Stella’s linear regression

Stella Zhang

2024-04-17

## Install or load the packages

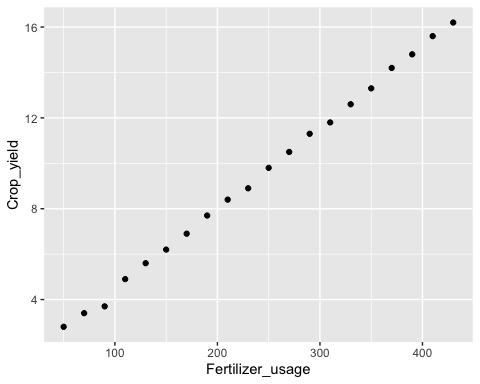
library(ggplot2)  
library(tidyverse)

## Read or give the data

Fertilizer\_usage=c(50, 70, 90, 110, 130, 150, 170, 190, 210, 230,  
250, 270, 290, 310, 330, 350, 370, 390, 410, 430)  
Crop\_yield=c(2.8, 3.4, 3.7, 4.9, 5.6, 6.2, 6.9, 7.7, 8.4, 8.9, 9.8,  
10.5, 11.3, 11.8, 12.6, 13.3, 14.2, 14.8, 15.6, 16.2)  
  
# scattrer plot  
my\_data1=data.frame(Fertilizer\_usage,Crop\_yield)

## Scotter plot for the data

ggplot(my\_data1,aes(x=Fertilizer\_usage,y=Crop\_yield))+  
 geom\_point()

 ## Simple linear regression

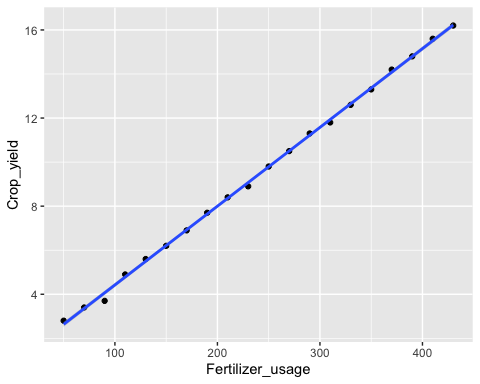
reg\_model=lm(Crop\_yield~Fertilizer\_usage)  
summary(reg\_model)

##   
## Call:  
## lm(formula = Crop\_yield ~ Fertilizer\_usage)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.36609 -0.03222 0.00925 0.08425 0.16429   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.8477444 0.0642300 13.2 1.07e-10 \*\*\*  
## Fertilizer\_usage 0.0357594 0.0002412 148.2 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1244 on 18 degrees of freedom  
## Multiple R-squared: 0.9992, Adjusted R-squared: 0.9991   
## F-statistic: 2.198e+04 on 1 and 18 DF, p-value: < 2.2e-16

Since the p-value is less than therefore the Fertilizer Usage is a significant variable for the Crop Yield.

my\_data1=data.frame(Fertilizer\_usage,Crop\_yield)  
ggplot(my\_data1,aes(x=Fertilizer\_usage,y=Crop\_yield))+  
 geom\_point()+  
 geom\_smooth(method=lm)

## `geom\_smooth()` using formula = 'y ~ x'

 ## R square

summary(reg\_model)$r.squared

## [1] 0.9991816

In conclusion, the Fertilizer Usage level can explain of the variability in the Crop Yield.

## Normality test

shapiro.test(reg\_model$residuals)

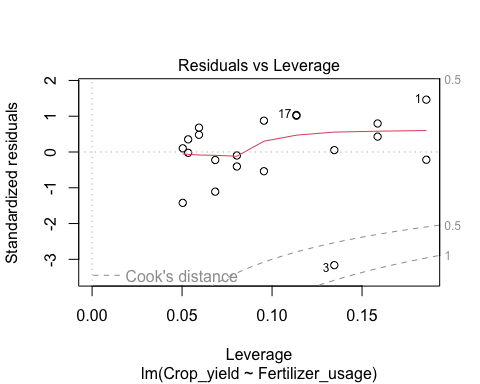
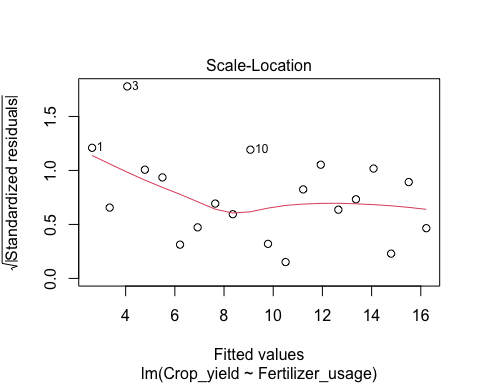
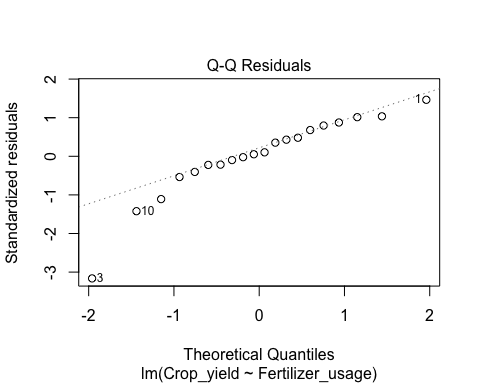
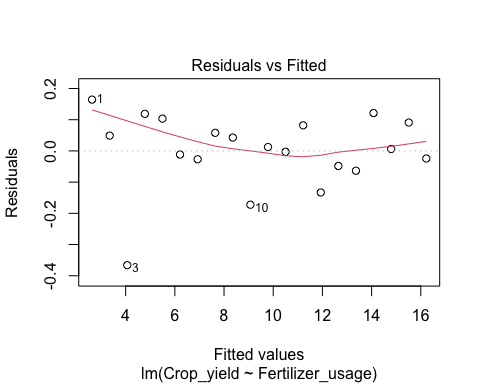
##   
## Shapiro-Wilk normality test  
##   
## data: reg\_model$residuals  
## W = 0.88829, p-value = 0.02502

ks.test(reg\_model$residuals,"pnorm")

##   
## Exact one-sample Kolmogorov-Smirnov test  
##   
## data: reg\_model$residuals  
## D = 0.43475, p-value = 0.0005912  
## alternative hypothesis: two-sided

Since the p-value for the Shapiro-Wilk test is , which is less than , and the p-value for the Kolmogorov-Smirnov test is , also less than , the normality assumption of the error terms is not confirmed. This suggests that the error terms do not follow a normal distribution, which is a violation of one of the key assumptions of the classical linear regression model.

plot(reg\_model)

 ## Constant variance Assumption

library(lmtest)  
res<-bptest(reg\_model)  
res

##   
## studentized Breusch-Pagan test  
##   
## data: reg\_model  
## BP = 2.5015, df = 1, p-value = 0.1137

Since the p-value is 0.1137, which is more than so the constant variance assumption is hold.

## Corrolation

to test is there is a significant correlation between two variables, we use the **cor.test()** function in R.

cor.test(my\_data1$Fertilizer\_usage,my\_data1$Crop\_yield)

##   
## Pearson's product-moment correlation  
##   
## data: my\_data1$Fertilizer\_usage and my\_data1$Crop\_yield  
## t = 148.24, df = 18, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9989413 0.9998418  
## sample estimates:  
## cor   
## 0.9995907

The correlation is with the p-value of . since the p-value is less than so we reject the null hypothesis. We accept that there is a significant linear association between Fertilizer Usage and Crop Yield.

## Install or load the packages

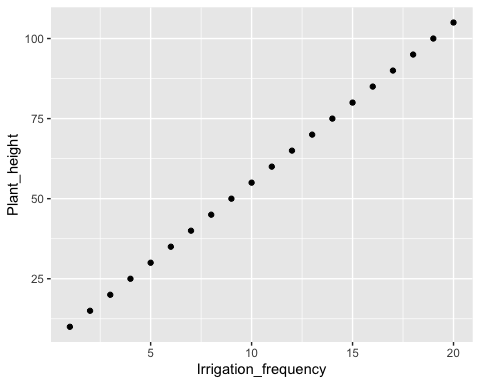
library(ggplot2)  
library(tidyverse)

## Read or give the data

Irrigation\_frequency=c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,  
14, 15, 16, 17, 18, 19, 20)  
Plant\_height=c(10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75,  
80, 85, 90, 95, 100, 105)  
  
# scottrer plot  
my\_data2=data.frame(Irrigation\_frequency,Plant\_height)

## Scatter plot for the data

ggplot(my\_data2,aes(x=Irrigation\_frequency,y=Plant\_height))+  
 geom\_point()

 ## Simple linear regression

reg\_model=lm(Plant\_height~Irrigation\_frequency)  
summary(reg\_model)

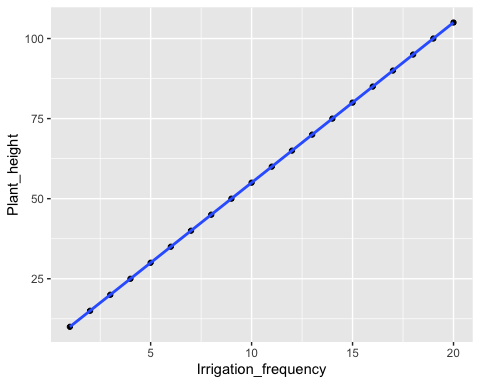
## Warning in summary.lm(reg\_model): essentially perfect fit: summary may be  
## unreliable

##   
## Call:  
## lm(formula = Plant\_height ~ Irrigation\_frequency)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.579e-14 -2.430e-15 -3.500e-17 6.640e-16 3.889e-14   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.000e+00 5.252e-15 9.52e+14 <2e-16 \*\*\*  
## Irrigation\_frequency 5.000e+00 4.384e-16 1.14e+16 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.131e-14 on 18 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 1.3e+32 on 1 and 18 DF, p-value: < 2.2e-16

since the p-value is less than therefore the Irrigation Frequency is a significant variable for the Plant Height.

my\_data2=data.frame(Irrigation\_frequency,Plant\_height)  
ggplot(my\_data2,aes(x=Irrigation\_frequency,y=Plant\_height))+  
 geom\_point()+  
 geom\_smooth(method=lm)

## `geom\_smooth()` using formula = 'y ~ x'

 ## R square

summary(reg\_model)$r.squared

## Warning in summary.lm(reg\_model): essentially perfect fit: summary may be  
## unreliable

## [1] 1

The Irrigation Frequency level can explain of the variability on the Plant Height.

## Normality test

shapiro.test(reg\_model$residuals)

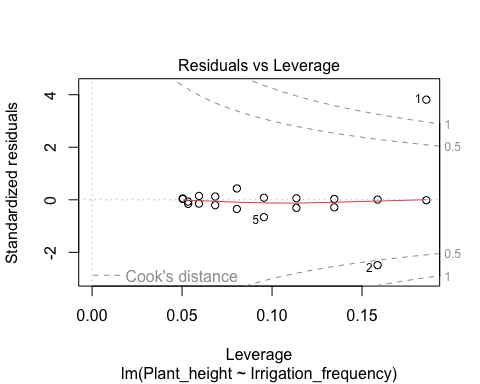
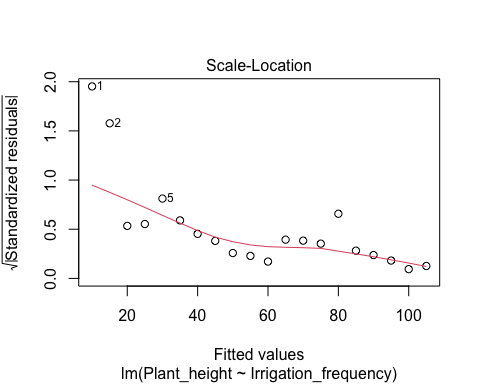
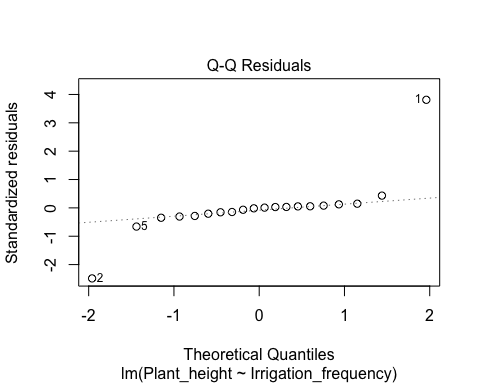
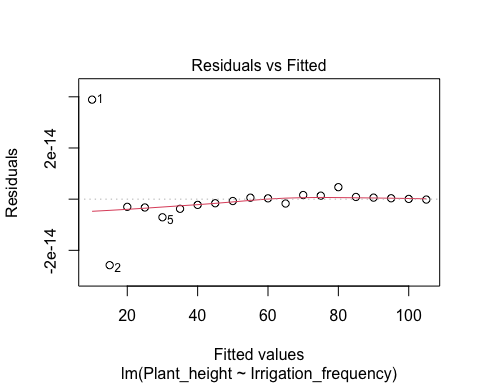
##   
## Shapiro-Wilk normality test  
##   
## data: reg\_model$residuals  
## W = 0.63404, p-value = 6.774e-06

ks.test(reg\_model$residuals,"pnorm")

##   
## Exact one-sample Kolmogorov-Smirnov test  
##   
## data: reg\_model$residuals  
## D = 0.5, p-value = 3.788e-05  
## alternative hypothesis: two-sided

Since the p-value for both the Shapiro-Wilk test and the Kolmogorov-Smirnov test are less than , we reject the null hypothesis. Therefore, the normality assumption of the error terms is not confirmed. This suggests that the error terms do not follow a normal distribution.

plot(reg\_model)

 ## Constant variance Assumption

library(lmtest)  
res<-bptest(reg\_model)  
res

##   
## studentized Breusch-Pagan test  
##   
## data: reg\_model  
## BP = 5.0702, df = 1, p-value = 0.02434

Since the p-value is , less than , we reject the null hypothesis of constant variance. This indicates that the assumption of constant variance (homoscedasticity) does not hold, suggesting the presence of heteroscedasticity in the residuals.

## Corrolation

to test is there is a significant correlation between two variables, we use the **cor.test()** function in R.

cor.test(my\_data2$Irrigation\_frequency,my\_data2$Plant\_height)

##   
## Pearson's product-moment correlation  
##   
## data: my\_data2$Irrigation\_frequency and my\_data2$Plant\_height  
## t = Inf, df = 18, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 1 1  
## sample estimates:  
## cor   
## 1

The correlation is with the p-value of less than . since the p-value is less than , so we reject the null hypothesis. We accept that there is a significant linear association between Irrigation Frequency and Plant Height.

## Install or load the packages

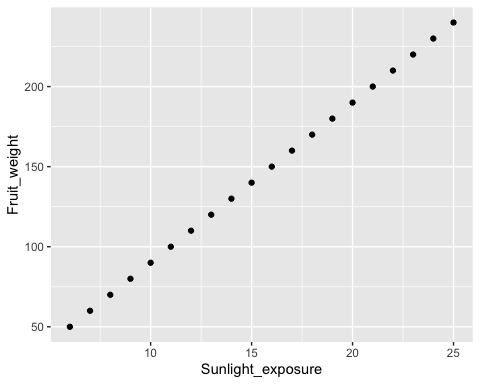
library(ggplot2)  
library(tidyverse)

## Read or give the data

Sunlight\_exposure=c(6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,  
18, 19, 20, 21, 22, 23, 24, 25)  
Fruit\_weight=c(50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150,  
160, 170, 180, 190, 200, 210, 220, 230, 240)  
  
# scottrer plot  
my\_data3=data.frame(Sunlight\_exposure,Fruit\_weight)

## Scatter plot for the data

ggplot(my\_data3,aes(x=Sunlight\_exposure,y=Fruit\_weight))+  
 geom\_point()

 ## Simple linear regression

reg\_model=lm(Fruit\_weight~Sunlight\_exposure)  
summary(reg\_model)

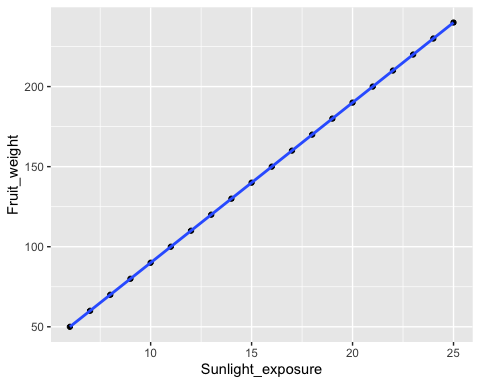
## Warning in summary.lm(reg\_model): essentially perfect fit: summary may be  
## unreliable

##   
## Call:  
## lm(formula = Fruit\_weight ~ Sunlight\_exposure)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.223e-14 -6.063e-15 -5.300e-17 1.920e-15 7.699e-14   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.000e+01 1.440e-14 -6.945e+14 <2e-16 \*\*\*  
## Sunlight\_exposure 1.000e+01 8.707e-16 1.149e+16 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.245e-14 on 18 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 1.319e+32 on 1 and 18 DF, p-value: < 2.2e-16

The p-value of less than , since the p-value is significantly less than , therefore the Sunlight Exposure is a significant variable for the Fruit Weight.

my\_data3=data.frame(Sunlight\_exposure,Fruit\_weight)  
ggplot(my\_data3,aes(x=Sunlight\_exposure,y=Fruit\_weight))+  
 geom\_point()+  
 geom\_smooth(method=lm)

## `geom\_smooth()` using formula = 'y ~ x'

 ## R square

summary(reg\_model)$r.squared

## Warning in summary.lm(reg\_model): essentially perfect fit: summary may be  
## unreliable

## [1] 1

The Sunlight Exposure level can explain of the variability on the Fruit Weight. Although this perfect fit suggests a strong linear relationship, it raises concerns about potential overfitting, meaning the model might not perform as well on new data.

## Normality test

shapiro.test(reg\_model$residuals)

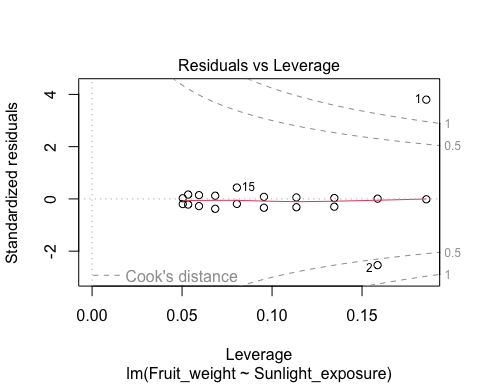
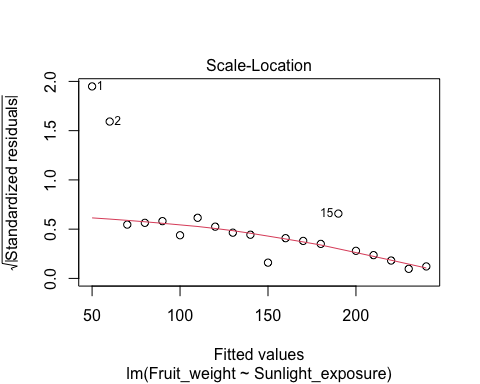
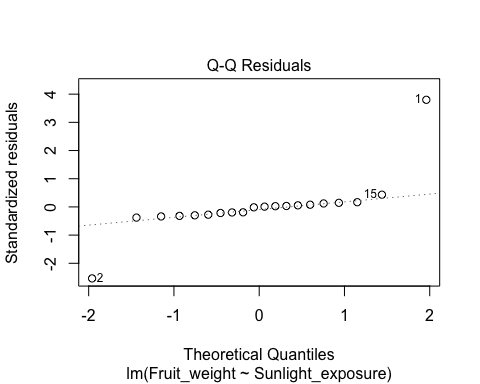
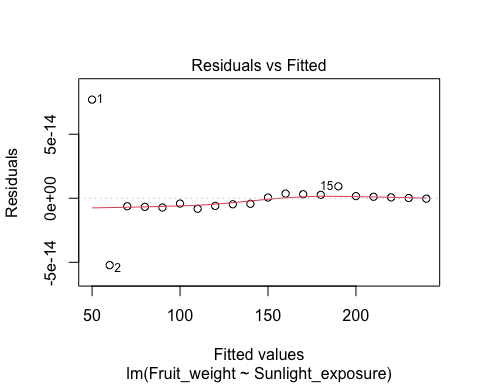
##   
## Shapiro-Wilk normality test  
##   
## data: reg\_model$residuals  
## W = 0.62526, p-value = 5.442e-06

ks.test(reg\_model$residuals,"pnorm")

##   
## Exact one-sample Kolmogorov-Smirnov test  
##   
## data: reg\_model$residuals  
## D = 0.5, p-value = 3.788e-05  
## alternative hypothesis: two-sided

since the p-value for both the Shapiro-Wilk test and the Kolmogorov-Smirnov test are less than , we reject the null hypothesis that the residuals are normally distributed. Therefore, the normality assumption of the error terms is not confirmed.

plot(reg\_model)

 ## Constant variance Assumption

library(lmtest)  
res<-bptest(reg\_model)  
res

##   
## studentized Breusch-Pagan test  
##   
## data: reg\_model  
## BP = 5.043, df = 1, p-value = 0.02473

Since the p-value is , less than , we reject the null hypothesis of constant variance. Therefore, the assumption of constant variance (homoscedasticity) does not hold.

## Corrolation

to test is there is a significant correlation between two variables, we use the **cor.test()** function in R.

cor.test(my\_data3$Sunlight\_exposure,my\_data3$Fruit\_weight)

##   
## Pearson's product-moment correlation  
##   
## data: my\_data3$Sunlight\_exposure and my\_data3$Fruit\_weight  
## t = Inf, df = 18, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 1 1  
## sample estimates:  
## cor   
## 1

The correlation is with the p-value of less than . since the p-value is significantly less than , we reject the null hypothesis. We accept that there is a perfect and significant linear association between Sunlight Exposure and Fruit Weight.

## Install or load the packages

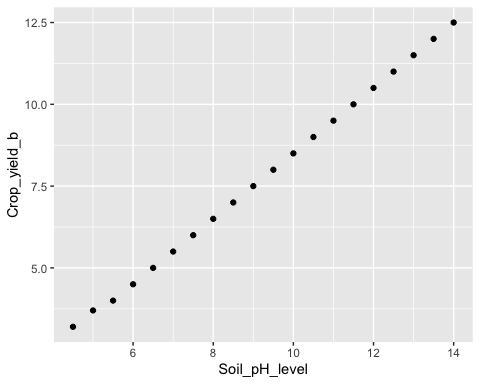
library(ggplot2)  
library(tidyverse)

## Read or give the data

Soil\_pH\_level=c(4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0, 8.5, 9.0, 9.5, 10.0, 10.5, 11.0, 11.5, 12.0, 12.5, 13.0, 13.5, 14.0)  
Crop\_yield\_b=c(3.2, 3.7, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0, 8.5, 9.0, 9.5, 10.0, 10.5, 11.0, 11.5, 12.0, 12.5)  
  
# scottrer plot  
my\_data4=data.frame(Soil\_pH\_level,Crop\_yield\_b)

## Scatter plot for the data

ggplot(my\_data4,aes(x=Soil\_pH\_level,y=Crop\_yield\_b))+  
 geom\_point()

 ## Simple linear regression

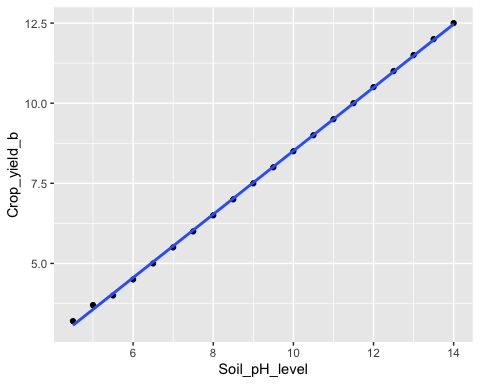
reg\_model=lm(Crop\_yield\_b~Soil\_pH\_level)  
summary(reg\_model)

##   
## Call:  
## lm(formula = Crop\_yield\_b ~ Soil\_pH\_level)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.060602 -0.034887 -0.009173 0.016541 0.133985   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.379850 0.040587 -34.0 <2e-16 \*\*\*  
## Soil\_pH\_level 0.989173 0.004189 236.1 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05401 on 18 degrees of freedom  
## Multiple R-squared: 0.9997, Adjusted R-squared: 0.9997   
## F-statistic: 5.576e+04 on 1 and 18 DF, p-value: < 2.2e-16

since the p-value is less than therefore the Soil pH Level is a significant variable for the Crop Yield.

my\_data4=data.frame(Soil\_pH\_level,Crop\_yield\_b)  
ggplot(my\_data4,aes(x=Soil\_pH\_level,y=Crop\_yield\_b))+  
 geom\_point()+  
 geom\_smooth(method=lm)

## `geom\_smooth()` using formula = 'y ~ x'

 ## R square

summary(reg\_model)$r.squared

## [1] 0.9996773

The Soil pH level can explain approximately of the variability on the Crop Yield.

## Normality test

shapiro.test(reg\_model$residuals)

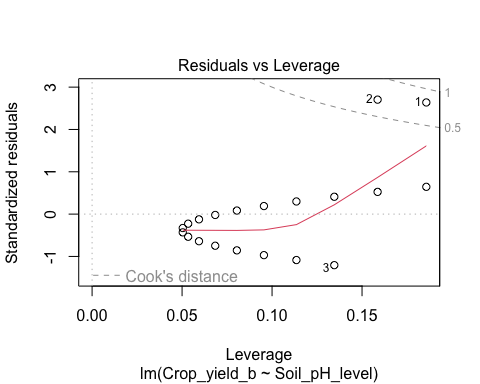
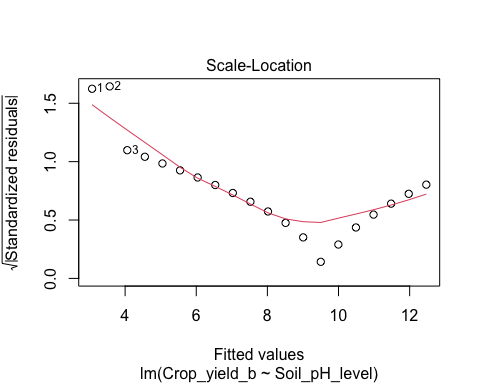
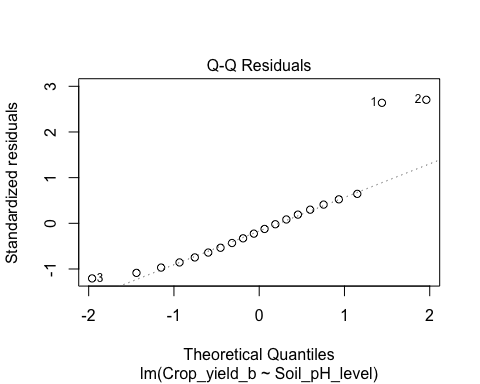
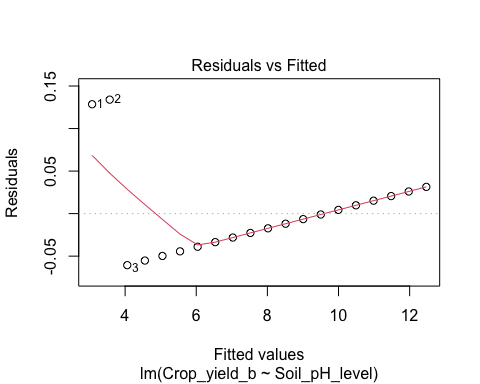
##   
## Shapiro-Wilk normality test  
##   
## data: reg\_model$residuals  
## W = 0.83777, p-value = 0.003348

ks.test(reg\_model$residuals,"pnorm")

##   
## Exact one-sample Kolmogorov-Smirnov test  
##   
## data: reg\_model$residuals  
## D = 0.47584, p-value = 0.0001108  
## alternative hypothesis: two-sided

since the p-value for both the Shapiro-Wilk test and the Kolmogorov-Smirnov test are less than , we reject the null hypothesis. Therefore, the normality assumption of the error terms is not confirmed. This suggests that the error terms do not follow a normal distribution.

plot(reg\_model)

 ## Constant variance Assumption

library(lmtest)  
res<-bptest(reg\_model)  
res

##   
## studentized Breusch-Pagan test  
##   
## data: reg\_model  
## BP = 8.2056, df = 1, p-value = 0.004176

Since the p-value is , less than , we reject the null hypothesis of constant variance. Therefore, the assumption of constant variance (homoscedasticity) does not hold.

## Corrolation

to test is there is a significant correlation between two variables, we use the **cor.test()** function in R.

cor.test(my\_data4$Soil\_pH\_level,my\_data4$Crop\_yield\_b)

##   
## Pearson's product-moment correlation  
##   
## data: my\_data4$Soil\_pH\_level and my\_data4$Crop\_yield\_b  
## t = 236.14, df = 18, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9995825 0.9999376  
## sample estimates:  
## cor   
## 0.9998386

The correlation is with the p-value of less than . since the p-value is less than so we reject the null hypothesis. We accept that there is a perfact and significant linear association between Soil pH Level and Crop Yield.

## Install or load the packages

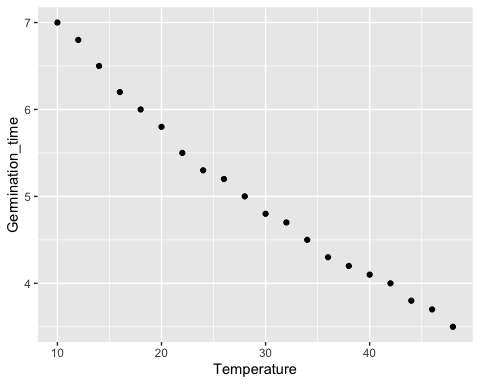
library(ggplot2)  
library(tidyverse)

## Read or give the data

Temperature=c(10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48)  
Germination\_time=c(7, 6.8, 6.5, 6.2, 6, 5.8, 5.5, 5.3, 5.2, 5, 4.8, 4.7, 4.5, 4.3, 4.2, 4.1, 4, 3.8, 3.7, 3.5)  
  
# scottrer plot  
my\_data5=data.frame(Temperature,Germination\_time)

## Scatter plot for the data

ggplot(my\_data5,aes(x=Temperature,y=Germination\_time))+  
 geom\_point()

 ## Simple linear regression

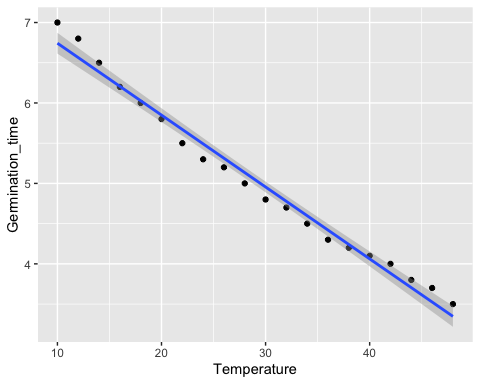
reg\_model=lm(Germination\_time~Temperature)  
summary(reg\_model)

##   
## Call:  
## lm(formula = Germination\_time ~ Temperature)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.19181 -0.11468 -0.03436 0.11511 0.25714   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.636466 0.086571 88.21 <2e-16 \*\*\*  
## Temperature -0.089361 0.002774 -32.22 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1431 on 18 degrees of freedom  
## Multiple R-squared: 0.983, Adjusted R-squared: 0.982   
## F-statistic: 1038 on 1 and 18 DF, p-value: < 2.2e-16

since the p-value is less than therefore the Temperature is a significant variable for the Germination Time.

my\_data5=data.frame(Temperature,Germination\_time)  
ggplot(my\_data5,aes(x=Temperature,y=Germination\_time))+  
 geom\_point()+  
 geom\_smooth(method=lm)

## `geom\_smooth()` using formula = 'y ~ x'

 ## R square

summary(reg\_model)$r.squared

## [1] 0.9829513

The Temperature level can explain of the variability on the Germination Time.

## Normality test

shapiro.test(reg\_model$residuals)

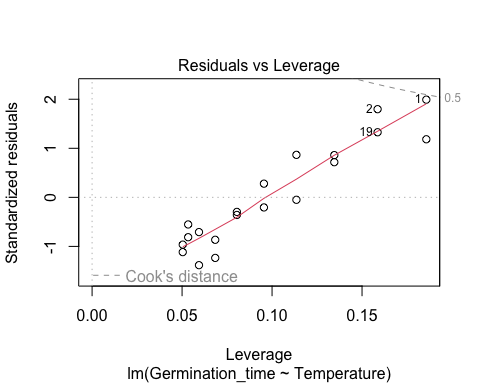
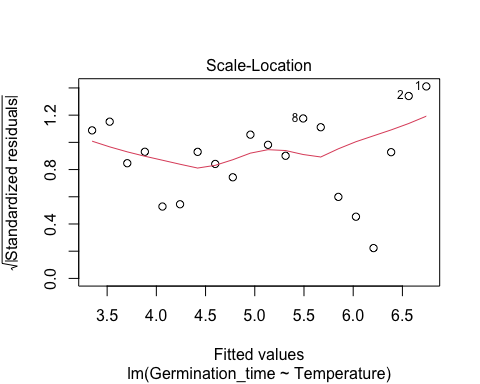
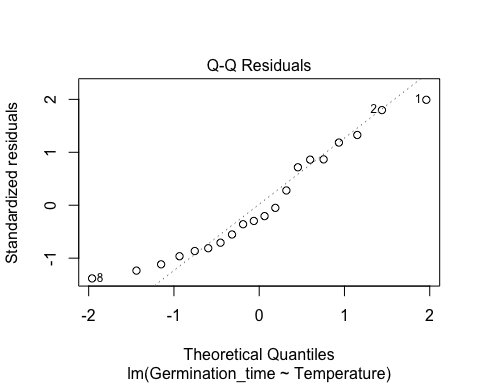
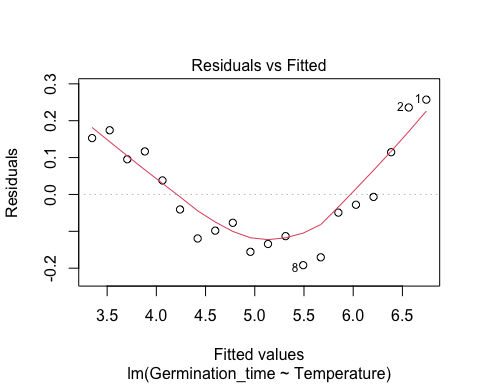
##   
## Shapiro-Wilk normality test  
##   
## data: reg\_model$residuals  
## W = 0.93676, p-value = 0.2081

ks.test(reg\_model$residuals,"pnorm")

##   
## Exact one-sample Kolmogorov-Smirnov test  
##   
## data: reg\_model$residuals  
## D = 0.42395, p-value = 0.0008911  
## alternative hypothesis: two-sided

since the p-value for the Shapiro-Wilk test is , higher than , this test does not reject the null hypothesis of normality. However, the Kolmogorov-Smirnov test’s p-value is , significantly less than , which rejects the null hypothesis of normality.Therefore, The normality of the residuals is not conclusively confirmed. There is evidence suggesting that the residuals may not be normally distributed.

plot(reg\_model)

 ## Constant variance Assumption

library(lmtest)  
res<-bptest(reg\_model)  
res

##   
## studentized Breusch-Pagan test  
##   
## data: reg\_model  
## BP = 2.3077, df = 1, p-value = 0.1287

Since the p-value is , more than , so the constant variance assumption is hold.

## Corrolation

to test is there is a significant correlation between two variables, we use the **cor.test()** function in R.

cor.test(my\_data5$Temperature,my\_data5$Germination\_time)

##   
## Pearson's product-moment correlation  
##   
## data: my\_data5$Temperature and my\_data5$Germination\_time  
## t = -32.215, df = 18, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.9966828 -0.9779973  
## sample estimates:  
## cor   
## -0.991439

The correlation is with the p-value of . since the p-value is less than so we reject the null hypothesis. We accept that there is a strong and significant negative linear association between Temperature and Germination Time.