Stats 101A_Final_Project

Jun Yu Chen

3/6/2023

Introduction

Background and Research Question

Bike-sharing systems have become popular due to their increasing role in addressing traffic, environmental, and health issues. Riding bikes is emission-free and boosts personal fitness. Therefore, analyzing the monitored data of bike rental and related information can provide valuable insights into this commercial sector and future trends. In this project, we are interested in exploring the various factors that affect bike rentals. Specifically, we aim to predict the day-to-day count of total rental bikes (cnt) based on relevant weather settings and time features, such as the type of weather and whether the given date is a holiday.

Source of data

We analyze the "Rental Bike Sharing Dataset" posted by Akash Patel on Kaggle. The core dataset is sourced from the two-year historical log of the Capital Bikeshare system corresponding to the years 2011 and 2012, which is publicly available on http://capitalbikeshare.com/system-data. Weather information is extracted from http://www.freemeteo.com and aggregated according to the date of the core dataset. Our research question is whether we can predict the count of bike rentals based on weather and time features, such as season, year, month, hour, holiday, weekday, working day, weather situation, temperature, feeling temperature, humidity, and wind speed.

Project Framework

The overall framework of this paper is as follows. First, we conduct data cleaning and create initial exploratory scatter and distribution plots to visualize the trends and patterns. Then, we produce a full model and utilize diagnostic plots to identify problems with constant variance and normality of error terms.We also apply different data transformations and select the best model that addresses the issues. After that, we conduct variable selection to choose the model with the optimal performance. We used multiple-linear regression models supported by lm() in R, because we used multiple features to predict the outcome of a numerical variable. We also experimented with neural network powered by the keras library as an innovative attempt. Lastly, we provide suggestions for possible limitations and future directions of the topic.

Data Description

```r
bike <- read.csv("day.csv", header=TRUE)
##Checks the dimension of the data frame and types of data
head(str(bike))
```

```
## 'data.frame':    731 obs. of  16 variables:
##  $ instant   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ dteday    : chr  "2011-01-01" "2011-01-02" "2011-01-03" "2011-01-04" ...
##  $ season    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ yr        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mnth      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ weekday   : int  6 0 1 2 3 4 5 6 0 1 ...
##  $ workingday: int  0 0 1 1 1 1 1 0 0 1 ...
##  $ weathersit: int  2 2 1 1 1 1 2 2 1 1 ...
##  $ temp      : num  0.344 0.363 0.196 0.2 0.227 ...
##  $ atemp     : num  0.364 0.354 0.189 0.212 0.229 ...
##  $ hum       : num  0.806 0.696 0.437 0.59 0.437 ...
##  $ windspeed : num  0.16 0.249 0.248 0.16 0.187 ...
##  $ casual    : int  331 131 120 108 82 88 148 68 54 41 ...
##  $ registered: int  654 670 1229 1454 1518 1518 1362 891 768 1280 ...
##  $ cnt       : int  985 801 1349 1562 1600 1606 1510 959 822 1321 ...

## NULL
```

```
##Checks the mean, mediann, and quartiles of the variables
summary(bike)
```

```
##     instant         dteday              season            yr        
##  Min.   :  1.0   Length:731         Min.   :1.000   Min.   :0.0000  
##  1st Qu.:183.5   Class :character   1st Qu.:2.000   1st Qu.:0.0000  
##  Median :366.0   Mode  :character   Median :3.000   Median :1.0000  
##  Mean   :366.0                      Mean   :2.497   Mean   :0.5007  
##  3rd Qu.:548.5                      3rd Qu.:3.000   3rd Qu.:1.0000  
##  Max.   :731.0                      Max.   :4.000   Max.   :1.0000  
##       mnth          holiday           weekday        workingday   
##  Min.   : 1.00   Min.   :0.00000   Min.   :0.000   Min.   :0.000  
##  1st Qu.: 4.00   1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.000  
##  Median : 7.00   Median :0.00000   Median :3.000   Median :1.000  
##  Mean   : 6.52   Mean   :0.02873   Mean   :2.997   Mean   :0.684  
##  3rd Qu.:10.00   3rd Qu.:0.00000   3rd Qu.:5.000   3rd Qu.:1.000  
##  Max.   :12.00   Max.   :1.00000   Max.   :6.000   Max.   :1.000  
##    weathersit         temp             atemp              hum        
##  Min.   :1.000   Min.   :0.05913   Min.   :0.07907   Min.   :0.0000  
##  1st Qu.:1.000   1st Qu.:0.33708   1st Qu.:0.33784   1st Qu.:0.5200  
##  Median :1.000   Median :0.49833   Median :0.48673   Median :0.6267  
##  Mean   :1.395   Mean   :0.49538   Mean   :0.47435   Mean   :0.6279  
##  3rd Qu.:2.000   3rd Qu.:0.65542   3rd Qu.:0.60860   3rd Qu.:0.7302  
##  Max.   :3.000   Max.   :0.86167   Max.   :0.84090   Max.   :0.9725  
##    windspeed           casual          registered        cnt      
##  Min.   :0.02239   Min.   :   2.0   Min.   :  20   Min.   :  22  
##  1st Qu.:0.13495   1st Qu.: 315.5   1st Qu.:2497   1st Qu.:3152  
##  Median :0.18097   Median : 713.0   Median :3662   Median :4548  
##  Mean   :0.19049   Mean   : 848.2   Mean   :3656   Mean   :4504  
##  3rd Qu.:0.23321   3rd Qu.:1096.0   3rd Qu.:4776   3rd Qu.:5956  
##  Max.   :0.50746   Max.   :3410.0   Max.   :6946   Max.   :8714  
```

```r
##chosen variables for the multiple linear regression model
head(bike[c(3,4,8,9,10,11,12,13,16)])
```

```
##   season yr workingday weathersit     temp    atemp      hum windspeed  cnt
## 1      1  0          0          2 0.344167 0.363625 0.805833 0.1604460  985
## 2      1  0          0          2 0.363478 0.353739 0.696087 0.2485390  801
## 3      1  0          1          1 0.196364 0.189405 0.437273 0.2483090 1349
## 4      1  0          1          1 0.200000 0.212122 0.590435 0.1602960 1562
## 5      1  0          1          1 0.226957 0.229270 0.436957 0.1869000 1600
## 6      1  0          1          1 0.204348 0.233209 0.518261 0.0895652 1606
```

Variable Description

The predictor variables are season, yr, workingday, temp, atemp, hum, and windspeed, while the response variable is cnt.

- season : season (1:springer, 2:summer, 3:fall, 4:winter)
- yr : year (0: 2011, 1:2012)
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit :
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
        - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
        - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
        - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp : Normalized temperature in Celsius. The values are divided to 41 (max)
- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- cnt: count of total rental bikes including both casual and registered

Ignored variables:

- For the scope of this project, I have decided to drop the categorical variables mnth, holiday, and weekday to reduce the number of categorical variables. Additionally, these categorical variables have many levels, which can make modeling more complex and lead to overfitting. Moreover, other variables in the dataset have clearer categories that account for the effect of these variables. For example, the variable "season" describes the seasonal settings with more concise categories than "mnth". Similarly, the effect of "holiday" and "weekday" on bike rental can be captured by the binary variable "workingday," which indicates 1 for non-working days (i.e., weekends and holidays) and 0 otherwise.
- I also dropped variables such as "dteday" and "instant" that are indexes and irrelevant to the task.

Data Cleaning

```r
# Identify rows with NA values
na_rows <- apply(bike, 1, function(x) any(is.na(x)))

# Print the rows with NA values
nrow(bike[na_rows, ])
```

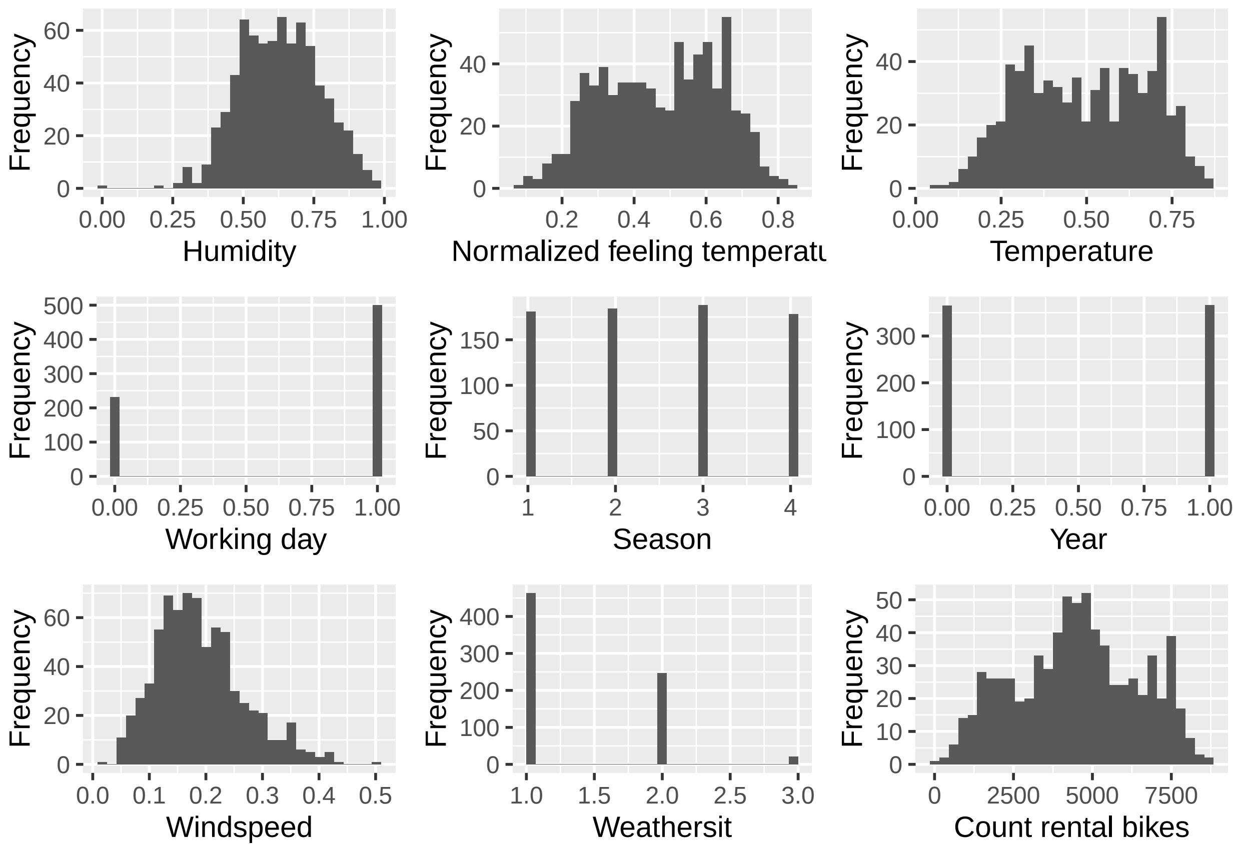
```
## [1] 0
```

```
# clean the dataset
bike <- na.omit(bike)
```

Histograms and Scatterplots

```
suppressMessages(library(dplyr))
suppressMessages(library(ggplot2))
suppressMessages(library(cowplot))
variables <- c("hum", "atemp", "temp", "workingday", "season", "yr", "windspeed", "weathersit", "cnt")
labels <- c("Humidity", "Normalized feeling temperature", "Temperature", "Working day", "Season", "Year

plots <- list()

for (i in 1:length(variables)) {
  p <- ggplot(bike, aes_string(x = variables[i])) +
    geom_histogram() +
    labs(x = labels[i], y = "Frequency")
  plots[[i]] <- p
}

plot_grid(plotlist = plots, nrow = 3)
```

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



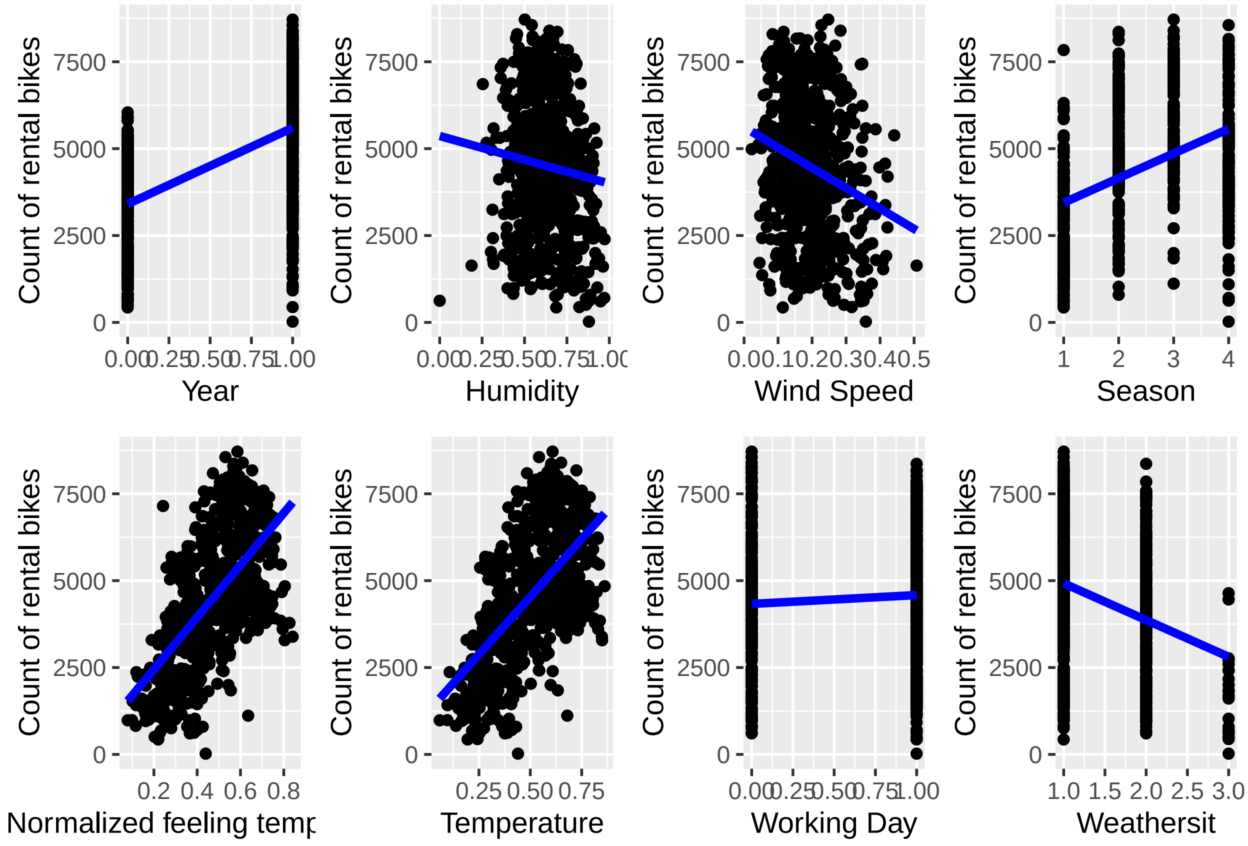
- During the initial exploration of the data, most graphs appeared to have a relatively normal distribution. The graph for windspeed, however, appeared to be slightly right-skewed, which is consistent with weather observations since wind speed is typically lower but can occasionally reach higher values during storms, which could be considered outliers.
- The distribution graph for temperature and normalized feeling temperature displayed a bimodal pattern, indicating that there may be issues with variance that will require further investigation.
- Regarding the frequency graphs for categorical variables, they all appeared to have a proportional distribution between categories. However, the frequency graph for working day showed a much higher frequency than the combination of holiday and weekend, which aligns with common sense.

```r
variables <- c("yr", "hum", "windspeed", "season", "atemp", "temp", "workingday", "weathersit")
labels <- c("Year", "Humidity", "Wind Speed", "Season", "Normalized feeling temperature", "Temperature"

plots <- list()

for (i in 1:length(variables)) {
  p <- ggplot(bike, aes_string(x = variables[i], y = "cnt")) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, color = "blue", size = 1.5) +
    labs(x = labels[i], y = "Count of rental bikes")
  plots[[i]] <- p
}
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
```

```
plot_grid(plotlist = plots, nrow = 2)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
cor(bike[c(4, 8, 10, 11, 12, 13, 16)])
```

```
##                      yr   workingday        temp       atemp         hum
## yr          1.000000000 -0.002012621  0.04760357  0.04610615 -0.11065104
## workingday -0.002012621  1.000000000  0.05265981  0.05218228  0.02432705
## temp        0.047603572  0.052659810  1.00000000  0.99170155  0.12696294
## atemp       0.046106149  0.052182275  0.99170155  1.00000000  0.13998806
## hum        -0.110651045  0.024327046  0.12696294  0.13998806  1.00000000
## windspeed  -0.011817060 -0.018796487 -0.15794412 -0.18364297 -0.24848910
## cnt         0.566709708  0.061156063  0.62749401  0.63106570 -0.10065856
##              windspeed         cnt
## yr          -0.01181706  0.56670971
## workingday  -0.01879649  0.06115606
## temp       -0.15794412  0.62749401
## atemp      -0.18364297  0.63106570
## hum        -0.24848910 -0.10065856
## windspeed   1.00000000 -0.23454500
## cnt        -0.23454500  1.00000000
```

```
bike %>%
  group_by(yr) %>%
  summarise(avg_count = mean(cnt))
```

```
## # A tibble: 2 x 2
##      yr avg_count
##   <int>     <dbl>
## 1     0     3406.
## 2     1     5600.
```

```
bike %>%
  group_by(workingday) %>%
  summarise(avg_count = mean(cnt))
```

```
## # A tibble: 2 x 2
##   workingday avg_count
##        <int>     <dbl>
## 1          0     4330.
## 2          1     4585.
```

- Based on the scatter plot and fitted slopes, most selected features, such as humidity, wind speed, season, and normalized feeling temperature, appear to have a linear correlation with the count of rental bikes. The correlation coefficient indicates that yr, temp, and atemp have a moderate to strong association, while windspeed, hum, and working day have a very weak to weak association. "atemp" and "temp" seem to be highly correlated as their correlation coefficient is 0.99, which will be addressed later.

- Specifically, the slope between working day and count of rental bikes appears to be flat, indicating a very weak relationship. This is further supported by the summary statistics, which show that the average bike rental between working day (4330.169) and non-working day (4584.820) is very similar. This finding is counterintuitive to the speculation that customers tend to rent bikes more frequently during weekends or holidays. However, we could now speculate that the bike sharing system incentivizes more people to use bikes as a daily transportation tool to their company, which reduces the difference in bike rental between working and non-working days.

- Another interesting pattern is that the average count of bike rentals in 2012 (5599.934) appears to be higher than in 2011 (3405.762). This may indicate a rise in the popularity of public sharing bikes in the city or other external factors.

Factor Multi-categorical Variables

```
bike$season_f <- factor(bike$season)
bike$weathersit_f <- factor(bike$weathersit)
```

Results and Interpretation

Full Modell

```
model1<-lm(cnt~windspeed+hum+atemp+temp+weathersit_f+workingday+season_f+yr,data=bike)
summary(model1)
```

```
##
## Call:
## lm(formula = cnt ~ windspeed + hum + atemp + temp + weathersit_f +
##     workingday + season_f + yr, data = bike)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3664.0  -386.7    81.9   477.6  3402.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1492.38     234.41   6.367 3.44e-10 ***
## windspeed     -2741.73     438.16  -6.257 6.72e-10 ***
## hum           -1343.16     297.80  -4.510 7.56e-06 ***
## atemp          1158.82    1533.78   0.756  0.45018
## temp           4050.72    1404.38   2.884  0.00404 **
## weathersit_f2  -421.34      81.89  -5.145 3.45e-07 ***
## weathersit_f3 -1891.12     209.14  -9.042  < 2e-16 ***
## workingday      174.80      66.22   2.640  0.00847 **
## season_f2      1150.07     114.33  10.060  < 2e-16 ***
## season_f3       872.80     151.57   5.758 1.26e-08 ***
## season_f4      1545.51      97.54  15.845  < 2e-16 ***
## yr             2013.40      62.20  32.370  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 828.1 on 719 degrees of freedom
## Multiple R-squared:   0.82,  Adjusted R-squared:  0.8173
## F-statistic: 297.8 on 11 and 719 DF,  p-value: < 2.2e-16
```

```
anova(model1)
```

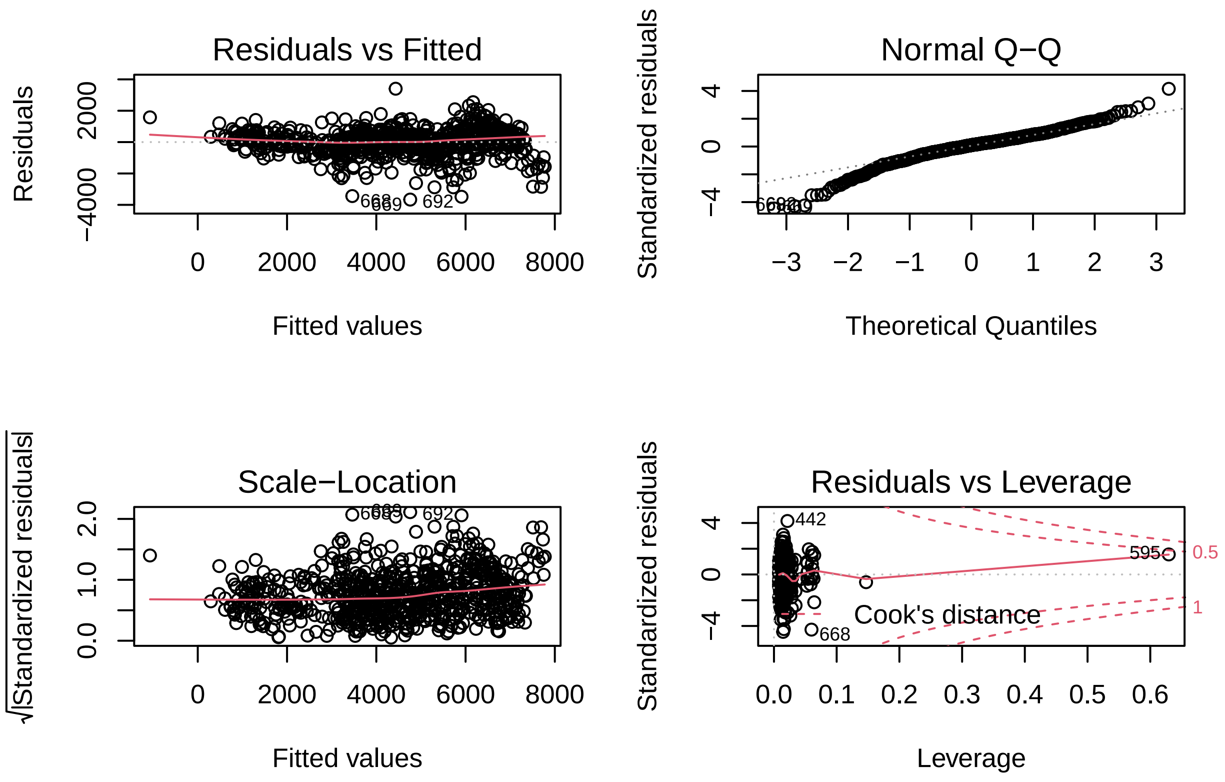
```
## Analysis of Variance Table
##
## Response: cnt
##               Df     Sum Sq    Mean Sq  F value    Pr(>F)
## windspeed      1  150705556  150705556  219.7789 < 2.2e-16 ***
## hum            1   73760811   73760811  107.5678 < 2.2e-16 ***
## atemp          1 1044485733 1044485733 1523.2081 < 2.2e-16 ***
## temp           1    1732390    1732390    2.5264   0.11239
## weathersit_f   2   50482221   25241110   36.8099 6.028e-16 ***
## workingday     1    3941102    3941102    5.7474   0.01677 *
## season_f       3  202898092   67632697   98.6310 < 2.2e-16 ***
## yr             1  718500847  718500847 1047.8136 < 2.2e-16 ***
## Residuals    719  493028640     685714
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Interpretation:

- The model summary and ANOVA analysis of the linear model are consistent, indicating that the model is a good fit for the data. All predictor variables except for "atemp" have p-values smaller than 0.05, suggesting that they are statistically significant.
- The F-statistic of 297.8 is large, indicating that the variance in the count of total bike rentals explained by the regression model is much larger than the residual variance, which suggests that the model has good predictive power. Additionally, the overall p-value is smaller than 2.2e-16, indicating that at least one of the 11 independent variables likely has a linear association with the count of total bike rentals.
- The adjusted R-squared value of 0.8172 is high, indicating that the predictor variables explain a large proportion of the variation in the count of total bike rentals. Specifically, 81.72% of the variation in the response variable can be explained by the predictor variables in the model.

Diagnostic Plots

```r
n<-dim(bike)[1]
p<-11
par(mfrow=c(2,2))
plot(model1)
```



```
2*(p+1)/n
```

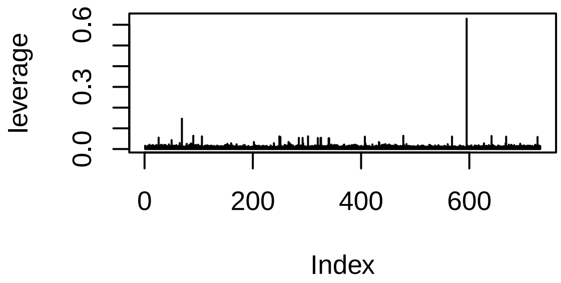
```
## [1] 0.03283174
```

```
# Calculate leverage for each observation
leverage <- hatvalues(model1)

# Find leverage points
high_leverage <- which(leverage > 2 * mean(leverage))
print(length(high_leverage))
```

```
## [1] 26
```

```
#plot leverage values for each observation
plot(leverage, type = 'h')

# Find the observations with the highest leverage
top10_leverage <- order(leverage, decreasing = TRUE)[1:10]
print(top10_leverage)
```

```
##  [1] 595  69  90 478 641 302 106 249 568 407
```

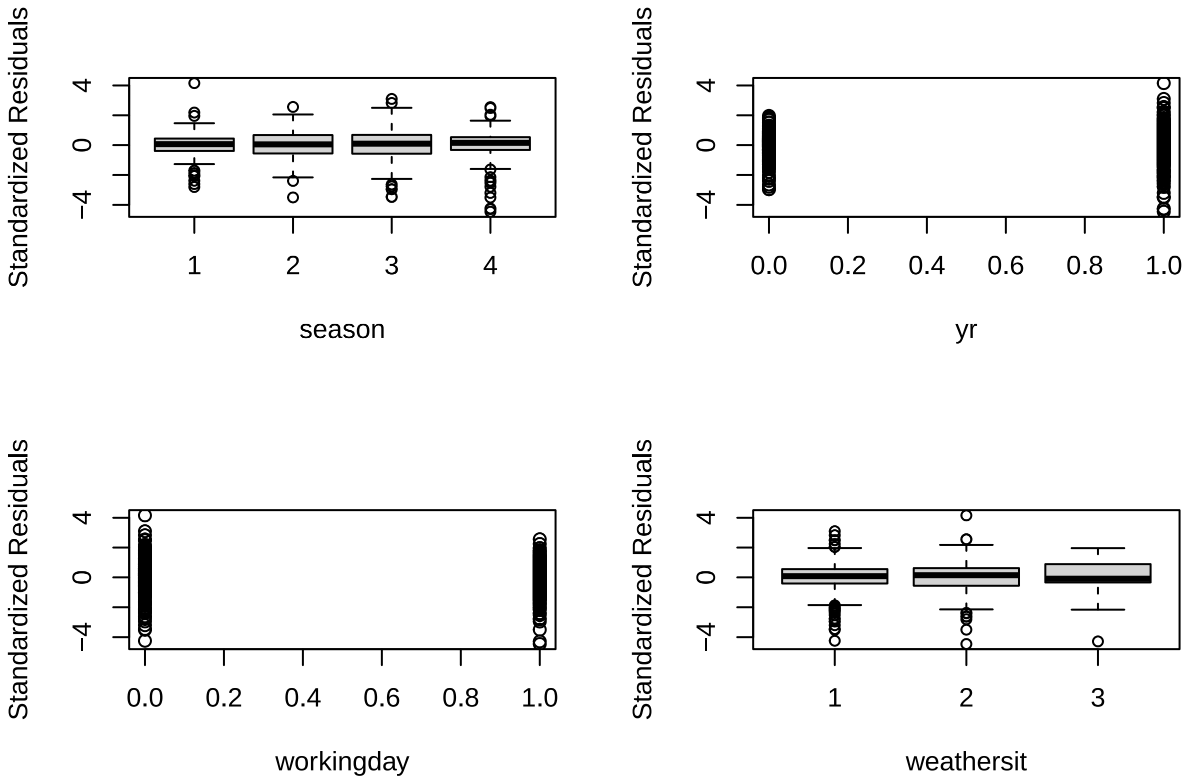


Plot interpretation:

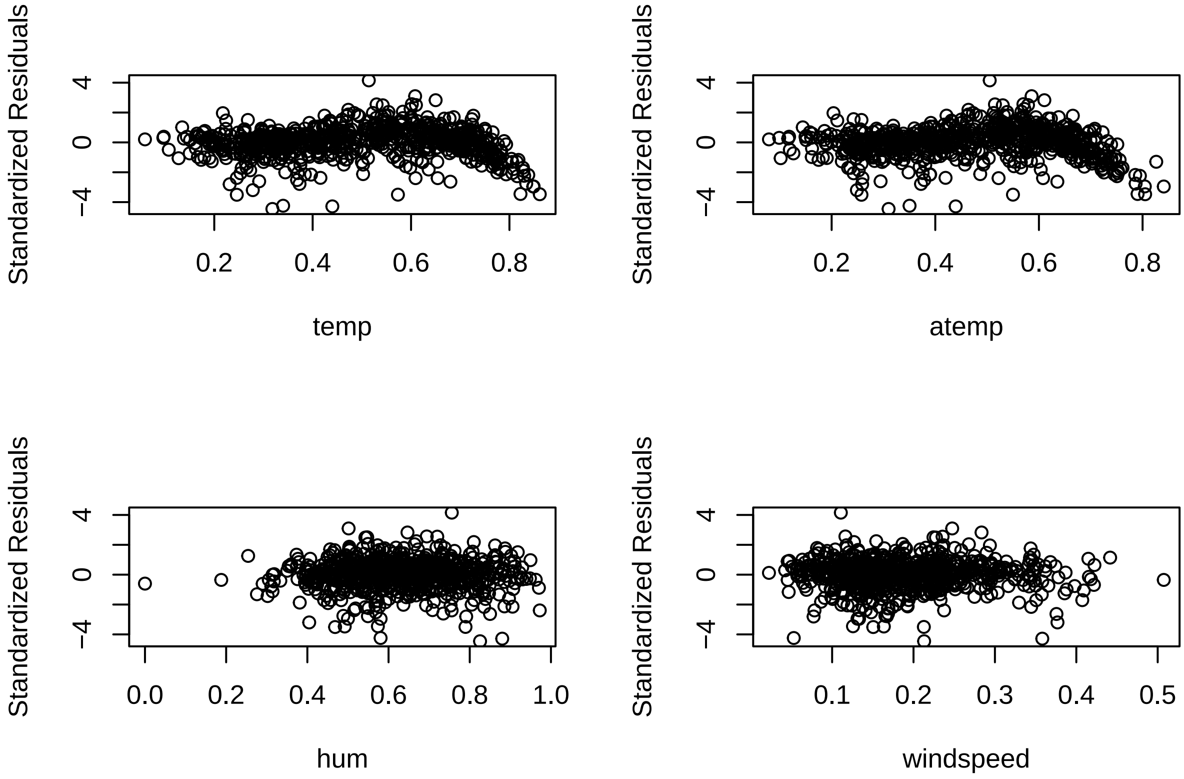
- The residual vs fitted values plot suggests a linear relationship between the predictor variables and the response variable, as the line is fairly straight with a slight curve. The errors are randomly scattered along the horizontal axis and have an average of 0, indicating a linear relationship and constant variance.
- The QQ plot shows that most of the points are aligned to the straight line. However, the error term seems to be somewhat left-skewed and heavy-tailed, with some data points more than 2 standard deviations away from the fitted line toward the negative side. This may indicate a violation of the normality assumption for the errors.
- The standardized residual vs fitted values plot shows no clear patterns, suggesting constant variance.
- The residual vs leverage plot reveals many leverage points with leverage values higher than 0.03283174. We identified 26 such points using the hatvalues() function. Additionally, we plotted the leverage values against the index of observations and found that data point 595 has a very high leverage value of 0.629510242, which is clearly visible on the graph and needs to be addressed

Individual Residual Plots

```
SR <- rstandard(model1)

par(mfrow=c(2,2))
plot1<-plot(factor(bike$season), SR, xlab = "season", ylab = "Standardized Residuals")
plot2<-plot(bike$yr, SR, xlab = "yr", ylab = "Standardized Residuals")
plot3<-plot(bike$workingday, SR, xlab = "workingday", ylab = "Standardized Residuals")
plot4<-plot(factor(bike$weathersit), SR, xlab = "weathersit", ylab = "Standardized Residuals")
```



```
par(mfrow=c(2,2))
plot5<-plot(bike$temp, SR, xlab = "temp", ylab = "Standardized Residuals")
plot6<-plot(bike$atemp, SR, xlab = "atemp", ylab = "Standardized Residuals")
plot7<-plot(bike$hum, SR, xlab = "hum", ylab = "Standardized Residuals")
plot8<-plot(bike$windspeed, SR, xlab = "windspeed", ylab = "Standardized Residuals")
```



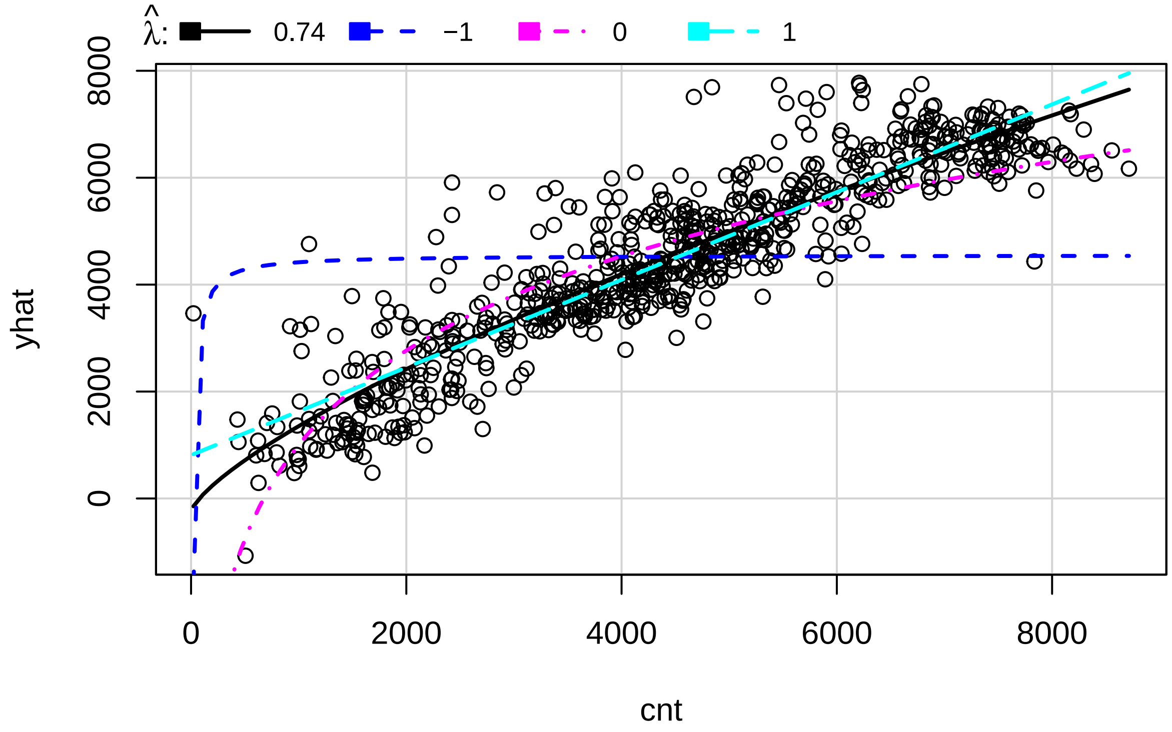
- for the categorical variables, the mean is all around 0 with values very scattered vertically, which indicates constant variance - for the continuous variables, while the shape and random scaterness seem appropriate for "hum" and "windspeed", temp" and "atemp" seem to present a slightly parabolic pattern, which may indicate some non-constant variance that needs to be addressed.

Data Transformation

```
suppressMessages(library(car))
```

```
## Warning: package 'car' was built under R version 4.1.2

## Warning: package 'carData' was built under R version 4.1.2
```

```
suppressMessages(library(MASS))

##show inverseResponsePlot
inverseResponsePlot(model1,key=TRUE)
```



```
##       lambda        RSS
## 1  0.7388611  390771996
## 2 -1.0000000 2208478385
## 3  0.0000000  635524781
## 4  1.0000000  404299274
```

```
attach(bike)

##find the optimal lambda value for each variable
summary(powerTransform(cbind(cnt,windspeed,atemp,temp)~1))
```

```
## bcPower Transformations to Multinormality
##             Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## cnt            0.7656        0.77       0.6544       0.8768
## windspeed      0.4517        0.50       0.3161       0.5874
## atemp          0.8533        0.85       0.7405       0.9661
## temp           0.6907        0.69       0.5867       0.7947
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                LRT df       pval
## LR test, lambda = (0 0 0 0) 610.1142  4 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                LRT df       pval
## LR test, lambda = (1 1 1 1) 147.3581  4 < 2.22e-16
```

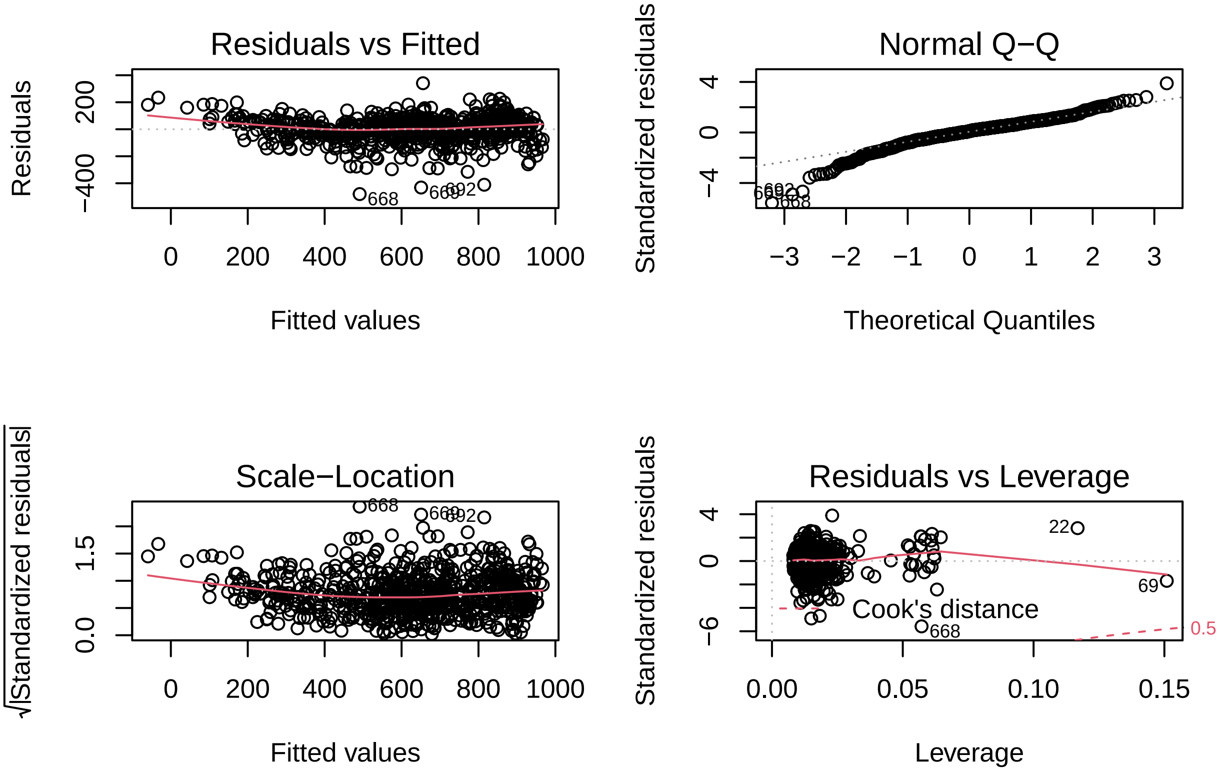
```
##transform the variables
bike$cnt_t<-bike$cnt^0.77
bike$windspeed_t<-bike$windspeed^0.5
bike$temp_t<-bike$temp^0.69
bike$atemp_t<-bike$temp^0.85
```

- To prepare the data for better modeling, we applied the Box-Cox transformation to the response variable and all numerical predictor variables simultaneously. This involved identifying the optimal lambda values for each variable, and then transforming the variables by raising them to the power of their respective lambda values.

```
##construct model 2
model2<-lm(cnt_t~windspeed_t+hum+atemp_t+temp_t+weathersit_f+season_f+workingday+yr, data = bike)
summary(model2)
```

```
##
## Call:
## lm(formula = cnt_t ~ windspeed_t + hum + atemp_t + temp_t + weathersit_f +
##     season_f + workingday + yr, data = bike)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -480.20  -41.82    8.88   51.71  340.78
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -214.406     72.247  -2.968  0.00310 **
## windspeed_t    -308.901     40.511  -7.625 7.73e-14 ***
## hum            -190.093     32.125  -5.917 5.07e-09 ***
## atemp_t       -4622.188    664.541  -6.955 7.91e-12 ***
## temp_t         5654.476    713.881   7.921 8.97e-15 ***
## weathersit_f2   -47.758      8.761  -5.451 6.88e-08 ***
## weathersit_f3  -244.055     22.328 -10.930  < 2e-16 ***
## season_f2       117.560     12.403   9.478  < 2e-16 ***
## season_f3       118.548     16.296   7.275 9.09e-13 ***
## season_f4       155.203     10.922  14.211  < 2e-16 ***
## workingday       20.620      7.089   2.909  0.00374 **
## yr              217.057      6.699  32.403  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88.62 on 719 degrees of freedom
## Multiple R-squared:  0.843,  Adjusted R-squared:  0.8406
## F-statistic: 350.9 on 11 and 719 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model2)
```



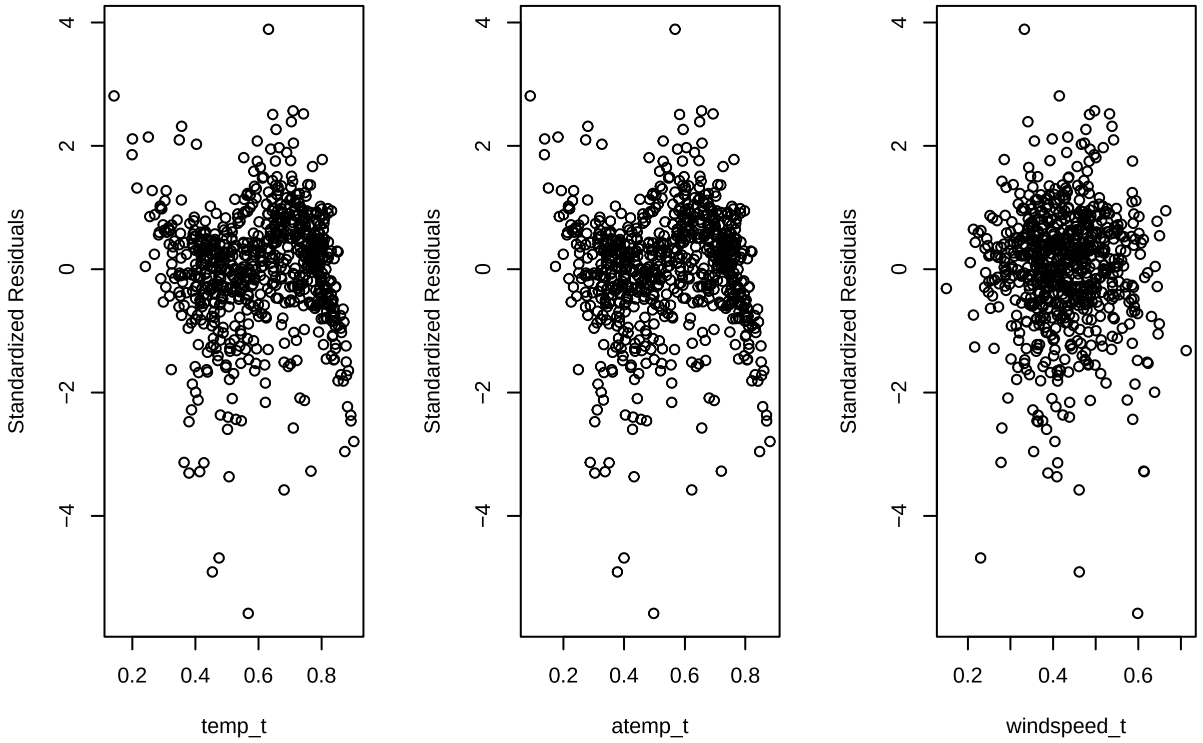
Model Interpretation:

- Model 2 employs box-cox transformation. The model summary indicates that the model has improved from the previous iteration. The adjusted R-squared value has increased from 0.8173 to 0.8406, which suggests that the model now explains more of the variance in the count of total bike rentals. Additionally, the variable “atemp” is now statistically significant with a p-value of 7.91e-12, indicating that it has a stronger association with the response variable with the presence of other variables.
- The F-statistic has also increased from 297.8 to 350.9, which further supports the notion that the regression model explains a larger proportion of the variance in the response variable. The overall p-value is still smaller than 2.2e-16, which suggests that the model is statistically significant.

Plot Diagnostics:

- The QQ-plot shows that the residuals are now less heavy-tailed, with more points within the 2 standard deviation bands. This indicates an improvement in the normality of the error term, which is a key assumption of linear regression. However, according to the residual VS leverage graph, there are still many leverage points, which needs to be addressed later.

```
SR <- rstandard(model2)
par(mfrow=c(1,3))
plot(bike$temp_t, SR, xlab = "temp_t", ylab = "Standardized Residuals")
plot(bike$atemp_t, SR, xlab = "atemp_t", ylab = "Standardized Residuals")
plot(bike$windspeed_t, SR, xlab = "windspeed_t", ylab = "Standardized Residuals")
```

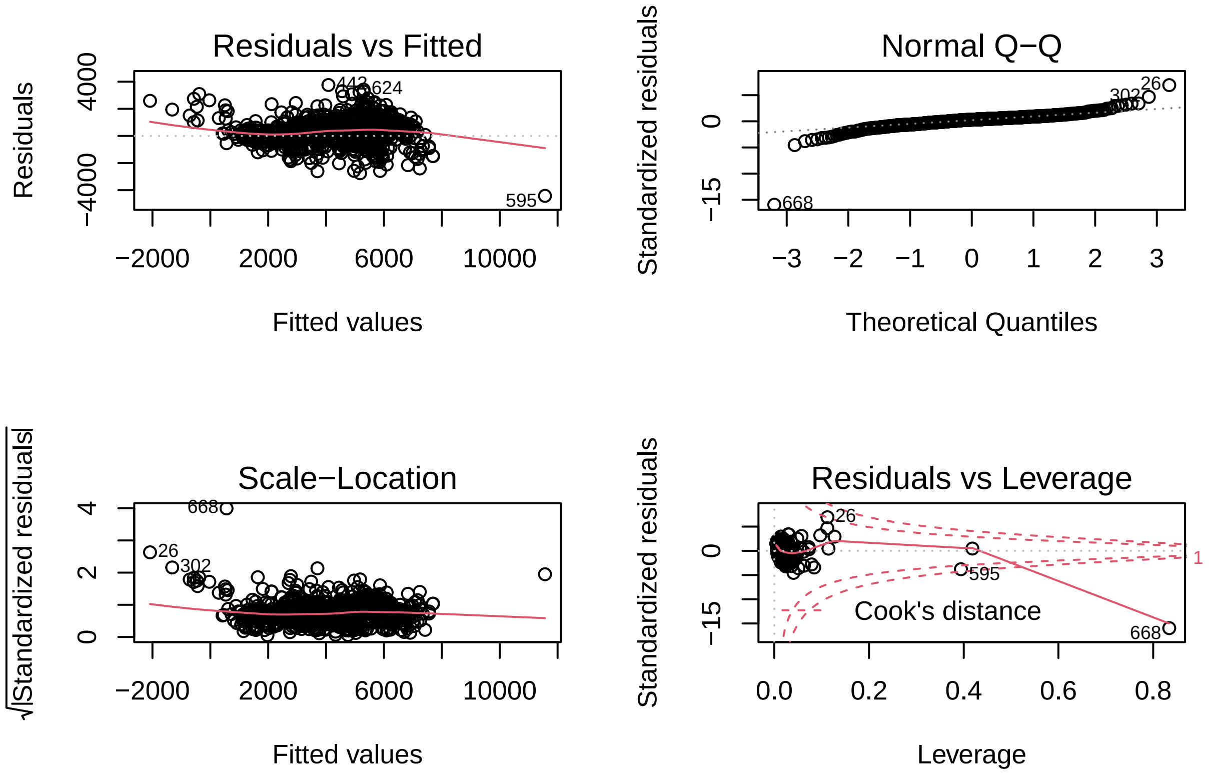


- In the individual residual plots, data points for "windspeed_t" are centered around 0 and scattered randomly. The data points for "atemp_t" and "temp_t" are more scattered ad concentrated towards 0. The curved patterns are also less noticeable, indicating an improvement in addressing non-constant variance.

```
model3<- lm(cnt~windspeed+hum+atemp+temp+weathersit_f+workingday+season_f+yr, data = bike, weight = 1/(
summary(model3)
```

```
##
## Call:
## lm(formula = cnt ~ windspeed + hum + atemp + temp + weathersit_f +
##     workingday + season_f + yr, data = bike, weights = 1/(cnt))
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -114.809   -4.589    5.424   12.253  115.120
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3149.77     234.24  13.447  < 2e-16 ***
## windspeed      -5919.49     486.21 -12.175  < 2e-16 ***
## hum            -2104.57     302.83  -6.950 8.22e-12 ***
## atemp         -12462.20    2175.85  -5.728 1.50e-08 ***
## temp           16397.96    2032.86   8.066 3.03e-15 ***
## weathersit_f2   -309.32      97.03  -3.188  0.00150 **
## weathersit_f3  -2859.84     158.84 -18.004  < 2e-16 ***
## workingday       151.88      79.07   1.921  0.05514 .
## season_f2       1122.55     134.31   8.358 3.29e-16 ***
## season_f3        644.94     198.15   3.255  0.00119 **
## season_f4       1003.02     113.46   8.840  < 2e-16 ***
## yr              1348.84      75.76  17.805  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.66 on 719 degrees of freedom
## Multiple R-squared:  0.826,  Adjusted R-squared:  0.8233
## F-statistic: 310.2 on 11 and 719 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model3)
```



Model Interpretation:

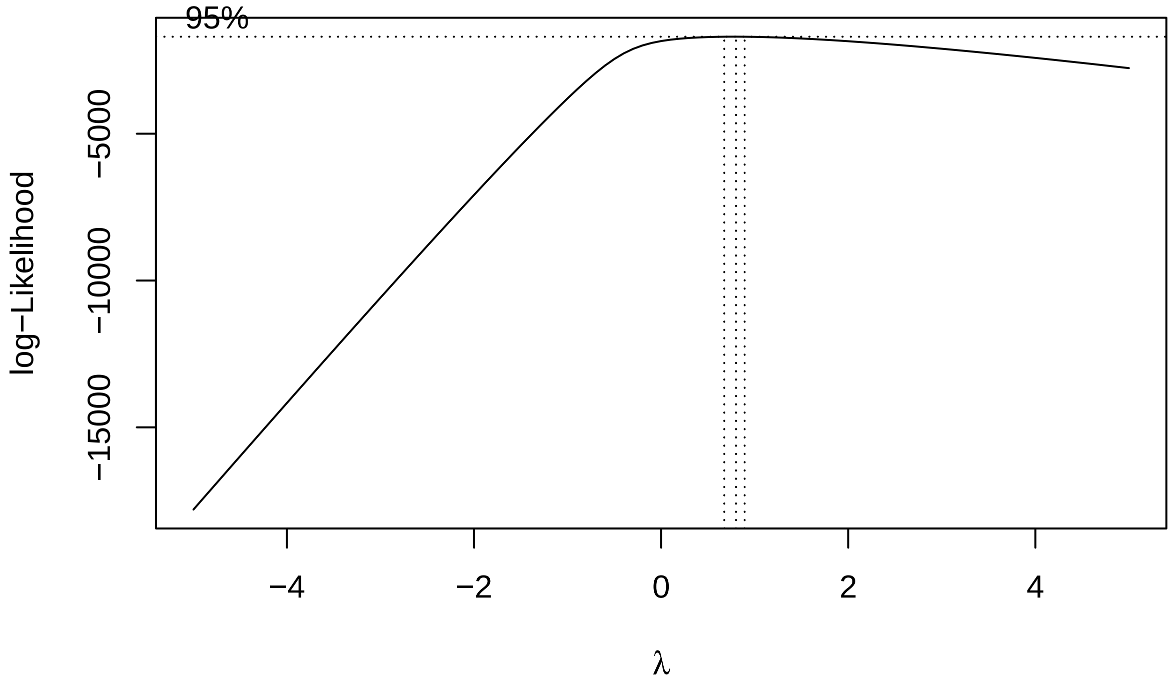
- Model 3 employs weighted least square regression model. The model summary indicates that model 3 has shown a slight improvement. The adjusted R-squared value has increased from 0.8173 to 0.8233, indicating that the model can now explain a slightly larger proportion of the variance in the count of total bike rentals. Additionally, the variable "atemp" is now statistically significant, while the statistical significance of the "workingday" variable decreased, with a p-value increasing to 0.05514.

- The F-statistic has also increased from 297.8 to 310.2, suggesting that the regression model explains a larger proportion of the variance in the response variable.

Plot Diagnostics:

- The Residual vs Fitted plot now shows a slightly curved pattern, with some points further apart from the center for the standardized residual plot.
- The QQ-plot shows that the residuals have not improved, with many points outside of 2 standard deviations, which still violates the normality of error terms.
- Overall, the improvement from the third model is not as significant as the full box-cox transformation, which may be due to the fact that weighted least squares regression is sensitive to the effects of outliers. As we have many influential points in our model, the weighted least squares regression may have increased the influence of outliers and negatively impacted the parameter estimation.

```
lambda <- boxcox(cnt~atemp+temp,data=bike,lambda = seq(-5, 5, 0.1))
```



```
m.null <- lm(cbind(atemp) ~ 1, data=bike)
bc_trans <- powerTransform(m.null)
bc_trans$lambda
```

```
##        Y1
## 0.9915727
```

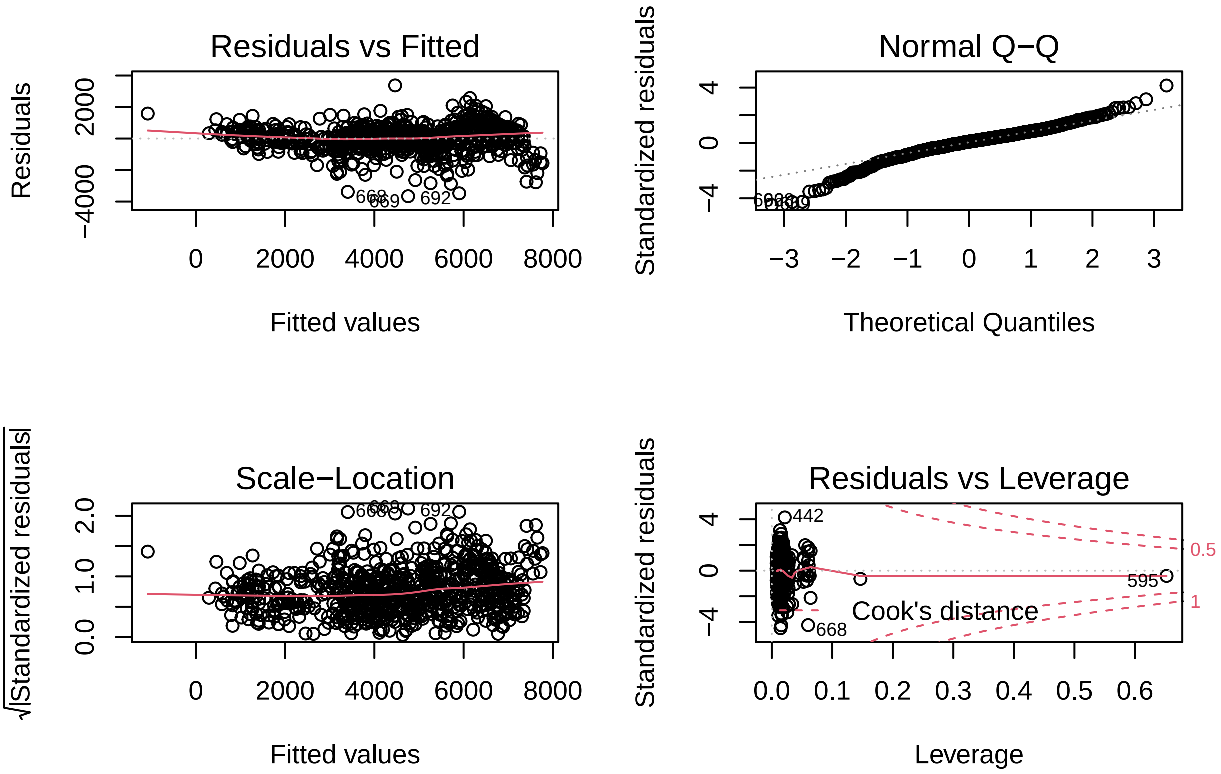
```
bike$atemp_t2 <- (bike$atemp^bc_trans$lambda)

m.null <- lm(cbind(temp) ~ 1, data=bike)
bc_trans <- powerTransform(m.null)
bc_trans$lambda
```

```
##       Y1
## 0.862932
```

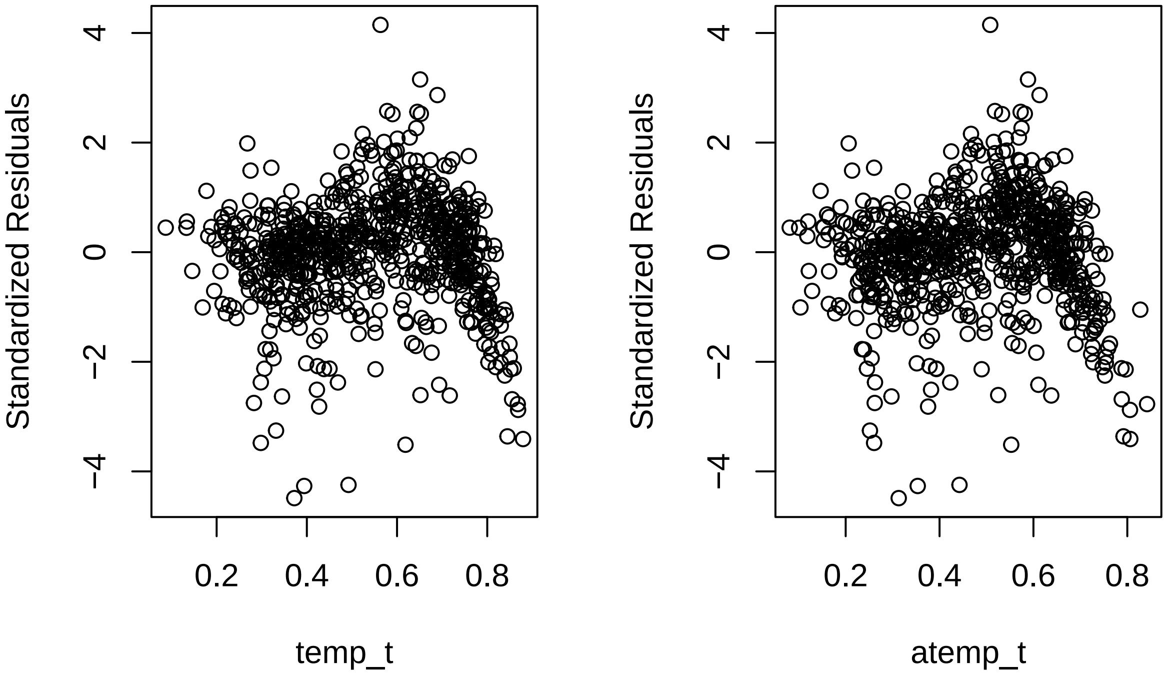
```
bike$temp_t2 <- (bike$temp^bc_trans$lambda)

model4<-lm(cnt~windspeed+hum+atemp_t2+temp_t2+weathersit_f+workingday+season_f+yr, data = bike)
summary(model4)
```

```
##
## Call:
## lm(formula = cnt ~ windspeed + hum + atemp_t2 + temp_t2 + weathersit_f +
##     workingday + season_f + yr, data = bike)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3659.9  -391.7    81.0   468.9  3370.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1197.67     232.03   5.162 3.17e-07 ***
## windspeed     -2906.27     436.97  -6.651 5.76e-11 ***
## hum           -1371.42     295.44  -4.642 4.10e-06 ***
## atemp_t2      -1175.03    1555.42  -0.755  0.45023
## temp_t2        6463.89    1473.55   4.387 1.32e-05 ***
## weathersit_f2  -422.25      81.25  -5.197 2.64e-07 ***
## weathersit_f3 -1901.74     207.51  -9.164  < 2e-16 ***
## workingday      171.65      65.70   2.613  0.00917 **
## season_f2      1113.40     113.91   9.775  < 2e-16 ***
## season_f3       830.09     149.33   5.559 3.83e-08 ***
## season_f4      1517.34      97.05  15.635  < 2e-16 ***
## yr             2004.18      61.77  32.445  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 821.6 on 719 degrees of freedom
## Multiple R-squared:  0.8228, Adjusted R-squared:  0.8201
## F-statistic: 303.6 on 11 and 719 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model4)
```



```
par(mfrow=c(1,2))
SR <- rstandard(model4)
plot(bike$temp_t2, SR, xlab = "temp_t", ylab = "Standardized Residuals")
plot(bike$atemp_t2, SR, xlab = "atemp_t", ylab = "Standardized Residuals")
```



Model Interpretation:

- As part of the experimentation process, I also tried applying box-cox transformation to only "atemp" and "temp" since their residual plots indicated the most non-constant variance. Model 4 has shown a slight improvement with an increased adjusted R-squared value from 0.8173 to 0.8201. However, the transformed version of "atemp", which is "atemp_t2", remains statistically insignificant.
- The F-statistic has also increased from 297.8 to 303.6, suggesting that the regression model explains a larger proportion of the variance in the response variable. The overall p-value is still smaller than 2.2e-16, indicating that the model remains statistically significant.

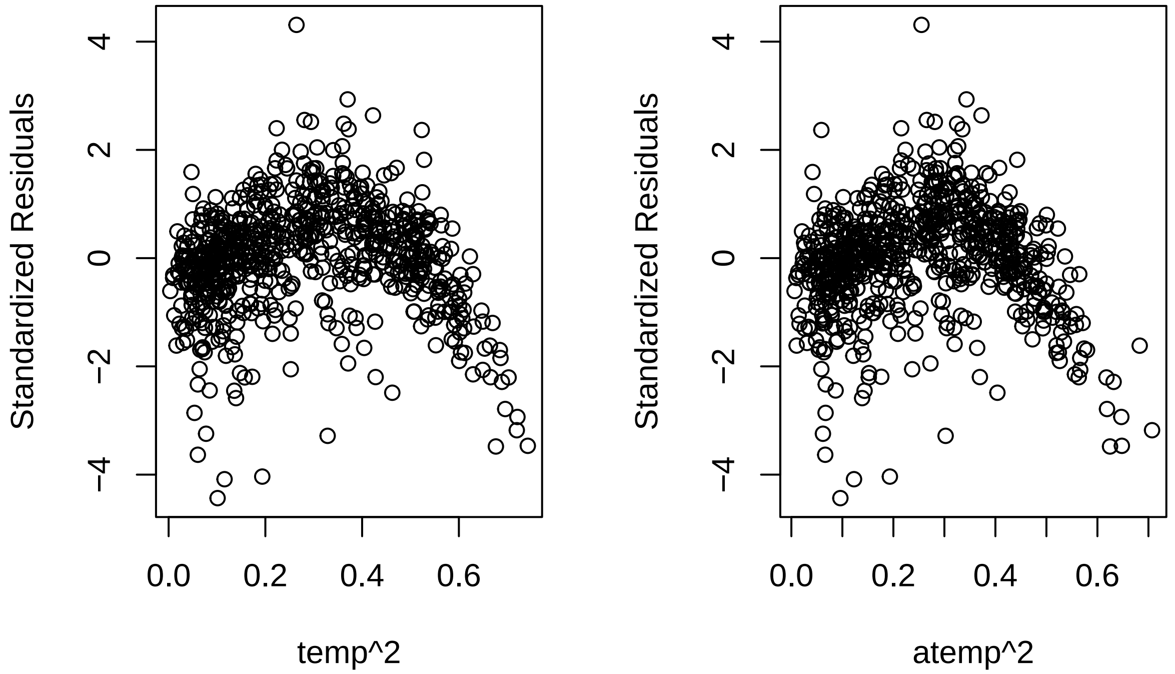
Plot Diagnostics:

- Most issues remain unresolved as the QQ-plot still shows many points outside of 2 standard deviations, which still violates the normality of error terms.Also, when you plot the individual residual plot for "atemp_t2" and "temp_t2", the curved pattern is still obvious. Moreover, there are still many leverage points in the residual vs leverage plot. Overall, the improvement from the third model is better than the weighted least square regression model, but not as significant as the full box-cox transformation.

```
model5<-lm(cnt~windspeed+hum+I(temp^2)+I(atemp^2)+weathersit_f+workingday+season_f+yr,data=bike)
summary(model5)
```

```
##
## Call:
## lm(formula = cnt ~ windspeed + hum + I(temp^2) + I(atemp^2) +
##     weathersit_f + workingday + season_f + yr, data = bike)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3886.6  -413.8    61.3   539.4  3766.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2314.83     235.99   9.809  < 2e-16 ***
## windspeed     -2549.98     461.22  -5.529 4.52e-08 ***
## hum            -967.56     315.06  -3.071  0.00221 **
## I(temp^2)      1659.09    1436.27   1.155  0.24842
## I(atemp^2)     2897.39    1659.81   1.746  0.08131 .
## weathersit_f2  -470.71      87.25  -5.395 9.31e-08 ***
## weathersit_f3 -1986.80     222.86  -8.915  < 2e-16 ***
## workingday      198.15      70.54   2.809  0.00511 **
## season_f2      1506.35     116.18  12.966  < 2e-16 ***
## season_f3      1277.57     162.13   7.880 1.21e-14 ***
## season_f4      1774.58     101.04  17.563  < 2e-16 ***
## yr             2061.33      66.10  31.184  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 882.7 on 719 degrees of freedom
## Multiple R-squared:  0.7955, Adjusted R-squared:  0.7924
## F-statistic: 254.3 on 11 and 719 DF,  p-value: < 2.2e-16
```

```r
SR <- rstandard(model5)
par(mfrow=c(1,2))
plot(bike$temp^2, SR, xlab = "temp^2", ylab = "Standardized Residuals")
plot(bike$atemp^2, SR, xlab = "atemp^2", ylab = "Standardized Residuals")
```



Model Interpretation:

- As part of the experimentation process, I also tried applying quadratic transformation to only "atemp" and "temp" since the shape of their residual plot looks somewhat quadratic. Model 5 has decreased in adjusted R-squared value from 0.8173 to 0.7924. The transformed version of "temp", which is "I(temp^2)", becomes statistically insignificant.
- The F-statistic has also decreased from 297.8 to 254.3, suggesting that the regression model explains an even smaller proportion of the variance in the response variable. The overall p-value is still smaller than 2.2e-16, indicating that the model remains statistically significant.

Plot Diagnostics:

- For the sake of space efficiency, I will not illustrate the diagnostic plots here. However, it is worth noting that the individual residual plot remains parabolic, and all the other prevalent problems of leverage points and normality of error terms also remain unsolved. Thus, we will not consider this model for later use.

Removing Outliers and Bad Leverage Points

- After building several transformed models and comparing their performance, I selected model 2 with the full box-cox transformation. Firstly, model 2 demonstrated the strongest improvement in F-statistics (350.9) and adjusted R-squared value (0.8406). Secondly, model 2 also slightly addresses the problem of normality of errors better than other models. However, one problem remains as all the transformed models still have many leverage points. Thus, I identified all the leverage points first. To identify all the bad leverage points, I selected those leverage points that have high residuals, defined as residuals bigger than 2 standard deviations away from the mean of residuals. After that, I removed those outlier data points from the dataset and constructed a new model.

```
# Find the observations with the highest leverage
high_leverage <- which(leverage > 2 * mean(leverage))


# Get the residuals for the observations with the highest leverage
resid_high_leverage <- resid(model1)[high_leverage]

# Calculate the mean and standard deviation of the residuals
mean_resid <- mean(abs(model1$residuals))
sd_resid <- sd(model1$residuals)

# Select the observations whose residuals are bigger than two standard deviation of mean residuals
outliers <- high_leverage[which(resid_high_leverage > mean_resid + 2*sd_resid | resid_high_leverage < me

cat("Mean residual:", mean_resid, "\n")
```

```
## Mean residual: 599.7953
```

```
cat("Standard deviation of residuals:", sd_resid, "\n")
```

```
## Standard deviation of residuals: 821.8161
```

```
cat("Observations with residuals bigger than two standard deviation from the mean:\n")
```

```
## Observations with residuals bigger than two standard deviation from the mean:
```

```
print(outliers)
```

```
## 202 266 478 668
## 202 266 478 668
```

```
##another method to filter outliers
outliers <- outlierTest(model1)
print(outliers)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 669 -4.517110         7.3276e-06    0.0053565
## 668 -4.337613         1.6467e-05    0.0120370
## 692 -4.293300         2.0022e-05    0.0146360
## 442  4.201299         2.9882e-05    0.0218440
```

```
# Remove outliers
bike_n <- bike[-c(668,669,692,202,266,478,442), ]
```

Refined Model

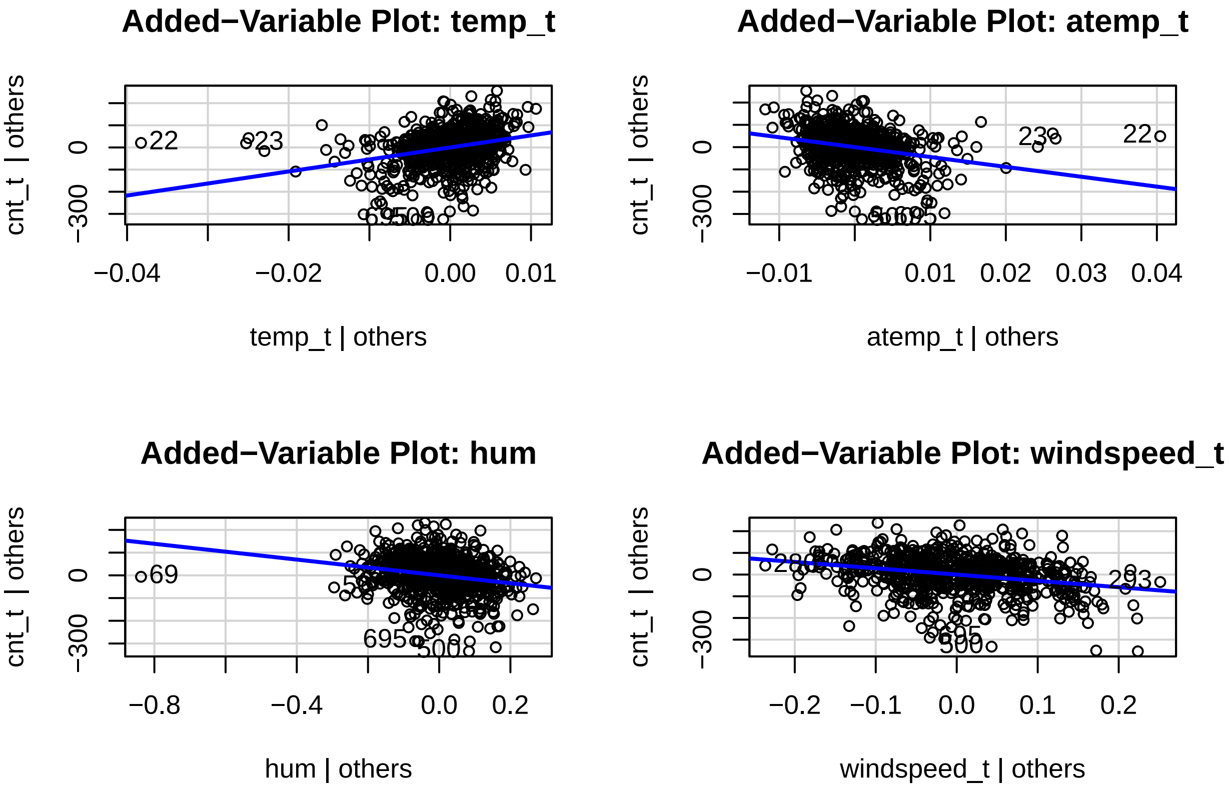
```
model_refined<-lm(cnt_t~windspeed_t+hum+atemp_t+temp_t+weathersit_f+season_f+workingday+yr, data = bike_
summary(model_refined)
```

```
##
## Call:
## lm(formula = cnt_t ~ windspeed_t + hum + atemp_t + temp_t + weathersit_f +
##     season_f + workingday + yr, data = bike_n)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -319.78  -42.77    8.21   50.58  228.12
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -207.564     67.104  -3.093  0.00206 **
## windspeed_t      -290.905     37.894  -7.677 5.39e-14 ***
## hum              -173.614     29.876  -5.811 9.36e-09 ***
## atemp_t         -4431.232    618.154  -7.168 1.90e-12 ***
## temp_t           5427.523    663.984   8.174 1.36e-15 ***
## weathersit_f2     -49.955      8.150  -6.129 1.46e-09 ***
## weathersit_f3    -215.053     21.431 -10.035  < 2e-16 ***
## season_f2         125.818     11.601  10.845  < 2e-16 ***
## season_f3         127.397     15.222   8.369 3.06e-16 ***
## season_f4         168.256     10.212  16.476  < 2e-16 ***
## workingday         21.781      6.611   3.295  0.00103 **
## yr                221.620      6.252  35.448  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.14 on 712 degrees of freedom
## Multiple R-squared:  0.8626, Adjusted R-squared:  0.8605
## F-statistic: 406.3 on 11 and 712 DF,  p-value: < 2.2e-16
```

Model Interpretation:

- The refined model remove leverage points with high residuals and a few outliers with the outlierTest() functuib. The model summary indicates that the model has improved from the selected model 2. The adjusted R-squared value has increased from 0.8406 to 0.8605, which suggests that the model now explains more of the variance in the count of total bike rentals without the influence of outliers.
- The F-statistic has also increased from 350.9 to 406.3, which is a greater step toward explaining the variance in the response variable. The overall p-value is still smaller than 2.2e-16, which suggests that the model is statistically significant.

```
suppressMessages(library(car))
par(mfrow=c(2,2))
avPlot(model_refined,variable="temp_t",ask=FALSE)
avPlot(model_refined,variable="atemp_t",ask=FALSE)
avPlot(model_refined,variable="hum",ask=FALSE)
avPlot(model_refined,variable="windspeed_t",ask=FALSE)
```



avPlot interpretation:

- As suggested by Professor Zanontian, the added variable plots were created for only the continuous variables. Based on the graphs, it appears that all variables contribute to predicting the total count of bike rentals, even after adjusting for the presence of other variables.
- However, there are some notable outliers that seem to have a large influence on certain regression coefficients. For instance, Case 69 appears to have a significant impact on the estimate for the "hum" coefficient, while Cases 23 and 233 seem to have an impact on "atemp_t" and "temp_t".
- Moreover, the trend for "atemp_t" is somewhat unexpected as it contradicts the positive correlation we expect to see with "temp_t". The added variable plot suggests that higher normalized feeling temperatures actually lead to lower bike rentals. One possible explanation for this finding is that the computation of normalized feeling temperature takes humidity into account. Therefore, after adjusting for other variables, higher humidity levels may lead to higher "atemp_t" values but lower incentives for people to rent bikes.

##Reduced model & Multi-collinearity

```
vif(model_refined)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## windspeed_t    1.199195  1        1.095078
## hum            1.920633  1        1.385869
## atemp_t     1253.528308  1       35.405202
## temp_t      1231.481828  1       35.092475
## weathersit_f    1.757920  2        1.151462
## season_f        4.743296  3        1.296224
## workingday      1.011756  1        1.005861
## yr              1.048448  1        1.023937
```

```
model_refined_reduced<-lm(cnt_t~windspeed_t+hum+temp_t+weathersit_f+season_f+workingday+yr, data = bike_
summary(model_refined_reduced)
```

```
##
## Call:
## lm(formula = cnt_t ~ windspeed_t + hum + temp_t + weathersit_f +
##     season_f + workingday + yr, data = bike_n)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -323.99  -39.98    9.78   50.73  252.75
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    223.838     30.718   7.287 8.43e-13 ***
## windspeed_t   -260.547     38.965  -6.687 4.61e-11 ***
## hum           -148.284     30.697  -4.831 1.67e-06 ***
## temp_t         674.222     35.774  18.847  < 2e-16 ***
## weathersit_f2  -48.406      8.430  -5.742 1.39e-08 ***
## weathersit_f3 -212.360     22.172  -9.578  < 2e-16 ***
## season_f2      134.095     11.944  11.227  < 2e-16 ***
## season_f3      104.372     15.396   6.779 2.54e-11 ***
## season_f4      187.449     10.197  18.383  < 2e-16 ***
## workingday      22.930      6.839   3.353 0.000842 ***
## yr             225.935      6.439  35.088  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 85 on 713 degrees of freedom
## Multiple R-squared:  0.8527, Adjusted R-squared:  0.8506
## F-statistic: 412.6 on 10 and 713 DF,  p-value: < 2.2e-16
```

```
anova(model_refined_reduced,model_refined)
```

```
## Analysis of Variance Table
##
## Model 1: cnt_t ~ windspeed_t + hum + temp_t + weathersit_f + season_f +
##     workingday + yr
## Model 2: cnt_t ~ windspeed_t + hum + atemp_t + temp_t + weathersit_f +
##     season_f + workingday + yr
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    713 5151179
## 2    712 4804429  1    346750 51.387 1.902e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Reduced Model and Partial F-test

- Due to high variance inflation factors for the variables “atemp_t” and “temp_t”, their slopes may be poorly estimated due to multicollinearity. In an attempt to address this issue, I tried reducing the full model by removing the less statistically significant variable of the two, which was “atemp_t”. However, the resulting ANOVA table output a p-value of 1.902e-12, which is much smaller than the typical significance level of 0.05. This indicates that we should reject the null hypothesis and accept the alternative hypothesis that the full model with both “atemp_t” and “temp_t” does improve the model more significantly.

- The summary of the reduced model reveals a slightly lower adjusted R-squared value of 0.8506, but a higher F-statistic of 412.6. This suggests that the reduced model may explain a greater portion of the variance in bike rental, but the full model still provides a better fit for the dataset. Given this dilemma, variable selection is another powerful method that can be used to determine the optimal model.

Variable Selection

```
suppressMessages(library(leaps))
backAIC <- step(model_refined, direction="backward", data=bike_n)
```

```
## Start:  AIC=6395.39
## cnt_t ~ windspeed_t + hum + atemp_t + temp_t + weathersit_f +
##     season_f + workingday + yr
##
##                Df Sum of Sq      RSS    AIC
## <none>                       4804429 6395.4
## - workingday    1     73242  4877671 6404.3
## - hum           1    227871  5032300 6426.9
## - atemp_t       1    346750  5151179 6443.8
## - windspeed_t   1    397666  5202094 6451.0
## - temp_t        1    450869  5255298 6458.3
## - weathersit_f  2    739243  5543671 6495.0
## - season_f      3   1845347  6649776 6624.7
## - yr            1   8479037 13283465 7129.7
```

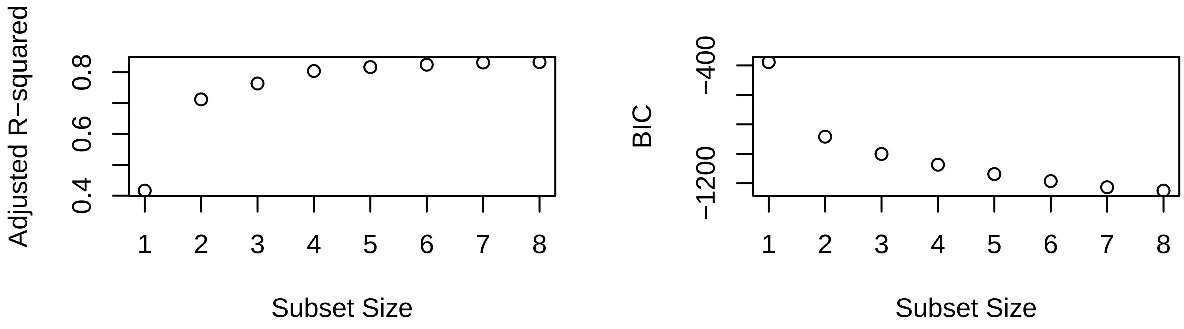
```
backBIC <- step(model_refined, direction="backward", data=bike_n, k= log(n))
```

```
## Start:  AIC=6450.52
## cnt_t ~ windspeed_t + hum + atemp_t + temp_t + weathersit_f +
##     season_f + workingday + yr
##
##                Df Sum of Sq      RSS    AIC
## <none>                       4804429 6450.5
## - workingday    1     73242  4877671 6454.9
## - hum           1    227871  5032300 6477.5
## - atemp_t       1    346750  5151179 6494.4
## - windspeed_t   1    397666  5202094 6501.5
## - temp_t        1    450869  5255298 6508.9
## - weathersit_f  2    739243  5543671 6540.9
## - season_f      3   1845347  6649776 6666.1
## - yr            1   8479037 13283465 7180.2
```

```r
par(mfrow=c(2,2))
attach(bike_n)
```

```
## The following objects are masked from bike:
##
##     atemp, casual, cnt, dteday, holiday, hum, instant, mnth,
##     registered, season, season_f, temp, weathersit, weathersit_f,
##     weekday, windspeed, workingday, yr
```

```r
X<-cbind(windspeed_t,hum,temp_t,atemp_t,weathersit_f,season_f,yr, workingday)
b<-regsubsets(as.matrix(X),cnt)
rs<-summary(b)
plot(1:8,rs$adjr2,xlab="Subset Size",ylab="Adjusted R-squared")

b1<-regsubsets(cnt~windspeed_t+hum+temp_t+atemp_t+weathersit_f+season_f+yr+workingday,data=bike_n)
rs<-summary(b1)
plot(1:8,rs$bic,xlab="Subset Size",ylab= "BIC")
```



Adjusted R-squared value, BIC, and AIC interpretation

- To identify the best subset of predictor variables for our regression model, we employed stepwise backward elimination with both AIC and BIC criteria. The summary results revealed that the full model had the lowest AIC and BIC scores, indicating that there was no need to consider simpler models over the full model.

- Additionally, we plotted the subset size against BIC and adjusted R-squared values and found that the full model had the highest adjusted R-squared value and the lowest BIC score. These findings further reinforced the decision to select the full model as the final model.
- Overall, after conducting the partial F-test, stepwise BIC and AIC backward elimination, and visualizing the subsets, we determined that the full model is likely to outperform other models. Although the high variance inflation factors for some variables may affect the precision of their estimates, the contribution of these variables to the overall improvement of the model may justify their inclusion in the full model.

Neural Network Experimentation

- I'm also interested in experimenting with neural networks though the concepts are beyond the scope of the class because they extend the basic concept of linear regression, relying on weighted sums of input features to model relationships between input and output variables. However, neural networks introduce non-linear activation functions and multiple layers of interconnected nodes, allowing them to capture more complex and nuanced relationships in the data. By exploring neural networks, we can gain a deeper understanding of tthe the potential for more accurate and sophisticated predictions.

```
# Load the required libraries
suppressMessages(library(keras))
```

```
## Warning: package 'keras' was built under R version 4.1.2
```

```
suppressMessages(library(rsq))
```

```
## Warning: package 'rsq' was built under R version 4.1.2
```

```
# Select the columns of interest
cols <- c("windspeed_t", "hum", "temp_t", "cnt_t","atemp_t")
bike_scaled <- bike_n[, cols]

# Normalize the data
bike_scaled <- scale(bike_scaled)

# Assign the normalized values back to the original dataframe
bike_n[, cols] <- bike_scaled

# Select the columns of interest
cols <- c("windspeed_t", "hum", "temp_t", "atemp_t", "weathersit", "workingday", "season", "yr", "cnt_t
bike_n <- bike_n[, cols]

# Split the data into training and testing sets
set.seed(121)
train_index <- sample(1:nrow(bike_n), round(0.8 * nrow(bike_n)))
train_data <- bike_n[train_index, ]
test_data <- bike_n[-train_index, ]

# Preprocess the input data
x_train <- as.matrix(train_data[, c(1:8)])
y_train <- train_data$cnt_t
x_test <- as.matrix(test_data[, c(1:8)])
```

```
# Build the neural network
model <- keras_model_sequential() %>%
  layer_normalization() %>%
  layer_dense(units = 32, activation = "relu", input_shape = c(ncol(x_train))) %>%
  layer_dense(units = 16, activation = "relu") %>%
  layer_dense(units = 1)

model %>% compile(
  optimizer = "adam",
  loss = "mse",
  metrics = c("mae")
)

# Train the model
history <- model %>% fit(
  x = x_train,
  y = y_train,
  epochs = 50,
  batch_size = 32,
  validation_split = 0.2
)


# Make predictions on the test data
predicted <- predict(model, x_test)

# Compute summary statistics
mse <- mean((predicted - test_data$cnt_t)^2)
mae <- mean(abs(predicted - test_data$cnt_t))

cat("Mean squared error:", mse, "\n")
```

```
## Mean squared error: 0.1265911
```

```
cat("Mean absolute error:", mae, "\n")
```

```
## Mean absolute error: 0.2657668
```

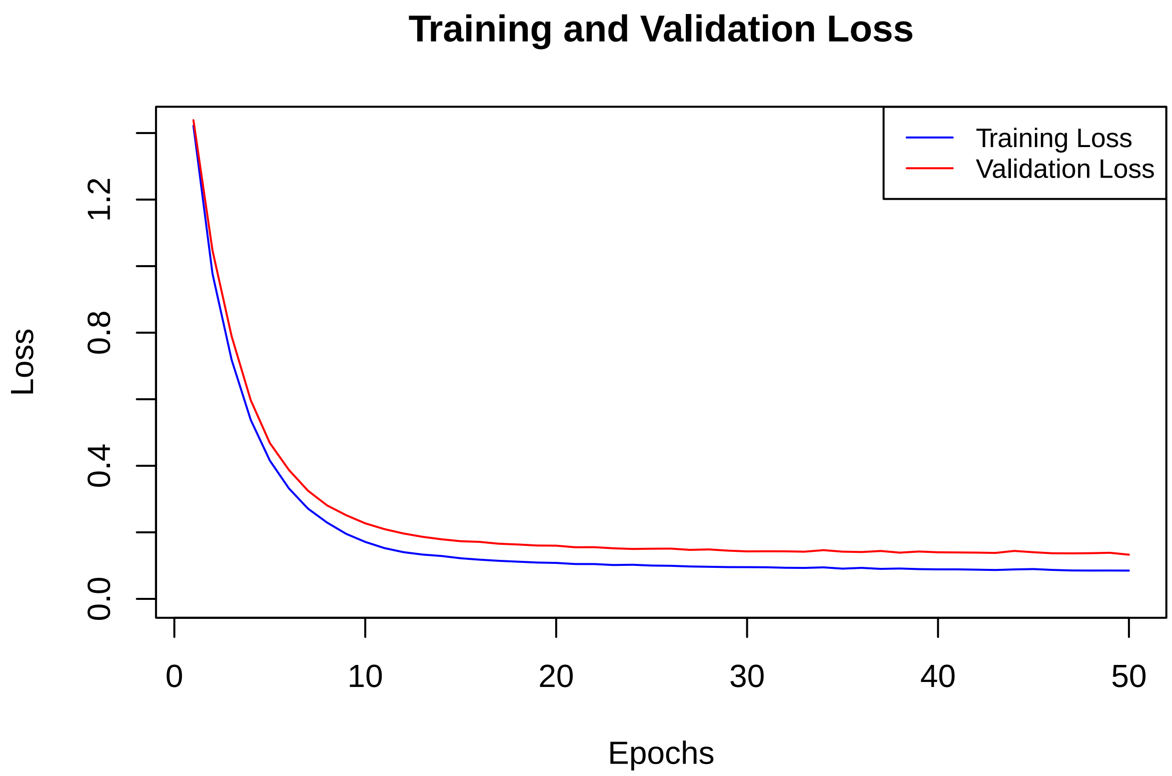
```
# Compute the R-squared value
r_squared <- 1 - sum((test_data$cnt_t - predicted)^2) / sum((test_data$cnt_t - mean(test_data$cnt_t))^2
cat("R-squared value:", r_squared, "\n")
```

```
## R-squared value: 0.8708112
```

```
# Compute the adjusted R-squared value
n <- nrow(train_data)
p <- ncol(x_train)
adj_r_squared <- 1 - ((1 - r_squared) * (n - 1)) / (n - p - 1)
cat("Adjusted R-squared value:", adj_r_squared, "\n")
```

```
## Adjusted R-squared value: 0.868998
```

```
plot(history$metrics$loss, type = "l", col = "blue", xlab = "Epochs", ylab = "Loss",
     main = "Training and Validation Loss", ylim = c(0, max(history$metrics$loss)))
lines(history$metrics$val_loss, col = "red")
legend("topright", legend = c("Training Loss", "Validation Loss"), col = c("blue", "red"), lty = 1, cex
```



Neural Network Interpretation

- The neural network model performs decently in predicting the total count of bike rentals, with a mean squared error of 0.1288364 and mean absolute error of 0.2683099, indicating a low level of prediction error. The loss graph shows that the error between predicted and true values decreases exponentially as the neural network learns the underlying patterns in the dataset, with similar trends observed in both the training and validation curves, suggesting that the model is not overfitting or underfitting the data. The adjusted R-squared value for the neural network model is 0.8666744, slightly better than the full model's value of 0.8605. However, since the full model was not split into training and testing data, it's unclear if the full model would perform similarly well on new data. Therefore, additional evaluation of the model's performance on new data would be necessary to determine the effectiveness of the full model.

Discussion

Summary

We began with the complete model and identified several issues in the diagnostic plots, including non-normality of error terms and an abundance of high-leverage points. To address these issues, we experimented with various data transformations such as full Box-Cox transformation, partial predictor Box-Cox transformation and squaring, and weighted least squares regression. Model 2, which utilized the full Box-Cox transformation, was chosen due to its superior performance in terms of adjusted R-squared value, statistical significance of coefficients, and F-statistic. Additionally, this model best addressed the normality of errors. However, the issue of high-leverage points remained unresolved.

To tackle the leverage points problem, we identified and removed the most influential high-leverage points with substantial residuals from the dataset. This resulted in the creation of the "bike_n" dataset, which demonstrated improvements in both F-statistic and adjusted R-squared values. Using this refined dataset, we developed the "model_refined" and proceeded with variable selection to address multicollinearity and model reduction. Despite utilizing F-partial tests, stepwise BIC and AIC backward elimination methods, and relevant graphs, the full model consistently outperformed reduced models.

Ultimately, we selected the full "model_refined", which achieved an adjusted R-squared value of 0.8605 and an F-statistic of 406.3, outperforming all alternative models. As a supplementary analysis, we constructed a neural network, which yielded a slightly improved adjusted R-squared value of 0.8666744. However, further evaluation of these models' performance on new data is required to determine the full model's effectiveness conclusively.

Real-World Application and Limitations

- Limited generalizability
  - The dataset used in this project only covers data from 2011 and 2012, which may limit the model's generalizability to other time periods. In the analysis, the variable "yr" played a crucial role in distinguishing external factors between the years, resulting in accurate predictions for 2011 and 2012. However, to predict bike rentals in the future or for a random year, the year should not be included as a variable.
  - Similar concerns apply to the model's generalizability across locations since the dataset focuses on bike rentals in a specific area. External factors, such as the availability of bike lanes and bike-accessible roads, could affect bike-sharing system usage. For example, bike-sharing might be less popular in Dallas, regardless of weather and time features, due to a lack of bike-friendly infrastructure.
- Other variables that may play important roles
  - The dataset does not consider factors that could influence bike-sharing demand, such as public transportation availability and infrastructure changes. For instance, the paper "Factors Influencing Travel Behaviors in Bikesharing" demonstrated that land-use factors positively influence bike-sharing. On non-rainy weekdays, commercial areas generated 15 times more rides than residential areas, while parks generated 3 to 5 times more rides than subway stations and schools.(Kim 2012, pp.13) Incorporating these variables could improve the model's real-world predictive accuracy.
- Weather data limitations:
  - The weather variables in the dataset are relatively simplistic, potentially failing to capture the full range of weather conditions that might affect bike-sharing usage. Including more granular and precise weather data, such as precipitation levels and visibility, could enhance the model's performance.
- Aggregated data
  - The current dataset is aggregated on a daily basis, which may obscure short-term fluctuations in bike-sharing demand. To gain a deeper understanding and predict bike rentals on an hourly basis, incorporating higher-resolution data, such as hourly changes, could offer more precise insights into the dynamics of bike-sharing usage.