(1, 2, 3, 4, 5, 6, 7, 8, 9), y = c(-2.08, -0.72, 0.28, 0.92, 1.20, 1.12,
Demodataset
```

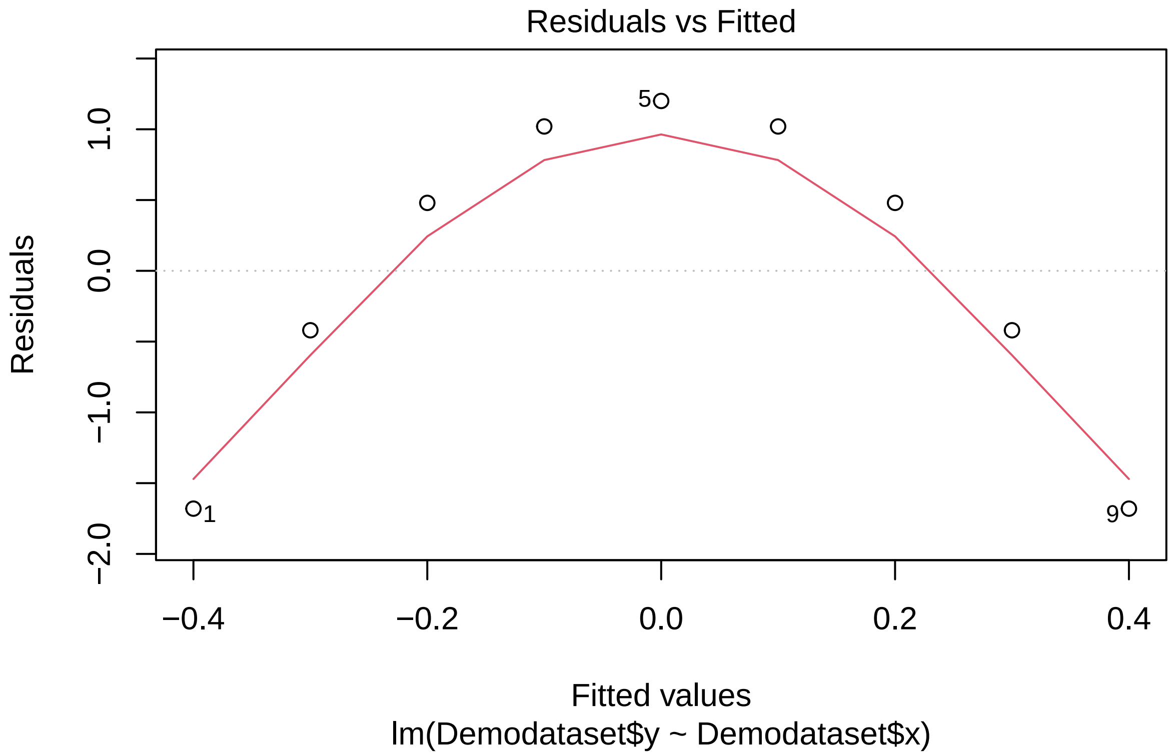
```
##   x     y
## 1 1 -2.08
## 2 2 -0.72
## 3 3  0.28
## 4 4  0.92
## 5 5  1.20
## 6 6  1.12
## 7 7  0.68
## 8 8 -0.12
## 9 9 -1.28
```

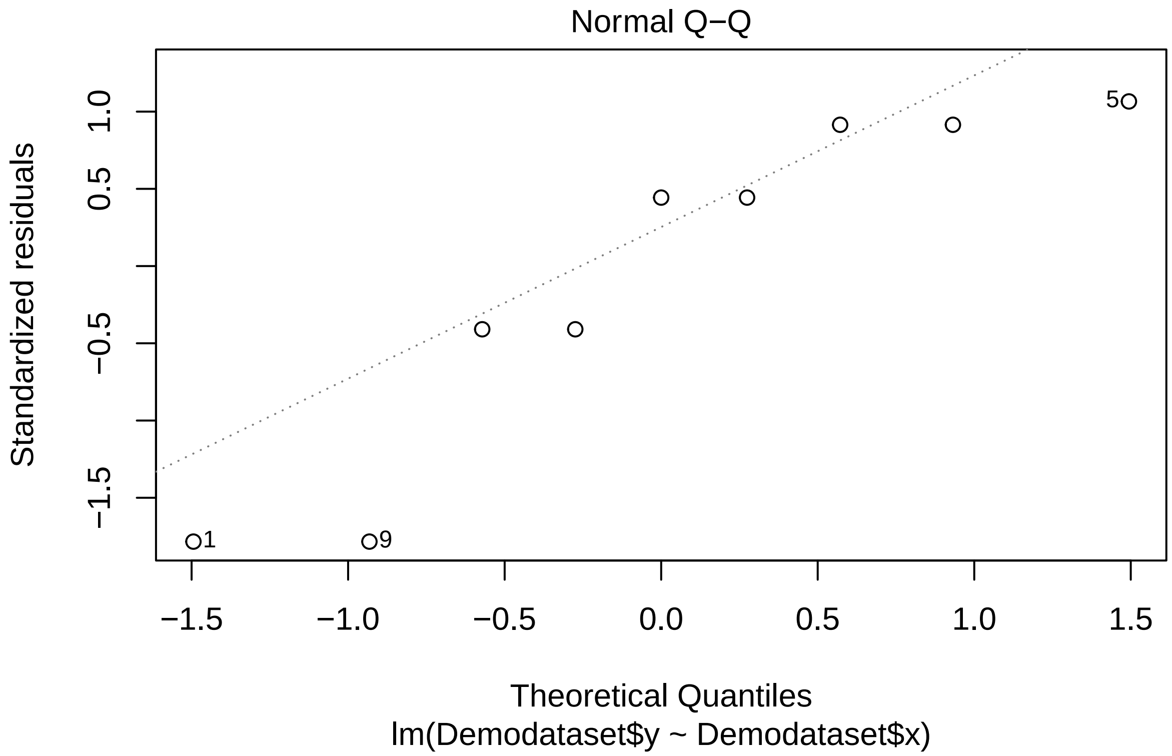
```
plot(Demodataset$x, Demodataset$y, xlab = "X", ylab = "Y")
```

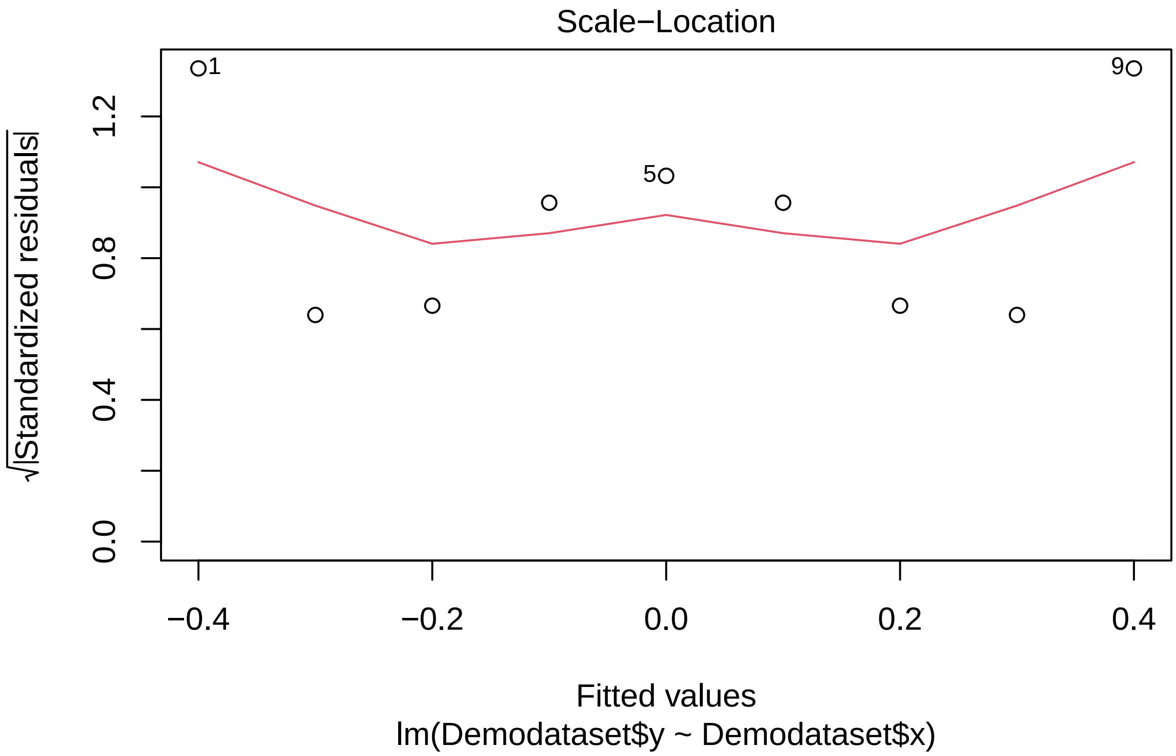


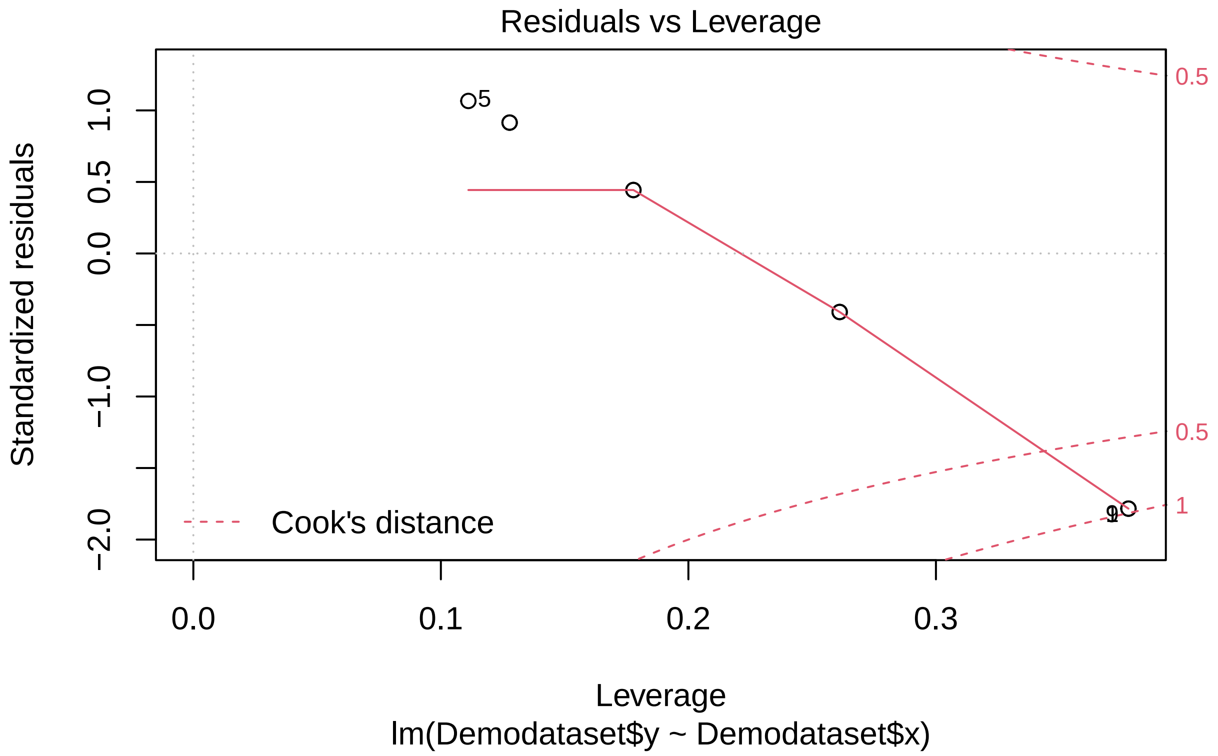
a

```
rawy = lm(Demodataset$y~Demodataset$x, data = Demodataset)
res = resid(rawy)
plot(rawy)
```


Residuals vs Fitted
Residuals
1.0
0.0
−1.0
−2.0
5
1
9
−0.4
−0.2
0.0
0.2
0.4
Fitted values
lm(Demodataset$y ~ Demodataset$x)


Normal Q−Q
Standardized residuals
1.0
0.5
−0.5
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles
lm(Demodataset$y ~ Demodataset$x)
1
9
5

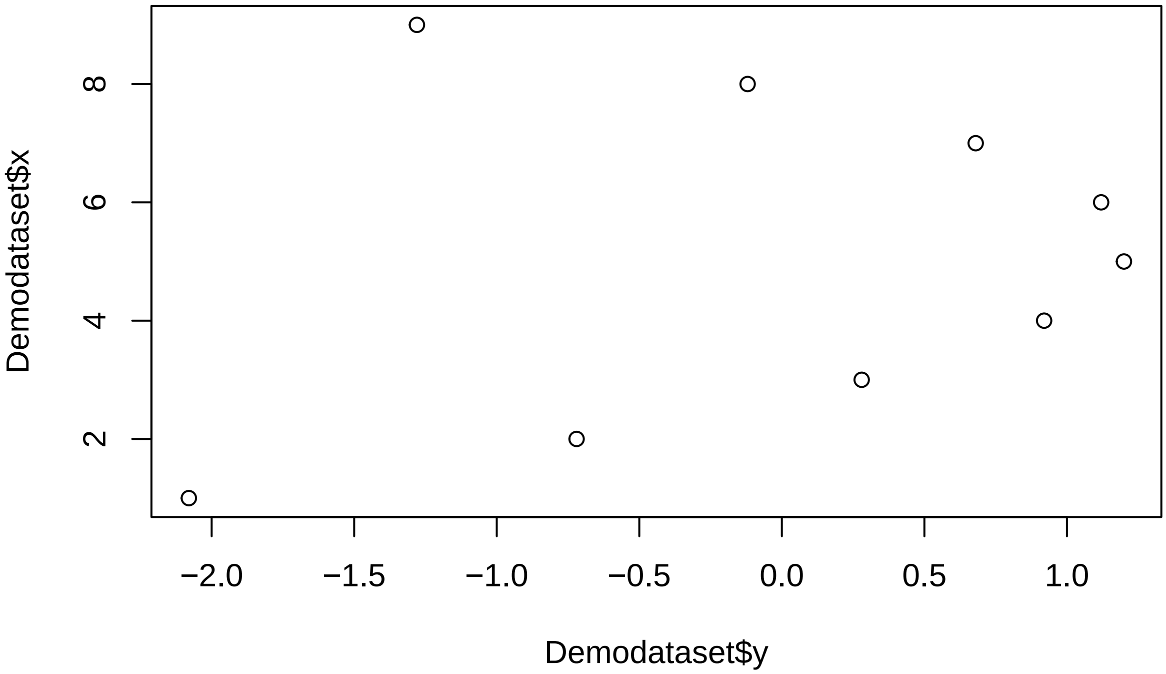

Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
−0.4
−0.2
0.0
0.2
0.4
Fitted values
lm(Demodataset$y ~ Demodataset$x)
1
9
5



```
res
```

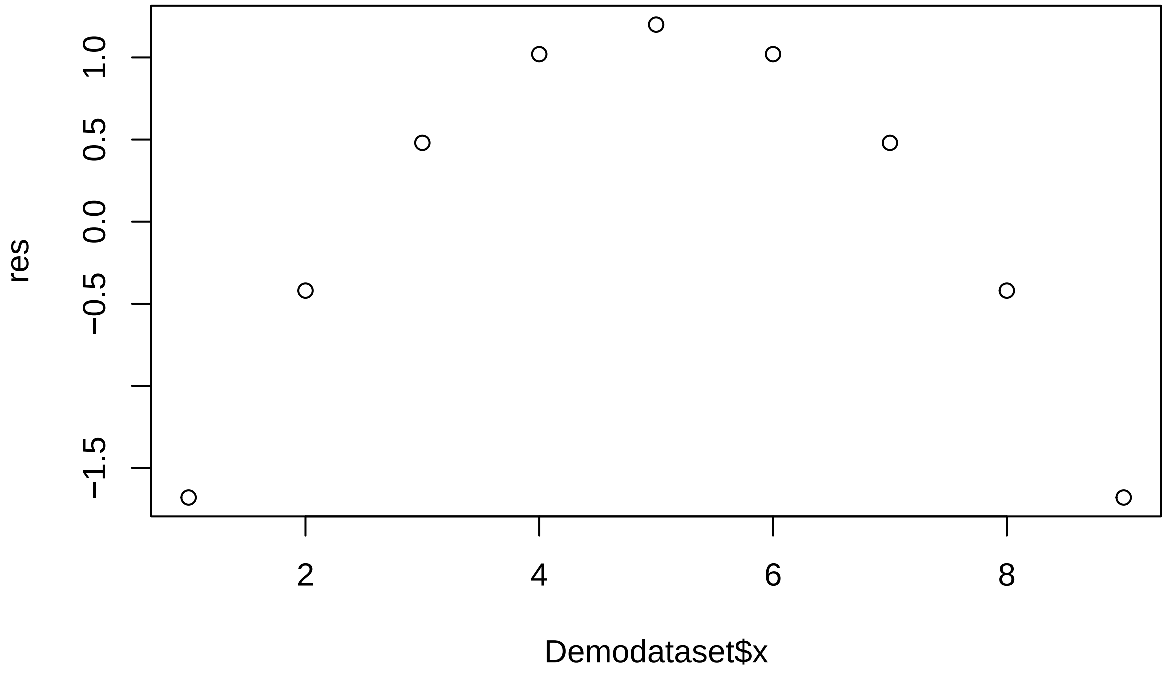
```
##     1     2     3     4     5     6     7     8     9
## -1.68 -0.42  0.48  1.02  1.20  1.02  0.48 -0.42 -1.68
```

```
plot(Demodataset$y, Demodataset$x)
```



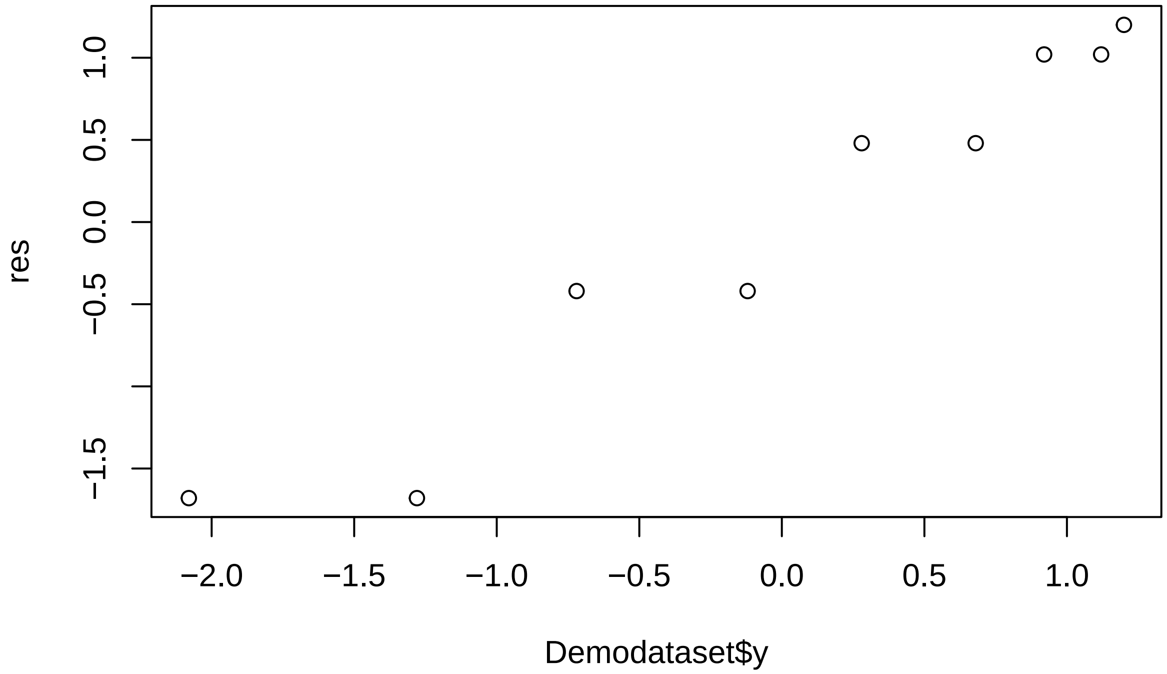
b

```
plot(Demodataset$x, res)
```



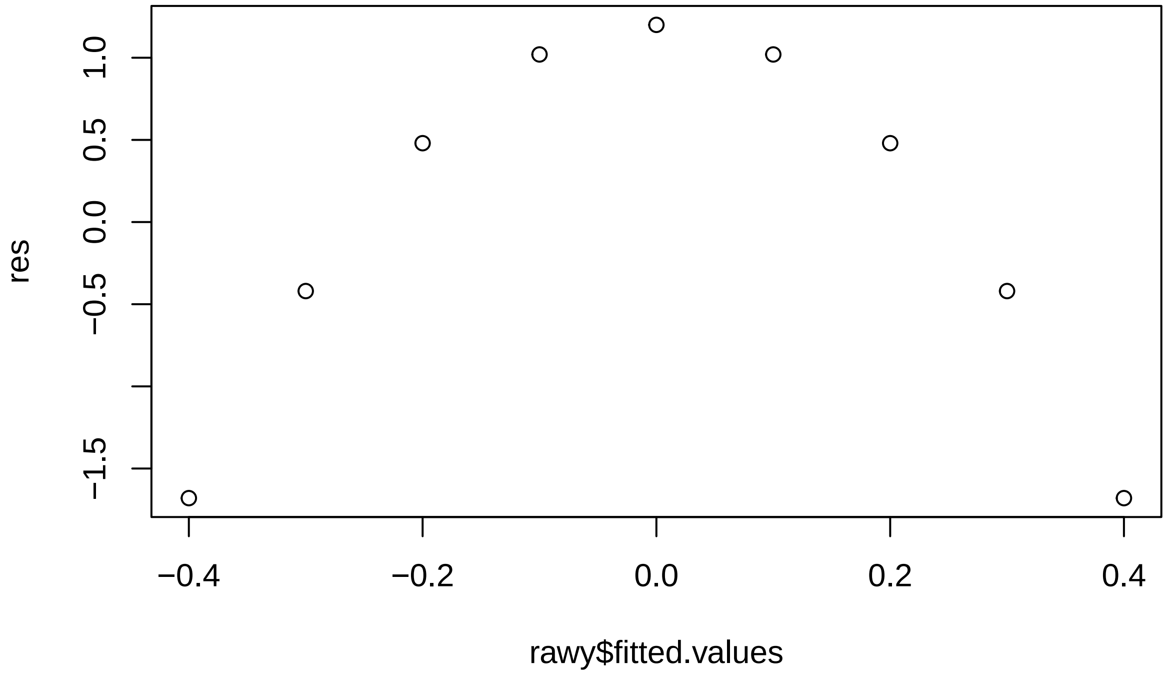
c

```
plot(Demodataset$y, res)
```



d

```
plot(rawy$fitted.values, res)
```


res
1.0
0.5
0.0
−0.5
−1.5
−0.4
−0.2
0.0
0.2
0.4
rawy$fitted.values