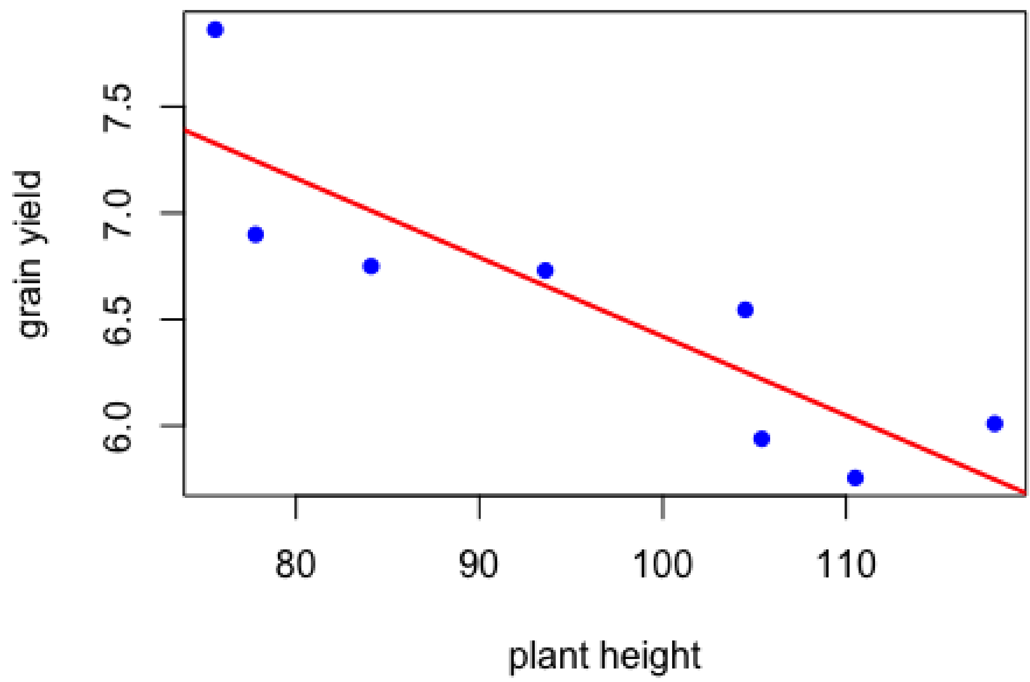
hw_06

Ahmed Al-Tohamy

11/15/2021

```
#creat the data which is plant height (predictor variable x) and grain yield
(response variable y)

data <- data.frame (x=c(110.5, 105.4, 118.1, 104.5, 93.6, 84.1, 77.8, 75.6),
y = c(5.755, 5.939, 6.010, 6.545, 6.730, 6.750, 6.899, 7.862))

#first of all we should plot our data
plot(data$x, data$y, xlab="plant height", ylab="grain yield", col ="blue",
pch = 16)


#creat the linear regression
fit_data <- lm(y ~ x, data = data)

#calculate Pearson correlation coefficient
cor(data$y, data$x)

## [1] -0.868707

#we can see there is a negative correlation between plant height and grain
yield

#Add a regression line

abline(fit_data, col="red", lwd=2)
```




```
#Q1
#a.the least squares estimate (^β1) of the slope β1
coef(fit_data)["x"]

##           x
## -0.03717469

#^β1 is the estimate of the measure of the steepness of the line that best
fit the data
#you have (best estimate of expected change in Y for unit increase in X), ^β
is using the least squares method

#b
#We can do that by different methods, one of them is to call summary function
and anova
summary(fit_data)

##
## Call:
## lm(formula = y ~ x, data = data)
##
## 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 ***
## x           -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
```

```
anova(fit_data)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## x          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
```

```
# from the tables we can conclude that
#T-test = 12.036
#F-test = 18.46
#So we can say that, we reject H0 and there is a strong evidence of a
relationship between
#grain yield and plant height.

#c
#by hand from the equation   ^β1 ± tn−2,α/2 × s.e.(^β1)
#^β1= -0.037175
#tn−2,α/2 = 2.44691
#s.e.(^β1) = 0.008653

#When we add all of them --> -0.037175 ± 0.0211

#The 95%  confidence interval is a range of values that
#you can be 95% confident contains the true mean of the population
qt(0.05/2, 6)
```

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

#d

#equation #y^=a+bx #^y = 10.13745532 + -0.03717469 x

```
#e
#from the summary table
summary(fit_data)
```

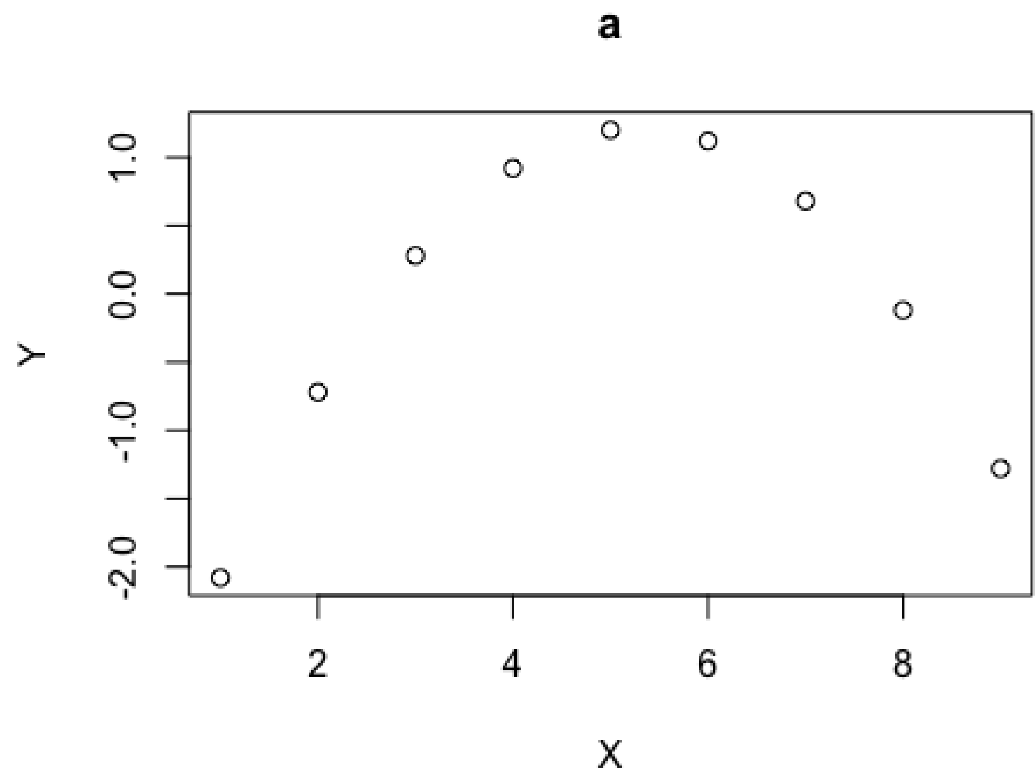
```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34626 -0.27605 -0.09448  0.27023  0.53495
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.137455   0.842265  12.036    2e-05 ***
## x           -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

#Residual standard error: 0.3624

#f
#Estimate the expected yield of a rice variety when x=100
predict(fit_data, newdata = data.frame(x = 100), interval = "confidence",
levels=0.95)

##        fit      lwr      upr
## 1 6.419986 6.096321 6.743651

#g
#predeict the expected yield of a rice variety when x=100
predict(fit_data, newdata = data.frame(x = 100), interval = "prediction",
levels=0.95)

##        fit      lwr      upr
## 1 6.419986 5.476038 7.363934

#clearly g is wider as it has lower value (5.476038) and upper value
(5.476038)
#in comparsion to f which has lower value (6.096321) and upper value
(6.743651)
# # confint(fit_data, level = 0.95)

#h
#again we can get the coefficient of determination R2 by using the summary
function
summary(fit_data)

##
## Call:
## lm(formula = y ~ x, data = data)
##
```

```
## 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 ***
## x           -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
```

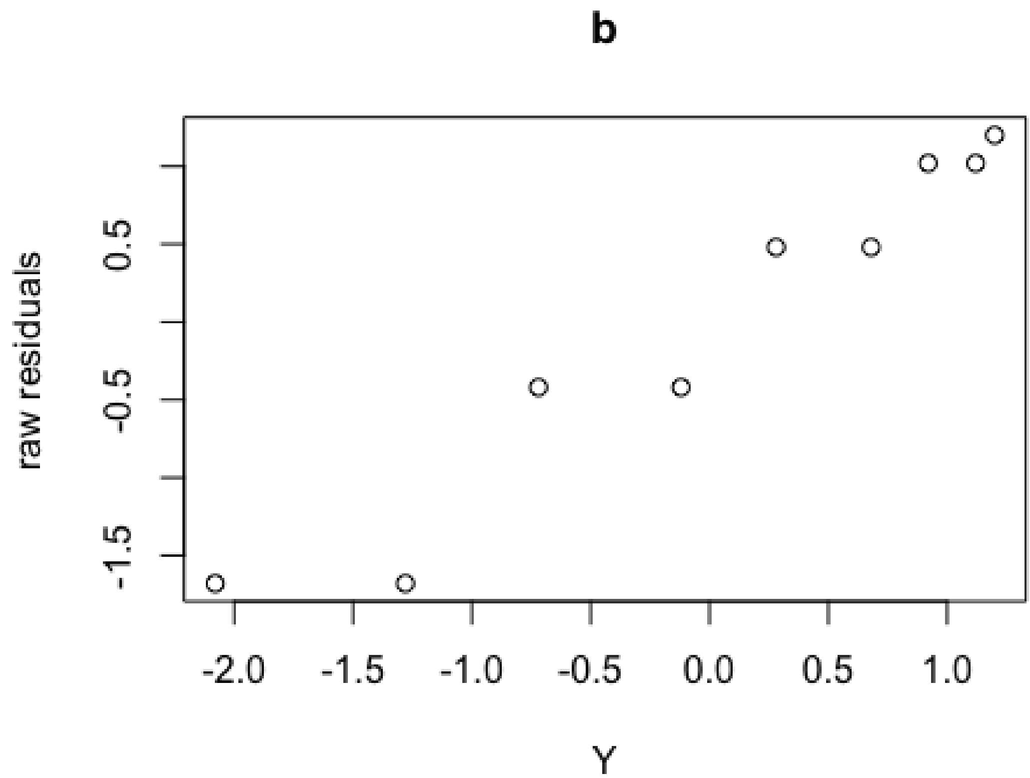
```
#coefficient of determination R2 --> 0.7
#That means 70% of the variation in grain yield can be predictable from the
plant height
```

#————————————————————————————————-

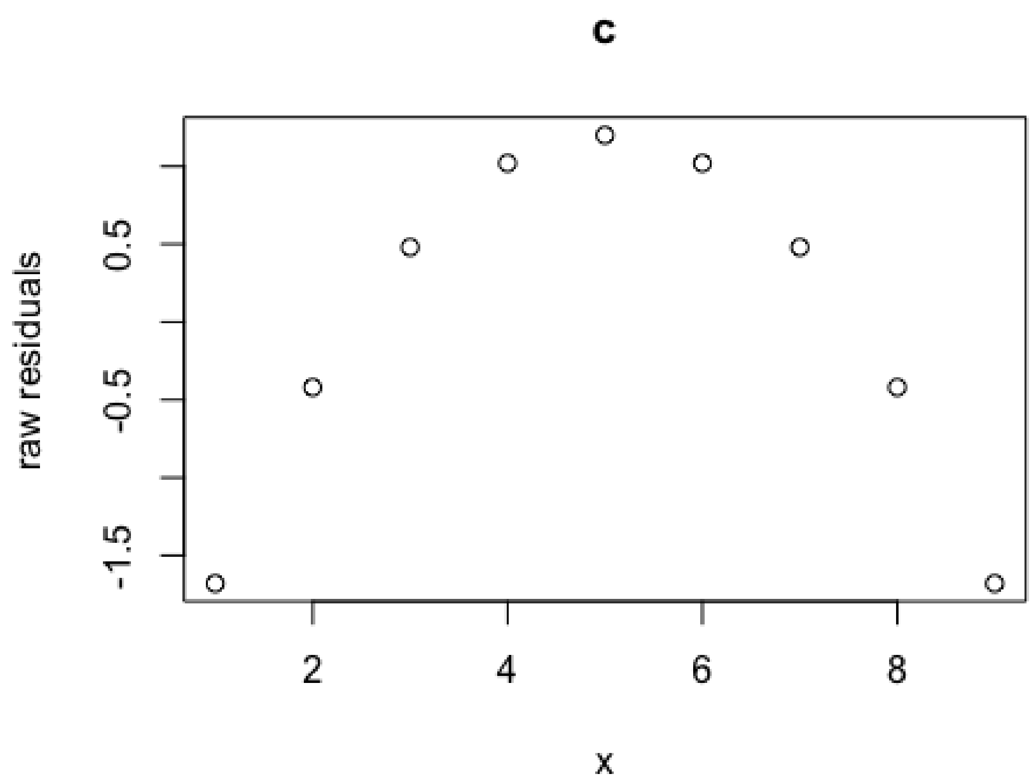
```
#Q2
d <- data.frame(x = c(1, 2, 3, 4, 5, 6, 7, 8, 9),
                y = c(-2.08, -0.72, 0.28, 0.92, 1.20, 1.12, 0.68, -0.12, -
1.28))

#creat the linear regression
fit_d = lm(y ~ x, data = d)

#a
#Plot y vs. x
plot(d$x, d$y, xlab="X", ylab="Y", main="a")
```



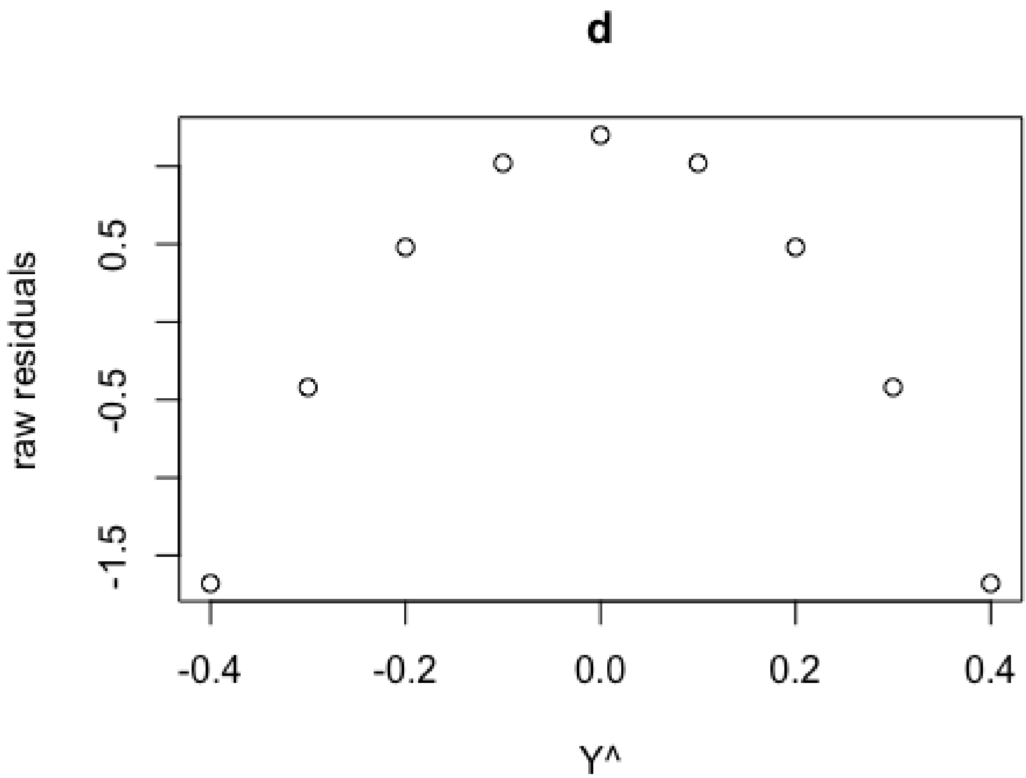

```
#b
#plot
d.res <-  resid(fit_d)
plot(d$y,d.res, xlab="Y", ylab="raw residuals", main="b")
```




```
#c
#Plot the raw residuals vs. x
plot(d$x,d.res,  xlab="x", ylab="raw residuals", main="c")
```




```
#d
#Plot the raw residuals vs. y^
plot(fitted(fit_d),d.res,  xlab="Y^", ylab="raw residuals", main="d")
```



#e #I cant see any meaningful difference between (c) and (d) #they both represent quadratic equation # d) gives a better indication of the lack of fit as it is obvious it does not #fit into a line(it is a quadratic equation)