Stat151A Midterm 2

Juliann Nguyen

4/10/2020

Introduction

Bike sharing systems are a service where individuals are able to rent a bike from a particular position and then return it at another position. This whole process including membership, rental, and return are all automatic. Some useful information would be the number of bike rentals in a given hour because this would allow bike sharing system owners to figure out estimates for revenue, what hours are best for repairs if needed, and a multitude of other things. In this report, we will explore environmental and seasonal conditions in relation with the number of bike rentals with the goal of predicting the number of bike rentals in a given hour given these conditions.

Data Description

The data that we will be using is from Capital Bikeshare System, Washington D.C. Bike sharing systems during the years of 2011 and 2012. The bike rentals from this system have an automatic process from membership, rental, and return, so users are easily able to rent and return bikes at different positions.

The dataset captures 17 variables:

1. **instant** : Unique observation number
2. **dteday** : Date
3. **season** : (1: Spring, 2: Summer, 3: Fall, 4: Winter)
4. **yr** : year (0: 2011, 1: 2012)
5. **mnth** : month (1, 2, . . . , 12) with numbers representing the nth month of the year
6. **hr** : hour of the day (0, 1, . . . , 23)
7. **holiday** : Indicator of whether the day is a holiday
8. **weekday** : Day of the week, numbered 0 (Sunday) through 6 (Saturday)
9. **workingday** : 1: if the day is neither a weekend nor holiday and 0 otherwise
10. **weathersit** : Takes four values
    a. 1 if the weather is Clear, Few Clouds, Partly Cloudy
    b. 2 if the weather is Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    c. 3 if the weather is Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    d. 4 if the weather is Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
11. **temp** : Normalized temperature in Celsisus, values are divided by 41 (the maximum temperature)
12. **atemp** : Normalized feeling temperature in Celsisus, values are divided by 50 (the maximum temperature)
13. **hum** : Normalized humidity, values are divided by 100 (the max humidity)
14. **windspeed** : Normalized wind speed, values are divided by 67 (the maximum wind speed)
15. **casual** : The number of bikes rented by casual (unregistered) users for that hour

16. **registered** : The number of bikes rented by registered users for that hour
17. **cnt** : The number of bikes rented by both casual and registered users for that hour (the sum of casual and registered)

In this analysis, we will be focusing on making predictions of **cnt** which is the number of bikes rented by all users for a given hour using the other variables.

Exploratory Data Analysis

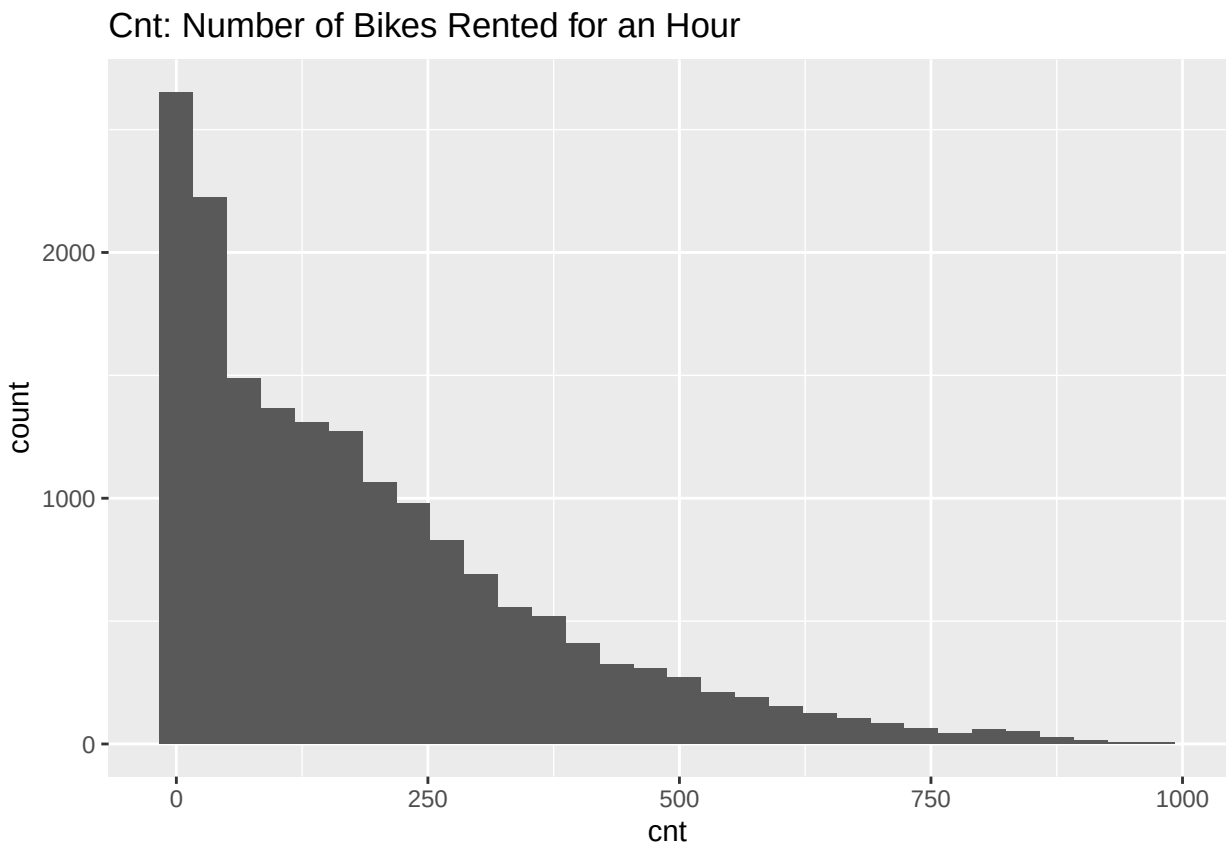
Before we begin our analysis, we will do some exploration into our variables, so that we may later justify our assumptions about our data when using certain models.

Response Variable

The response variable in this case is coded as **cnt** in the dataframe and represents the number of bike rentals in an hour.

```
##       cnt
##  Min.   :  1.0
##  1st Qu.: 40.0
##  Median :142.0
##  Mean   :189.5
##  3rd Qu.:281.0
##  Max.   :977.0
```

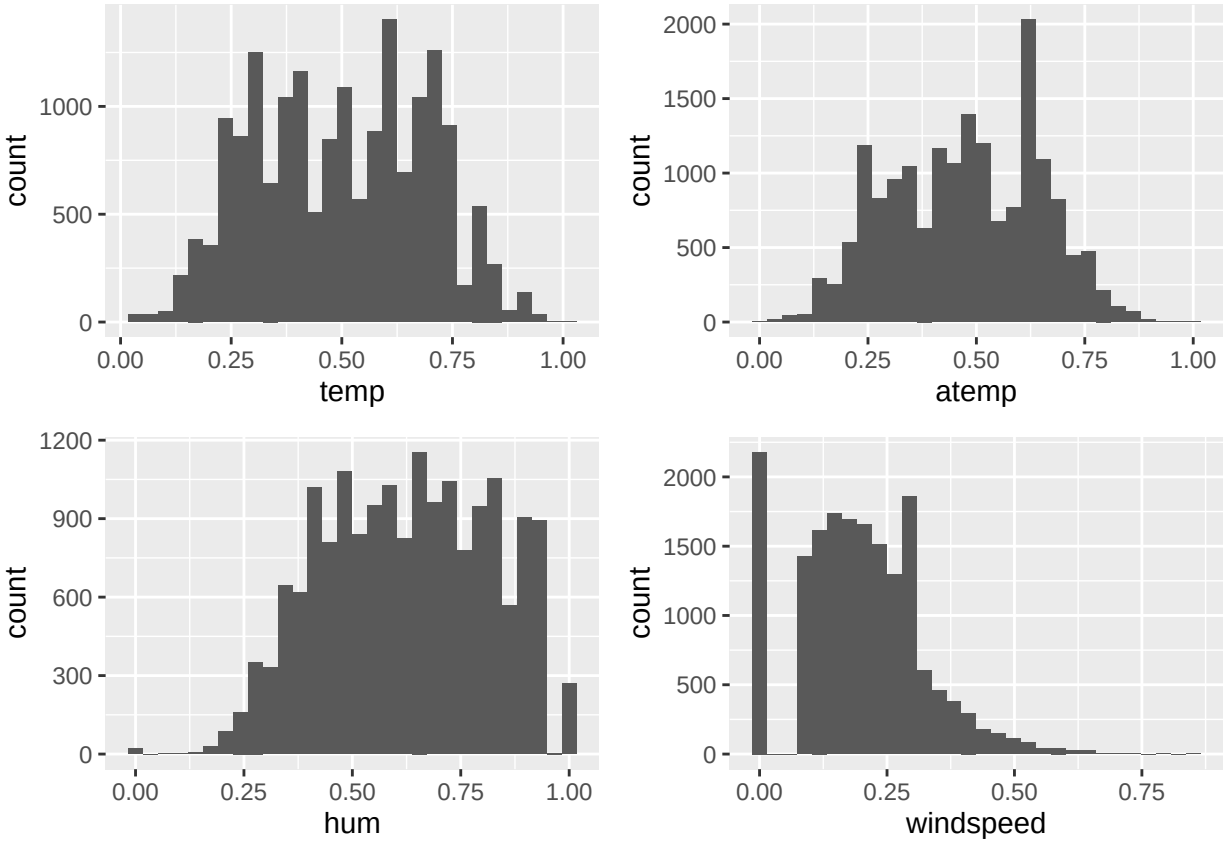
Looking at the summary statistics, we can see that the data is skewed since the mean is quite a bit higher than the median and the difference between the 75th percentile and the max is much greater than the minimum and the 25th percentile. Calculating the variance of **cnt** which is 3.2901461×10^4, it is also clear that the number of rentals is also quite varible. We will plot the distribution below to get a better look.

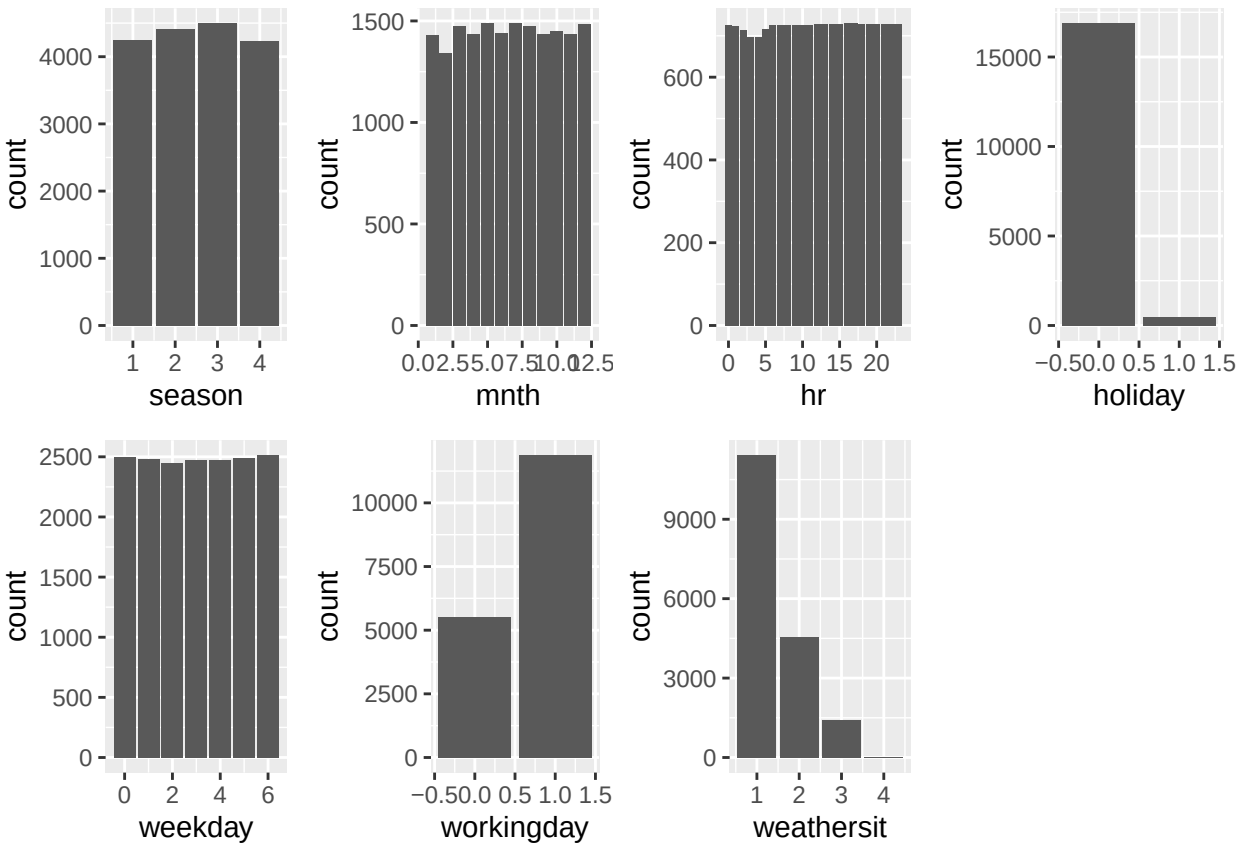


The distribution of **cnt**, the number of bike rentals for an hour, is skewed right, we could consider transforming the data, but it would depend on whether relationships with our explanatory variables end up being linear or not. Some transformations that might help if our explanatory variables are not skewed are log or a box-cox transformation.

We are now going to shift into visualizing the distribution of continous explanatory variables in histograms and discrete explanatory variables in barcharts. This will allow us to see whether some of the continous values follow a normal or skewed distribution and whether discrete values are fairly uniform or whether some values are much more common than others. The variables of **casual** and **registered** are being left out as explanatory variables although they are not our response variable **cnt** the sum of the two makes our variable **cnt** and also the variable of **yr** is going to be left out because we hope to make predictions about year and the difference between 2011 and 2012 does not provide enough information to help us determine values for future years that makes sense, so it does not make sense to use them as explanatory variables.

Univariate Distributions of Explanatory variables



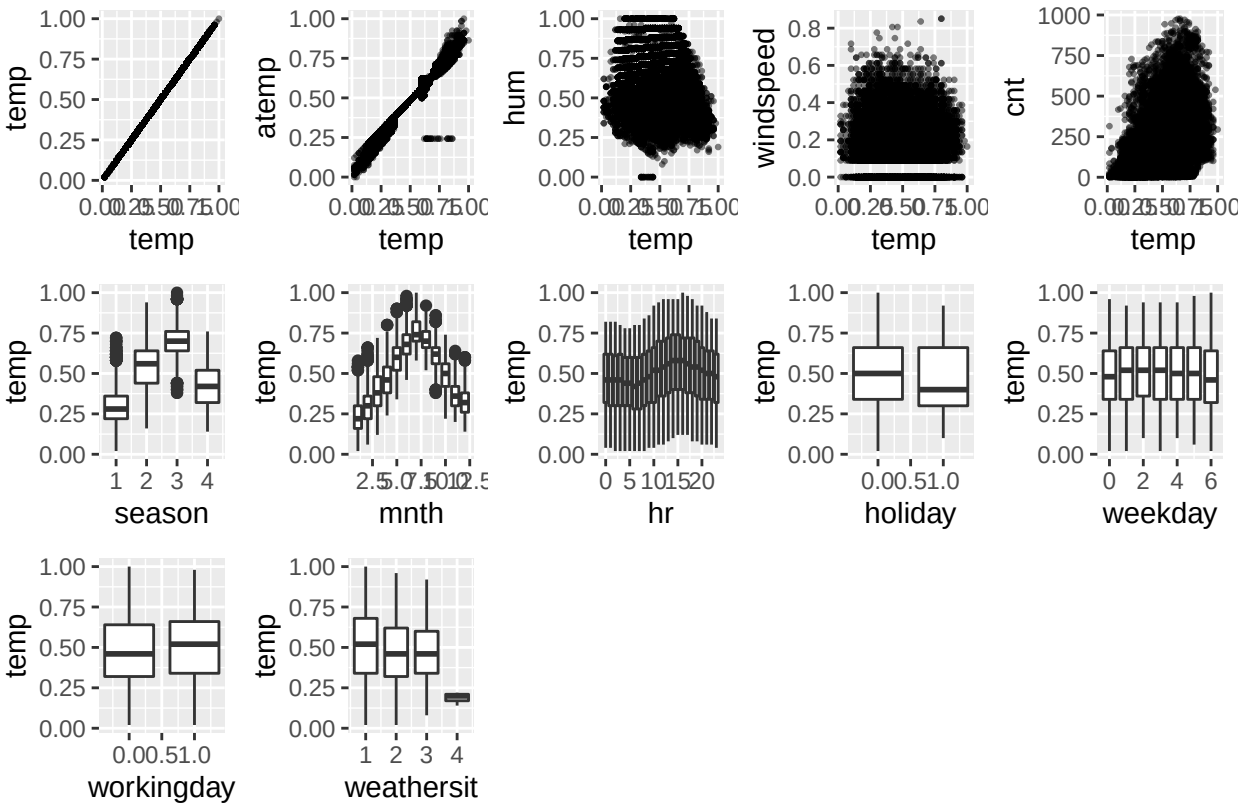


Looking at the plots, temp, atemp, and hum have a variety of values with an extremely rough possibly normal distributions. While windspped has more of a positively skewed distribution, this would make sense as weather variables such as wind sometimes face extremes accounting for the right skew. There was a decision to leave out plotting both Instant and dteday since these should have no interesting distribution since the days should be fairly uniform after analyzing month and year and hour; Instant is simply an unique observation number and the numbers are all different. The barplots for the discrete variables are as we expected fairly uniform with season, month, hour, and weekday while holiday, workingday, and weathersit have a majority value.

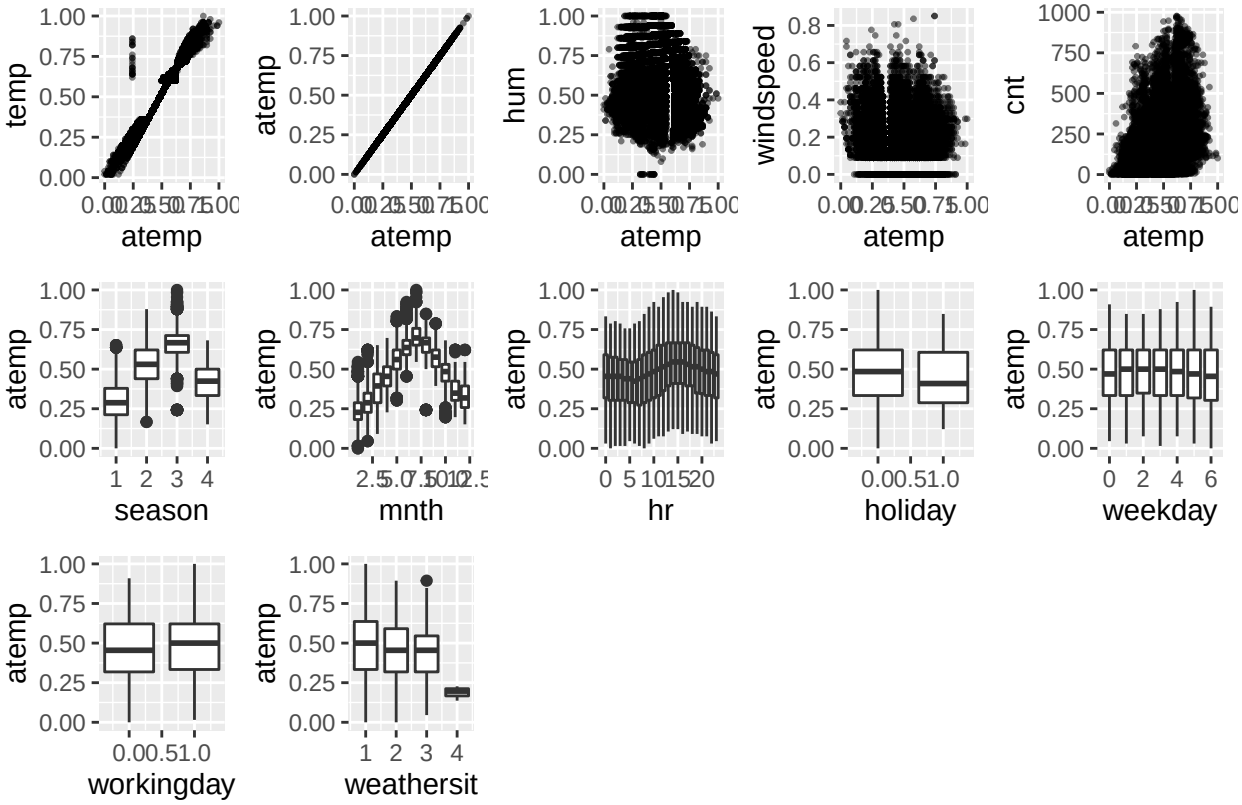
Bivariate Analysis

Analyzing the relationships between the explanatory variables as well as their relationship with our response variable will allow us to vizualize whether collinearity will be a problem with highly correlated explanatory variables and whether transformations for either explanatory variables or our resposnse variable would be helpful to linearize the relationship better.

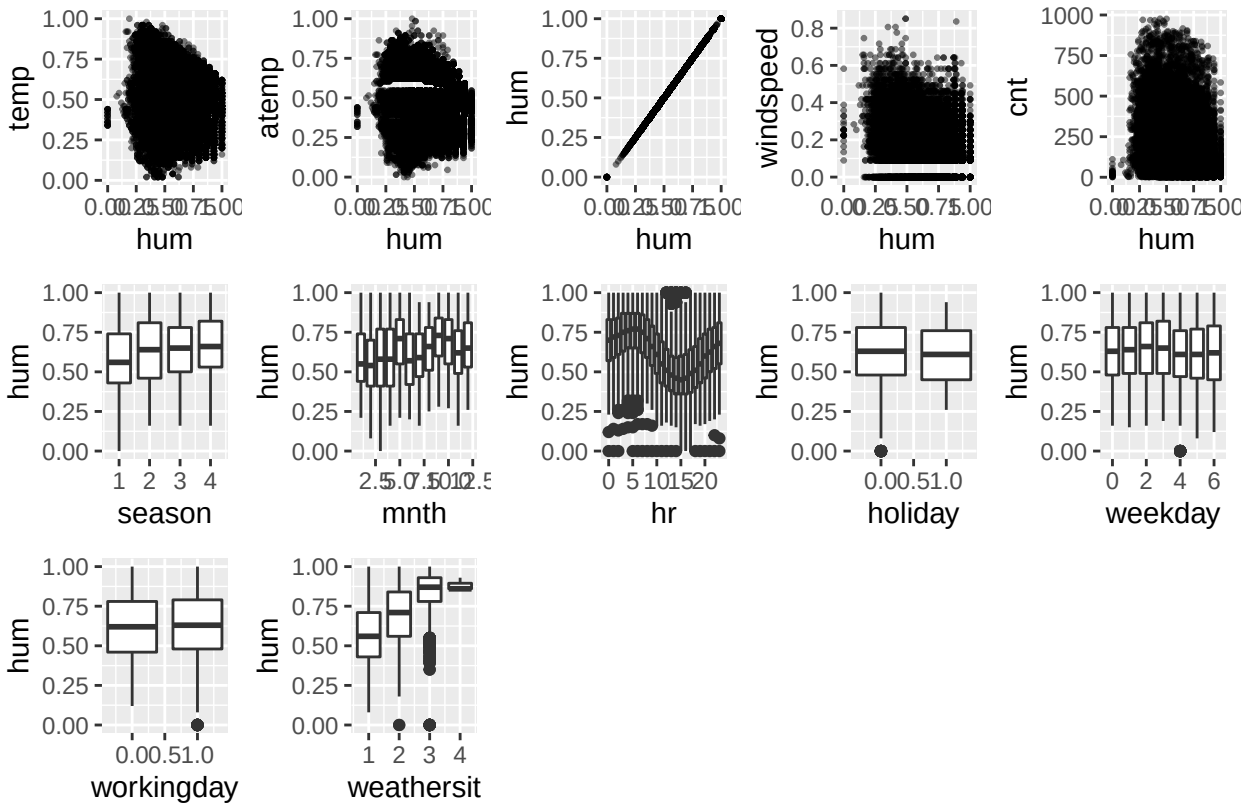
Scatterplots and Boxplots for temp



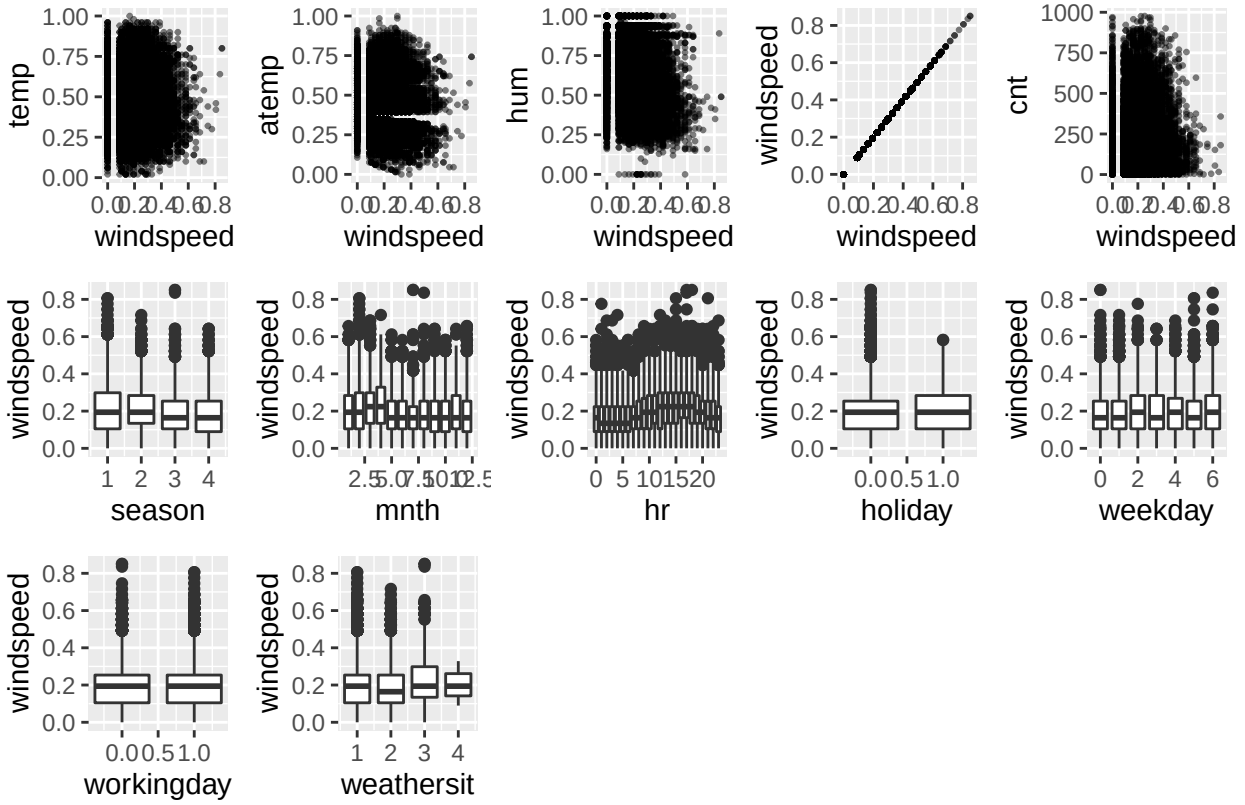
Scatterplots and Boxplots for atemp



Scatterplots and Boxplots for hum



Scatterplots and Boxplots for windspeed



From these plots, it is clear that there is fairly high correlation between **temp** and **atemp** which makes sense since they are both measures of temperature, so using both in the final model could cause some problems due to collinearity. There also appears to be some relationship between **hum** and **weathersit**, **hum** and **hour**, **season**, **month**, **weathersit** with both **temp** and **atemp**. From the scatterplots involving our reponse, **cnt**, there doesn't seem to be a very clear linear relationship with any of the variables, so a transformation of our response variable might be helpful in this situation, but since there are so many points it is difficult to clearly see, in which case a correlation matrix could be helpful.

	season	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed
season	1.00	0.83	-0.01	-0.01	0.00	0.01	-0.01	0.31	0.32	0.15	-0.15
mnth	0.83	1.00	-0.01	0.02	0.01	0.00	0.01	0.20	0.21	0.16	-0.14
hr	-0.01	-0.01	1.00	0.00	0.00	0.00	-0.02	0.14	0.13	-0.28	0.14
holiday	-0.01	0.02	0.00	1.00	-0.10	-0.25	-0.02	-0.03	-0.03	-0.01	0.00
weekday	0.00	0.01	0.00	-0.10	1.00	0.04	0.00	0.00	-0.01	-0.04	0.01
workingday	0.01	0.00	0.00	-0.25	0.04	1.00	0.04	0.06	0.05	0.02	-0.01
weathersit	-0.01	0.01	-0.02	-0.02	0.00	0.04	1.00	-0.10	-0.11	0.42	0.03
temp	0.31	0.20	0.14	-0.03	0.00	0.06	-0.10	1.00	0.99	-0.07	-0.02
atemp	0.32	0.21	0.13	-0.03	-0.01	0.05	-0.11	0.99	1.00	-0.05	-0.06
hum	0.15	0.16	-0.28	-0.01	-0.04	0.02	0.42	-0.07	-0.05	1.00	-0.29
windspeed	-0.15	-0.14	0.14	0.00	0.01	-0.01	0.03	-0.02	-0.06	-0.29	1.00
cnt	0.18	0.12	0.39	-0.03	0.03	0.03	-0.14	0.40	0.40	-0.32	0.09

From this matrix it does hold that there is high correlation between **atemp** and **temp** as well as some slight relationships between **cnt** and some of the explanatory variables such as **atemp**, **temp**, **hr**, and **hum**. We will now compare the correlations between the explanatory variables and a transformation of **cnt** using a square root to adjust for the skewness.

	cnt	sqrt_cnt
season	0.1780557	0.1826257
mnth	0.1206378	0.1238334
hr	0.3940715	0.4956388
holiday	-0.0309273	-0.0310286
weekday	0.0268999	0.0292771
workingday	0.0302844	0.0238823
weathersit	-0.1424261	-0.1418882
temp	0.4047723	0.4207743
atemp	0.4009293	0.4172280
hum	-0.3229107	-0.3521589
windspeed	0.0932338	0.1093095

For some of the highly correlated values, the correlation coefficient magnitude increases with the transformed value of **cnt**, so moving forward we will use **sqrt_cnt** as our response variable since our goal is prediction and accuracy matters.

Modeling

To predict **cnt**, the number of bike rentals for an hour, we are going to consider 4 methodologies (1) Model Selection using Adjusted R2 with an exhaustive method, (2) Ridge Regression, (3) Lasso Regression, and (4) Lasso Regression for Feature Selection. So that we may test the validity of our model better in the future we only be using a subset of the data as training data and leave the rest as testing data.

Model Selection Adjusted R2

We are going to run two model selections, one using all of the explanatory variables not taking into account **atemp** and one not taking into account **temp** since these were so highly correlated, running and OLS model after selection of variables of this method, the collinearity could cause problems when we fit the model. We also decided to use a maximum of 6 variables to predict **sqrt_cnt**, because using this method we believe that not all of the variables really matter in the prediction and we didn't want to allow for all variables.

The adjusted R2 value for the model not using temp: 0.434

The adjusted R2 value for the model not using atemp: 0.432

Since the adjusted R2 value for not using **temp** is higher that is the one we will test against the other models.

Variables to use after Model Selection:

	variable
1	(Intercept)
2	season
4	hr
9	atemp
10	hum

Ridge Regression

Ridge regression would be useful tool in this case because it is a regression technique that helps with analyzing multiple regression data that may suffer from collinearity. By biasing the size of the regression coefficients, the standard errors are reduced and the large variance caused by collinearity could be reduced. We are not quite sure whether all of the variables have nonzero coefficients in the true model, so ridge in the case only shrinks coefficients, instead of making them zero as we did in model selection above.

Something that ridge regression requires however is a lambda value, this value is something that a person must determine before fitting the model, but this value can be determined using cross validation, so that is what was done. We ended up using a value of 0.332.

Coefficients of Optimal Model

variable	coefficient
(Intercept)	12.0030607
season	0.7574733
mnth	0.0810393
hr	2.4691600
holiday	-0.1254925
weekday	0.1423697
workingday	0.0117273
weathersit	-0.0487430
temp	0.9215504
atemp	1.1926110
hum	-1.5394410
windspeed	0.1679709

LASSO Regression

Lasso regression may be helpful in fitting a model in the case as well because it is a method for model shrinkage by setting some coefficients to 0. This helps since less variables leave less room for collinearity problems and it is also helpful in feature selection, because the dataset does provide a relatively large amount

of variables that may lead to overfitting and a simplier model would help interpretation.

Similar to ridge regression, lasso also requires a lambda function, which was calculated using cross validation and was set to be 0.017.

Coefficients of Optimal Model

variable	coefficient
(Intercept)	12.0030607
season	0.8474007
mnth	0.0000000
hr	2.5587613
holiday	-0.1136873
weekday	0.1378606
workingday	0.0000000
weathersit	0.0000000
temp	0.3617729
atemp	1.7649214
hum	-1.6085380
windspeed	0.1522102

LASSO for Feature Selection

Features using LASSO for Feature Selection

variable
(Intercept)
season
mnth
hr
holiday
weekday
workingday
weathersit
temp
atemp
hum
windspeed

Choosing a Final Model

In order to choose a final model, we are going to be looking at the MSE of these four models and whether their assumptions are met to determine which is the best model.

Compare MSE

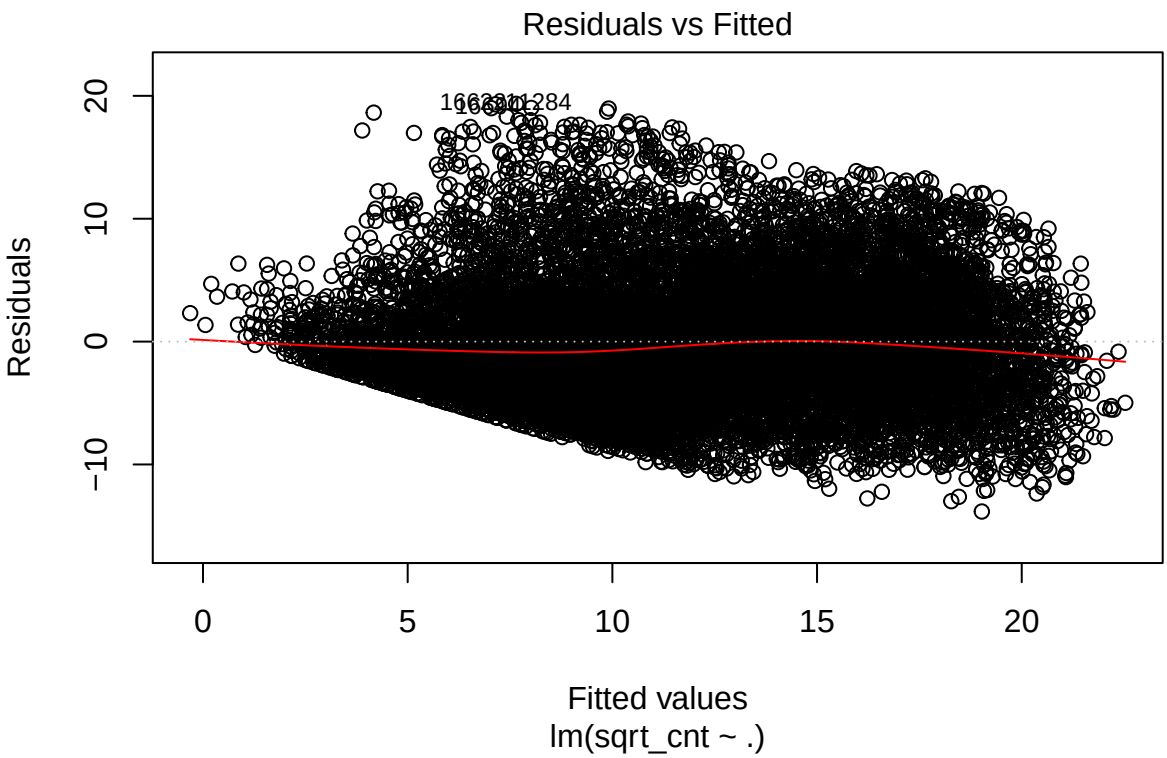
	selection	ridge	lasso	lasso feature
k1	25.24795	25.17708	25.14249	25.13681
k2	25.73406	25.75323	25.69979	25.69245
k3	25.91817	25.98834	25.93070	25.93557
k4	24.50819	24.44777	24.41264	24.39402
k5	26.70229	26.61396	26.62405	26.62603

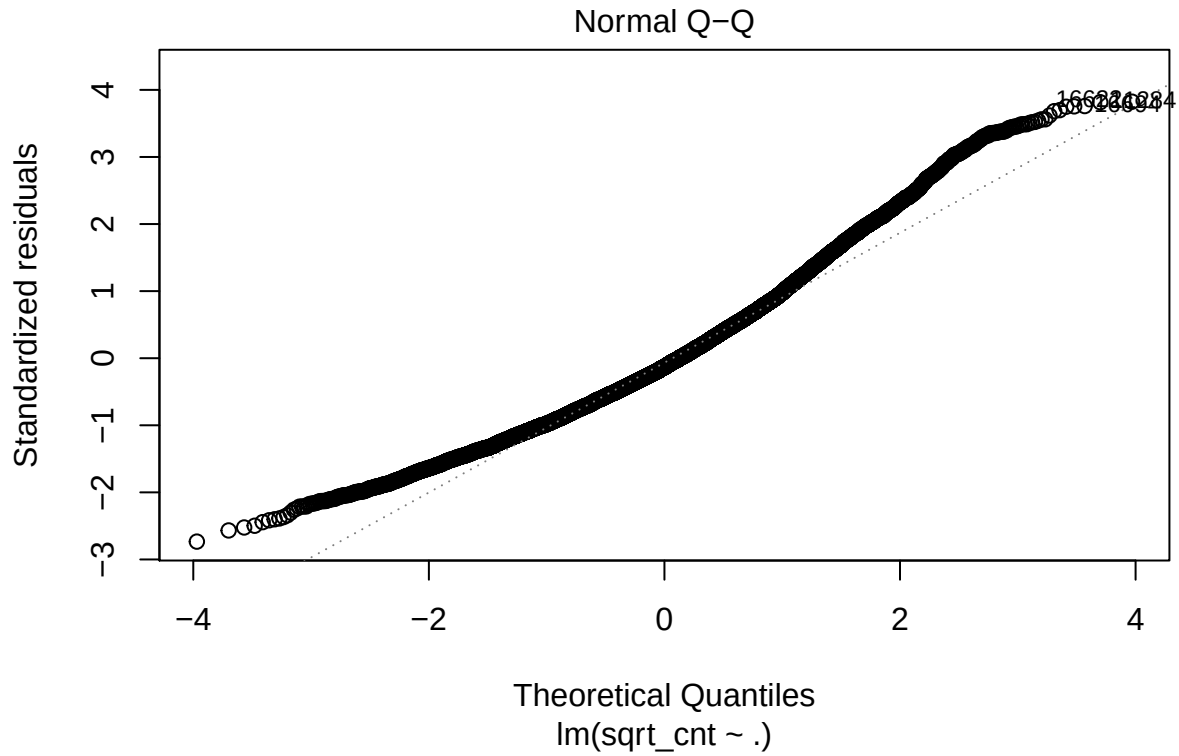
	selection	ridge	lasso	lasso feature
average MSE across ks	25.62213	25.59608	25.56193	25.55698

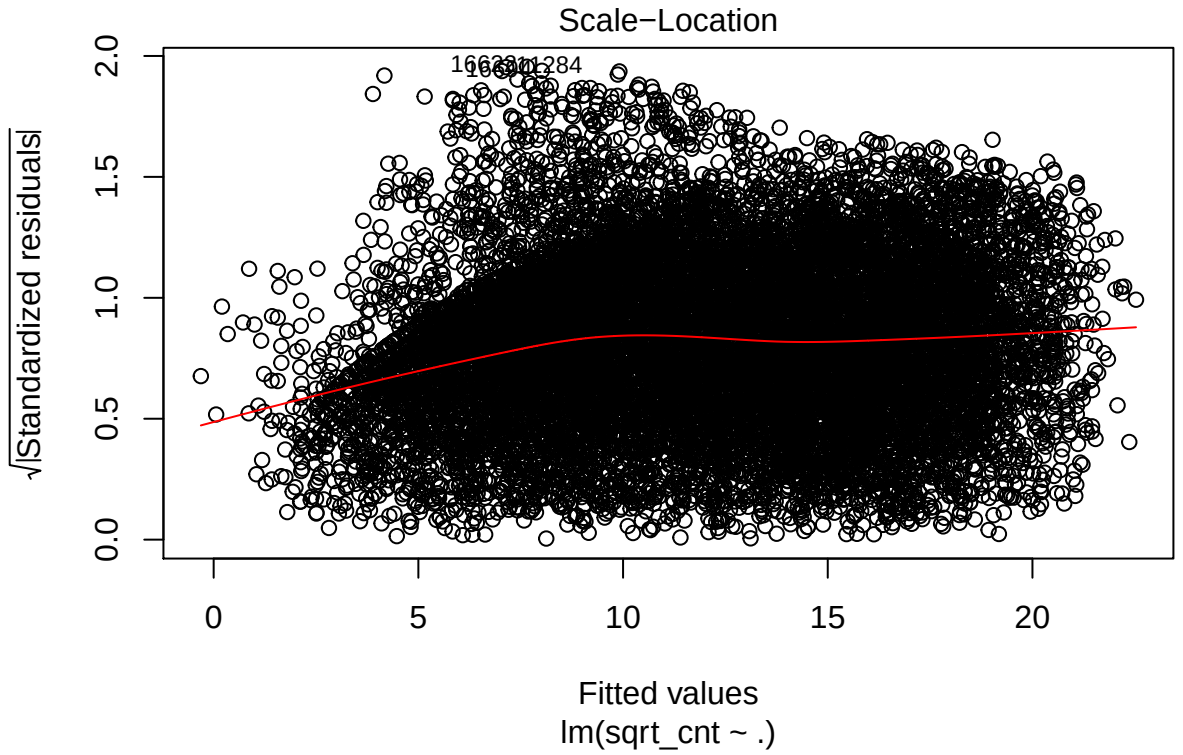
Using K-fold cross validation, we found that the average MSE for all four of the models are really similar, so our choice of model will depend mainly on whether assumptions are met. Ridge and lasso have less model assumptions and less guarantees as well having a more difficult interpretation, so analyzing the models fit on features selected through an exhaustive subset method and on lasso features would be better models

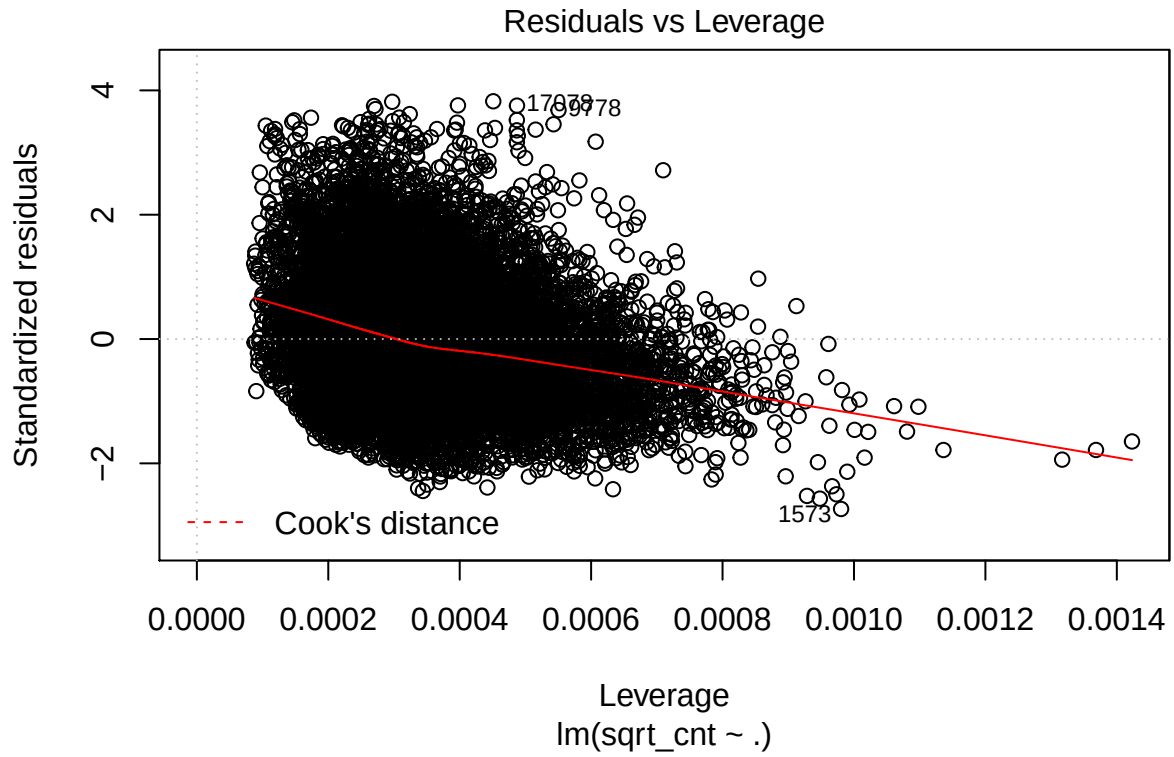
Model Diagnostics

Selection using Adjusted R2

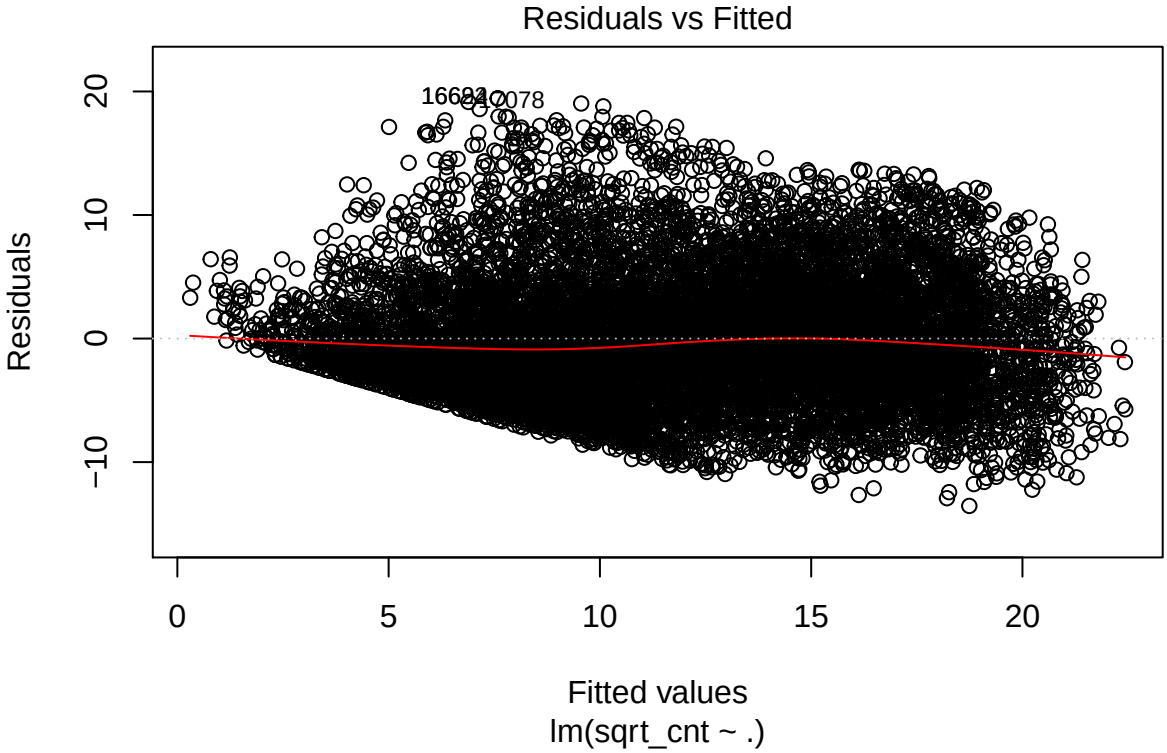


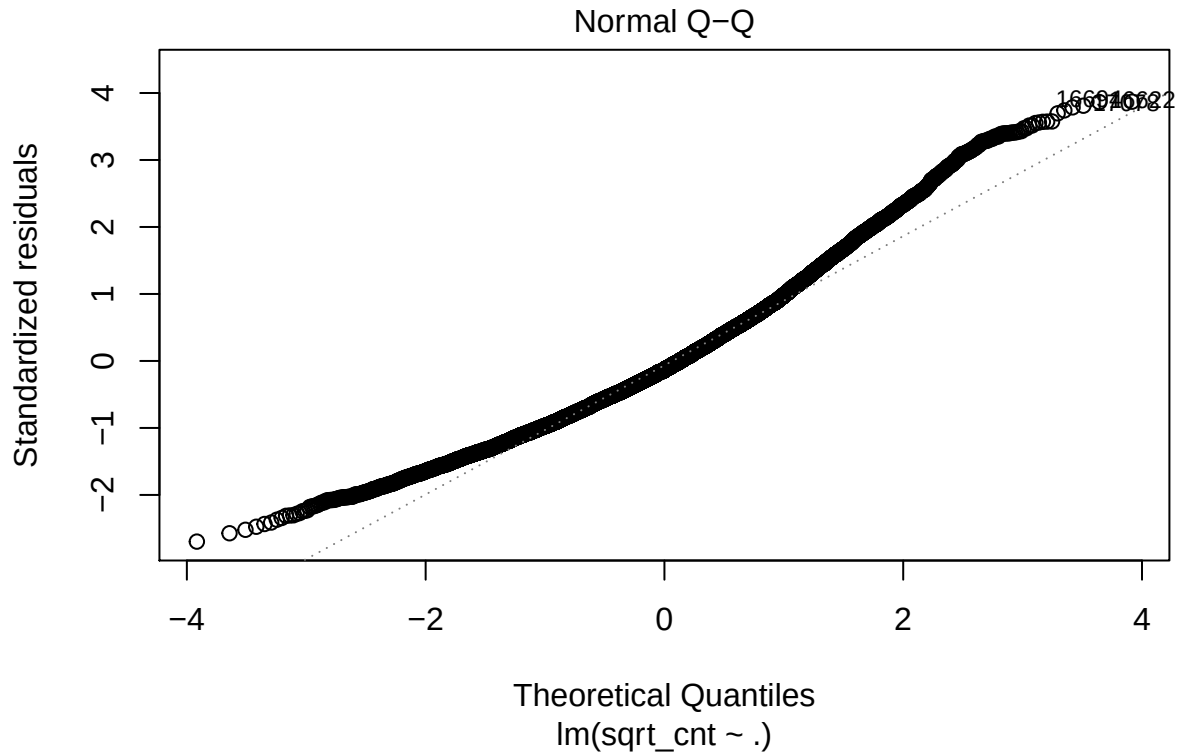

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(sqrt_cnt ~ .)

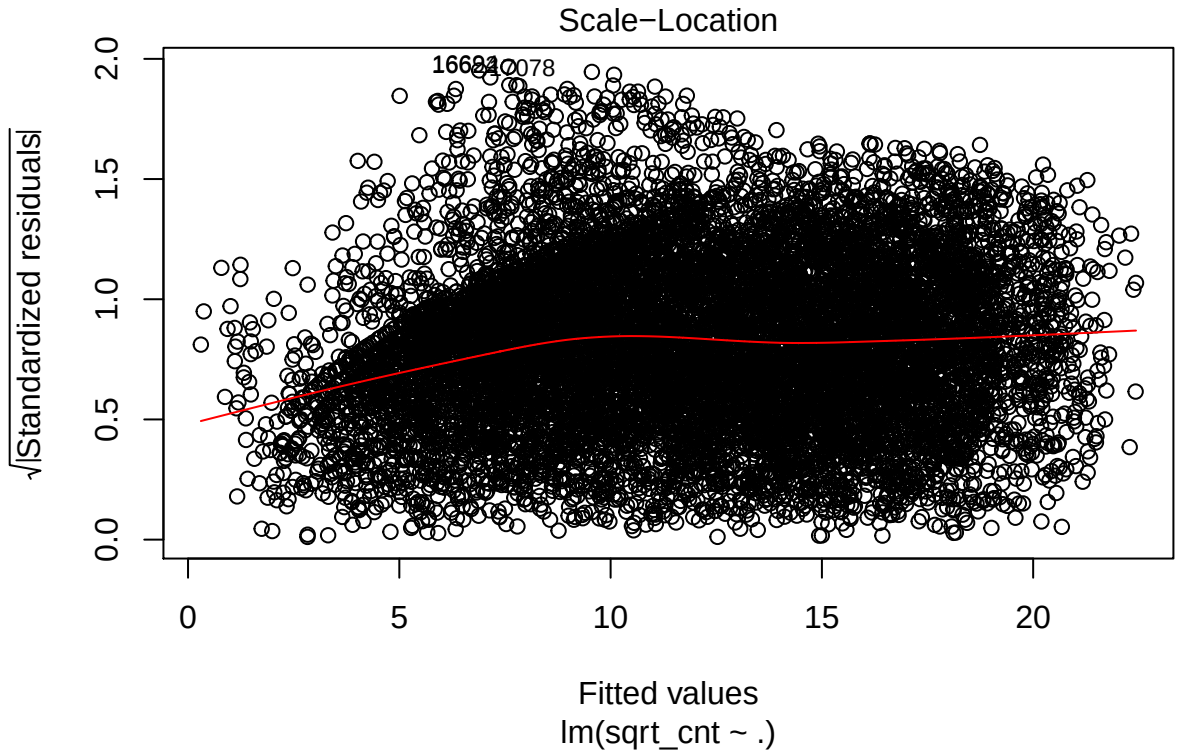

Scale−Location
√|Standardized residuals|
Fitted values
lm(sqrt_cnt ~ .)

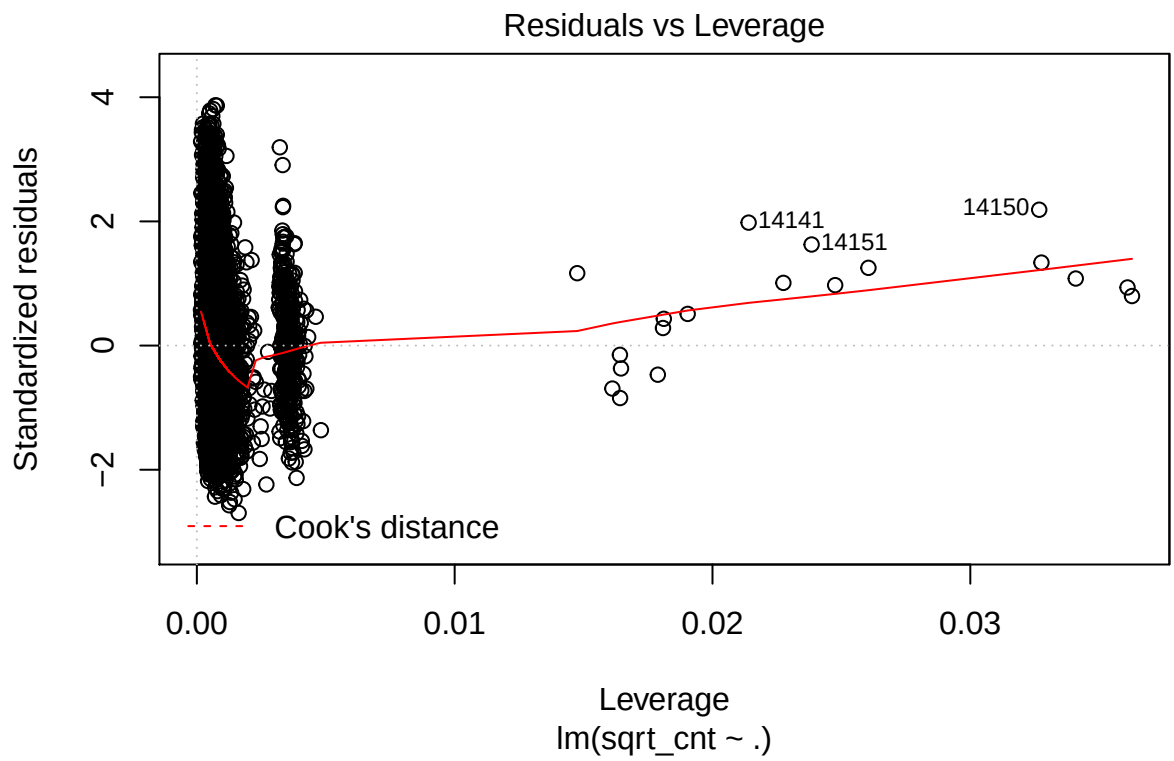


Selection using Lasso Features




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(sqrt_cnt ~ .)


Scale−Location
√|Standardized residuals|
Fitted values
lm(sqrt_cnt ~ .)



We can further compare these two models using diagnostic plots. The first plot shown for each model is plotting the residuals vs the fitted values. This plot for both models are quite similar and don't have any patterns that would suggest that the assumption of linearity is badly broken. The second plot which is a normal QQ plot for the standardized residuals look similar and although the residuals do not follow a normal distribution exactly since there is some inconsistencies with the extreme values, overall the residuals appear to be fairly normal, suggesting that the assumption of normal errors is held . The third plot to consider is the scale-location plot which also seems similar for the two models. These plots allows us to analyze whether the residuals are equally spread to check the assumption of constant variance. There does seem to be a little less variance for smaller fitted values but, overall the variance is fairly constant. The last plot allows us to check for any highly influential points and to analyze the leverage values for the explanatory variables. Both of the plots don't appear to have any highly influential points, but the leverage values for the observations using the features selected using adjusted R2 are much lower. The features used for this model probably have less extremes, which would make it more appealing to use as a predictor since we could possibly expect future datapoints to have low leverage in comparison to our training data. For that reason, choosing the model using features selected from adjusted R2 is our final and best model.

Conclusion

	value
(Intercept)	5.3664428
season	0.7675233
hr	0.3732520
atemp	12.3580910
hum	-8.6937637

The MSE on the test set of the data is: 25.708 which is quite close to the MSE on training models which gives some reassurance that the model did not grossly overfit to the training data.

Overall the model that was chosen as the best in this analysis was an OLS model that was trained using features selected with an exhaustive subset model selection with adjusted R2. The features that were chosen as predictive values for our transformed response variable, the square root of the number of bike rentals for a given an hour, are **season**, **hr**, **atemp**, and **hum**. This model was chosen for it's interpretability over models with similar accuracy due to it being an OLS model with assumptions for the model that appear to be hold, providing it with a better guarantee than using LASSO or Ridge Regression that may have less assumptions, but less guarantees. Using this model, we can interpret it as an increase in 1 unit of **season** is associated with an increase in 0.764 units of the square root of **cnt** if the variables **hr**, **atemp**, and **hum** are held constant, an increase in 1 unit of **hr** is associated with an increase in 0.375 units of the square root of **cnt** if the variables **season**, **atemp**, and **hum** are held constant, an increase in 1 unit of **atemp** is associated with an increase of 12.058 units of the square root of **cnt** if the variables **season**, **hr**, and **hum** are held constant, and an increase in 1 units of **hum** is associated with a decrease in 8.682 units of the square root of **cnt**. If we analyze the p-values for each of these variables, there is a small probability that they do not exist in the true model given a t-test, and the probability that all of these variables do not exist in the true model is also very low given an F-test on the model. However, the adjusted R2, an indicator of fit, is not particularly high, so I would be wary of the magnitude of associated relationships of the variables with the response and due to the nature of the data since it was only collected from one bike sharing system. I would also be extremely wary of using any of these coefficients as causal effects since location, company, and many other things could very important variables in the true model, but were not represented here. The model was also personally impacted by me as well as I chose to only look at models in the model selection phase with a maximum of 5 predictors. An assumption that could not hold very true is independence as well, although the data spanded 2 years, which is quite variable, there could be some underlying conditions that make the observations very dependent on each other that would also make me wary of using this model for causal inference. However, the model does provide some insight into the variables that impact the number of bike rentals in a given hour listed again as **season**, **hr**, **atemp**, and **hum** that could be taken into account if a rough estimate for the number of bike rentals for a given hour is needed. These variables compared to other explanatory variables that were available to be part of the model also appear to be more precise measurements compared to **weathersit** for example that coded many different weather possibilities into just 4 categories or **holiday** that was a simple indicator variable. There are definitely many other approaches to this analysis that may produce different results, but using model diagnostics and cross validation, the model does provide some insight into prediciting the number of bike rentals for an hour given some weather conditions.

Code

```
# Load Packages
library(gpairs)
library(ggplot2)
library(gridExtra)
library(lattice)
library(leaps)
library(glmnet)
library(knitr)
# Load Data
rentals <- read.csv('BikeSharingDataset.csv')
# Summary of response variable
summary(rentals['cnt'])
# Variance of response variable
variance_cnt <- var(rentals[['cnt']])
# Histogram for response variable
ggplot(rentals, aes_string(x = 'cnt')) + geom_histogram()
+ ggtitle('Cnt: Number of Bikes Rented for an Hour')
# Histograms for continous explanatory variables
grid.arrange(ggplot(rentals, aes_string(x = 'temp')) + geom_histogram(),
```

```r
             ggplot(rentals, aes_string(x = 'atemp')) + geom_histogram(),
             ggplot(rentals, aes_string(x = 'hum')) + geom_histogram(),
             ggplot(rentals, aes_string(x = 'windspeed')) + geom_histogram(),
             nrow = 2, ncol = 2)
# Barplots for discrete explanatory variables
grid.arrange(ggplot(rentals, aes_string(x = 'season')) + geom_bar(),
             ggplot(rentals, aes_string(x = 'mnth')) + geom_bar(),
             ggplot(rentals, aes_string(x = 'hr')) + geom_bar(),
             ggplot(rentals, aes_string(x = 'holiday')) + geom_bar(),
             ggplot(rentals, aes_string(x = 'weekday')) + geom_bar(),
             ggplot(rentals, aes_string(x = 'workingday')) + geom_bar(),
             ggplot(rentals, aes_string(x = 'weathersit')) + geom_bar(),
             nrow = 2, ncol = 4)
# Function that plots scatter plots for a continous variable against other
#continous variables in the data and side by side boxplots for comparison
#against discrete variables
plot_scatter <- function(main) {
  grid.arrange(ggplot(rentals, aes_string(x = main, y = 'temp')) +
                 geom_point(size = 0.5, alpha = 0.5),
               ggplot(rentals, aes_string(x = main, y = 'atemp')) +
                 geom_point(size = 0.5, alpha = 0.5),
               ggplot(rentals, aes_string(x = main, y = 'hum')) +
                 geom_point(size = 0.5, alpha = 0.5),
               ggplot(rentals, aes_string(x = main, y = 'windspeed')) +
                 geom_point(size = 0.5, alpha = 0.5),
               ggplot(rentals, aes_string(x = main, y = 'cnt')) +
                 geom_point(size = 0.5, alpha = 0.5),
               ggplot(rentals, aes_string(x = 'season', y = main, group = 'season'))
               + geom_boxplot(),
               ggplot(rentals, aes_string(x = 'mnth', y = main, group = 'mnth')) +
                 geom_boxplot(),
               ggplot(rentals, aes_string(x = 'hr', y = main, group = 'hr'))
               + geom_boxplot(),
               ggplot(rentals, aes_string(x = 'holiday', y = main, group = 'holiday')) +
                 geom_boxplot(),
               ggplot(rentals, aes_string(x = 'weekday', y = main, group = 'weekday')) +
                 geom_boxplot(),
               ggplot(rentals, aes_string(x = 'workingday', y = main, group = 'workingday')) +
                 geom_boxplot(),
               ggplot(rentals, aes_string(x = 'weathersit', y = main, group = 'weathersit')) +
                 geom_boxplot(),
               ncol = 5, nrow = 3, top = paste('Scatterplots and Boxplots for', main))
}
# Using the function above to create exploratory plots
plot_scatter('temp')
plot_scatter('atemp')
plot_scatter('hum')
plot_scatter('windspeed')
# create correlation matrix for variables and print as html table
explanatory <- c("season", "mnth", "hr", "holiday", "weekday",
                 "workingday","weathersit" ,"temp", "atemp", "hum", "windspeed", "cnt" )
cor_vec <- c()
for (i in explanatory) {
```

```r
  for (j in explanatory) {
    cor_vec <- c(cor_vec, cor(rentals[[i]], rentals[[j]]))
  }
}
cor_mat <- matrix(cor_vec, byrow = TRUE, ncol = length(explanatory))
cor_df <- data.frame(cor_mat, row.names = explanatory)
names(cor_df) <- explanatory
kable(cor_df)
# transfrom response variable and add to table, then create table comparing previous
# correlation with response variable to transformed variable
rentals['sqrt_cnt'] <- sqrt(rentals[['cnt']])
log_comp_cor <- c()
for (i in explanatory[-12]) {
  for (j in c('cnt', 'sqrt_cnt')) {
    log_comp_cor <- c(log_comp_cor, cor(rentals[[i]], rentals[[j]]))
  }
}
log_comp_mat <- matrix(log_comp_cor, byrow = TRUE, ncol = 2)
log_comp_df <- data.frame(log_comp_mat, row.names = explanatory[-12])
names(log_comp_df) <- c('cnt', 'sqrt_cnt')
kable(log_comp_df)
# setting seed for reproducibility and splitting data into train and test
set.seed(333)
explanatory_transf <- c(explanatory[-12], 'sqrt_cnt')
rentals_explanatory <- rentals[, explanatory_transf]
rentals_explanatory_shuffled <- rentals_explanatory
/[sample(1:nrow(rentals_explanatory),/
           size = nrow(rentals_explanatory), replace = FALSE),  ]
train_num <- round(nrow(rentals_explanatory) * 0.8)
rentals_explanatory_train <- rentals_explanatory_shuffled[1:train_num, ]
rentals_explanatory_test <- rentals_explanatory_shuffled
/[(train_num + 1):nrow(rentals_explanatory_shuffled), ]
# exhaustive subset feature selection using adjusted R2
rentals_explanatory_wo_atemp <- rentals_explanatory_train[, explanatory_transf[-9]]
rentals_explanatory_wo_temp <- rentals_explanatory_train[, explanatory_transf[-8]]
outs_atemp <- regsubsets(sqrt_cnt ~ ., data = rentals_explanatory_wo_atemp,
                         nbest = 5, nvmax = 4, method = 'exhaustive')
outs_atemp_summary <- summary(outs_atemp)
adjr2_exh_atemp <- outs_atemp_summary$which[which.max(outs_atemp_summary$adjr2),]
outs_temp <- regsubsets(sqrt_cnt ~ ., data = rentals_explanatory_wo_temp,
                        nbest = 5, nvmax = 4, method = 'exhaustive')
outs_temp_summary <- summary(outs_temp)
adjr2_exh_temp <- outs_temp_summary$which[which.max(outs_temp_summary$adjr2),]
# creation of variables chosen
mod_sel_df_coef <- data.frame(names(adjr2_exh_temp), stringsAsFactors = FALSE)
names(mod_sel_df_coef) <- c('variable')
mod_sel_df_coef['in'] <-  adjr2_exh_temp
mod_sel_df_coef[mod_sel_df_coef[['in']] == TRUE, ]['variable']
# Using CV to find best lambda value for ridge regression for this data
lambda <- c(seq(0.0001, 1, length = 100), seq (1, 50, length = 100))
explan.mat <- scale(as.matrix(rentals_explanatory_train[, explanatory[-12]]))
ridge.fit <- glmnet(explan.mat, rentals_explanatory_train[['sqrt_cnt']],
                    alpha = 0, lamdas = lamda)
```

```r
cv.fit.ridge <- cv.glmnet(explan.mat, rentals_explanatory_train[['sqrt_cnt']],
                          alpha = 0, lambdas = lambda)
optimal.ridge.lambda <- cv.fit.ridge$lambda.min
# Using CV to find best lambda value for lasso regression for this data
lasso.fit <- glmnet(explan.mat, rentals_explanatory_train[['sqrt_cnt']],
                    alpha = 1, lamdas = lamda)
cv.fit.lasso <- cv.glmnet(explan.mat, rentals_explanatory_train[['sqrt_cnt']],
                          alpha = 1, lambdas = lambda)
optimal.lasso.lambda <- cv.fit.lasso$lambda.min
# Creating coefficient table
mod.ridge <- glmnet(explan.mat, rentals_explanatory_train[['sqrt_cnt']], alpha = 0, lambdas = optimal.ri
ridge_coef_smat <- coef(mod.ridge, s = optimal.ridge.lambda)
ridge_rn <- rownames(ridge_coef_smat)
ridge_coef_df <- data.frame(matrix(ridge_coef_smat))
names(ridge_coef_df) <- c('coefficient')
ridge_coef_df['variable'] <- ridge_rn
ridge_coef_df[, c('variable', 'coefficient')]
# Using optimal lambda value for lasso regression above, perform feature selection
mod.lasso <- glmnet(explan.mat, rentals_explanatory_train[['sqrt_cnt']],
                    alpha = 1, lambdas = optimal.lasso.lambda)
lasso_coef_smat <- coef(mod.lasso, s = optimal.lasso.lambda)
lasso_rn <- rownames(lasso_coef_smat)
lasso_coef_df <- data.frame(matrix(lasso_coef_smat))
names(lasso_coef_df) <- c('coefficient')
lasso_coef_df['names'] <- lasso_rn
lasso_featsel <- lasso_coef_df[lasso_coef_df$coefficient != 0,][['names']][-1]
lasso_featsel_rep <- c(lasso_featsel, 'sqrt_cnt')
# define mse function
mse <- function(predictions, actual) {
  return(mean((predictions - actual) ^ 2))
}
# run k fold CV on the four models, calculate MSE,
#average MSE across folds and show in a table
k1 <- 1:2780
k2 <- 2781:(2 * 2780)
k3 <- (2 * 2780 + 1): (3 * 2780)
k4 <- (3 * 2780 + 1): (4 * 2780)
k5 <- (4 * 2780 + 1):nrow(rentals_explanatory_train)
all <- 1:nrow(rentals_explanatory_train)
list_of_ks <- list()
list_of_ks[[1]] <- k1
list_of_ks[[2]] <- k2
list_of_ks[[3]] <- k3
list_of_ks[[4]] <- k4
list_of_ks[[5]] <- k5
mses <- c()
for (i in 1:5) {
  val <- list_of_ks[[i]]
  train <- all[-val]
  train_df <- rentals_explanatory_wo_temp[train, ]
  test_df <- rentals_explanatory_wo_temp[val, ]
  train_df2 <- rentals_explanatory_train[train, ]
  test_df2 <- rentals_explanatory_train[val, ]
```

```r
  train_mat <- explan.mat[train, ]
  test_mat <- explan.mat[val, ]
  mod.selec <- lm(sqrt_cnt ~ ., train_df[, adjr2_exh_temp[-1]])
  mses <- c(mses, mse(predict(mod.selec, test_df[, adjr2_exh_temp[-1]]),
                      test_df[['sqrt_cnt']]))
  mod.rid <- glmnet(train_mat, train_df[['sqrt_cnt']], alpha = 0,
                    lambdas = optimal.ridge.lambda)
  mses <- c(mses, mse(predict(mod.rid, test_mat, s = optimal.ridge.lambda,
                              type = 'response'),  test_df[['sqrt_cnt']]))
  mod.lass <- glmnet(train_mat, train_df[['sqrt_cnt']], alpha = 1,
                     lambdas = optimal.lasso.lambda)
  mses <- c(mses, mse(predict(mod.lass, test_mat, s = optimal.lasso.lambda, type = 'response'),
                      test_df[['sqrt_cnt']]))
  mod.lassfeat <- lm(sqrt_cnt ~ ., train_df2[, lasso_featsel_rep])
  mses <- c(mses, mse(predict(mod.lassfeat, test_df2[, lasso_featsel_rep]),
                      test_df[['sqrt_cnt']]))
}
mat_mse <- matrix(c(mses, NA, NA, NA, NA), ncol = 4, byrow = TRUE)
mat_mse[6, ] <- c(mean(mat_mse[, 1], na.rm = TRUE),
                  mean(mat_mse[, 2], na.rm = TRUE), mean(mat_mse[, 3], na.rm = TRUE),
                  mean(mat_mse[, 4], na.rm = TRUE))
mse_df <- data.frame(mat_mse,
                     row.names = c('k1', 'k2', 'k3', 'k4', 'k5', 'average MSE across ks'))
names(mse_df) <- c('selection', 'ridge', 'lasso', 'lasso feature')
kable(mse_df)
# Fitting model on all training data for features selected using adjusted r2
#and plotting diagnostic plots for assumptions
mod.selection <- lm(sqrt_cnt ~ .,
                    rentals_explanatory_wo_temp[, adjr2_exh_temp[-1]])
mse.selection <- mse(predict(mod.selection,
                             rentals_explanatory_wo_temp[, adjr2_exh_temp[-1]]),
                     rentals_explanatory_wo_temp[['sqrt_cnt']])
plot(mod.selection)
# Fitting model on all training data for features selected
#using lasso and plotting diagnostic plots for assumptions
mod.lassofeat <- lm(sqrt_cnt ~ ., data = rentals_explanatory_train[, lasso_featsel_rep])
plot(mod.lassfeat)
# final coefficient table
final_mod_coef <- data.frame(mod.selection$coefficients)
names(final_mod_coef) <- c('value')
kable(final_mod_coef)
# Summary of model for analysis of F and T tests
sum_of_fin_mod <- summary(mod.selection)
# calculation of final model test mse
mse_final <- mse(predict(mod.selection, rentals_explanatory_test),
                 rentals_explanatory_test$sqrt_cnt)
```