HW 11

Priya Shaji

11/3/2019

Using R, build a regression model for data that interests you. Conduct residual analysis. Was the linear model appropriate? Why or why not?

Dataset: Air Quality built-in dataset of R

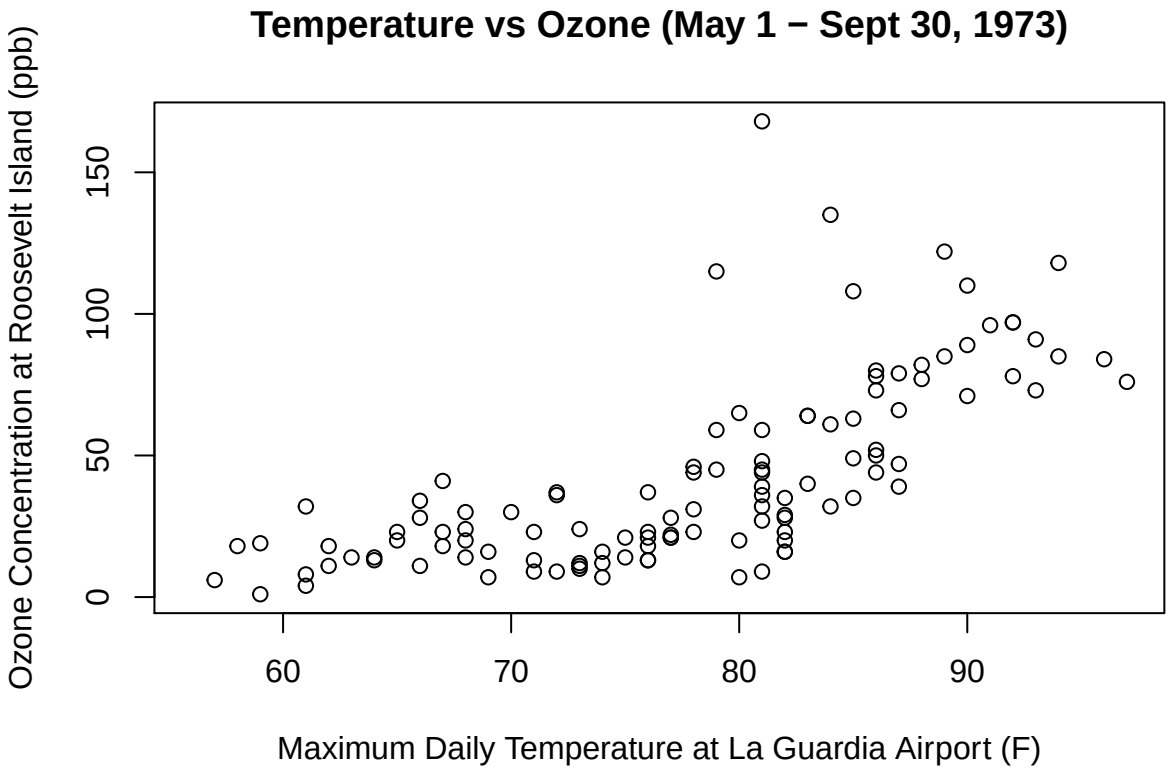
"Air Quality" dataset contains daily air quality measurements in New York recorded from May to September 1973. The model compares Ozone (mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island) with Temp (maximum daily temperature in degrees Fahrenheit at La Guardia Airport).

Step 1) load the dataset

```
data("airquality")
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

```
plot(airquality$Temp, airquality$Ozone,
     xlab='Maximum Daily Temperature at La Guardia Airport (F)',
     ylab='Ozone Concentration at Roosevelt Island (ppb)',
     main='Temperature vs Ozone (May 1 - Sept 30, 1973)')
```



THe above model seems to have a correlation.

Step 2) Let's build a simple linear regression model

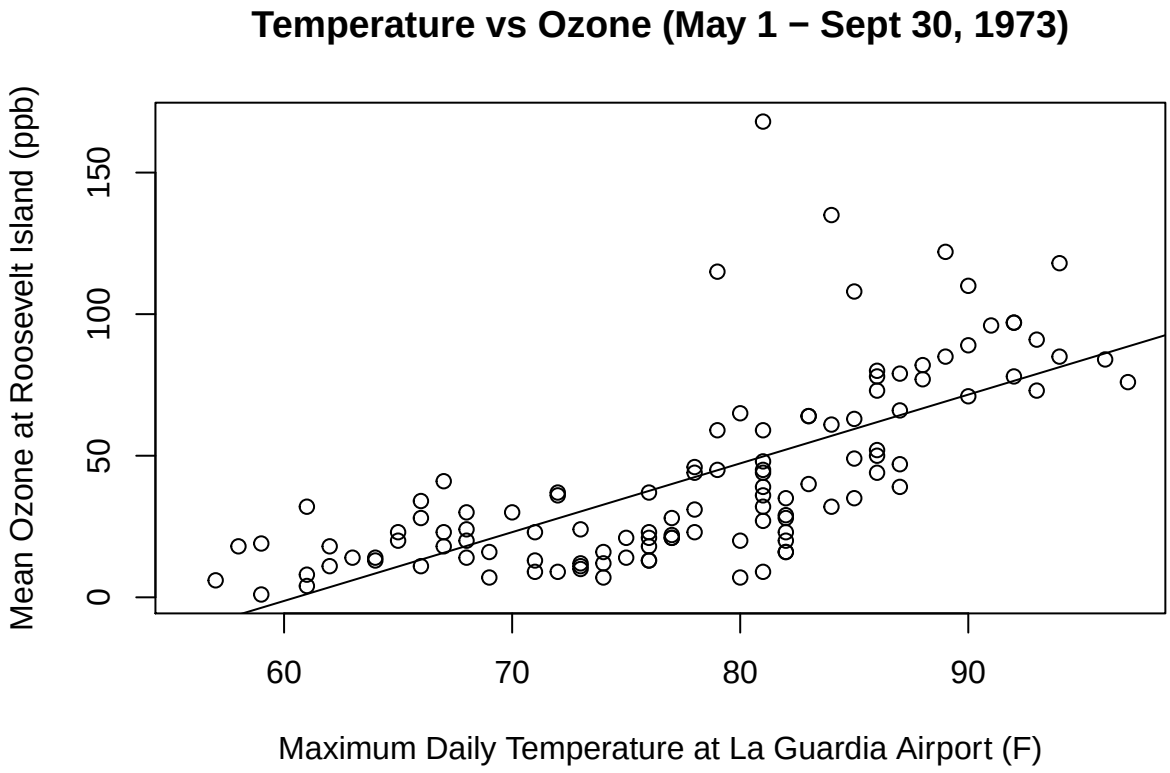
```
linear_regression <- lm(airquality$Ozone ~ airquality$Temp)
summary(linear_regression)
```

```
##
## Call:
## lm(formula = airquality$Ozone ~ airquality$Temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.729 -17.409  -0.587  11.306 118.271
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -146.9955    18.2872  -8.038 9.37e-13 ***
## airquality$Temp    2.4287     0.2331  10.418  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.71 on 114 degrees of freedom
##   (37 observations deleted due to missingness)
## Multiple R-squared:  0.4877, Adjusted R-squared:  0.4832
## F-statistic: 108.5 on 1 and 114 DF,  p-value: < 2.2e-16
```

R-squared value = 0.48. Therefore, our model shows approx 50% varibility.

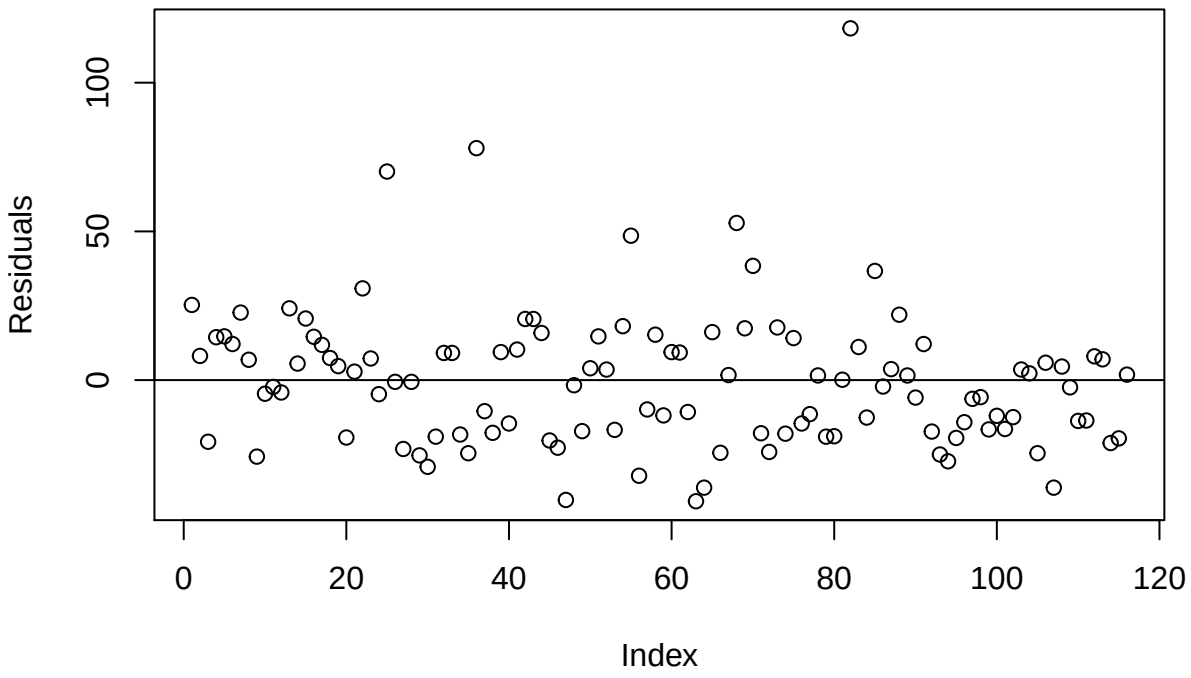
Step 3) Plotting regression line

```
plot(airquality$Ozone ~ airquality$Temp,
     xlab='Maximum Daily Temperature at La Guardia Airport (F)',
     ylab='Mean Ozone at Roosevelt Island (ppb)',
     main='Temperature vs Ozone (May 1 - Sept 30, 1973)')
abline(linear_regression)
```



Step 3) Now, plot the residuals

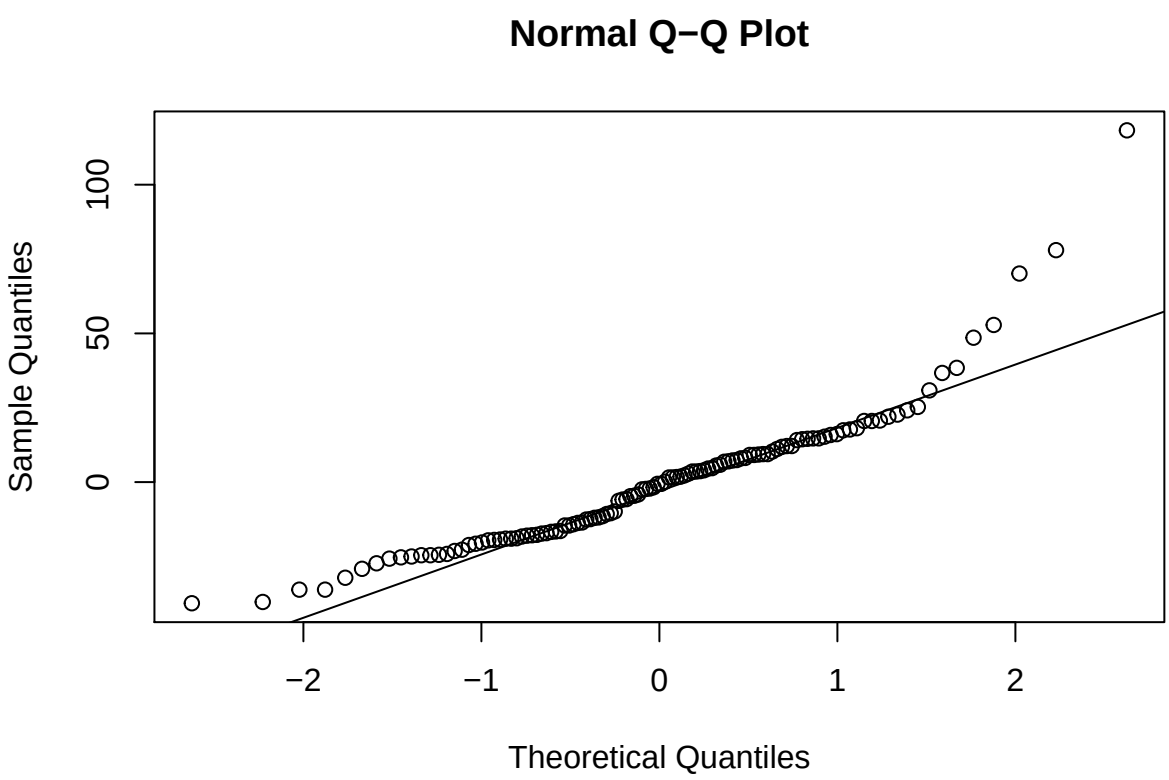
```
plot(linear_regression$residuals, ylab='Residuals')
abline(a=0, b=0)
```



No. as we go through the residuals plot , varaibility seems constant with no significant pattern.

Step 4) Plot the Q-Q plot(quantile-quntile)

```
qqnorm(linear_regression$residuals)
qqline(linear_regression$residuals)
```



The Q-Q plot looks even except noth tails at either side.

Conclusion

The R-squared value is approx 50%, therefore model describes almost 50% of variability.Residual plot shows that variability is constant with no pattern. The Q-Q plot looks even except noth tails at either side.

Therefore, as per the above insights, Temperature is statistically significant variable to predict ozone concentration at all levels. Looking at the plots, there are few outliers, Non-linear model can fit the data better