Polynomial regression model for inbuilt dataset ‘pressure’

Nithya S

2024-01-08

Objective:

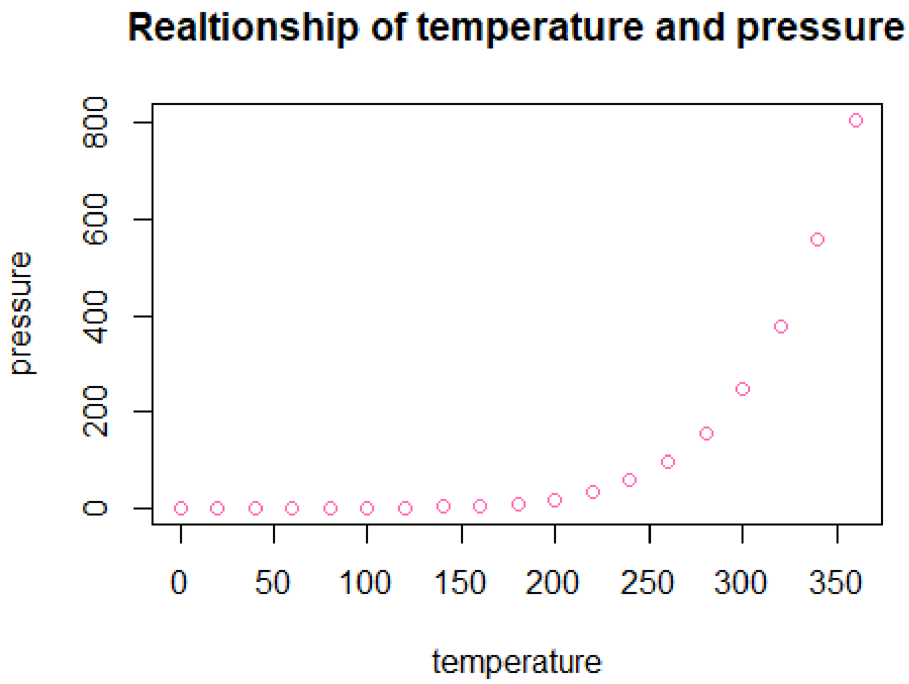
1. By using the inbuilt data set “pressure” in R, fit a suitable linear regression model that relates pressure and temperature.
2. Also, perform residual analysis and comment about the possibility of modelling non linear regression model for the data set.

First we call for the data and plot the same.

```
data=pressure
data

##    temperature pressure
## 1            0   0.0002
## 2           20   0.0012
## 3           40   0.0060
## 4           60   0.0300
## 5           80   0.0900
## 6          100   0.2700
## 7          120   0.7500
## 8          140   1.8500
## 9          160   4.2000
## 10         180   8.8000
## 11         200  17.3000
## 12         220  32.1000
## 13         240  57.0000
## 14         260  96.0000
## 15         280 157.0000
## 16         300 247.0000
## 17         320 376.0000
## 18         340 558.0000
## 19         360 806.0000

plot(data$temperature,data$pressure,main="Realtionship of temperature and 
pressure",col='hotpink',xlab="temperature",ylab="pressure")
```



The above graph shows that there could be a linear relationship but of a polynomial degree. So instead of proceeding for a linear regression model, we fit a polynomial regression model and proceed for residual analysis. Also the graph has one bend, so our model could be

$$Y[Pressure] = B_0 + B_1X[temperature] + B_2X^2 + E$$

To fit a polynomial regression model, we proceed as follows:

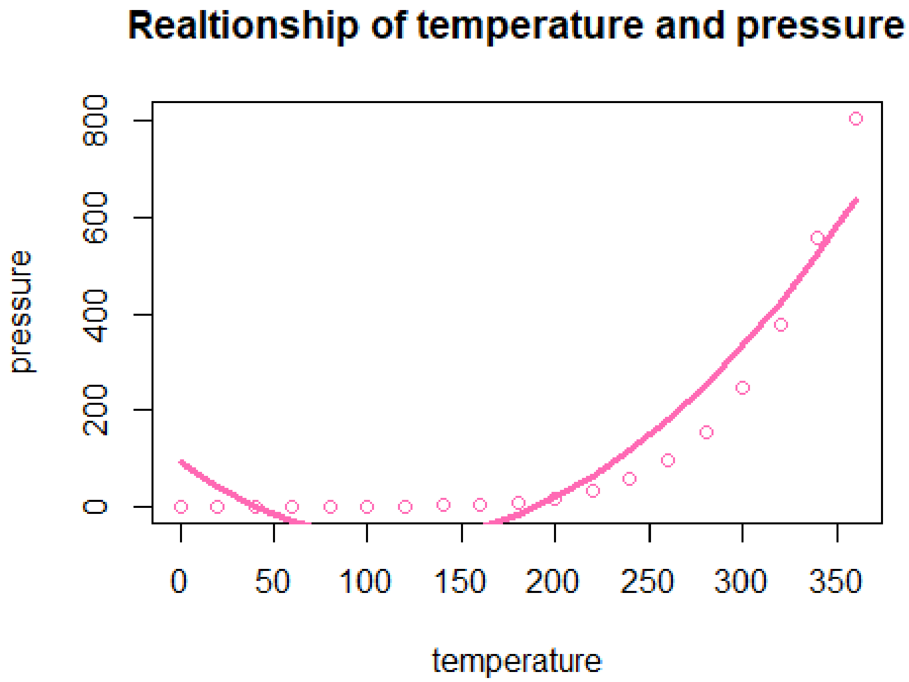
```
model=lm(pressure~temperature+I(temperature^2),data=data)
model

##
## Call:
## lm(formula = pressure ~ temperature + I(temperature^2), data = data)
##
## Coefficients:
##      (Intercept)       temperature  I(temperature^2)
##         91.15438          -2.70617           0.01172

summary(model)

##
## Call:
## lm(formula = pressure ~ temperature + I(temperature^2), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -95.142 -54.391  -1.353  48.238 170.374 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      91.154379  46.262513   1.970 0.066354 .  
## temperature      -2.706167   0.595775  -4.542 0.000333 ***
## I(temperature^2)  0.011718   0.001597   7.336 1.67e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 74.42 on 16 degrees of freedom
## Multiple R-squared:  0.9024, Adjusted R-squared:  0.8902 
## F-statistic:    74 on 2 and 16 DF,  p-value: 8.209e-09
```

```
plot(data$temperature,data$pressure,main="Realtionship of temperature and pressure",col='hotpink',xlab="temperature",ylab="pressure")
lines(smooth.spline(data$temperature,predict(model)),col="hotpink",lwd=3)
```



We validate the model as

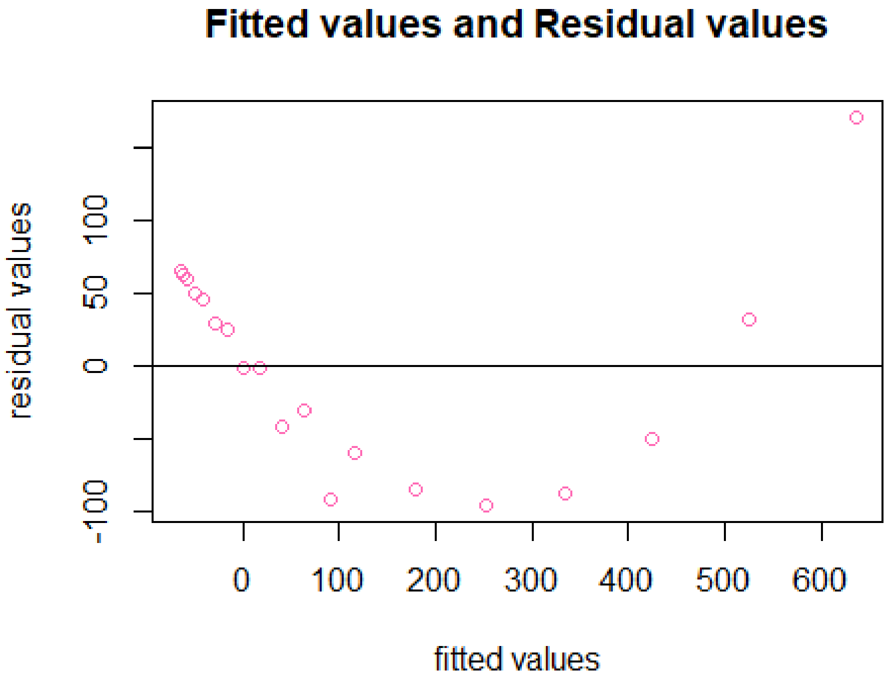
p-values of the regression coefficients are lesser[almost zero] than the significance level [5%]. So we say that the regression coefficients are significant. Also, the R-squared value is 0.9024, whic says that 90.24% of the variation in the data is explained by temperature for pressure.

i) First, we get the residuals of the model and plot the obtained residuals with fitted values in order to understand the linearity, constant variance and outliers of the model.

```
fitted_values=fitted.values(model)
fitted_values

##          1          2          3          4          5          6
7
##  91.154379  41.718359   1.656977 -29.029768 -50.341876 -62.279346 -
64.842179
##          8          9         10         11         12         13
14
## -58.030375 -41.843934 -16.282855  18.652862  62.963215 116.648206
179.707834
##         15         16         17         18         19
## 252.142100 333.951003 425.134543 525.692720 635.625535

plot(fitted_values,resid(model),col="hotpink",main="Fitted values and
Residual values",xlab="fitted values",ylab="residual values")
abline(0,0)
```



The plot resembles a u-shaped upward open funnel. The plot throws light on the linearity of the variables. Here we say that X and Y variables are not linear. This graph does not infer anything about the variance of the error term.

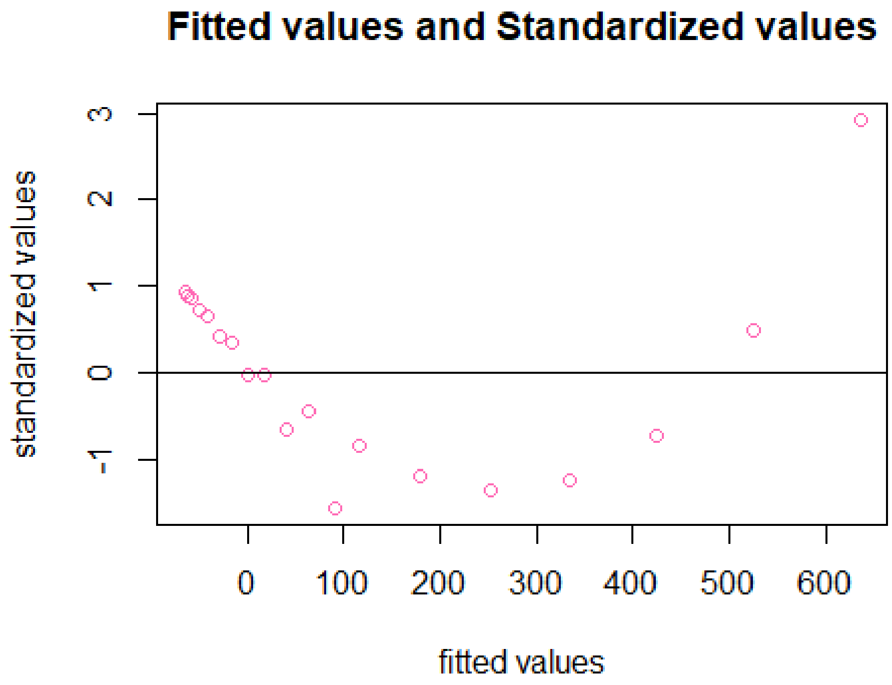
Now, to see if there are any outliers or influential points in the model: We check the same by using two methods

1. Standardized residuals

2. Studentized residuals

Let us use standardized residuals. The procedure is as follows:

```
plot(fitted_values,rstandard(model),col="hotpink",main="Fitted values and
Standardized values",xlab="fitted values",ylab="standardized values")
abline(0,0)
```

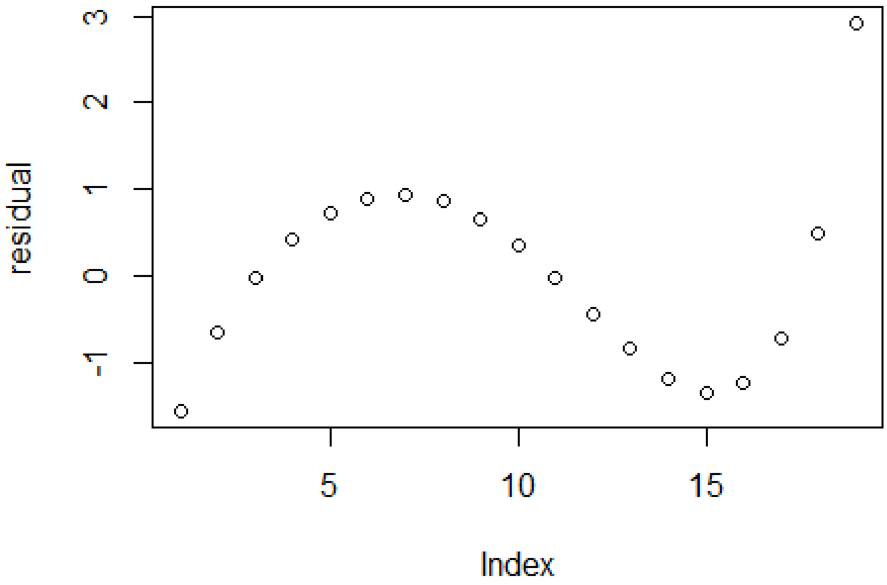


To be very precise we can even check in the residual values for outliers as follows:

```
residual=rstandard(model)
residual

##           1           2           3           4           5           6 
## -1.56380948 -0.64736155 -0.02428165  0.41590439  0.71368868  0.88361058 
##           7           8           9          10          11          12 
##  0.92956879  0.85268591  0.65821226  0.35909478 -0.01933957 -0.43948670 
##          13          14          15          16          17          18 
## -0.84533112 -1.18250841 -1.34640718 -1.24444570 -0.72264345  0.50134024 
##          19 
##  2.92288522

plot(residual)
```



Now we have numerically seen that there are no outliers in the dataset with the maximum value being 2.92288522 Since, there are no outliers in the model, we proceed further.

ii) Now, we check if the normality assumption is satisfied by using

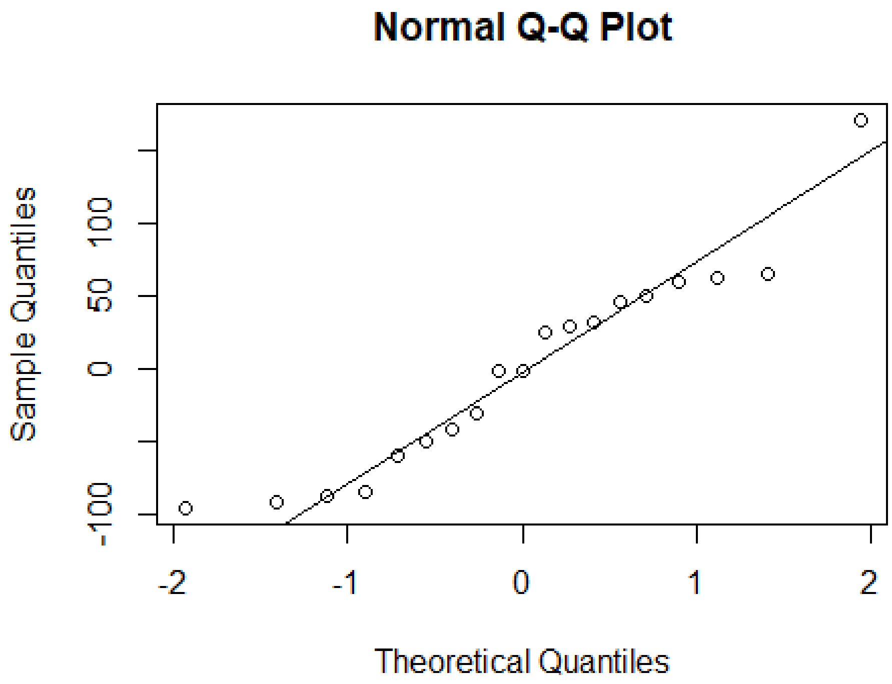
1. QQ plot
2. Shapiro test

The QQ plot is as follows:

```
resi=resid(model)

qqnorm(resi) # QQ plot-of residual

qqline(resi) # plots the points
```



If not all, majority of the points fall on the line thus the quartile of normal and residual are almost same, hence it indicates that the residuals follow a normal distribution. However the assumption of normality has to be verified by using a statistical test.(but to know if the deviation of the points lying away from the line, we use the test to further confirm the normality.)

The Shapiro test is as follows:

Hypothesis to testing for normality:

H_0: Errors follow normal distribution.

v/s

H_1: Errors do not follow normal distribution.

```
shapiro.test(residual)

##
##  Shapiro-Wilk normality test
##
## data:  residual
## W = 0.92626, p-value = 0.1477
```

At 0.05 level of significance the p value is 0.1477 which is greater than 0.05, thus we accept null and and say that the residuals follow normal distribution. Hence one of the assumption of errors is satisfied. The assumption of normality is satisfied.

Hypothesis testing for constant variance:

To test if the errors have constant variance.

H_0: Errors have constant variance

v/s

H_1: Errors have no constant variance

```
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

bptest(model)

##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 7.4533, df = 2, p-value = 0.02407
```

At 0.05 level of significance the p value is 0.02407 which is lesser than 0.05, thus we reject null and and say that the errors do not have constant variance.

There is no problem with normality as all errors follow a normal distribution. But the assumption of constant variance is not validated. We can fix this problem by performing a transformation or by fitting the non-linear regression model.

Conclusion

A suitable linear regression model for the inbuilt data set in R called ‘pressure’ was built.

$$Y[Pressure] = B_0 + B_1X[temperature] + B_2X^2 + E$$

is the model.

The estimated model is

$$Y[Pressure]^{hat} = 91.15438 - 2.70617X[temperature] + 0.01172X^2$$

The model had R-squared value as 0.9024, which says that 90.24% of the variation in the data is explained by temperature for pressure. Residual analysis was performed. The normality condition was satisfied, the assumption of constant variance was not satisfied, which lead to the idea of building a non-linear regression model.

Characteristics of non-linear regression model is as follows. A statistical technique for modeling relationships between variables that aren't strictly linear. It fits a nonlinear function to the data, capturing more complex patterns than linear regression.

Choosing the appropriate nonlinear function is crucial for accurate results. Examining residuals for patterns can reveal model inadequacy.