Linear Regression on Bike Sharing Dataset

Saksham Rana

April 14, 2019

---

1 Introduction

The purpose of this exercise is to study the validation of assumptions of simple linear regression and draw conclusions as to what all parameters do we need to look at in order to decide whether linear regression is the right way to go.

So we will be observing a dataset in which there is violation of assumptions and reporting our observations based data analysis and residual analysis.

2 Importing Libraries

```
library(readr)
library(dplyr)
library(ggplot2)
library(plotly)
library(caTools)
```

3 Importing the Data

```
bike_data <- read_csv("hour.csv", progress = show_progress())
```

4 Data Exploration

4.1 Some observations of Data

```
head(bike_data)
```

```
## # A tibble: 6 x 17
##   instant dteday     season    yr  mnth    hr holiday weekday workingday
##     <dbl> <date>      <dbl> <dbl> <dbl> <dbl>   <dbl>   <dbl>      <dbl>
## 1       1 2011-01-01      1     0     1     0       0       6          0
## 2       2 2011-01-01      1     0     1     1       0       6          0
## 3       3 2011-01-01      1     0     1     2       0       6          0
## 4       4 2011-01-01      1     0     1     3       0       6          0
## 5       5 2011-01-01      1     0     1     4       0       6          0
## 6       6 2011-01-01      1     0     1     5       0       6          0
## # ... with 8 more variables: weathersit <dbl>, temp <dbl>, atemp <dbl>,
## #   hum <dbl>, windspeed <dbl>, casual <dbl>, registered <dbl>, cnt <dbl>
```

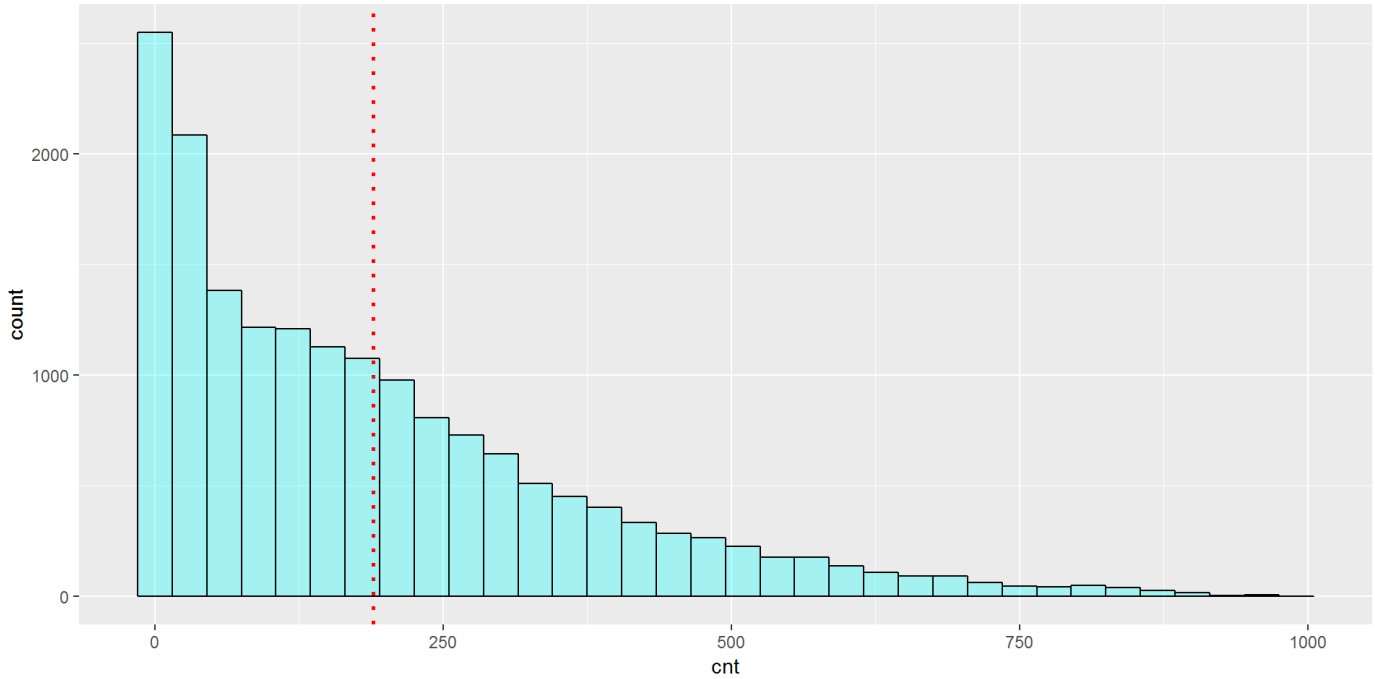
4.2 Check for any NA values

```
sapply(bike_data, function(x)(sum(is.na(x))))
```

```
##    instant     dteday     season         yr       mnth         hr
##          0          0          0          0          0          0
##    holiday    weekday workingday weathersit       temp      atemp
##          0          0          0          0          0          0
##        hum  windspeed     casual registered        cnt
##          0          0          0          0          0
```

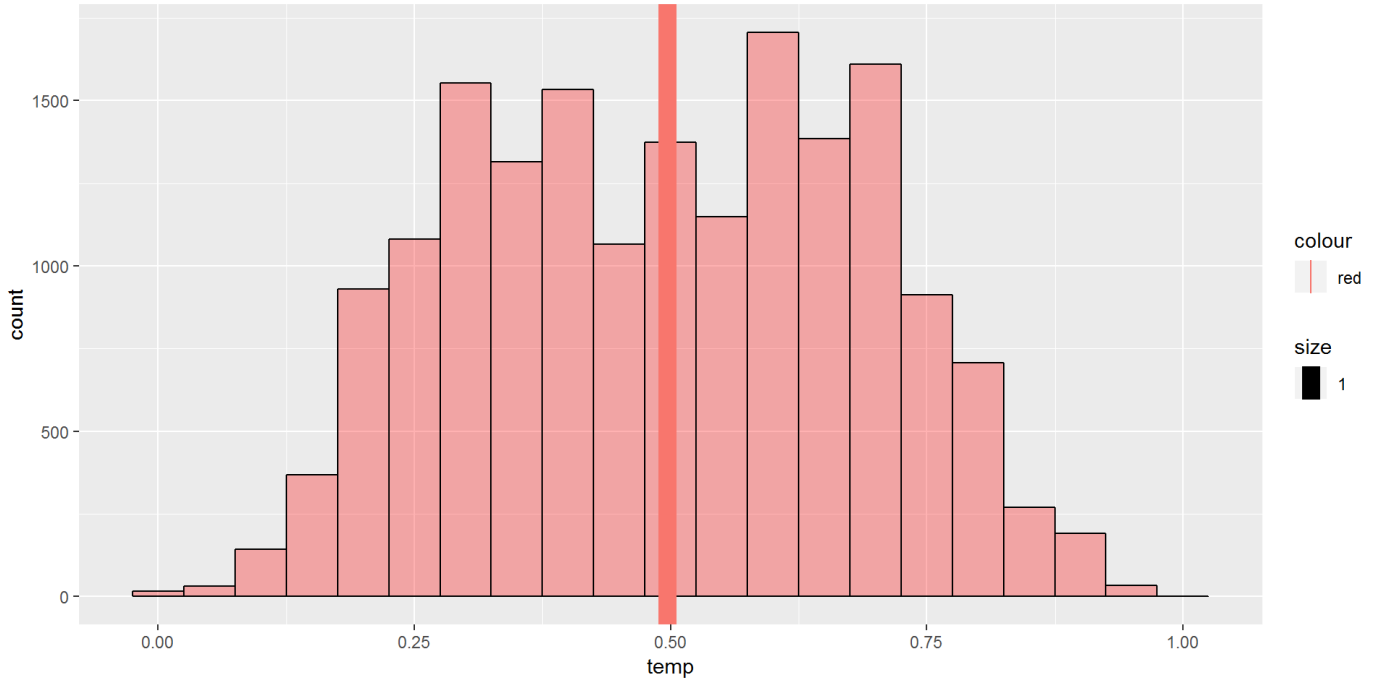
4.3 Check for normality in response and predictor variables

```
ggplot(bike_data, aes(x = cnt)) +
  geom_histogram(position = "identity", color = "black", fill = "cyan", alpha = 0.3, binwidth = 30) +
  geom_vline(aes(xintercept = mean(cnt)), colour = "red", linetype = "dotted", size = 1)
```



It is observed that the data is normally distributed and validates the assumption of regression. However the data seems to be skewed towards large distribution tending towards the left of mean.

```
ggplot(bike_data, aes(x = temp)) +
  geom_histogram(color = "black", fill = "red", alpha = 0.3, binwidth = 0.05) +
  geom_vline(aes(xintercept = mean(temp), colour = "red", size = 1))
```



It is observed that distribution is skewed implying that assumption of linear regression is not validated; this may result in a faulty/inaccurate model.

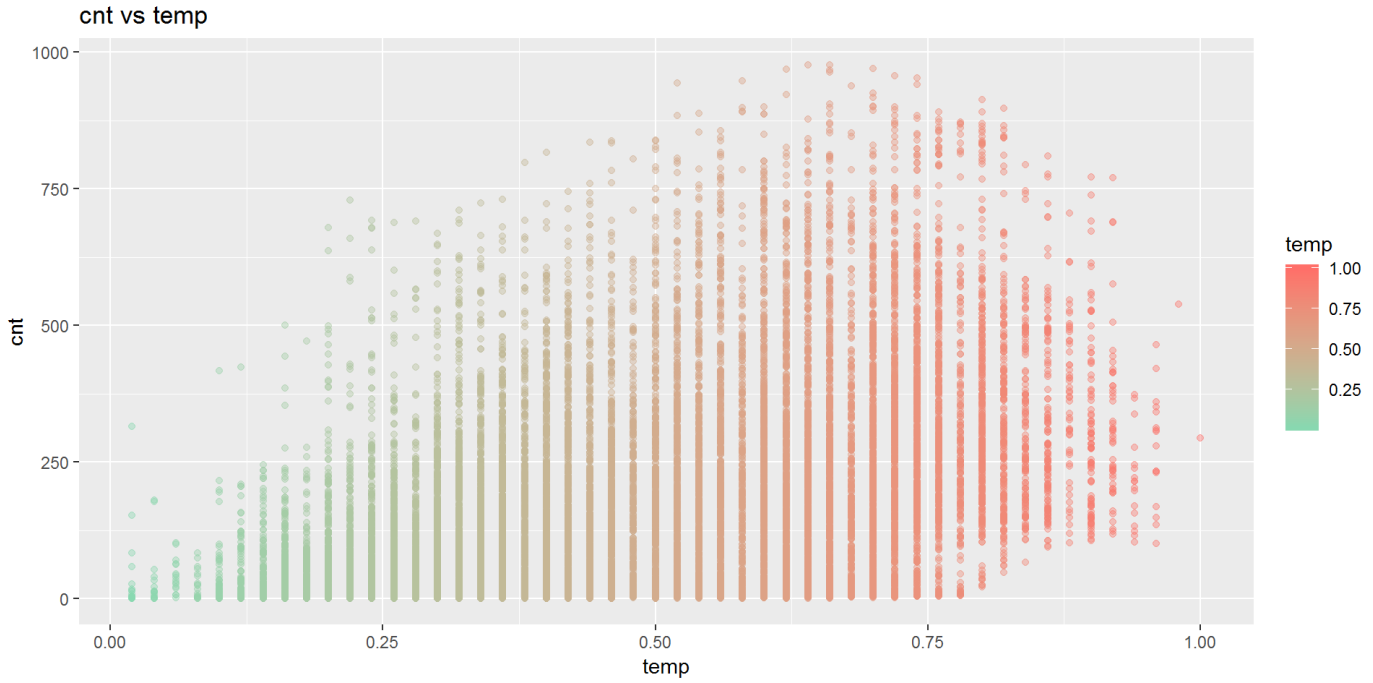
4.4 Check correlation among variables

The Pearson coefficient indicates that there is strong linear relationship between the predictor and response variable, which is a desired to design a linear model.

The correlation between “bikes rented” and “temperature” is 0.4047723.

It is observed that there is little positive linear relationship between “cnt” and “temp”. We will observe scatter plot will have high heteroskedasticity in data.

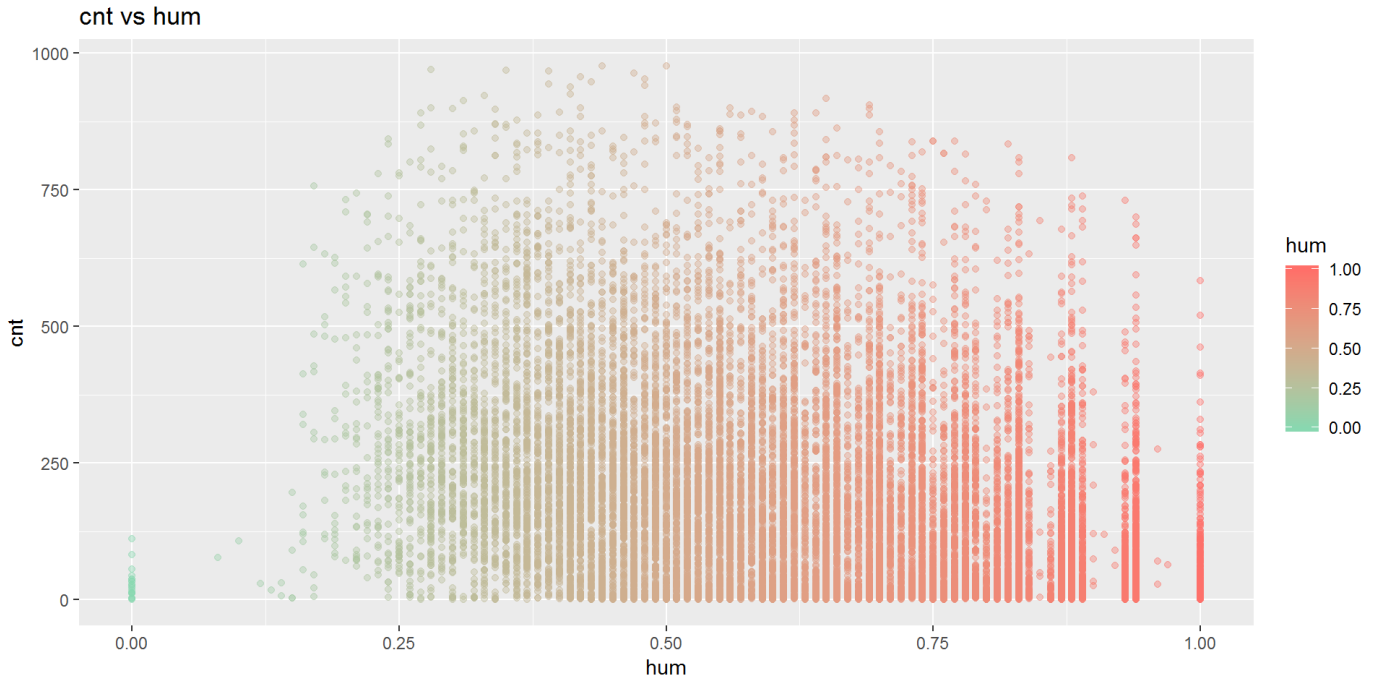
```
ggplot(bike_data, aes(x = temp, y = cnt)) +
  geom_point(aes(color = temp), alpha = 0.4) +
  scale_color_gradient(low = "#88d8b0", high = "#ff6f69") +
  ggtitle(label = "cnt vs temp")
```



It is observed that as temperature increases the number of bikes rented increases, however relation is not strong. Also looking at the spreading nature of scatter plot we can say that there is heteroskedasticity in data. Thus linear regression will not produce an accurate model.

Now let us check some more scatter plots to see if we identify a variable that shows high linearity with response variable.

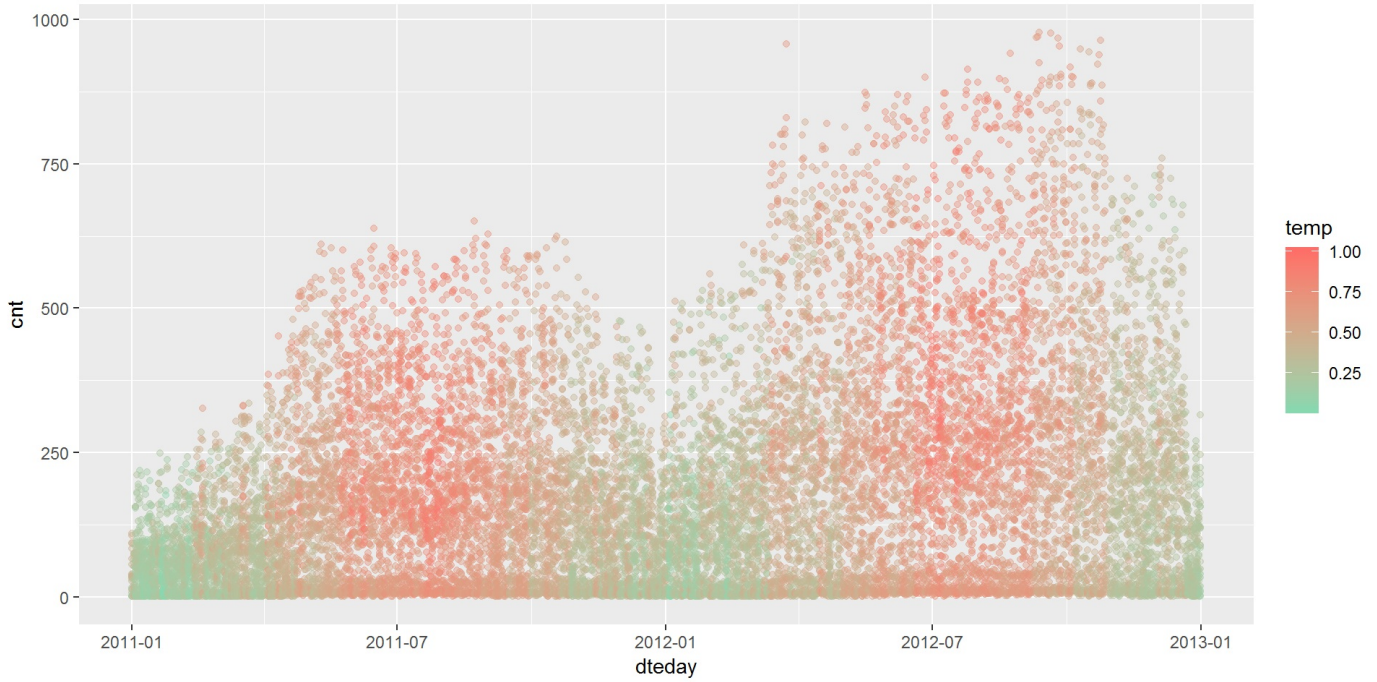
```
ggplot(bike_data, aes(x = hum, y =  cnt)) +
  geom_point(aes(color = hum), alpha = 0.4) +
  scale_color_gradient(low = "#88d8b0", high = "#ff6f69") +
  ggtitle(label = "cnt vs hum")
```



Here also spread is high and there is no signs if linear relationship. Also Pearson Coefficient is -0.3229107. This shows there is negative weak linear relation between humidity and count; indicationg linear regression is not the right way to go for prediction.

Let us now check how rentals are affected throughout the year.

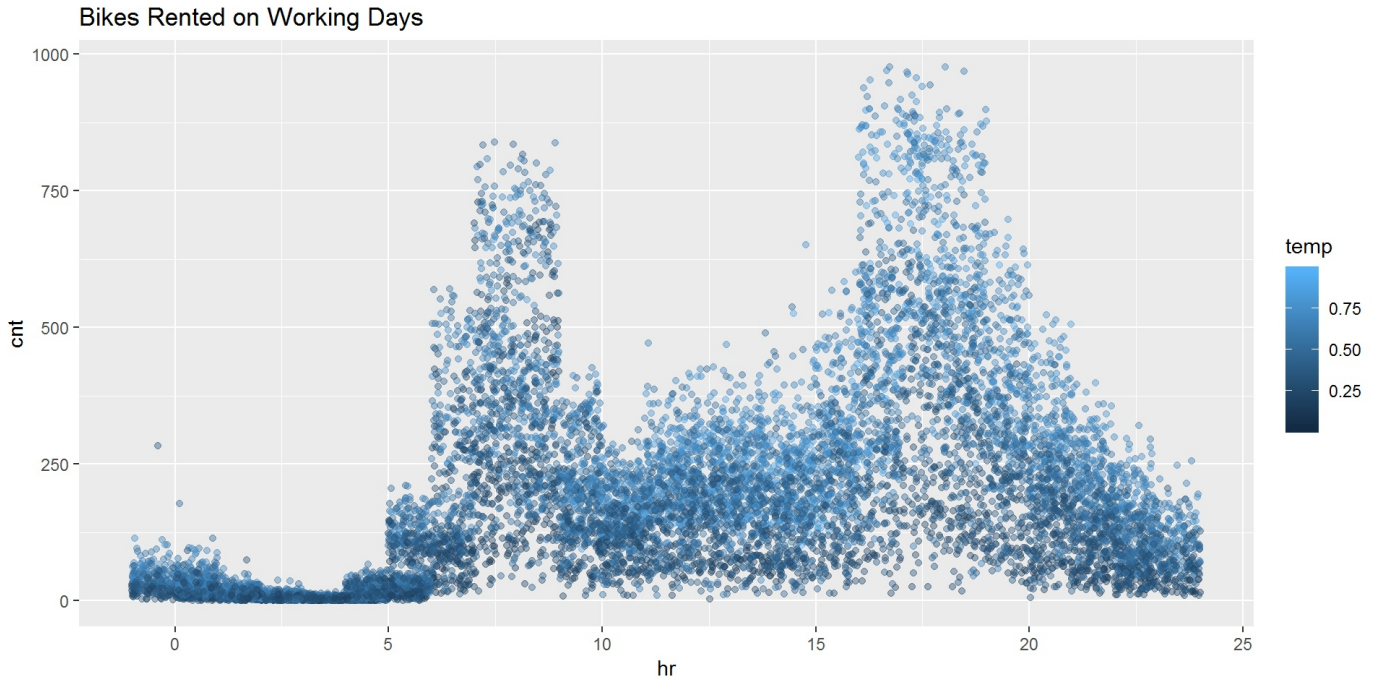
```
ggplot(bike_data, aes(x = dteday, y = cnt)) +
  geom_point(aes(color = temp), alpha = 0.4) +
  scale_color_gradient(low = "#88d8b0", high = "#ff6f69")
```



It is observed that bike rentals increase during summers as compared. Also overall the bike rentals have increased.

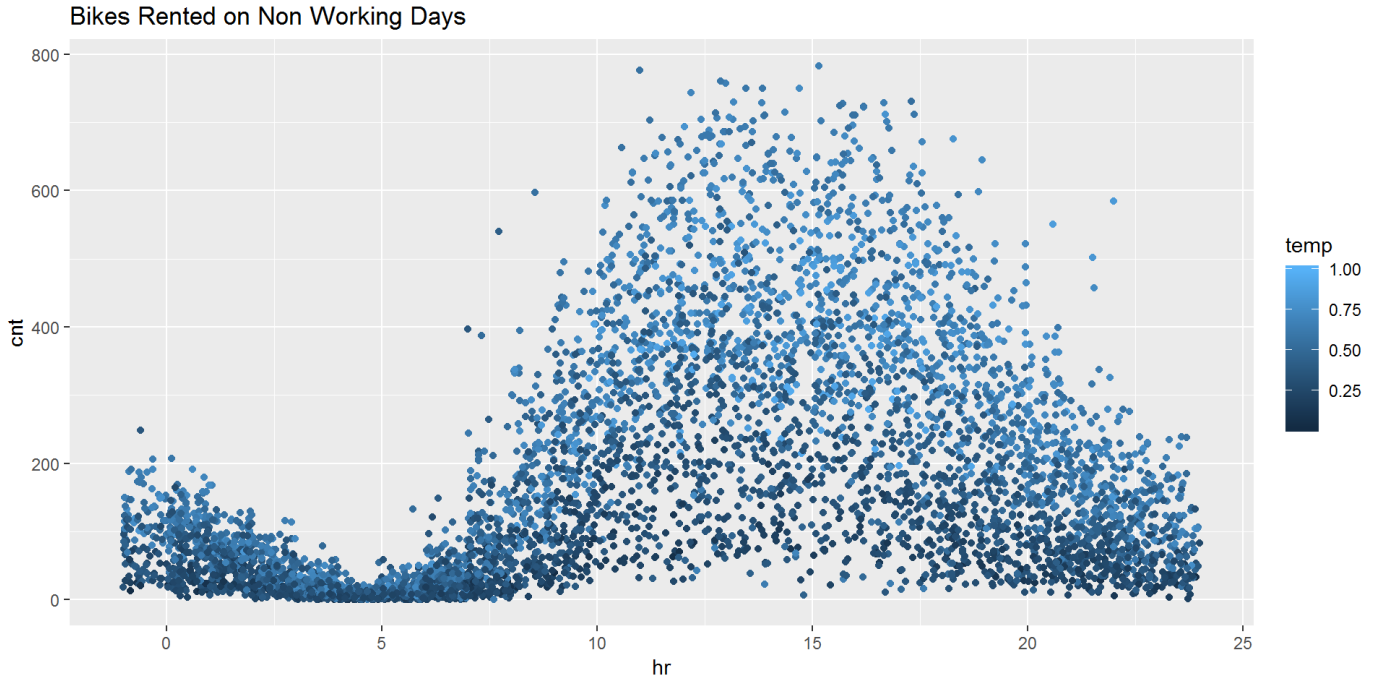
Let us now see how rental numbers are impacted if we analyze working days or non working days solely.

```
ggplot(bike_data[bike_data$workingday == 1, ], aes(x = hr, y = cnt)) +
  geom_point(aes(color = temp), alpha = 0.4, position = position_jitter(w = 1, h = 0)) +
  ggtitle(label = "Bikes Rented on Working Days")
```



It is observed that bike rentals on working days tend to peak during 8A.M. and 5 P.M.. This maybe as people hire bikes for exercise purposes or transit to office.

```
ggplot(bike_data[bike_data$workingday == 0, ], aes(x = hr, y = cnt)) +
  geom_point(aes(color = temp), position = position_jitter(w = 1, h = 0)) +
  ggtitle(label = "Bikes Rented on Non Working Days")
```



It is observed that bike rental activity escalates during noon.

Based on our rudimentary analysis of data we can say none of the variables show high linear relationship with bikes rented. Let us however build a model taking “cnt” and “temp” to see how we can later analyse the residuals to conclude that the model is not accurate.

5 Build Linear Model for Bikes Rented

5.1 Split to trainging and test set

We split the data into training and test set in order to build our test how accurate our model predictions are with respect to actual values in the test set. The model is trained on training data.

```
training_set <- subset(bike_data, as.numeric(format(as.Date(bike_data$dteday), "%d")) <= 20)
test_set <- subset(bike_data, as.numeric(format(as.Date(bike_data$dteday), "%d")) > 20)
```

5.2 Linear Model

```
model <- lm(cnt ~ temp, training_set)
summary(model)
```

```
##
## Call:
## lm(formula = cnt ~ temp, data = training_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -293.90 -112.86  -33.51   78.88  741.15
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.741      4.360   1.317    0.188
## temp         377.130      8.246  45.737   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 166.9 on 11458 degrees of freedom
## Multiple R-squared:  0.1544, Adjusted R-squared:  0.1543
## F-statistic:  2092 on 1 and 11458 DF,  p-value: < 2.2e-16
```

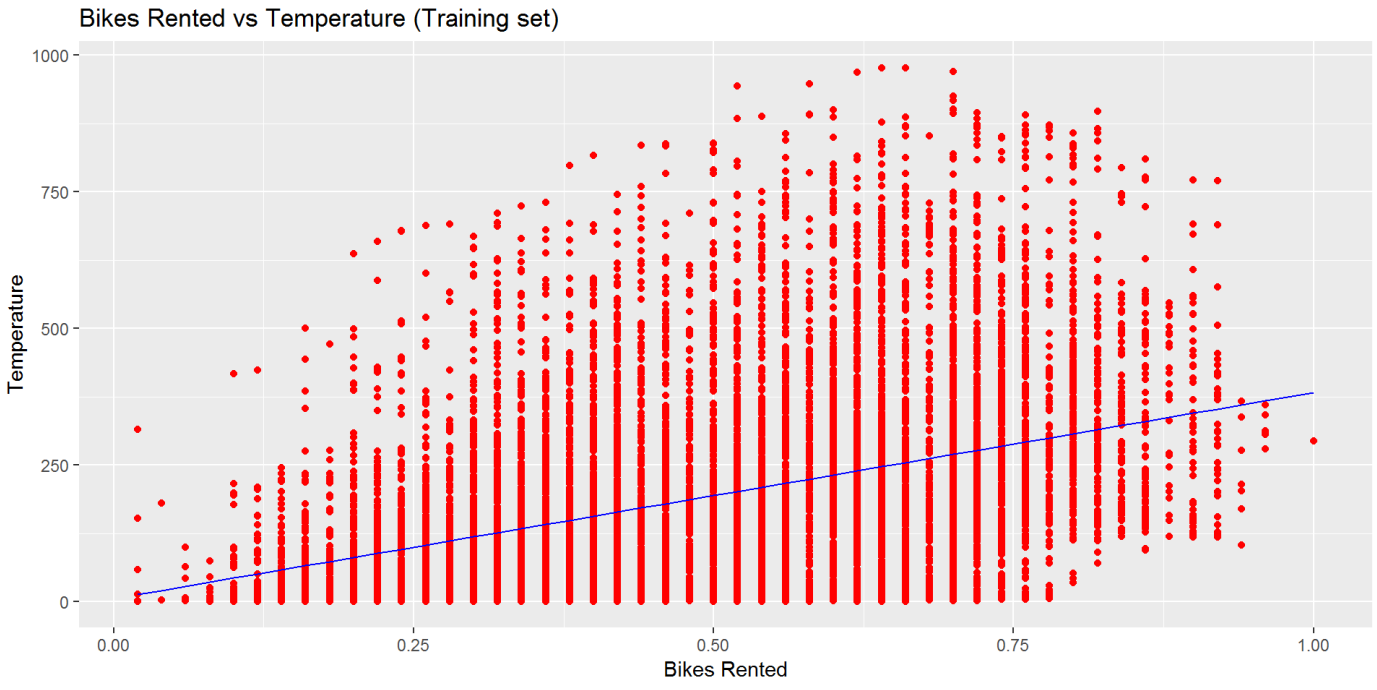
From the summary we can see that slope is 377.1297201 and intercept is 5.7410352.

```
y_pred <- predict(model, newdata = test_set)
```

5.3 Graphical representation of Actual Data and Predicted Data

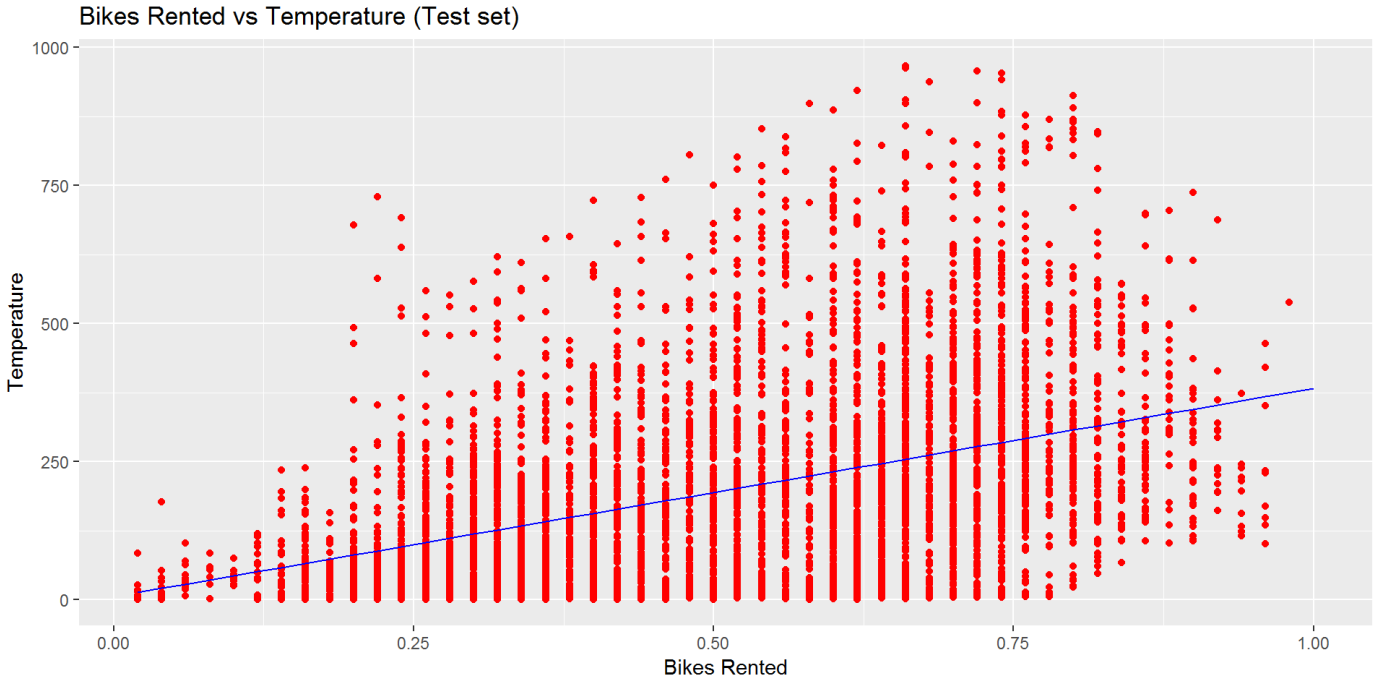
Below is the graph between actual values and predicted values for training data

```
ggplot() +
  geom_point(aes(x = training_set$temp, y = training_set$cnt),
             colour = 'red') +
  geom_line(aes(x = training_set$temp, y = predict(model, newdata = training_set)),
            colour = 'blue') +
  ggtitle('Bikes Rented vs Temperature (Training set)') +
  xlab('Bikes Rented') +
  ylab('Temperature')
```



Below is the graph between actual values and predicted values for test data

```
ggplot() +
  geom_point(aes(x = test_set$temp, y = test_set$cnt),
             colour = 'red') +
  geom_line(aes(x = training_set$temp, y = predict(model, newdata = training_set)),
            colour = 'blue') +
  ggtitle('Bikes Rented vs Temperature (Test set)') +
  xlab('Bikes Rented') +
  ylab('Temperature')
```



From the above two graphs we can see that the model does not accurately predict values for given temperature.This means that the most of the variation in “Bikes Rented” cannot be explained with variation in “Temperature”. This is also evident from the \(R^2 = 0.1544\).

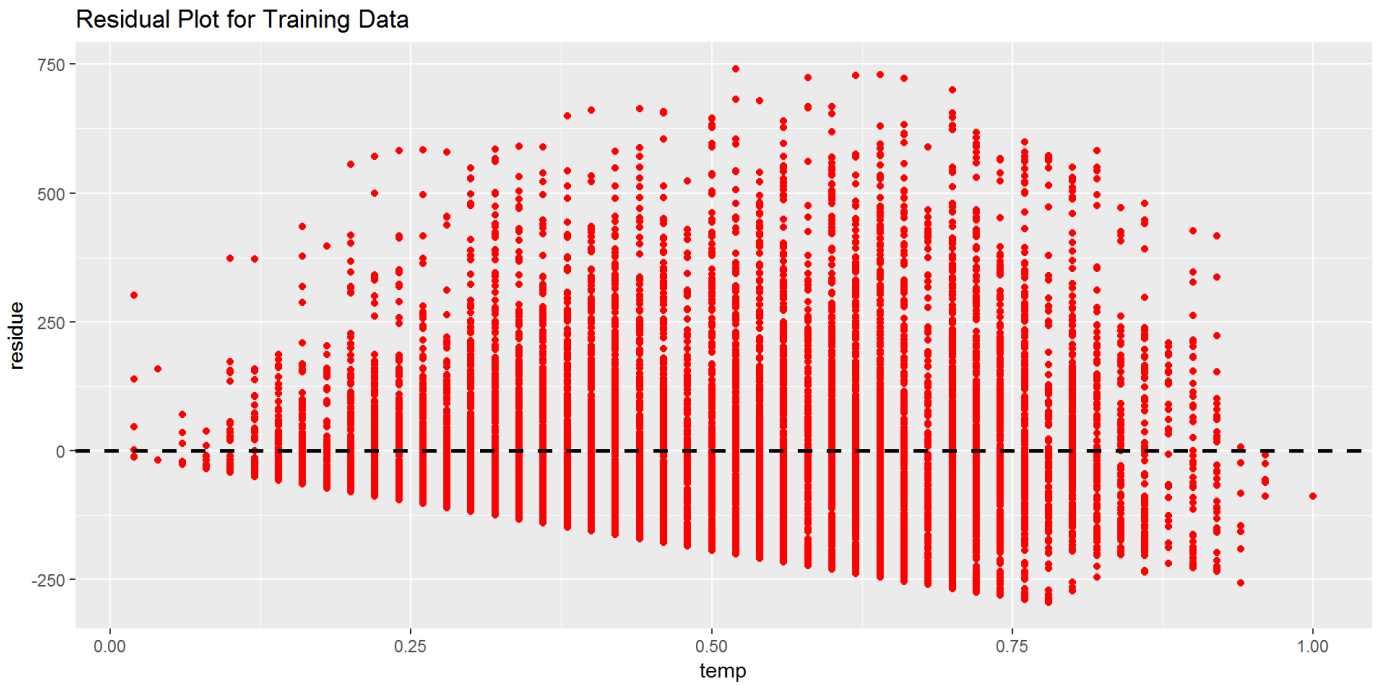
A low value of coefficient of determination indicates that bad fit and the model is not accurate.

5.4 Residual Plots

Another way to gauge the goodness of fit for the model is to look at the residue plots. Let us look at some residue plots to justify that in this case selecting a linear model is a bad choice.

Residual plot should show constant variance/spread. Y axis = residue; X axis = temp there should be random spread in the graph that indicated constant variance.

```
ggplot() +
  geom_point(data = training_set, aes(x = training_set$temp, y = resid(model)), colour = "red") +
  geom_hline(aes(yintercept = mean(resid(model))), linetype = "dashed", size = 1) +
  ggtitle(label = "Residual Plot for Training Data") +
  xlab("temp") + ylab("residue")
```

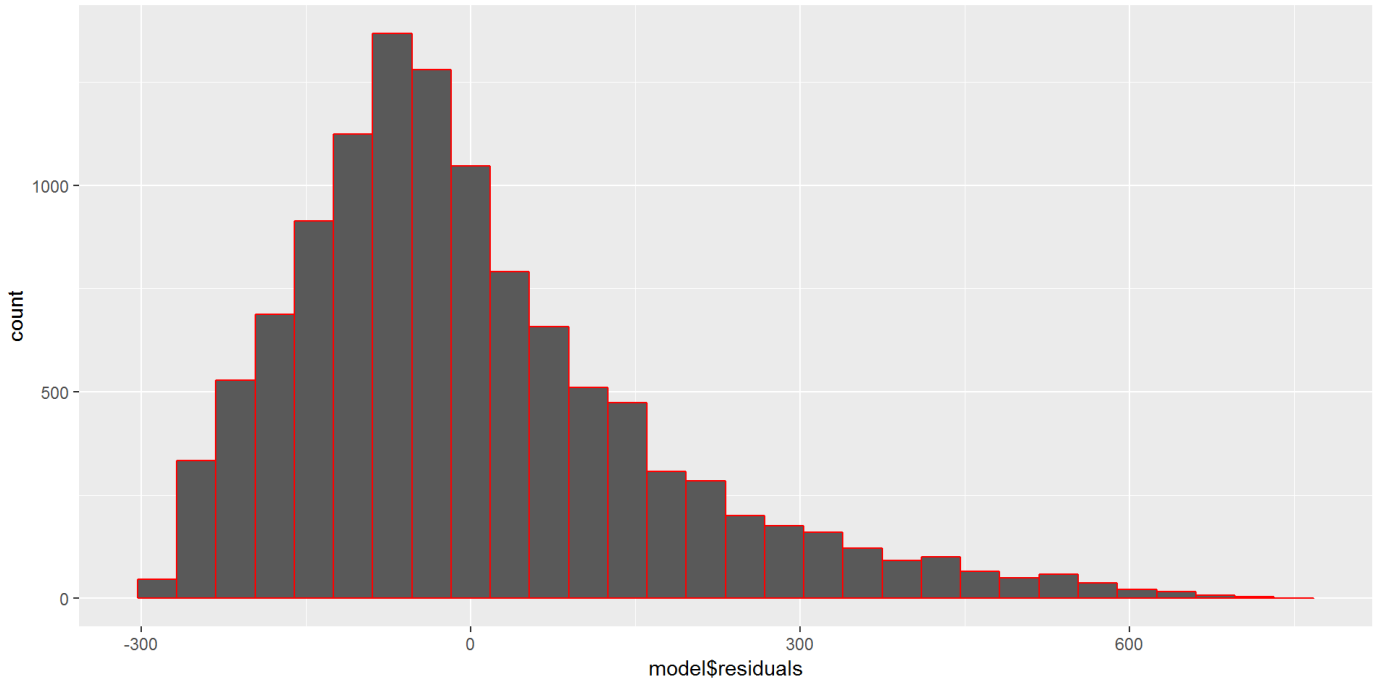


It is observed that the residuals increase as temperture increases. This means that variance in predicted values and actual values increases. This indicates that the model does not have constant variance. Also this phenomenon is known as heteroskedasticity.

Below graph of residuals shows skewness in normal distribution

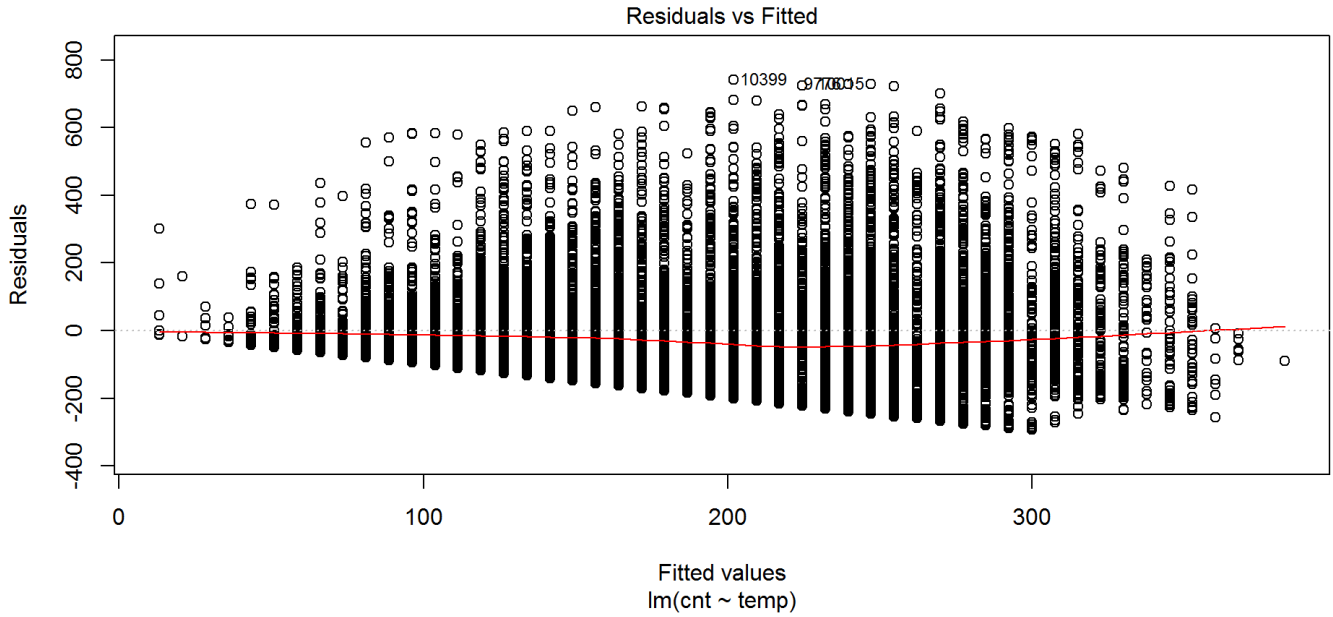
```
ggplot() +
  geom_histogram(aes(x = model$residuals), color = "red")
```

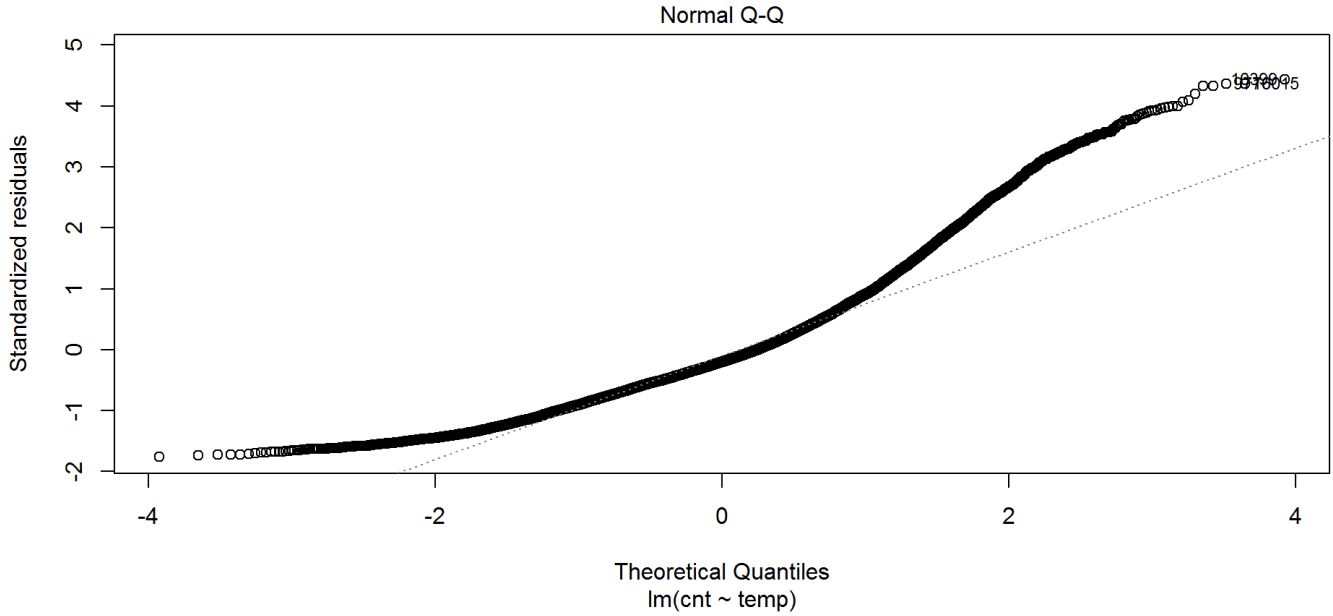
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

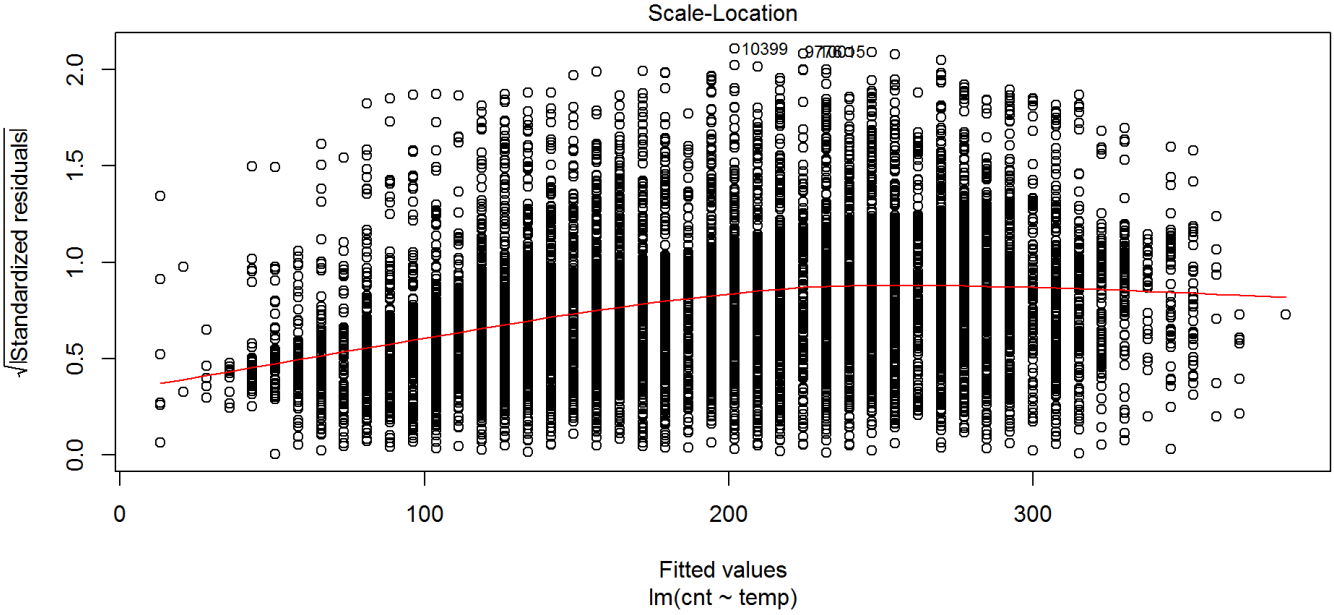


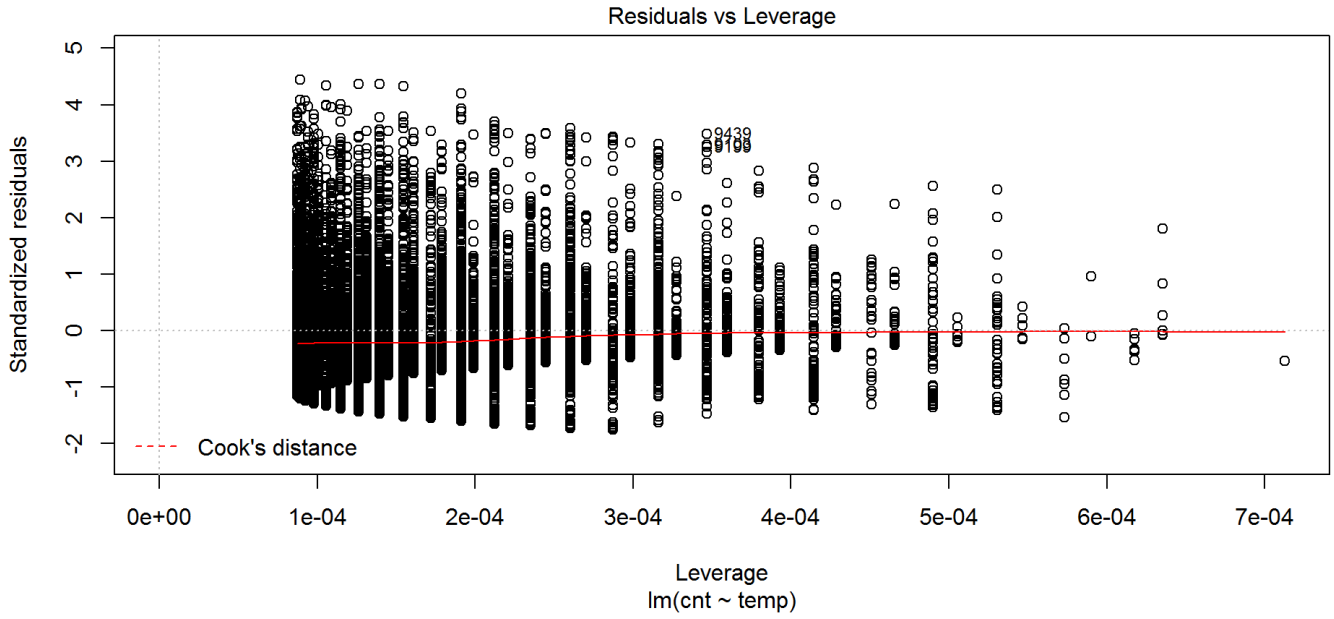
Let us look at some generic graphs for residual plots to check distribution and errors.

```
plot(model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(cnt ~ temp)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(cnt ~ temp)


Scale-Location
√|Standardized residuals|
Fitted values
lm(cnt ~ temp)



6 Conclusion

- Temperature has weak linear relationship with bikes rented.
- Coefficient of Determination is very low, thereby supporting that model does not explain variation in temp with variation in y.
- Residual plots also do not show constant variance, there is increase in error
- From the above Q-Q plot we can see that the residual distribution is also skewed indicating violation of assumption.