R Notebook

Final Project

variable selection (stepwise)

```
step(lm(casual~windspeed+hum+atemp+temp+weathersit+workingday+weekday+holiday
+mnth+season, training),direction="both",k=log(365))
```

```
## Call:
## lm(formula = casual ~ windspeed + temp + weathersit + workingday,
##     data = training)
##
## Coefficients:
## (Intercept)    windspeed         temp   weathersit   workingday
##       653.0       -727.5       1705.0       -151.3       -658.3
```

```
# final model given by stepwise with the BIC is that:

casual ~ windspeed + temp + weathersit + workingday
```

```
step(lm(casual~windspeed+hum+atemp+temp+weathersit+workingday+weekday+holiday
+mnth+season, training),direction="both",k=2)
```

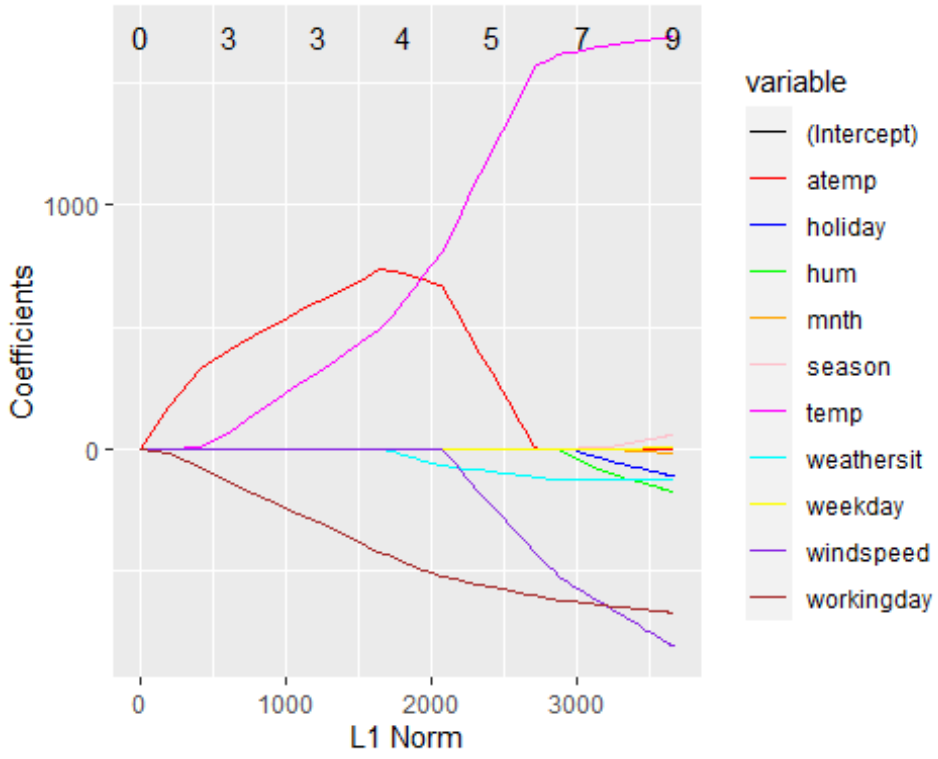
```
## Call:
## lm(formula = casual ~ windspeed + temp + weathersit + workingday +
##     mnth + season, data = training)
##
## Coefficients:
## (Intercept)    windspeed         temp   weathersit   workingday         mn
th
##      649.26      -730.66      1670.49      -155.52      -658.82       -20.
33
##      season
##       64.07
```

```
# final model given by stepwise with the Aic is that:

casual ~ windspeed + temp + weathersit + workingday + mnth + season
```

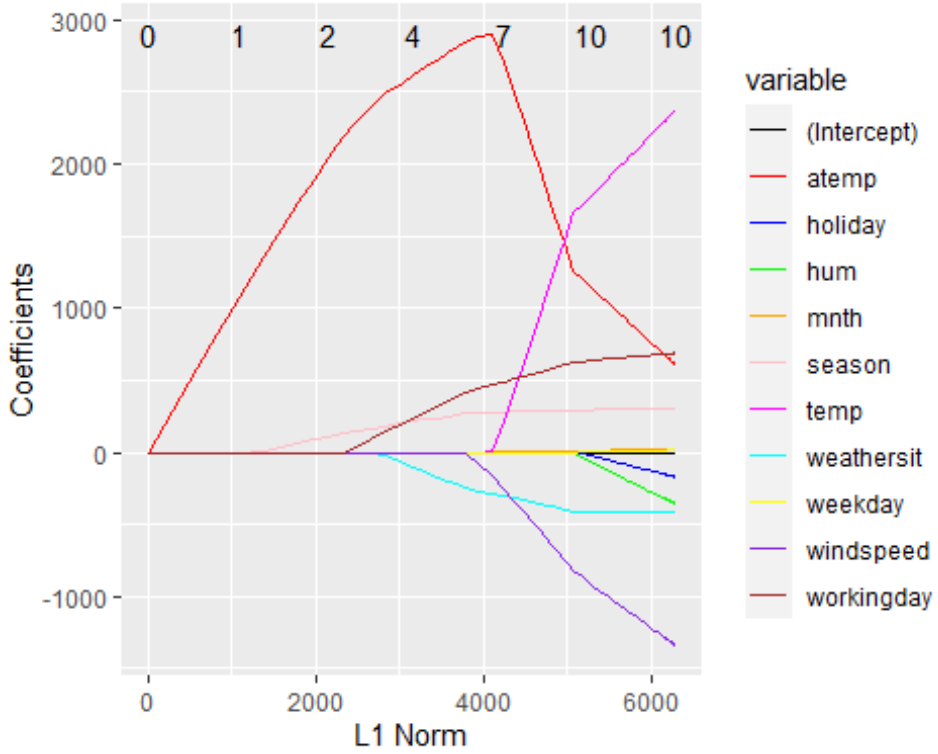
```
step(lm(registered~windspeed+hum+atemp+temp+weathersit+workingday+weekday+hol
iday+mnth+season, training),direction="both",k=log(365))
```

```
## Call:
## lm(formula = registered ~ windspeed + temp + weathersit + workingday +
##     season, data = training)
##
## Coefficients:
## (Intercept)    windspeed         temp   weathersit   workingday       seas
on
##       885.4      -1235.5       2866.8       -478.1        714.3        350
.1

# final model given by stepwise with the bic is that:

registered ~ windspeed + temp + weathersit + workingday + season

step(lm(registered~windspeed+hum+atemp+temp+weathersit+workingday+weekday+hol
iday+mnth+season, training),direction="both",k=2)

## Call:
## lm(formula = registered ~ windspeed + temp + weathersit + workingday +
##     weekday + season, data = training)
##
## Coefficients:
## (Intercept)    windspeed         temp   weathersit   workingday      weekd
ay
##      829.26     -1267.29      2874.93      -480.76       712.71        21.
44
##      season
##      349.58

# final model given by stepwise with the aic is that:

registered ~ windspeed + temp + weathersit + workingday + weekday + season
```

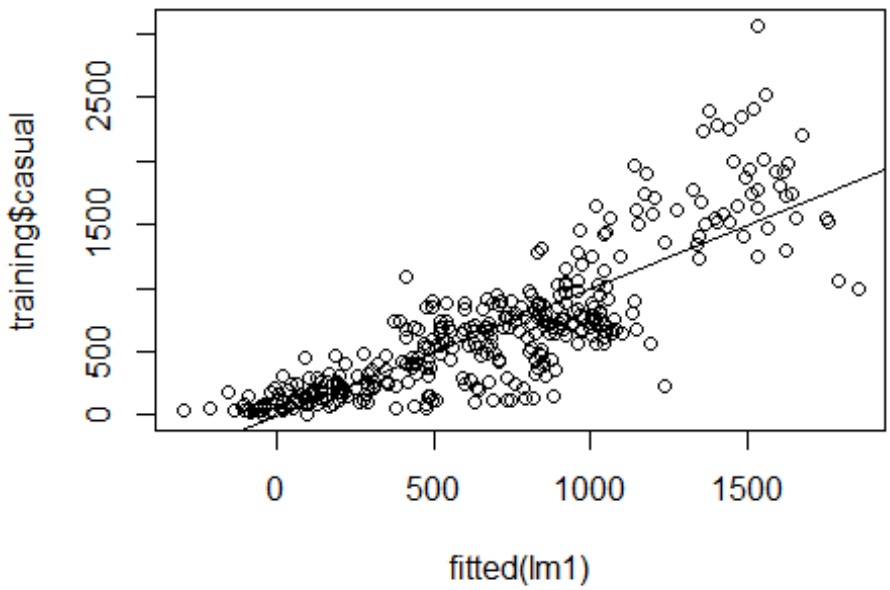
variable selection (lasso)

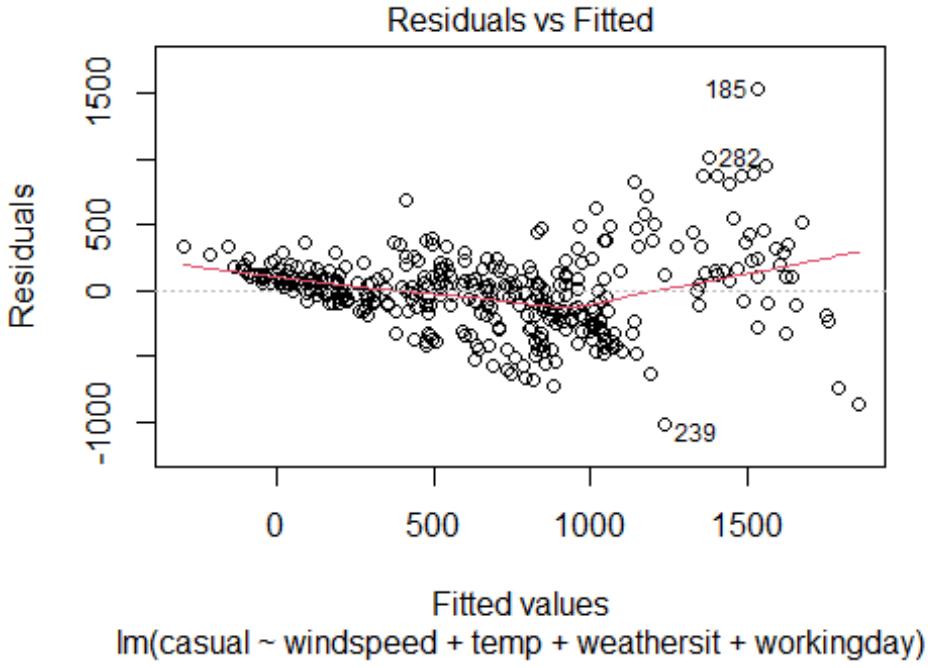


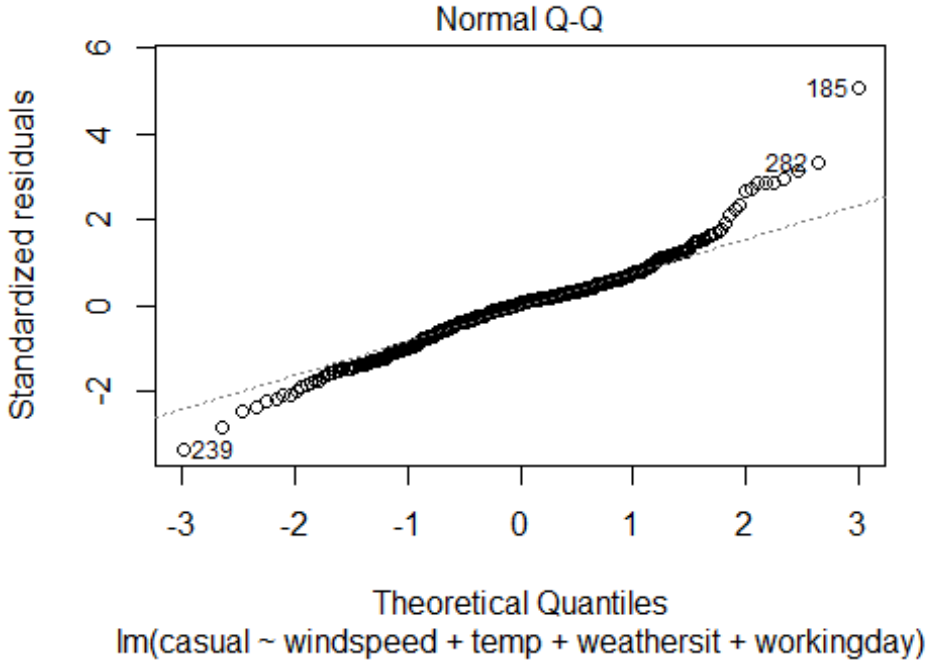

```
# lasso result of casual
```

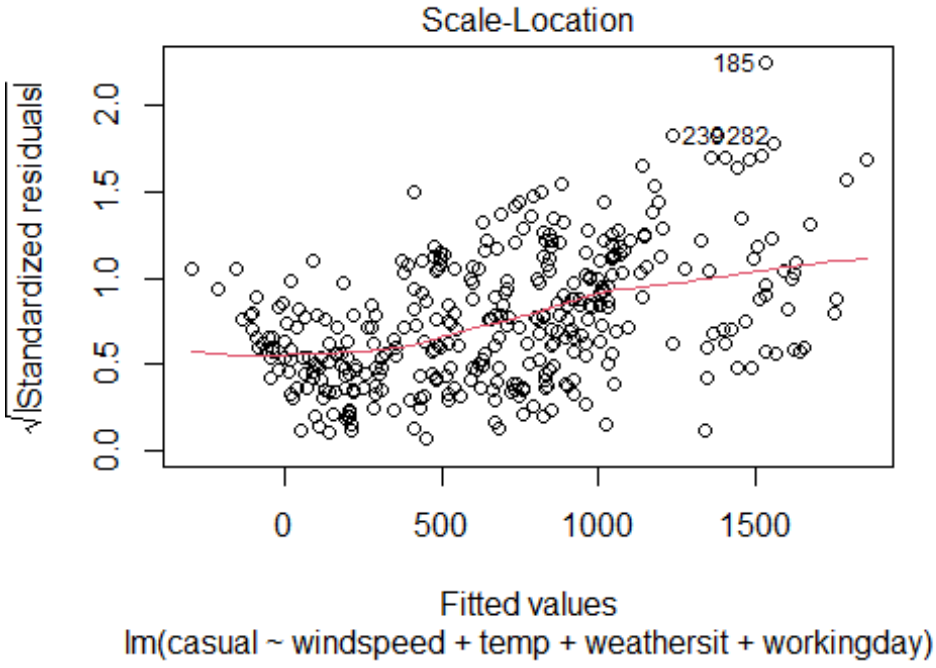


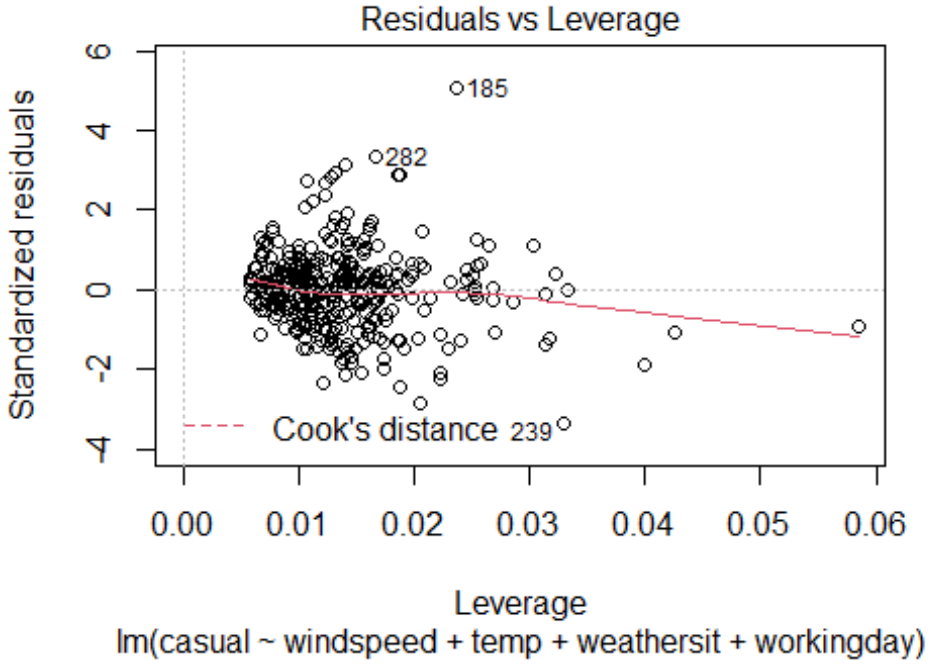
```r
# lasso result of registered
lm1 <- lm(casual ~ windspeed + temp + weathersit + workingday,training)

# model from stepwise with the bic


lm2 <- lm(casual ~ windspeed + temp + weathersit + workingday + mnth + season
,training)

# model from stepwise with the aic
```

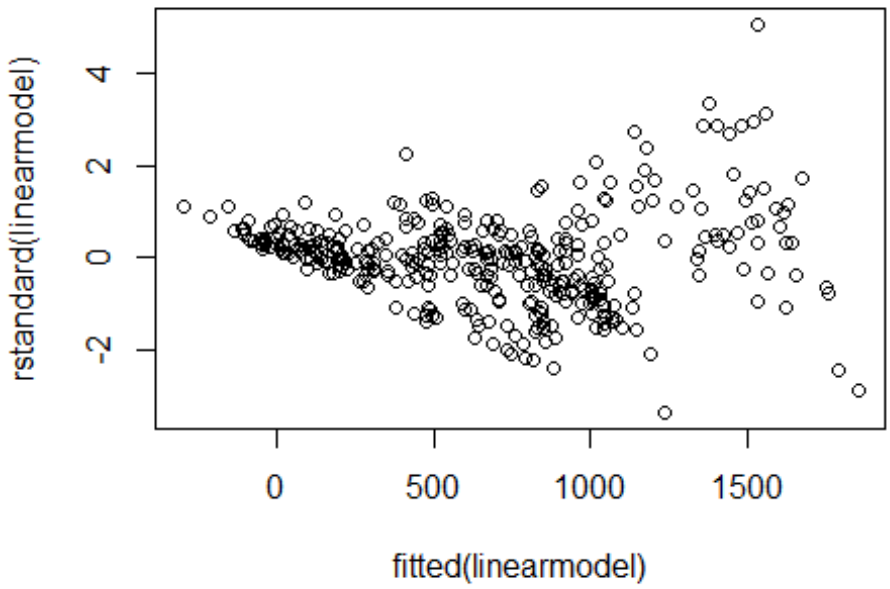


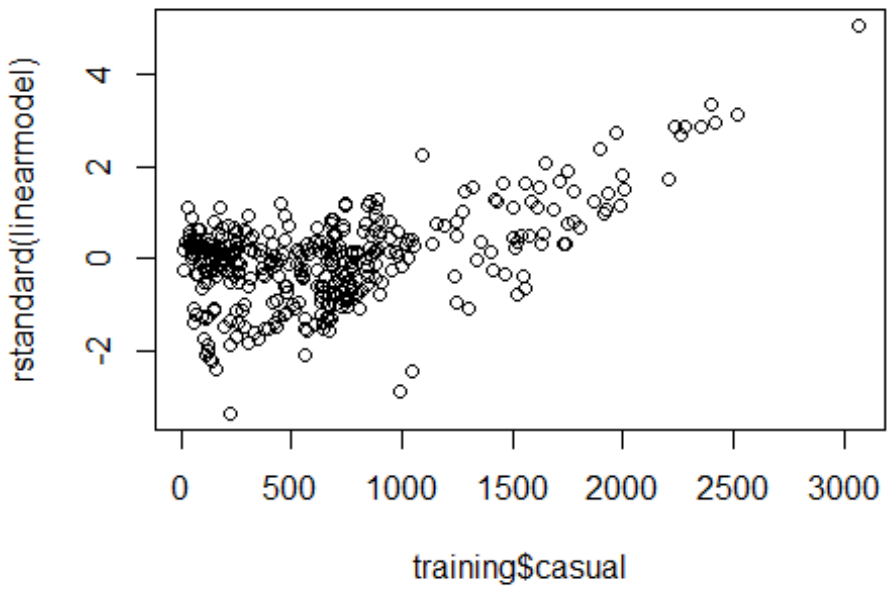

Residuals vs Fitted
Residuals
Fitted values
lm(casual ~ windspeed + temp + weathersit + workingday)
185
282
239
1500
500
0
-1000
0
500
1000
1500

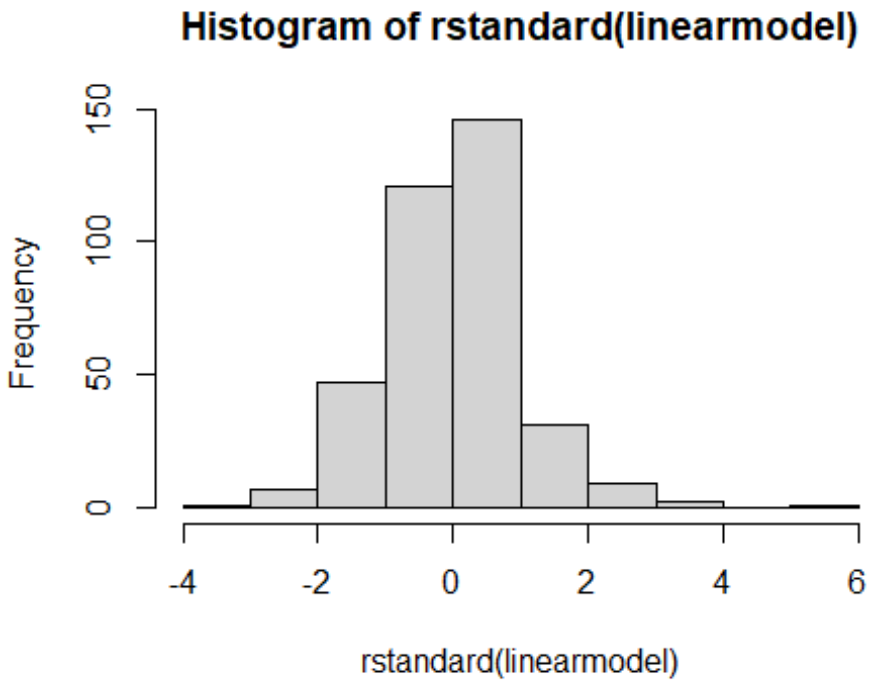

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(casual ~ windspeed + temp + weathersit + workingday)
185
282
239
6
4
2
0
-2
-3
-2
-1
0
1
2
3


Scale-Location
185
239
282
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
500
1000
1500
Fitted values
lm(casual ~ windspeed + temp + weathersit + workingday)

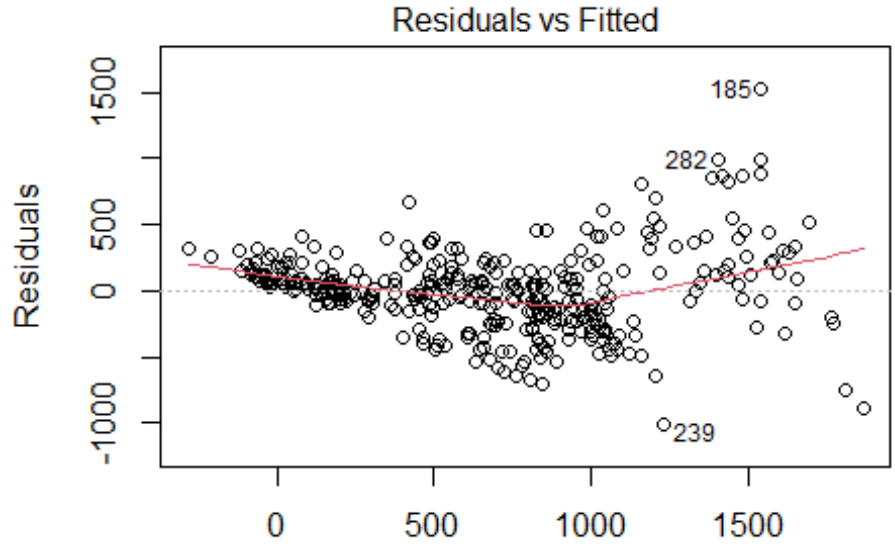

Residuals vs Leverage
185
282
Cook's distance
239
Standardized residuals
-4
-2
0
2
4
6
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Leverage
lm(casual ~ windspeed + temp + weathersit + workingday)

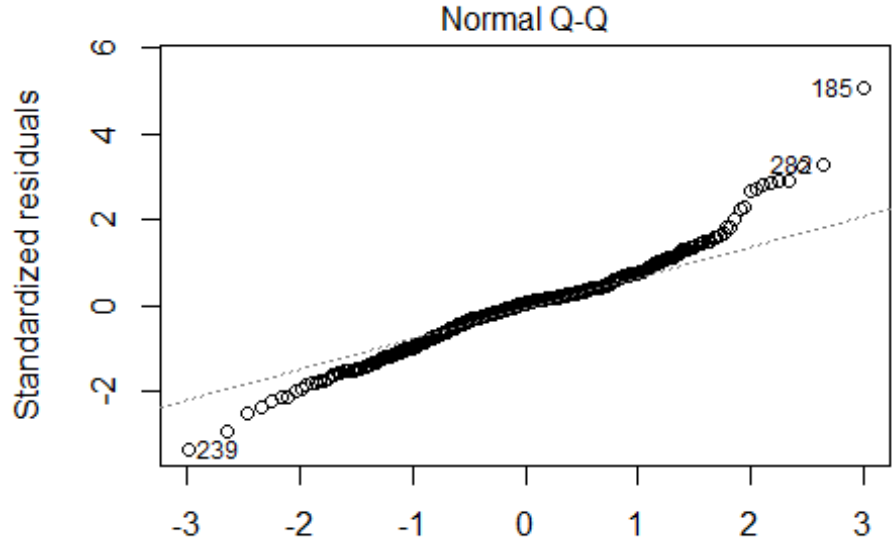

rstandard(linearmodel)
4
2
0
-2
0
500
1000
1500
fitted(linearmodel)

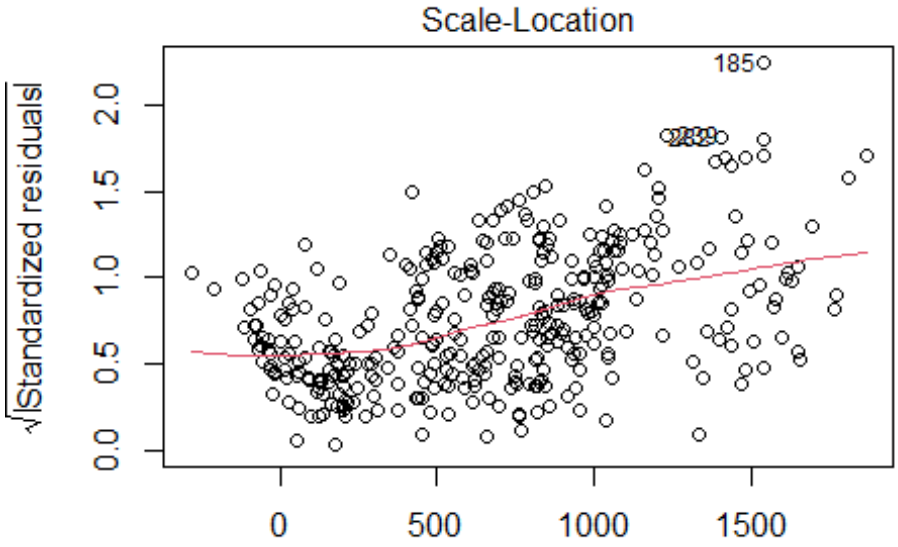

rstandard(linearmodel)
4
2
0
-2
0
500
1000
1500
2000
2500
3000
training$casual

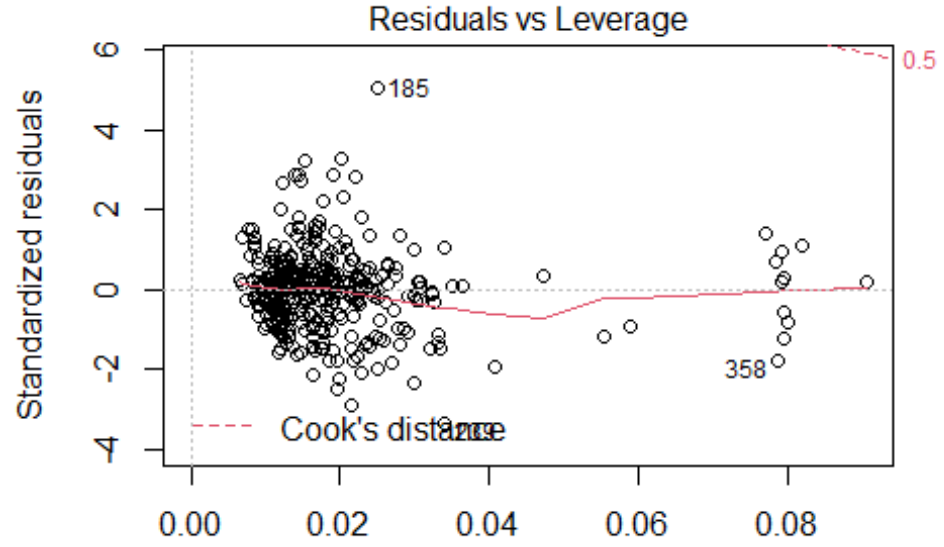


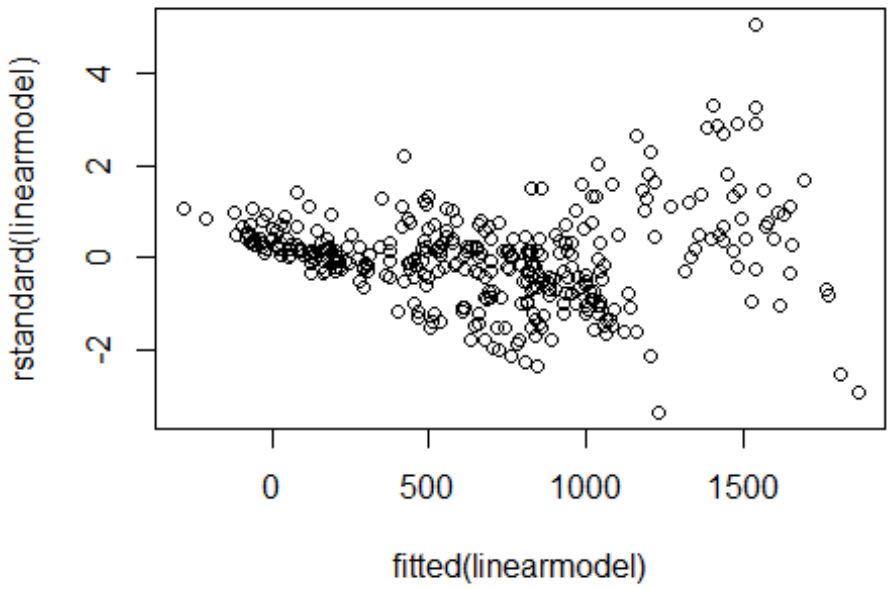

```
# plots and pictures from linear model 1
```

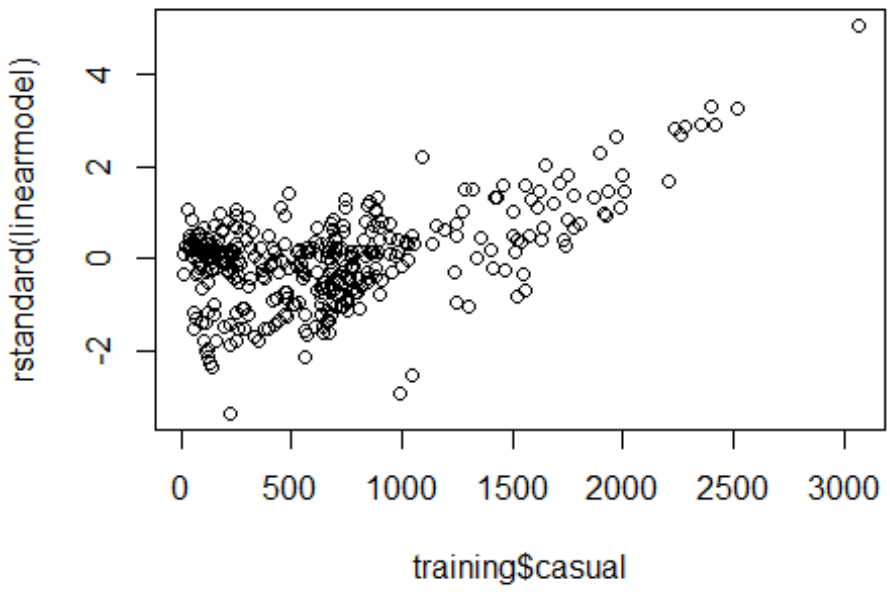

Residuals vs Fitted
Residuals
Fitted values
lm(casual ~ windspeed + temp + weathersit + workingday + mnth + se
185
282
239
1500
500
0
-1000
0
500
1000
1500

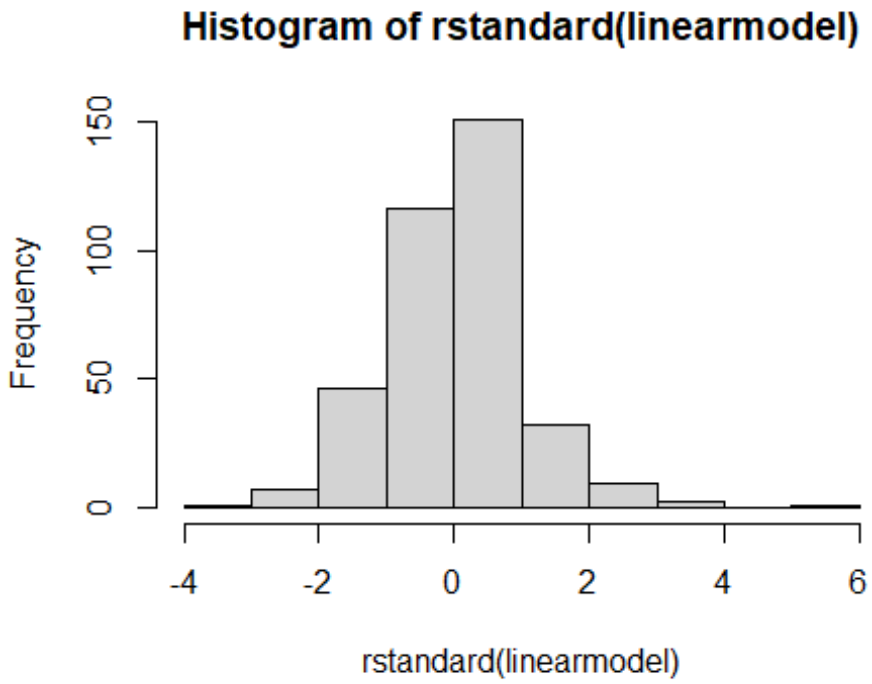

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(casual ~ windspeed + temp + weathersit + workingday + mnth + se
185
282
239
6
4
2
0
-2
-3
-2
-1
0
1
2
3


Scale-Location
185
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
500
1000
1500
Fitted values
lm(casual ~ windspeed + temp + weathersit + workingday + mnth + se


Residuals vs Leverage
185
358
0.5
Standardized residuals
6
4
2
0
-2
-4
Cook's distance
0.00
0.02
0.04
0.06
0.08
Leverage
lm(casual ~ windspeed + temp + weathersit + workingday + mnth + se

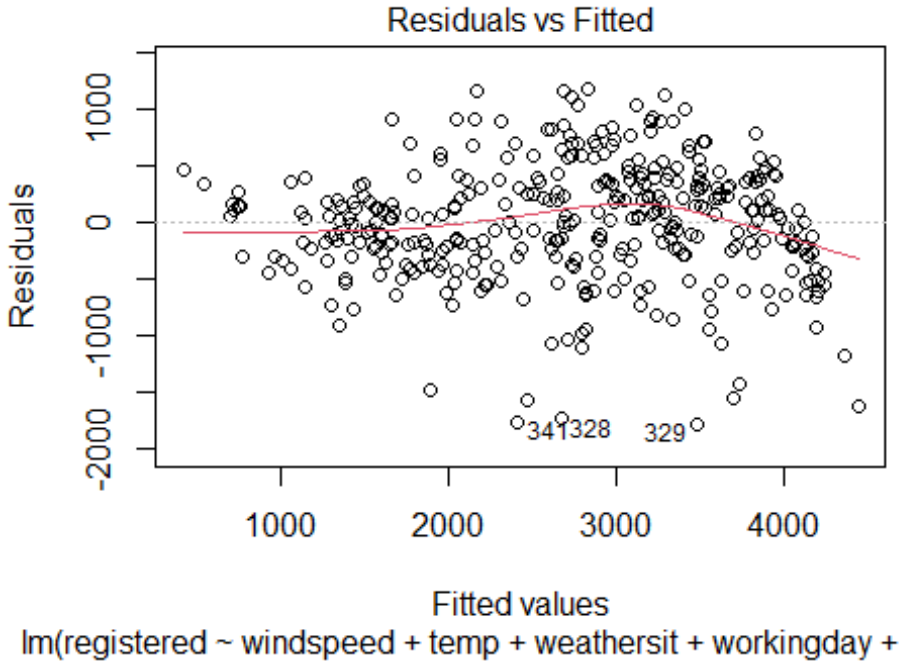

rstandard(linearmodel)
-2
0
2
4
0
500
1000
1500
fitted(linearmodel)

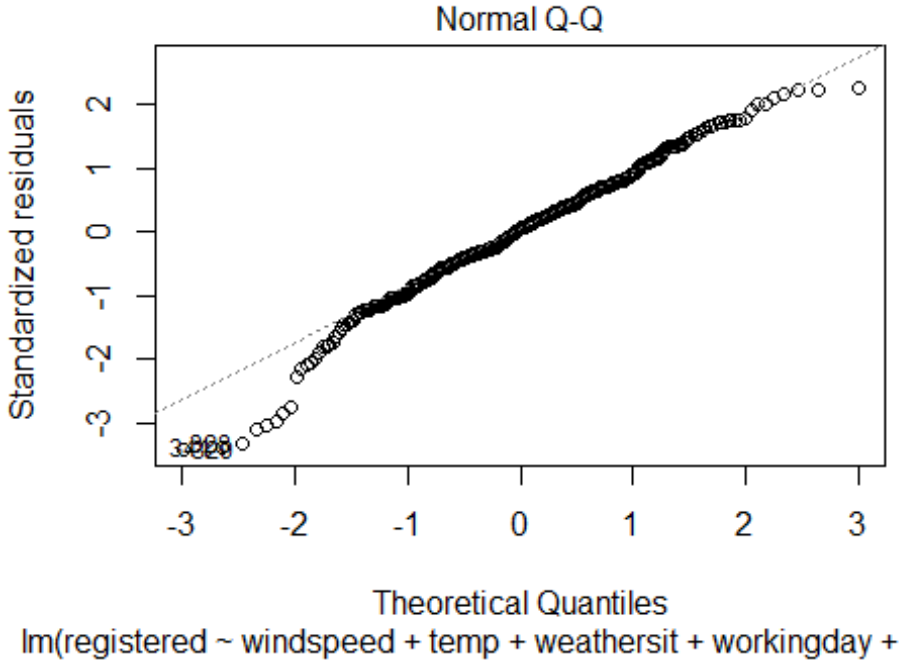

rstandard(linearmodel)
-2
0
2
4
0
500
1000
1500
2000
2500
3000
training$casual

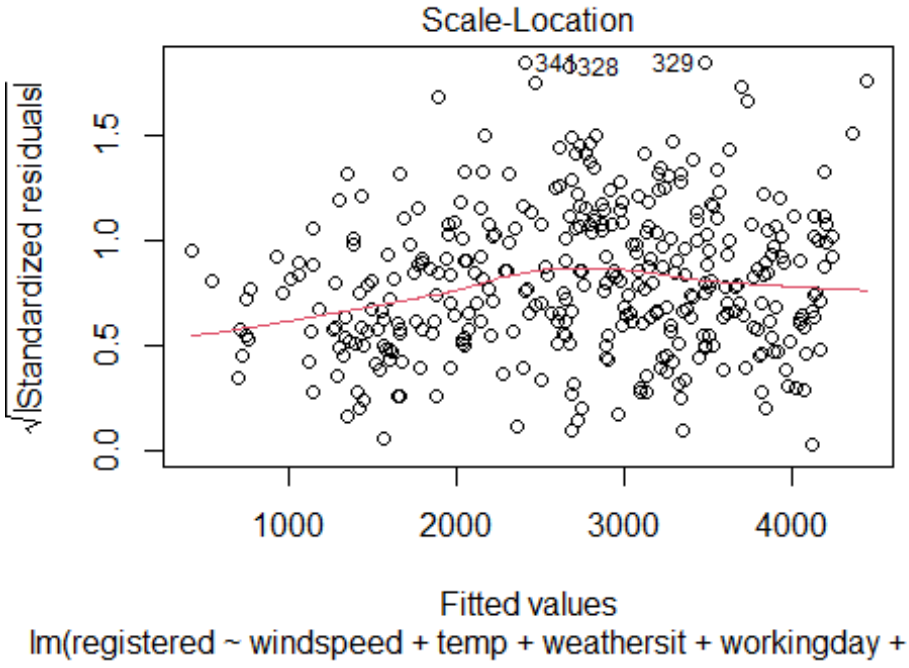


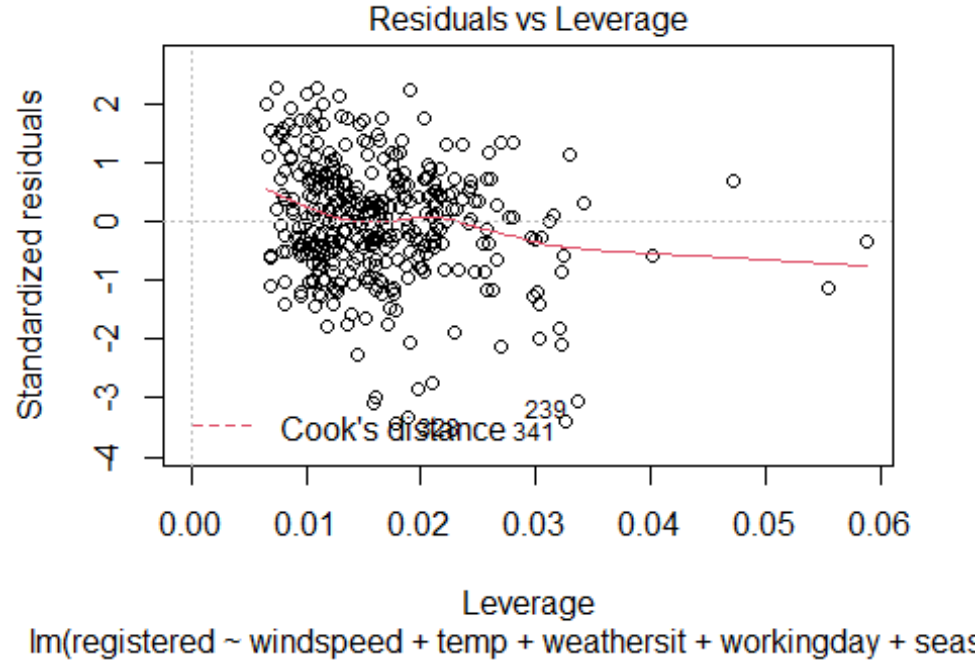

```
# plots and pictures from linear model 2
```

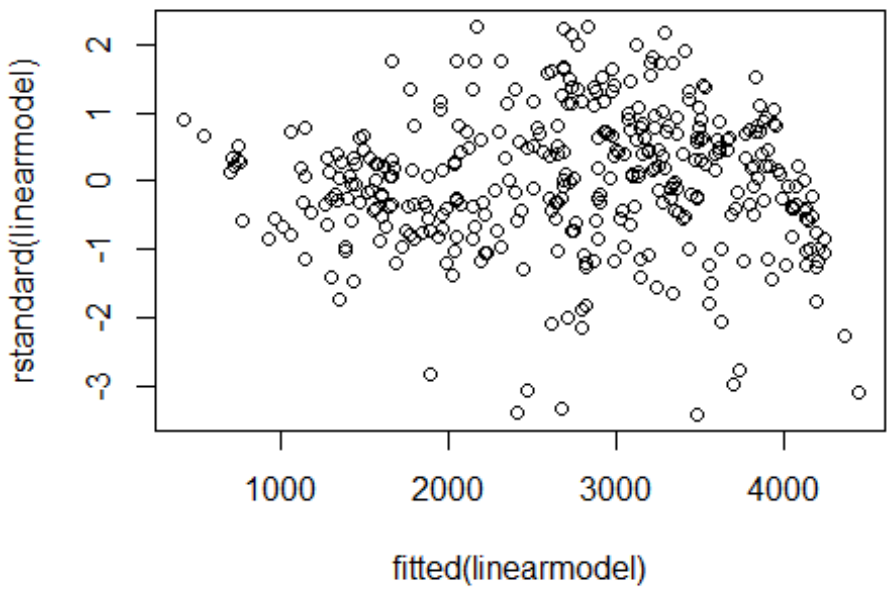
```
lm3 <- lm(registered ~ windspeed + temp + weathersit + workingday + season,tr
aining)

# model from stepwise with the bic

lm4 <- lm(registered ~ windspeed + temp + weathersit + workingday + weekday +
season,training)
# model from stepwise with the bic
```

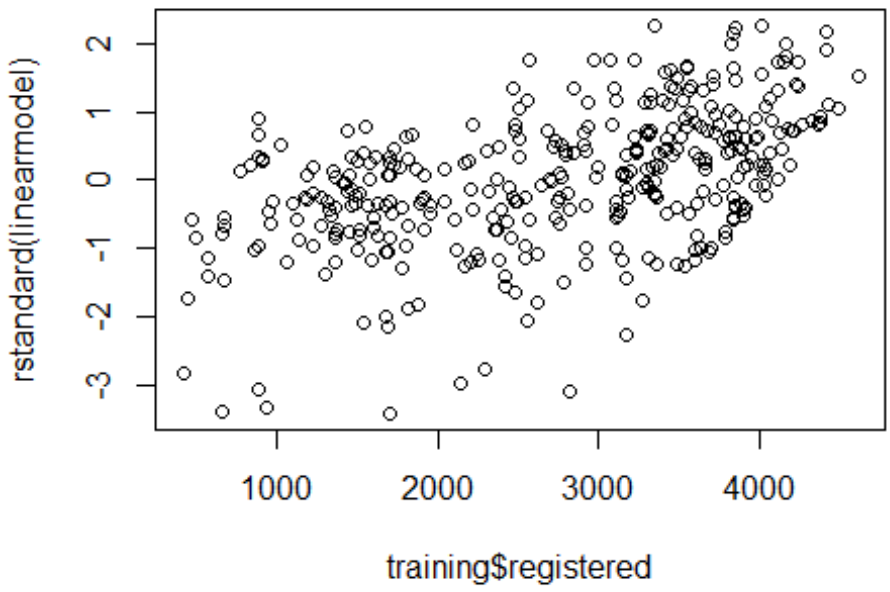

Residuals vs Fitted
Residuals
1000
0
-1000
-2000
341
328
329
1000
2000
3000
4000
Fitted values
lm(registered ~ windspeed + temp + weathersit + workingday + seas

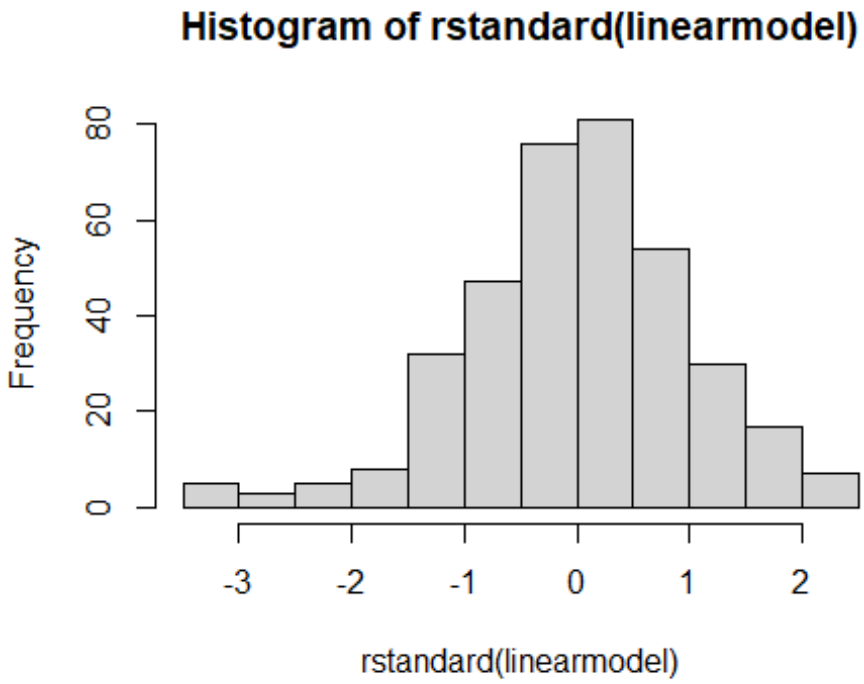

Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(registered ~ windspeed + temp + weathersit + workingday + seas


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1000
2000
3000
4000
Fitted values
lm(registered ~ windspeed + temp + weathersit + workingday + seas

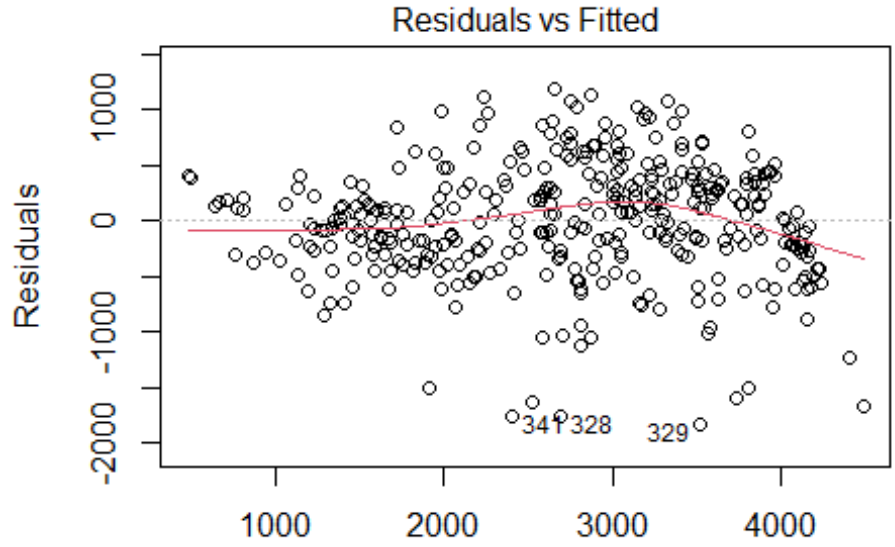

Residuals vs Leverage
Standardized residuals
-4
-3
-2
-1
0
1
2
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Leverage
lm(registered ~ windspeed + temp + weathersit + workingday + seas

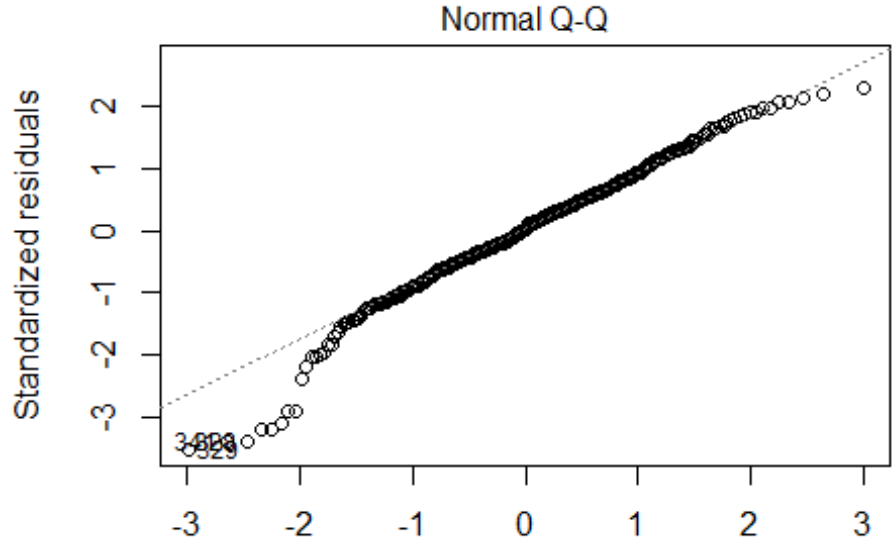

rstandard(linearmodel)
-3
-2
-1
0
1
2
1000
2000
3000
4000
fitted(linearmodel)

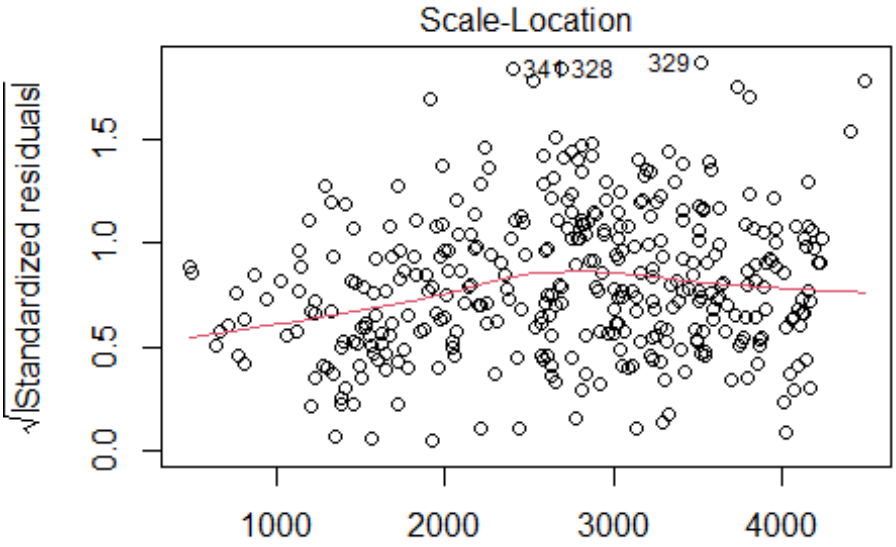

rstandard(linearmodel)
-3
-2
-1
0
1
2
1000
2000
3000
4000
training$registered

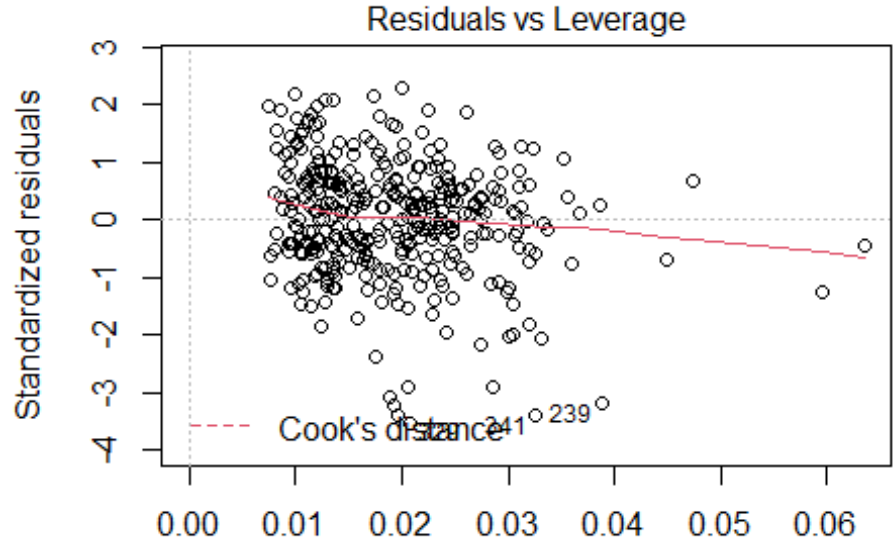


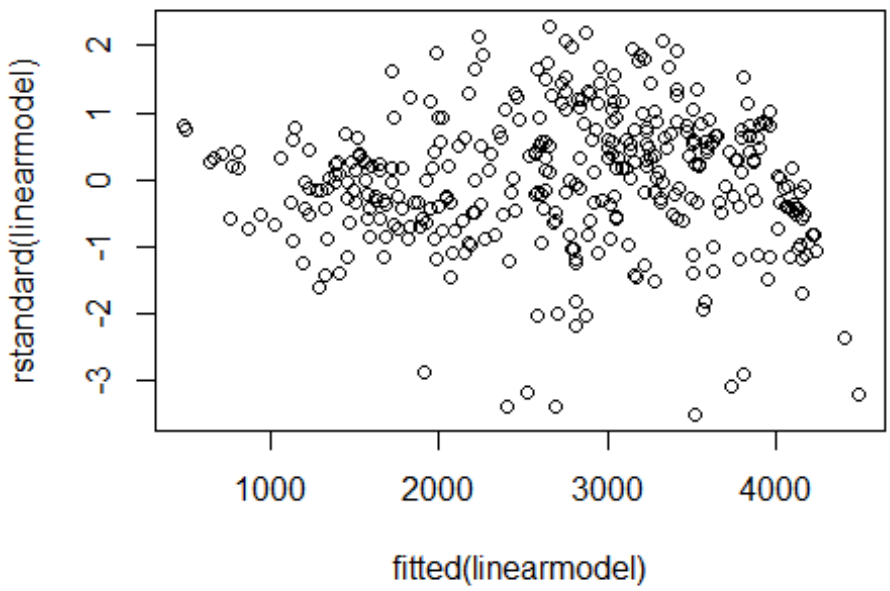

```
# plots and pictures from linear model 3
```

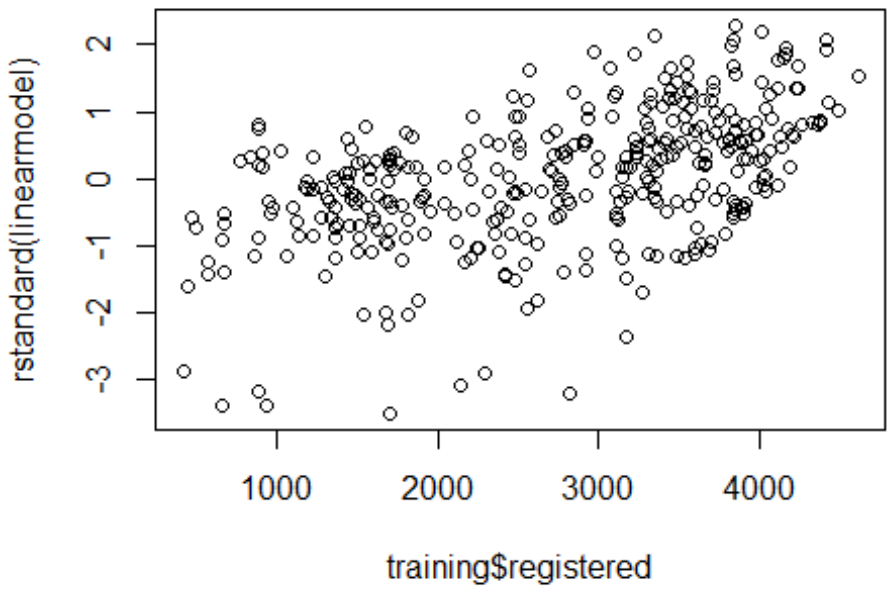

Residuals vs Fitted
Residuals
1000
0
-1000
-2000
341
328
329
1000
2000
3000
4000
Fitted values
egistered ~ windspeed + temp + weathersit + workingday + weekday

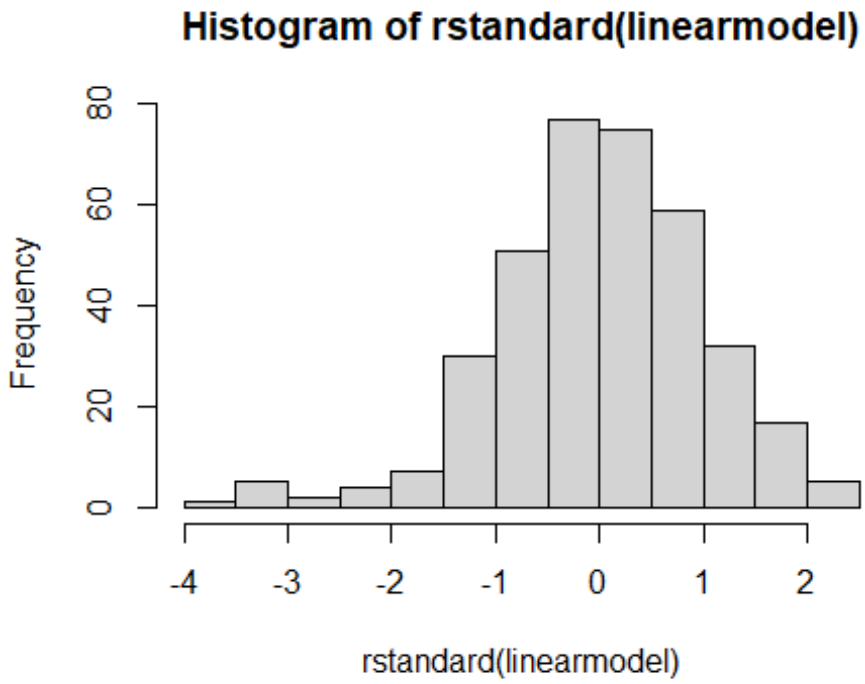

Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
Theoretical Quantiles
egistered ~ windspeed + temp + weathersit + workingday + weekday


Scale-Location
√|Standardized residuals|
341
328
329
0.0
0.5
1.0
1.5
1000
2000
3000
4000
Fitted values
egistered ~ windspeed + temp + weathersit + workingday + weekday


Residuals vs Leverage
Standardized residuals
3
2
1
0
-1
-2
-3
-4
Cook's distance
241
239
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Leverage
egistered ~ windspeed + temp + weathersit + workingday + weekday

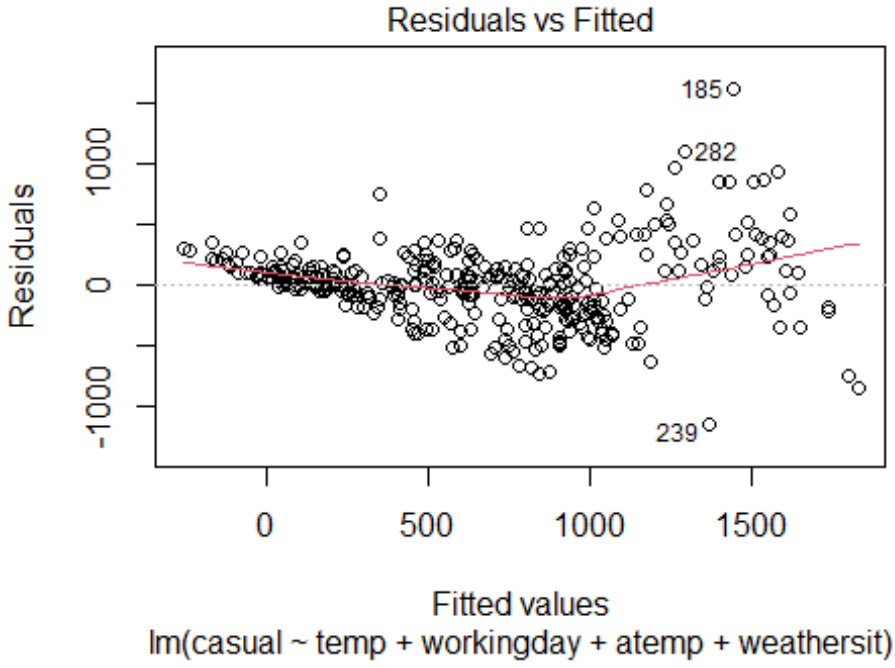

rstandard(linearmodel)
-3
-2
-1
0
1
2
1000
2000
3000
4000
fitted(linearmodel)

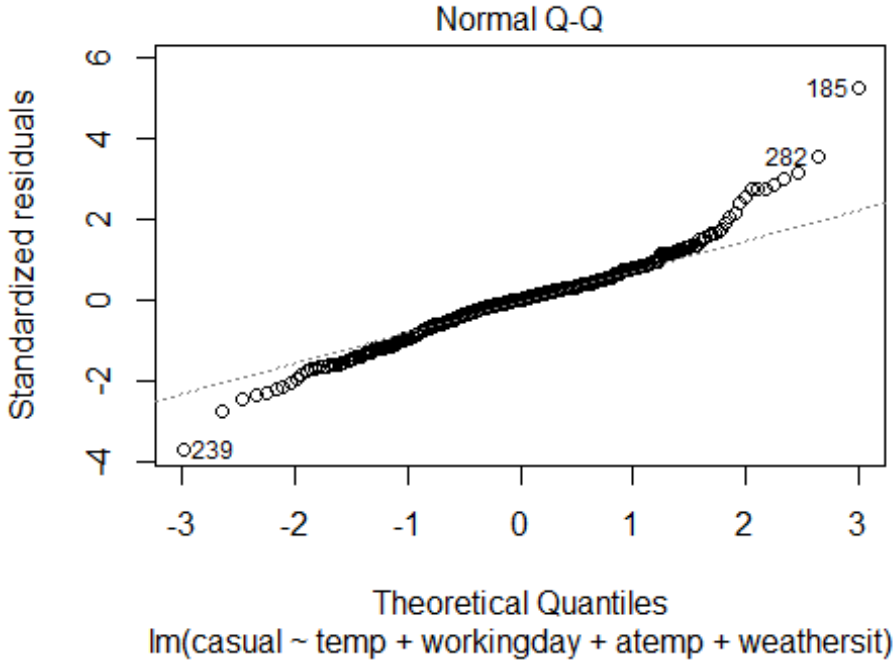

rstandard(linearmodel)
-3
-2
-1
0
1
2
1000
2000
3000
4000
training$registered

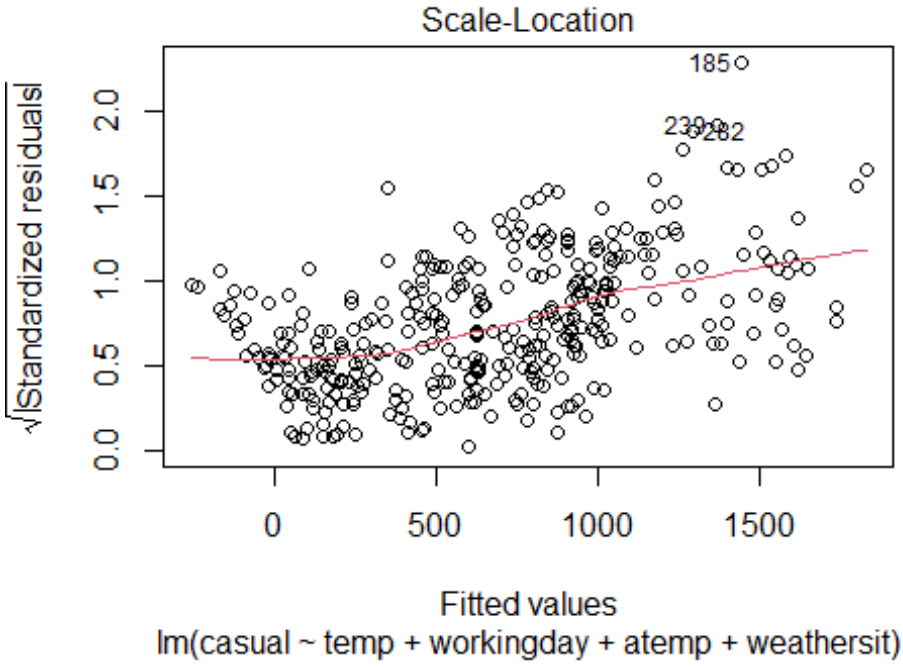


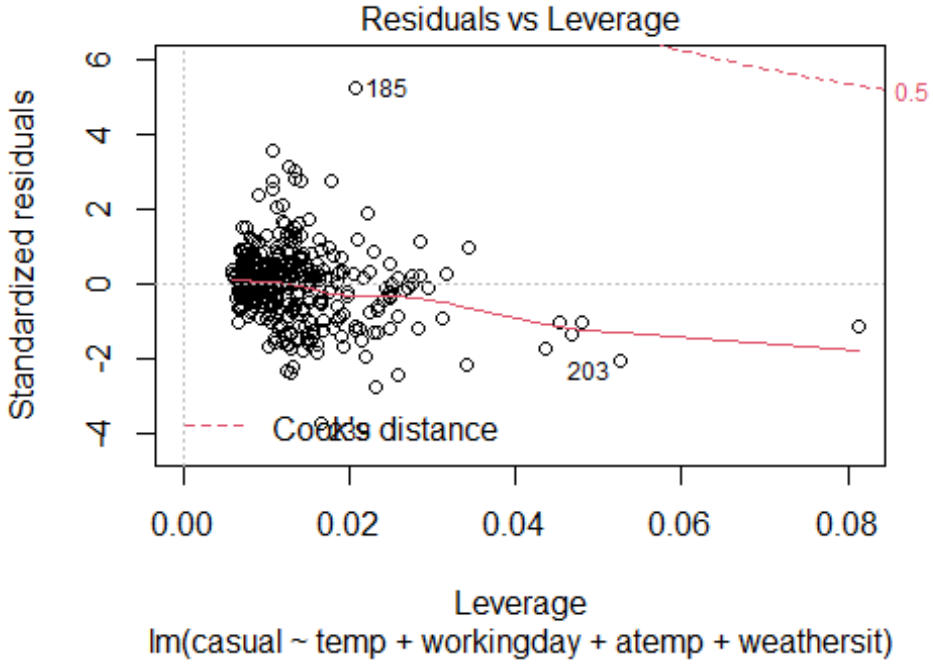

```
# plots and pictures from linear model 4
```

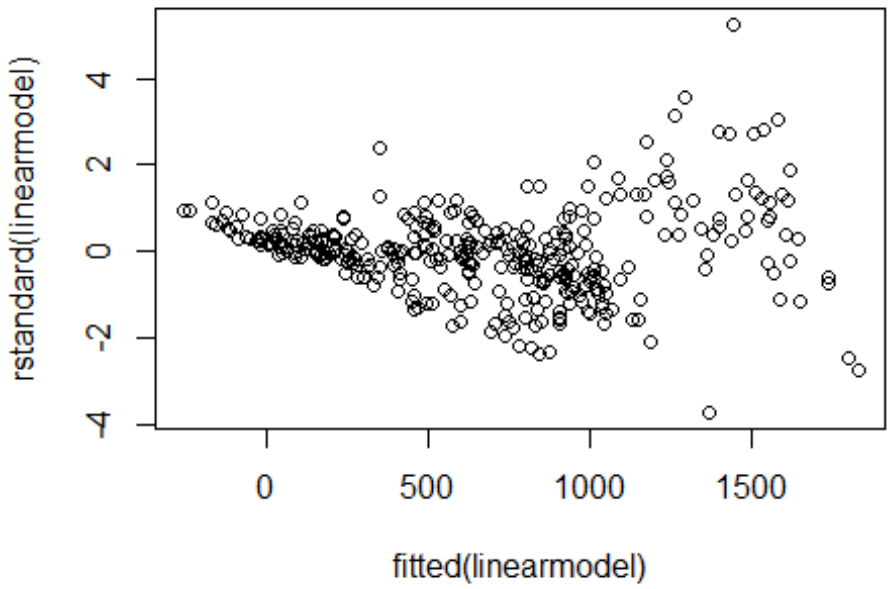
```
lm5 <- lm(casual ~ temp + workingday + atemp + weathersit, training)
# linear model of lasso of casual when choosing 4 predictors

lm55 <- lm(casual ~ temp + workingday + atemp, training)
# linear model of lasso of casual when choosing 3 predictors
```

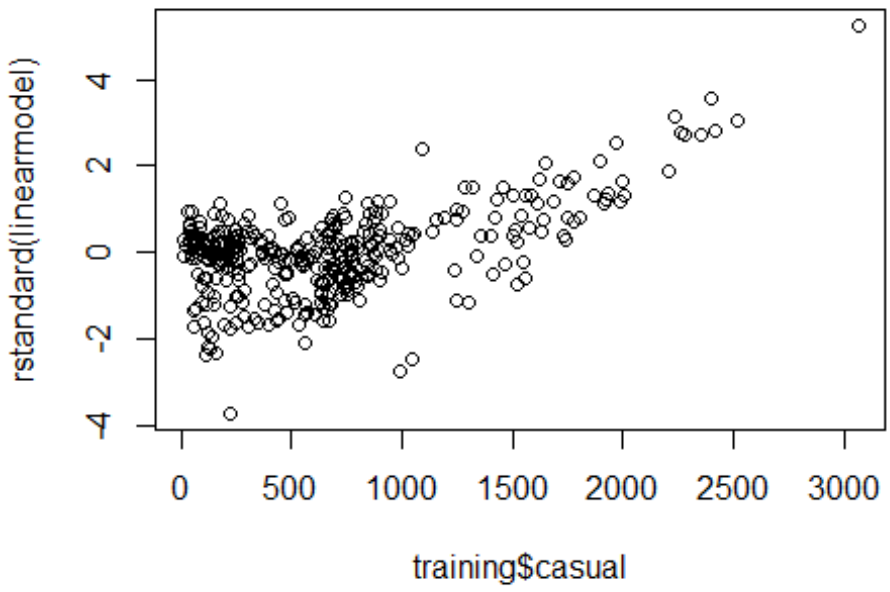

Residuals vs Fitted
Residuals
Fitted values
lm(casual ~ temp + workingday + atemp + weathersit)
185
282
239
1000
0
-1000
0
500
1000
1500

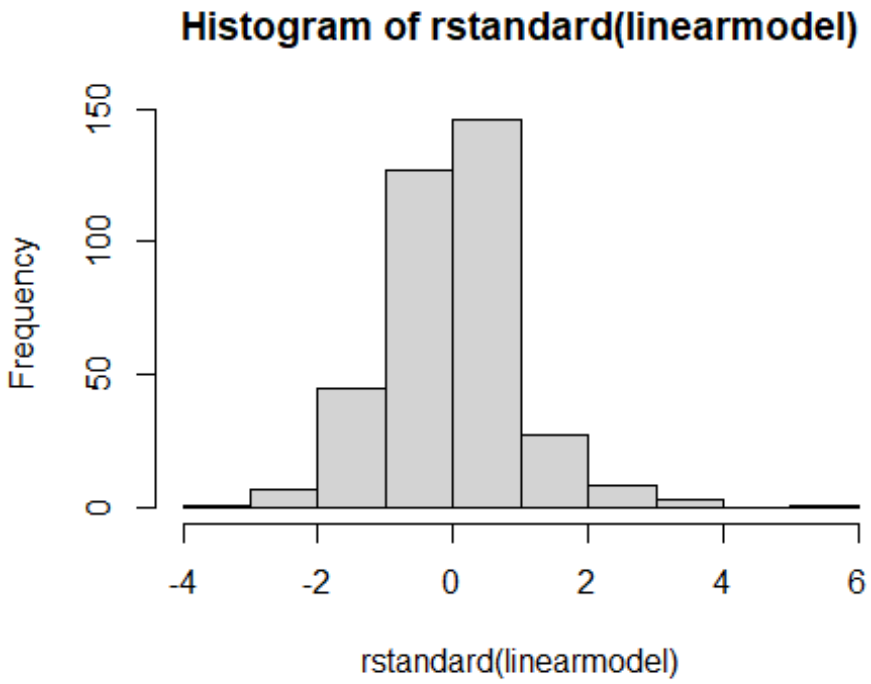

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(casual ~ temp + workingday + atemp + weathersit)
185
282
239
6
4
2
0
-2
-4
-3
-2
-1
0
1
2
3


Scale-Location
185
239
282
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
500
1000
1500
Fitted values
lm(casual ~ temp + workingday + atemp + weathersit)

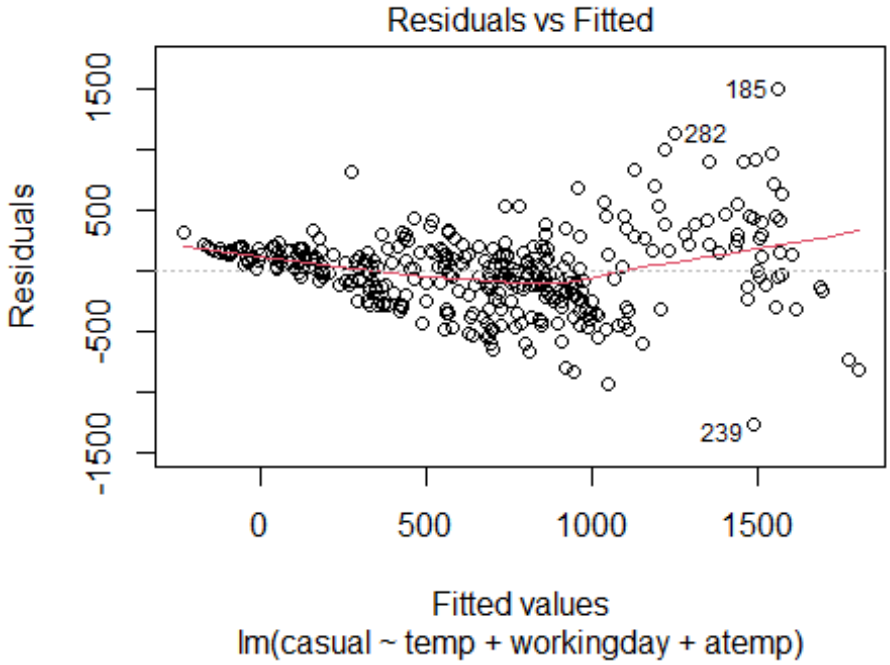

Residuals vs Leverage
185
203
0.5
Cook's distance
Standardized residuals
-4
-2
0
2
4
6
0.00
0.02
0.04
0.06
0.08
Leverage
lm(casual ~ temp + workingday + atemp + weathersit)

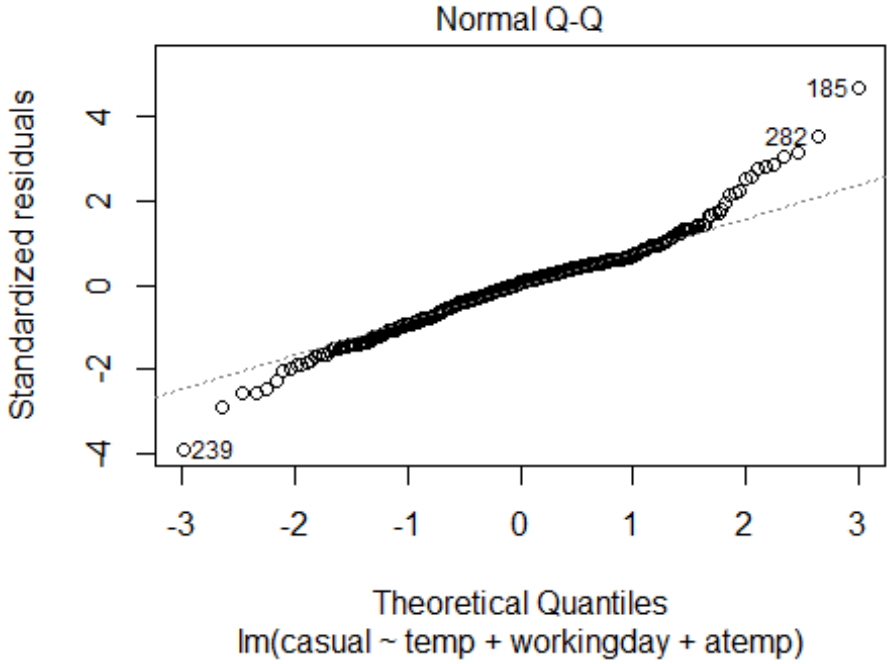

rstandard(linearmodel)
4
2
0
-2
-4
0
500
1000
1500
fitted(linearmodel)

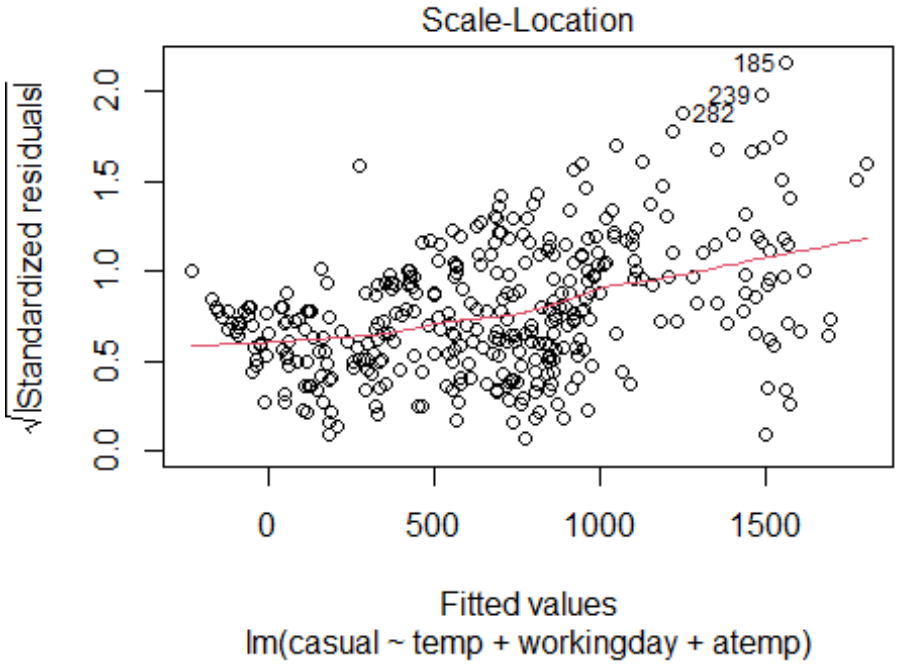

rstandard(linearmodel)
4
2
0
-2
-4
0
500
1000
1500
2000
2500
3000
training$casual

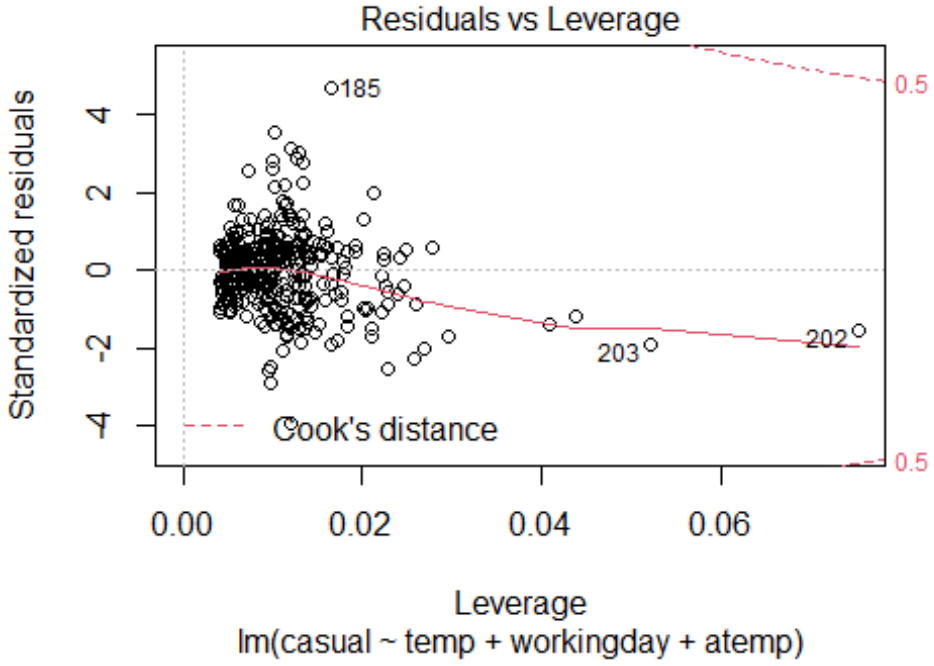


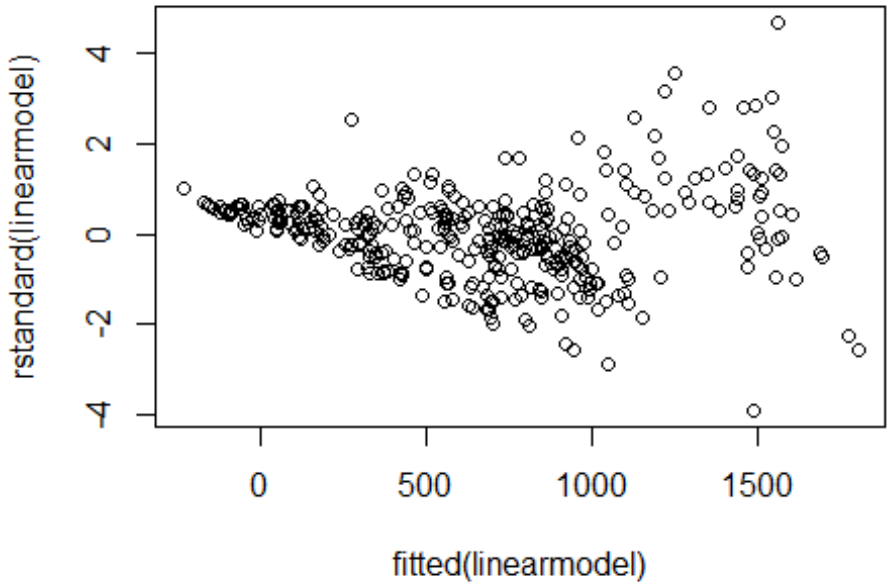

```
# plots and pictures from linear model 5
```

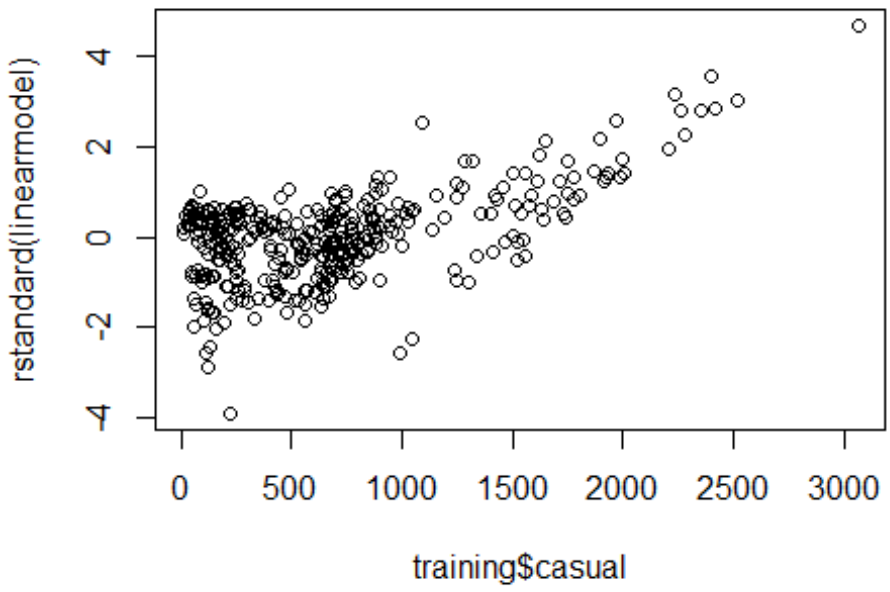

Residuals vs Fitted
Residuals
Fitted values
lm(casual ~ temp + workingday + atemp)
185
282
239
1500
500
-500
-1500
0
500
1000
1500

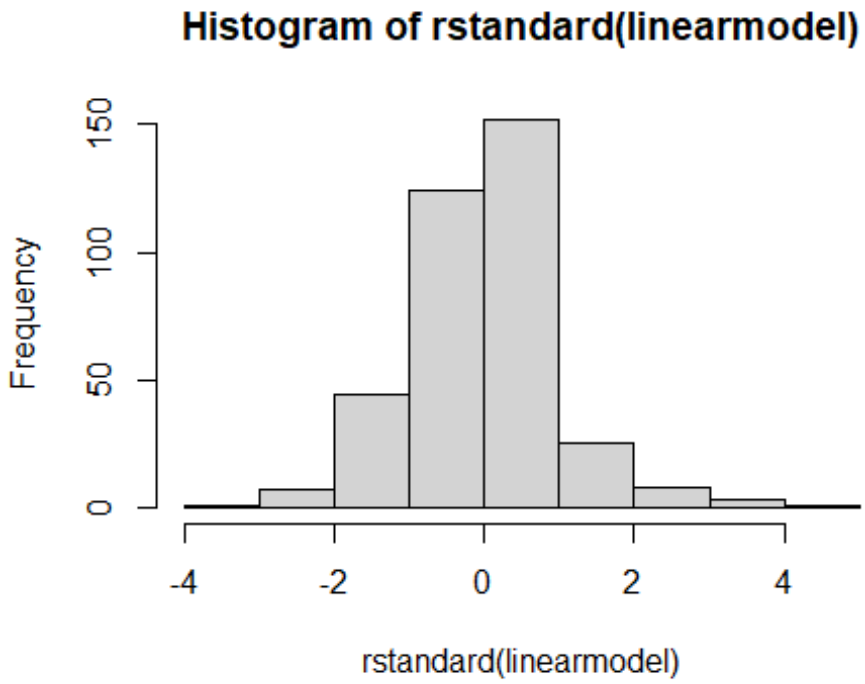

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(casual ~ temp + workingday + atemp)
185
282
239
4
2
0
-2
-4
-3
-2
-1
0
1
2
3


Scale-Location
185
239
282
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
500
1000
1500
Fitted values
lm(casual ~ temp + workingday + atemp)


Residuals vs Leverage
185
203
202
Cook's distance
0.5
0.5
Standardized residuals
-4
-2
0
2
4
0.00
0.02
0.04
0.06
Leverage
lm(casual ~ temp + workingday + atemp)

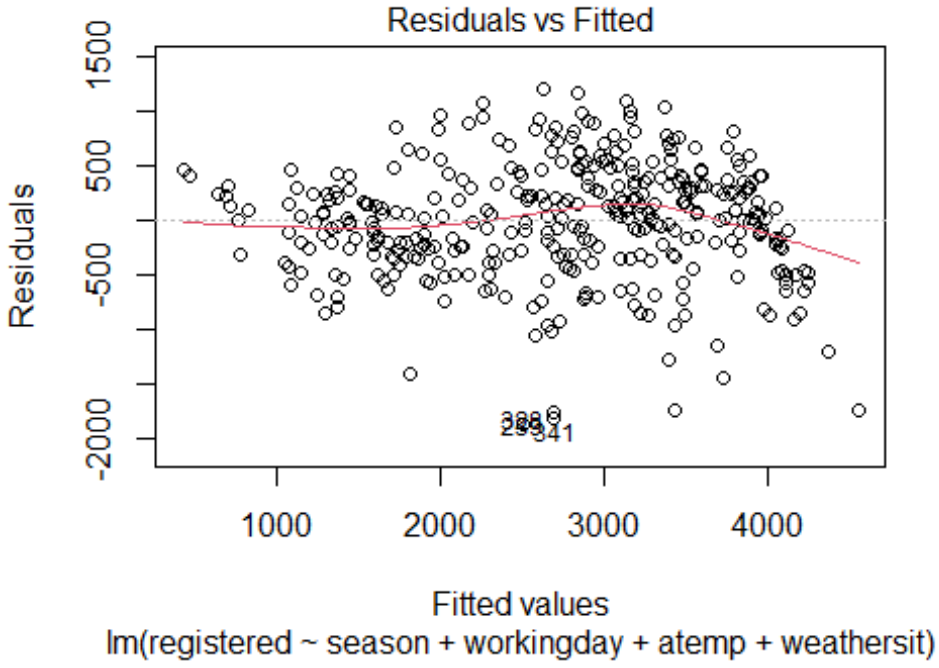

rstandard(linearmodel)
4
2
0
-2
-4
0
500
1000
1500
fitted(linearmodel)

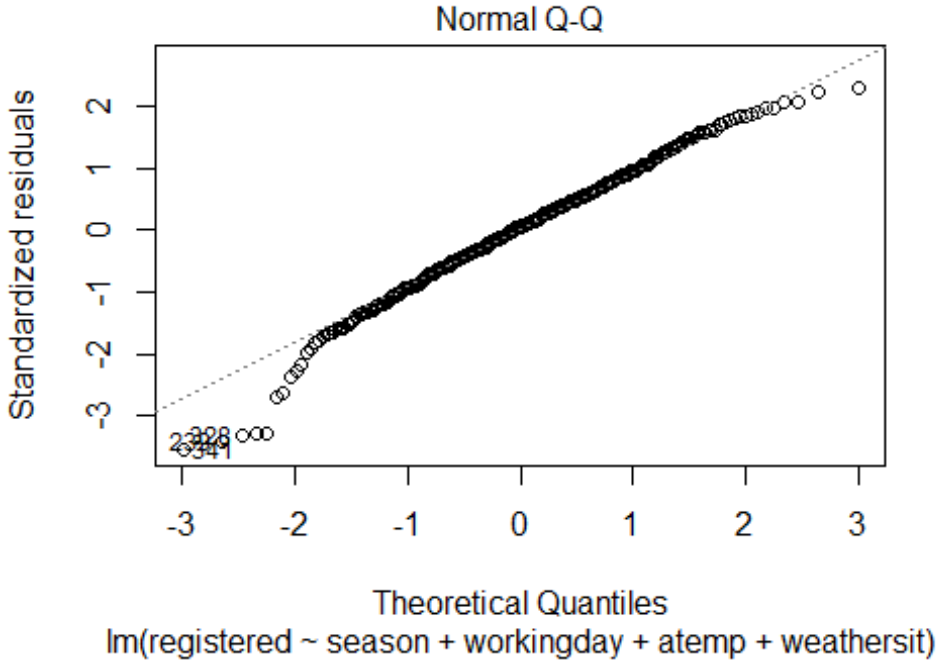

rstandard(linearmodel)
4
2
0
-2
-4
0
500
1000
1500
2000
2500
3000
training$casual

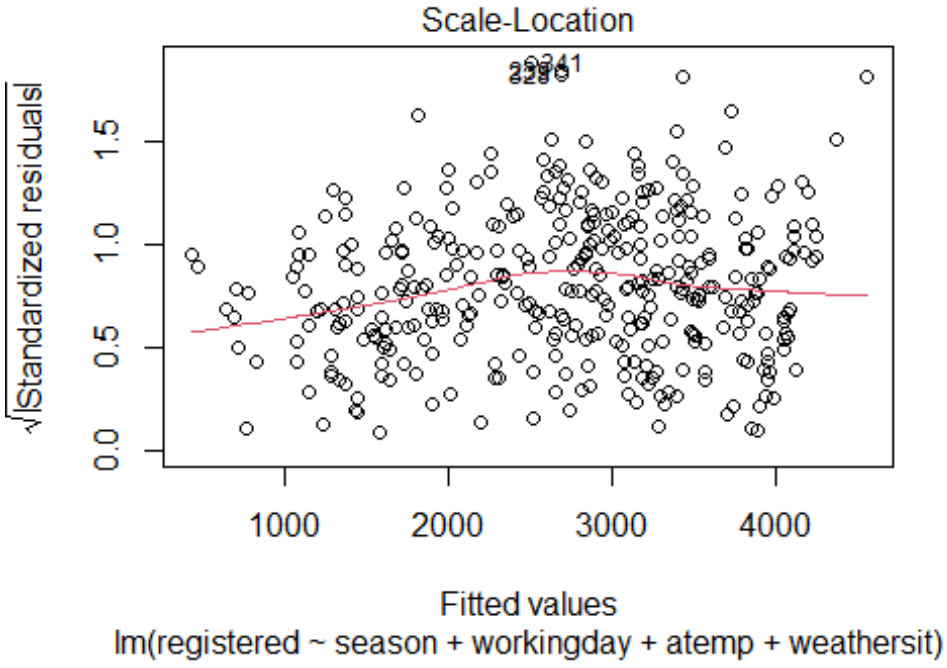


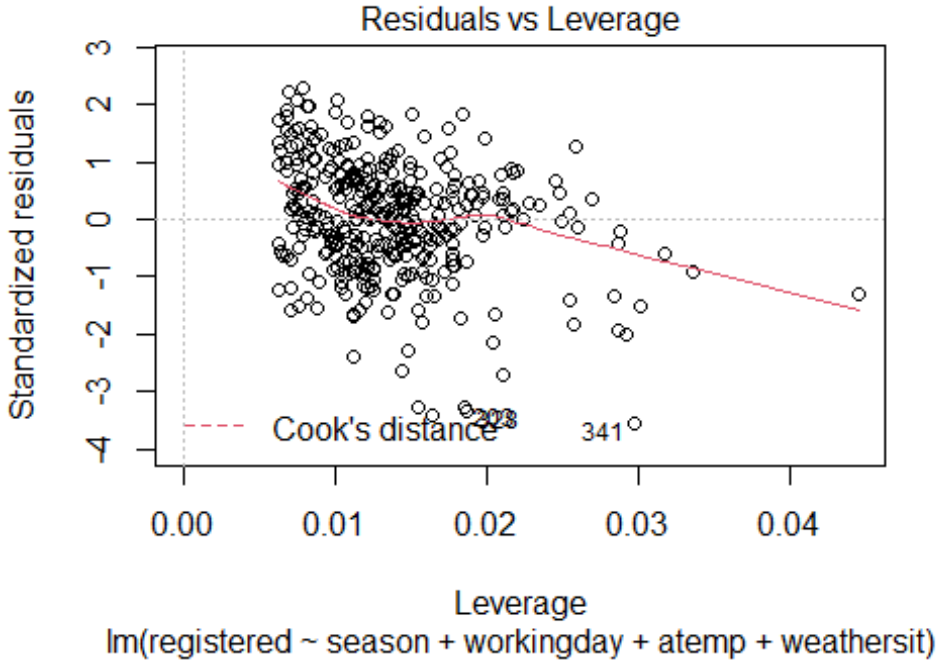

```
# plots and pictures from linear model 55
```

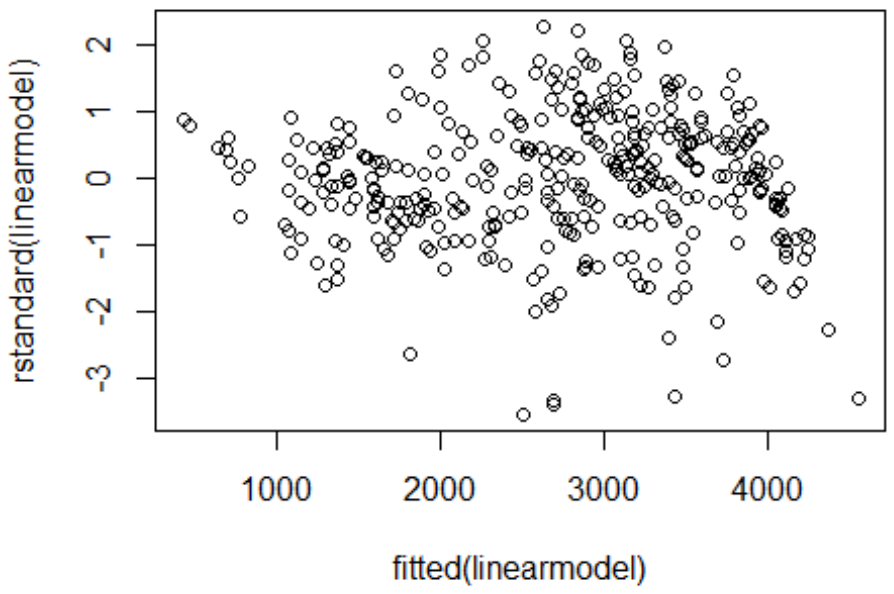
```
lm6 <- lm(registered ~ season + workingday + atemp + weathersit, training)
# linear model of lasso of registered when choosing 4 predictors

lm7 <- lm(registered ~ atemp, training)
# linear model of lasso of registered when choosing 4 predictors
```

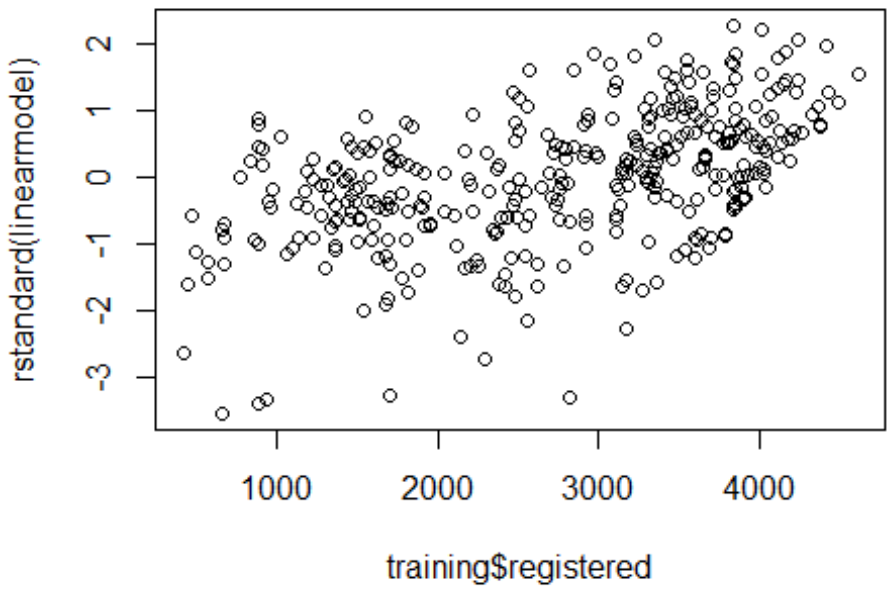

Residuals vs Fitted
Residuals
1500
500
-500
-2000
1000
2000
3000
4000
Fitted values
lm(registered ~ season + workingday + atemp + weathersit)

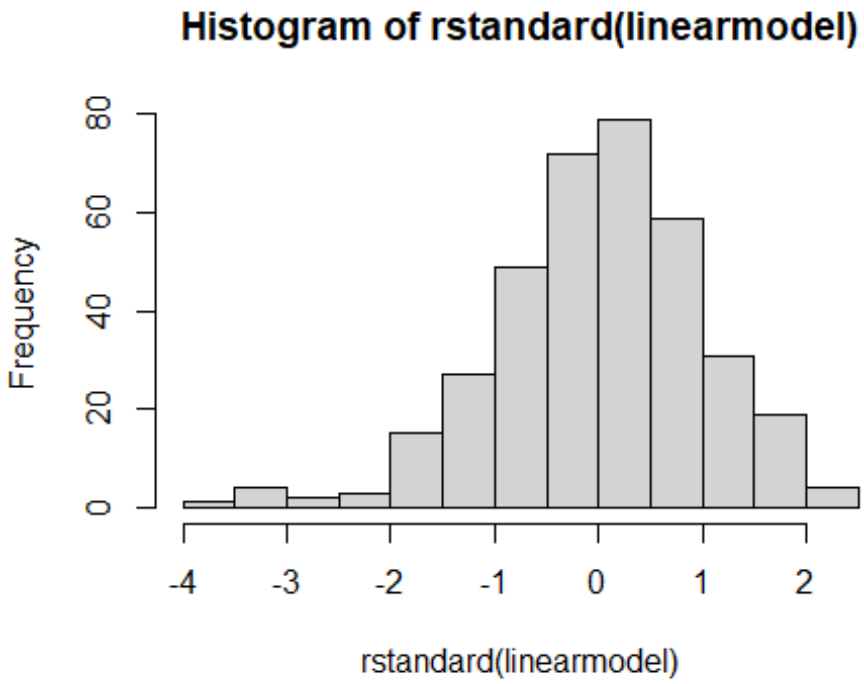

Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(registered ~ season + workingday + atemp + weathersit)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1000
2000
3000
4000
Fitted values
lm(registered ~ season + workingday + atemp + weathersit)

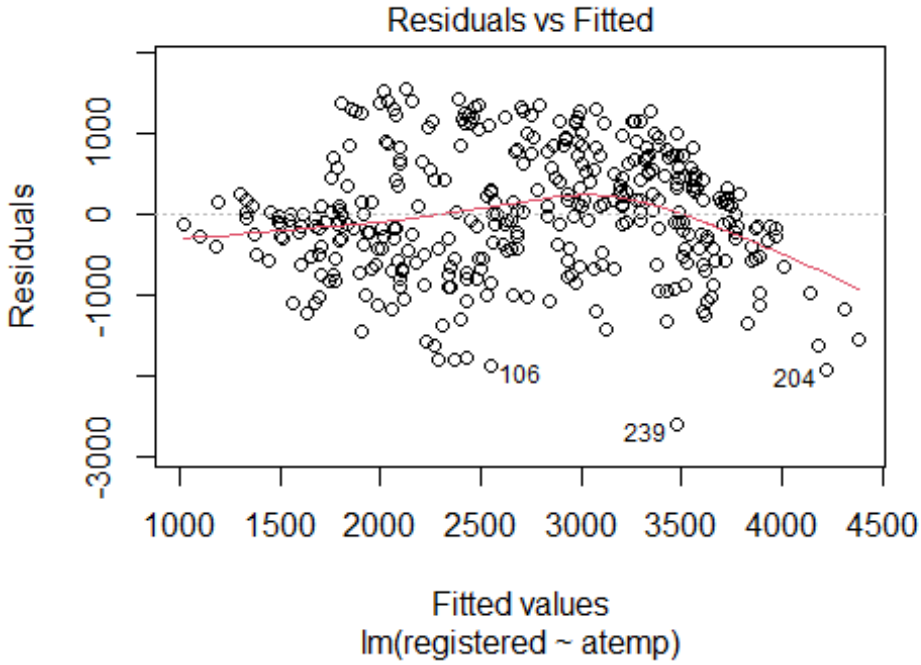

Residuals vs Leverage
Standardized residuals
3
2
1
0
-1
-2
-3
-4
Cook's distance
341
0.00
0.01
0.02
0.03
0.04
Leverage
lm(registered ~ season + workingday + atemp + weathersit)

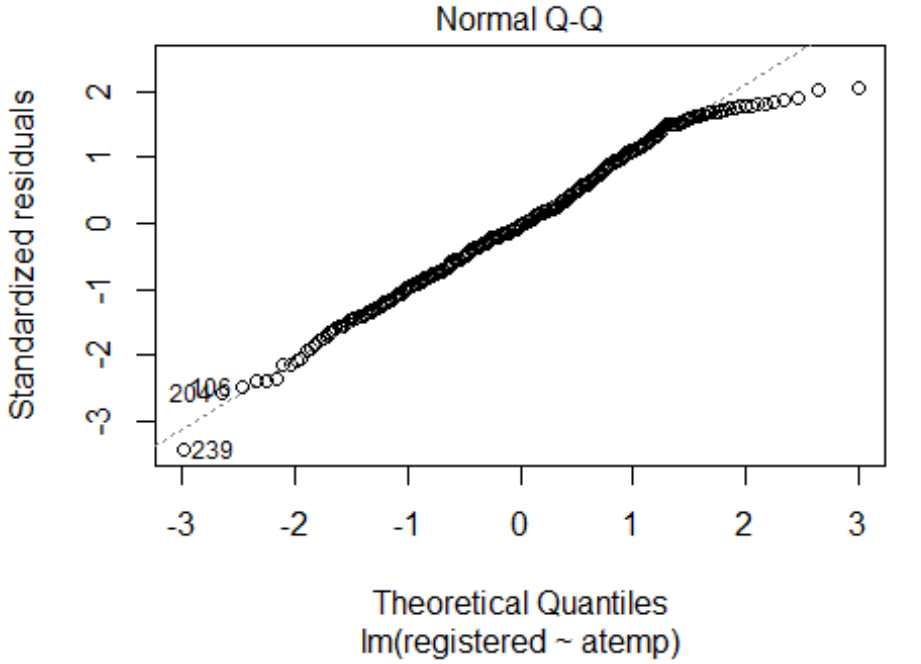

rstandard(linearmodel)
-3
-2
-1
0
1
2
1000
2000
3000
4000
fitted(linearmodel)

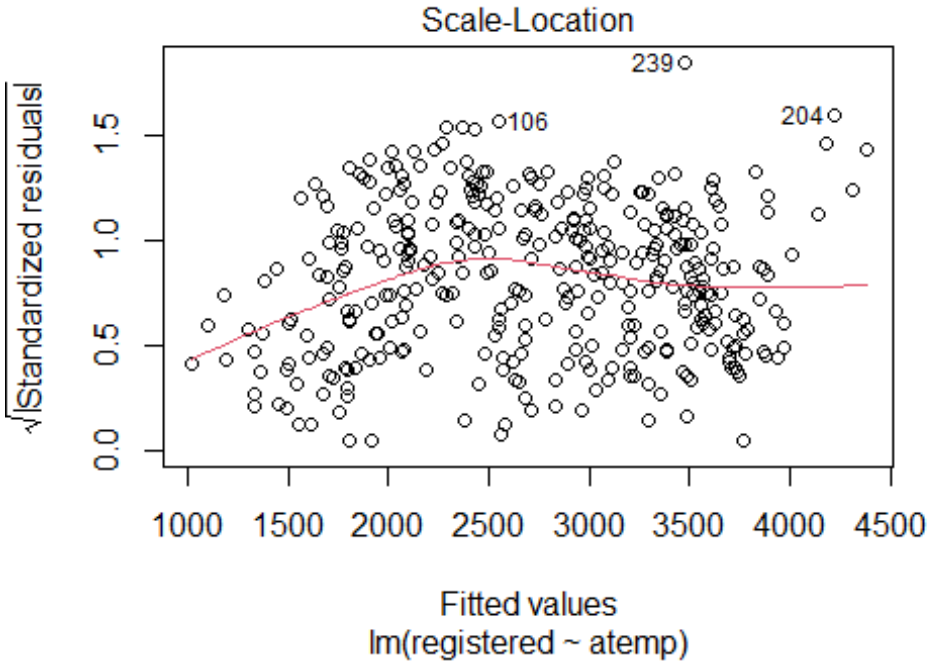

rstandard(linearmodel)
-3
-2
-1
0
1
2
1000
2000
3000
4000
training$registered

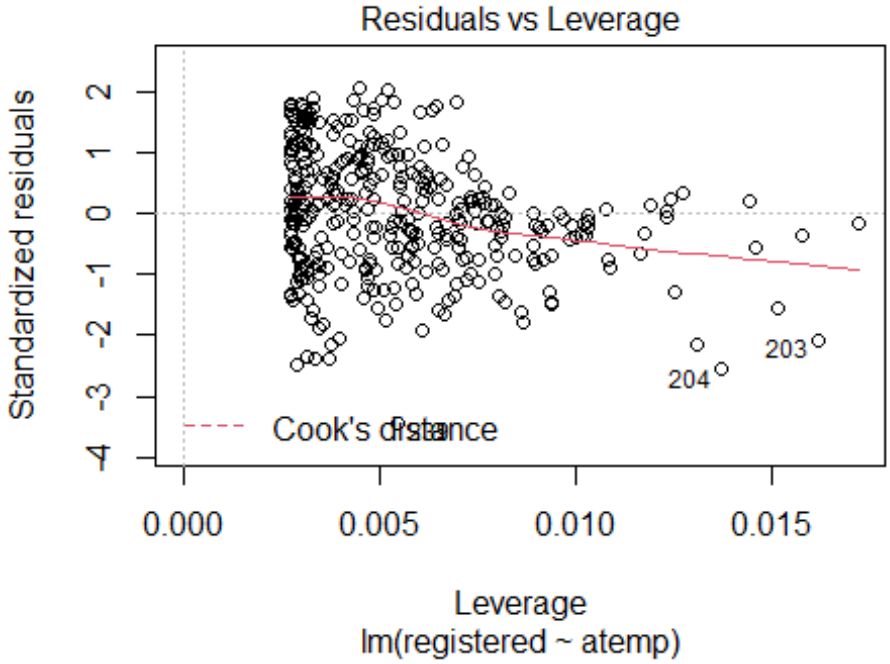


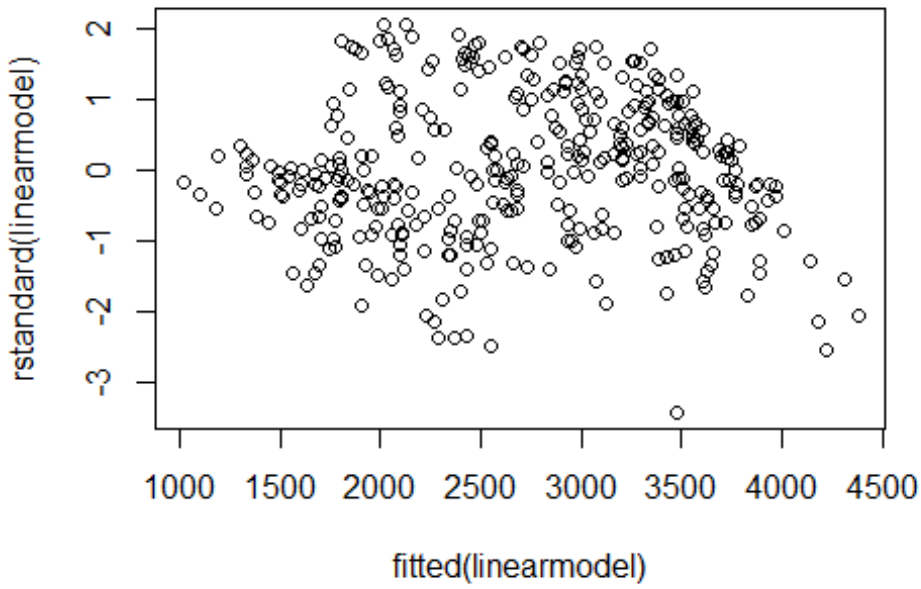

```
# plots and pictures from linear model 6
```

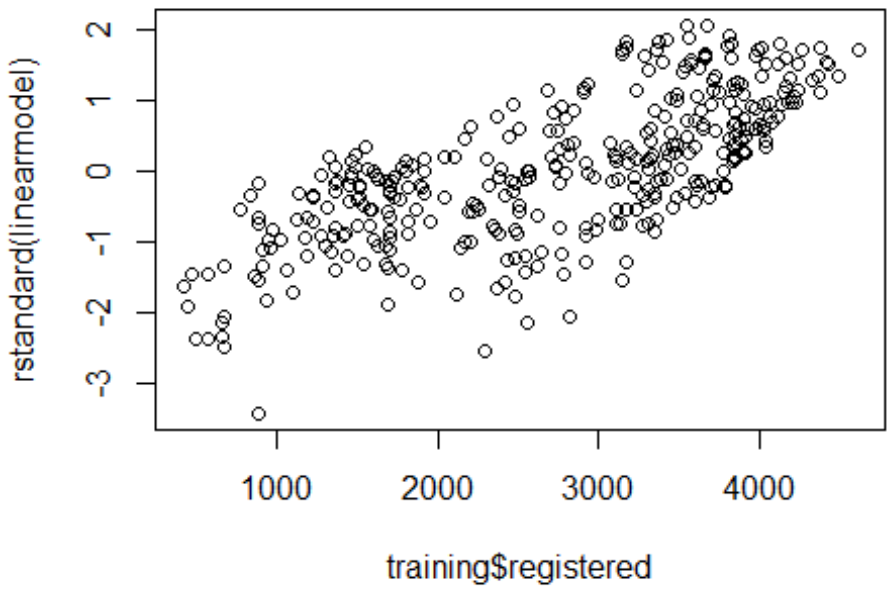

Residuals vs Fitted
Residuals
1000
0
-1000
-3000
106
204
239
1000
1500
2000
2500
3000
3500
4000
4500
Fitted values
lm(registered ~ atemp)

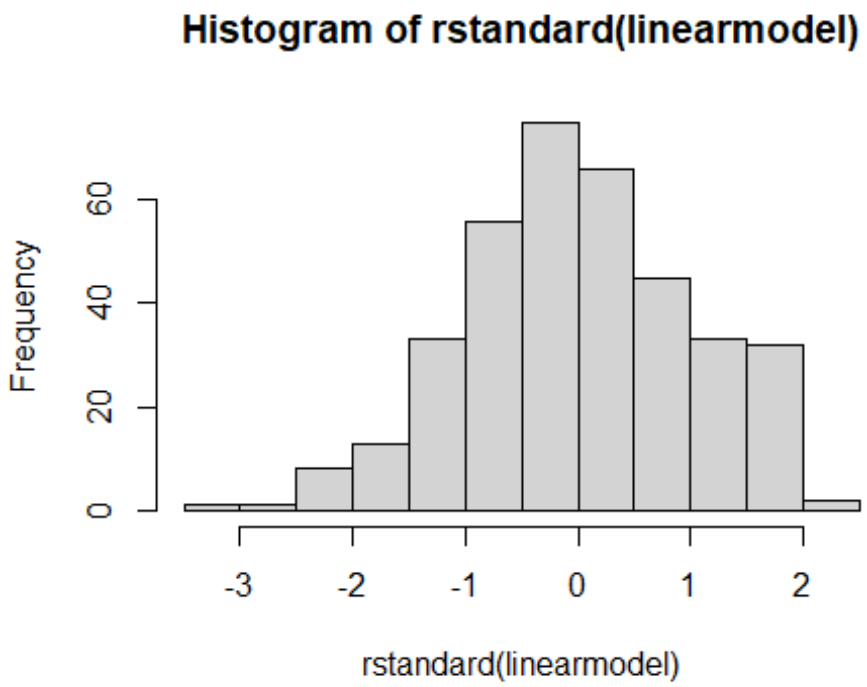

Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-3
204
106
239
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(registered ~ atemp)


Scale-Location
√|Standardized residuals|
239
106
204
0.0
0.5
1.0
1.5
1000
1500
2000
2500
3000
3500
4000
4500
Fitted values
lm(registered ~ atemp)


Residuals vs Leverage
Standardized residuals
2
1
0
-1
-2
-3
-4
204
203
Cook's distance
0.000
0.005
0.010
0.015
Leverage
lm(registered ~ atemp)


rstandard(linearmodel)
2
1
0
-1
-2
-3
1000
1500
2000
2500
3000
3500
4000
4500
fitted(linearmodel)


rstandard(linearmodel)
2
1
0
-1
-2
-3
1000
2000
3000
4000
training$registered




```
# plots and pictures from linear model 7


summary(lm1)

##
## Call:
## lm(formula = casual ~ windspeed + temp + weathersit + workingday,
##     data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1010.43  -175.64    16.89   149.10  1535.01
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   652.97      76.51   8.535 3.98e-16 ***
## windspeed    -727.54     211.87  -3.434 0.000664 ***
## temp         1705.03      85.91  19.848  < 2e-16 ***
## weathersit   -151.34      28.60  -5.292 2.11e-07 ***
## workingday   -658.31      34.88 -18.874  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 307.2 on 360 degrees of freedom
## Multiple R-squared:  0.6983, Adjusted R-squared:  0.695
## F-statistic: 208.3 on 4 and 360 DF,  p-value: < 2.2e-16
```

```r
# model 1 summary


summary(lm2)

##
## Call:
## lm(formula = casual ~ windspeed + temp + weathersit + workingday +
##     mnth + season, data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1003.27  -161.38    16.78   131.14  1522.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  649.262     81.803   7.937 2.69e-14 ***
## windspeed   -730.659    216.751  -3.371 0.000831 ***
## temp        1670.489     91.854  18.186  < 2e-16 ***
## weathersit  -155.522     28.587  -5.440 9.88e-08 ***
## workingday  -658.822     34.676 -18.999  < 2e-16 ***
## mnth         -20.331      8.371  -2.429 0.015644 *
## season        64.071     26.910   2.381 0.017793 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 305.3 on 358 degrees of freedom
## Multiple R-squared:  0.7037, Adjusted R-squared:  0.6987
## F-statistic: 141.7 on 6 and 358 DF,  p-value: < 2.2e-16

# model 2 summary


summary(lm3)

##
## Call:
## lm(formula = registered ~ windspeed + temp + weathersit + workingday +
##     season, data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1777.97  -291.38    35.72   339.33  1181.95
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   885.37     140.11   6.319  7.8e-10 ***
## windspeed   -1235.53     371.68  -3.324 0.000978 ***
## temp         2866.85     157.76  18.172  < 2e-16 ***
## weathersit   -478.10      49.13  -9.732  < 2e-16 ***
## workingday    714.32      59.62  11.982  < 2e-16 ***
```

```
## season        350.12      27.49  12.735  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 525 on 359 degrees of freedom
## Multiple R-squared:  0.7581, Adjusted R-squared:  0.7547
## F-statistic:   225 on 5 and 359 DF,  p-value: < 2.2e-16
```

```
# model 3 summary
```

```
summary(lm4)
```

```
##
## Call:
## lm(formula = registered ~ windspeed + temp + weathersit + workingday +
##     weekday + season, data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1822.52  -294.50     8.93   327.50  1186.89
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   829.26     144.38   5.744 1.98e-08 ***
## windspeed   -1267.29     371.50  -3.411  0.00072 ***
## temp         2874.93     157.53  18.250  < 2e-16 ***
## weathersit   -480.76      49.06  -9.799  < 2e-16 ***
## workingday    712.71      59.51  11.977  < 2e-16 ***
## weekday        21.44      13.74   1.561  0.11940
## season        349.58      27.44  12.739  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 524 on 358 degrees of freedom
## Multiple R-squared:  0.7597, Adjusted R-squared:  0.7557
## F-statistic: 188.7 on 6 and 358 DF,  p-value: < 2.2e-16
```

```
# model 4 summary
```

```
summary(lm5)
```

```
##
## Call:
## lm(formula = casual ~ temp + workingday + atemp + weathersit,
##     data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1144.67  -170.16     2.17   147.45  1618.23
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   496.74      78.91   6.295 8.94e-10 ***
## temp         1357.62    1028.26   1.320    0.188
## workingday   -659.08      35.44 -18.597  < 2e-16 ***
## atemp         426.82    1155.32   0.369    0.712
## weathersit   -160.26      28.99  -5.527 6.25e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 312.1 on 360 degrees of freedom
## Multiple R-squared:  0.6886, Adjusted R-squared:  0.6851
## F-statistic:   199 on 4 and 360 DF,  p-value: < 2.2e-16

# model 5 summary



summary(lm55)

##
## Call:
## lm(formula = casual ~ temp + workingday + atemp, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1260.2  -187.4    20.4   166.1  1505.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   245.31      67.06   3.658 0.000292 ***
## temp          996.40    1067.36   0.934 0.351175
## workingday   -681.29      36.62 -18.603  < 2e-16 ***
## atemp         886.44    1198.56   0.740 0.460033
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 324.7 on 361 degrees of freedom
## Multiple R-squared:  0.6621, Adjusted R-squared:  0.6593
## F-statistic: 235.8 on 3 and 361 DF,  p-value: < 2.2e-16

# model 55 summary


summary(lm6)

##
## Call:
## lm(formula = registered ~ season + workingday + atemp + weathersit,
##     data = training)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1855.09  -313.64    28.82   332.22  1210.13
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   511.10     120.20   4.252  2.7e-05 ***
## season        364.91      27.26  13.385  < 2e-16 ***
## workingday    713.68      60.42  11.813  < 2e-16 ***
## atemp        3245.01     180.43  17.985  < 2e-16 ***
## weathersit   -491.05      49.48  -9.924  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 532 on 360 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7481
## F-statistic: 271.3 on 4 and 360 DF,  p-value: < 2.2e-16
```

```
# model 6 summary


summary(lm7)
```

```
##
## Call:
## lm(formula = registered ~ atemp, data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2584.51  -523.86   -26.45   538.59  1541.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    666.6      116.3   5.733 2.08e-08 ***
## atemp         4416.5      234.3  18.853  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 754.6 on 363 degrees of freedom
## Multiple R-squared:  0.4947, Adjusted R-squared:  0.4934
## F-statistic: 355.4 on 1 and 363 DF,  p-value: < 2.2e-16
```

```
# model 7 summary

rmse.casual(lm1)
```

```
## [1] 0.2888159
```

```
rmse.casual(lm2)
```

```
## [1] 0.2850549
```

```
rmse.casual(lm5)
```

```
## [1] 0.2950554
```

```
rmse.casual(lm55)
```

```
## [1] 0.30775
```

```
rmse.reg(lm3)
```

```
## [1] 0.06992535
```

```
rmse.reg(lm4)
```

```
## [1] 0.0696465
```

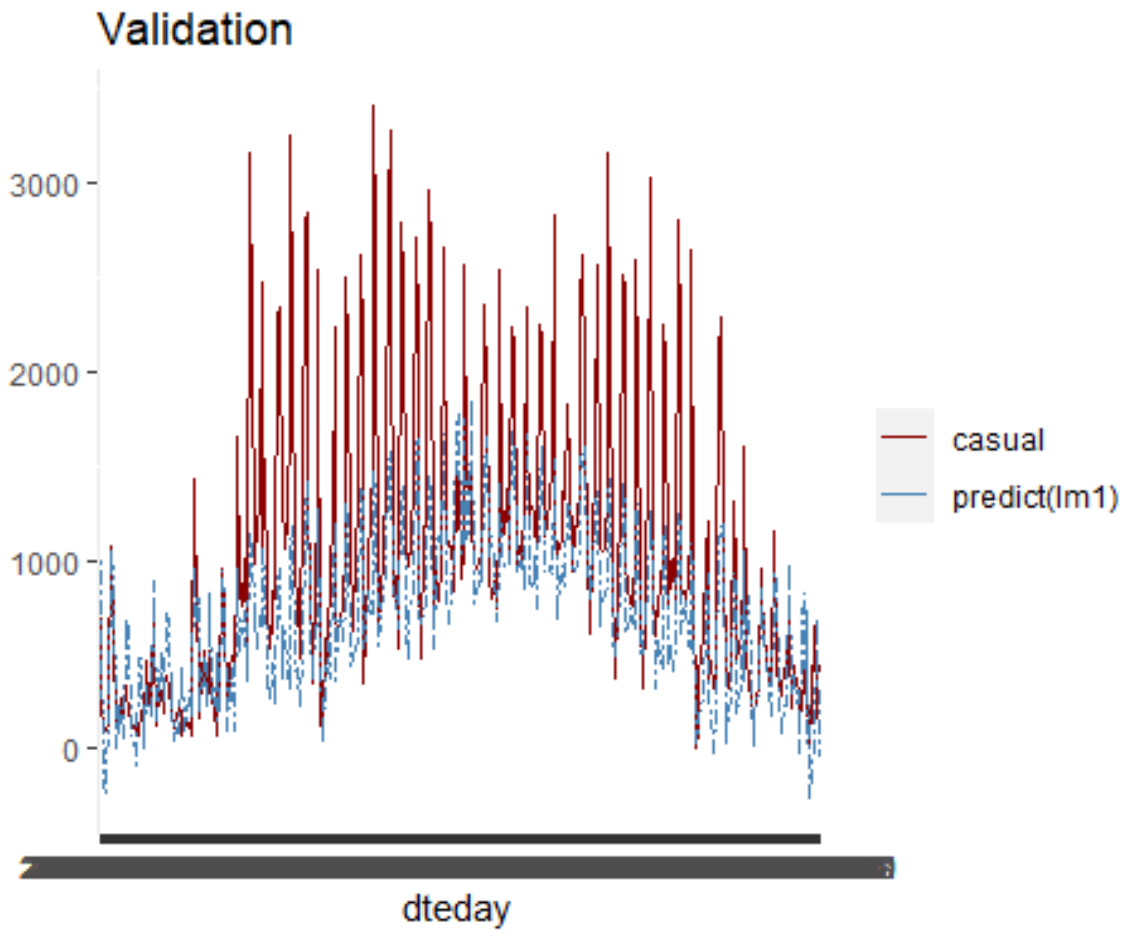
```
rmse.reg(lm6)
```

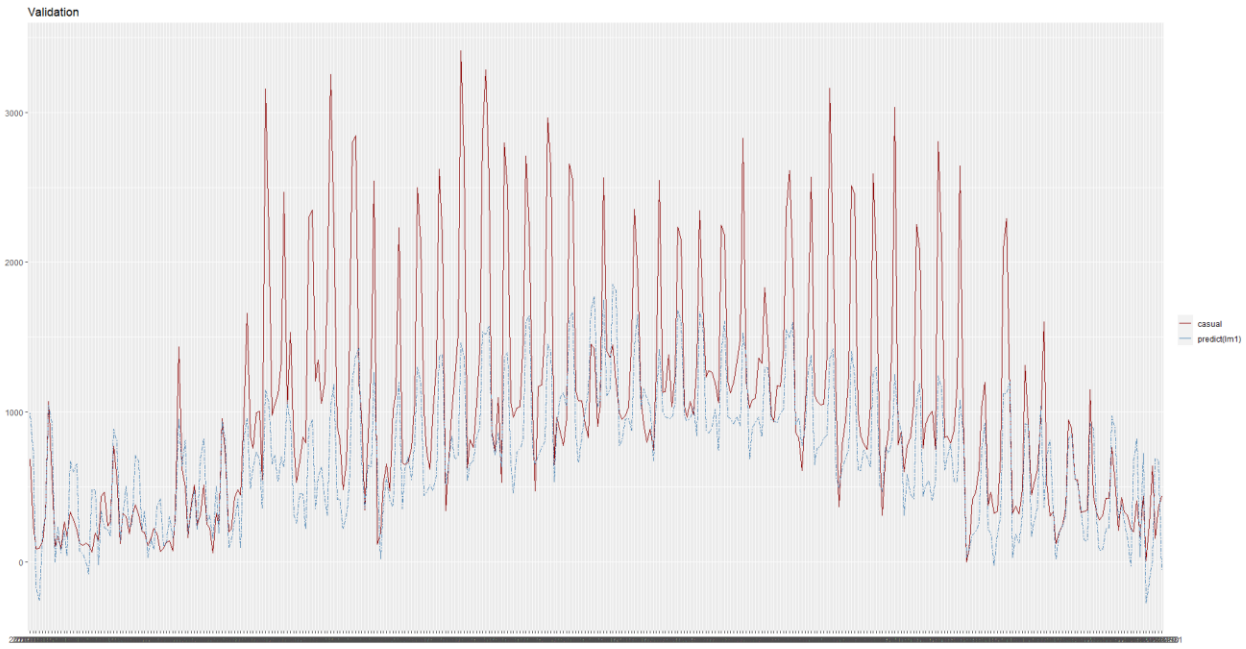
```
## [1] 0.07163259
```

```
rmse.reg(lm7)
```

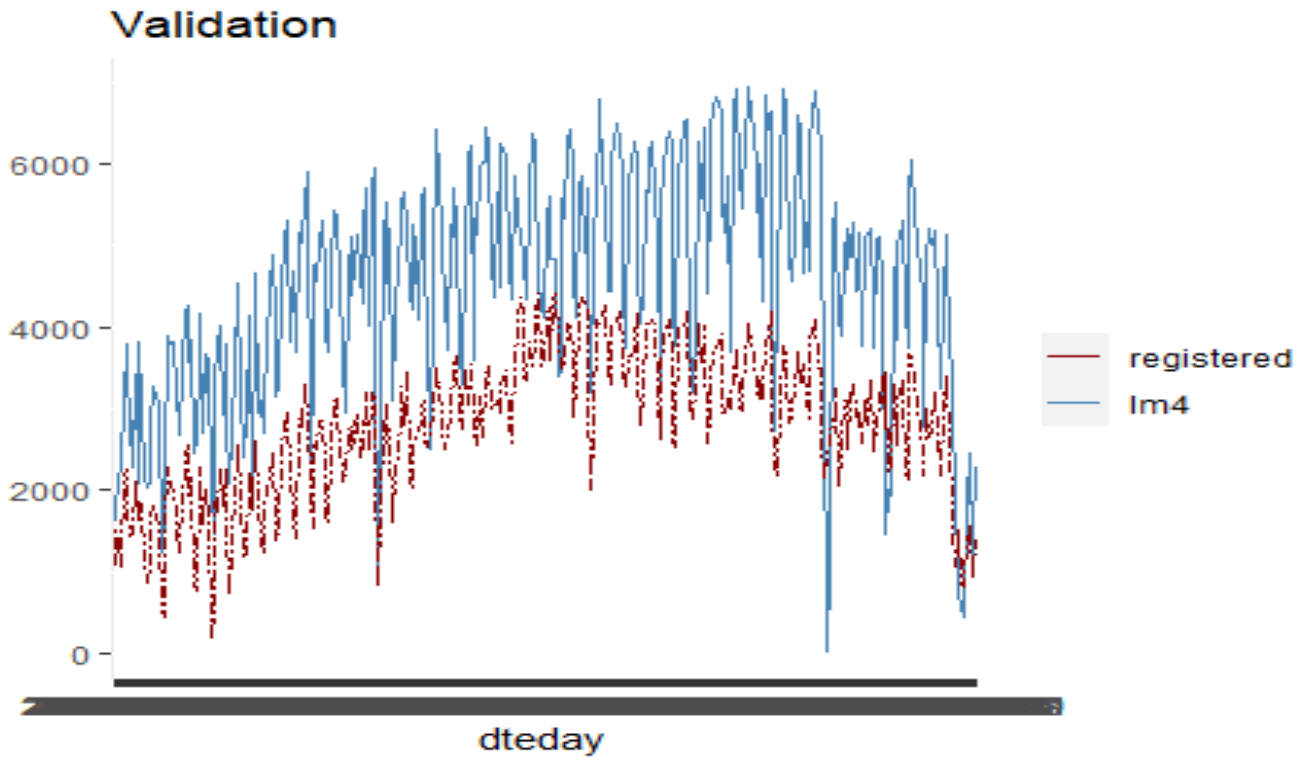
```
## [1] 0.126267
```

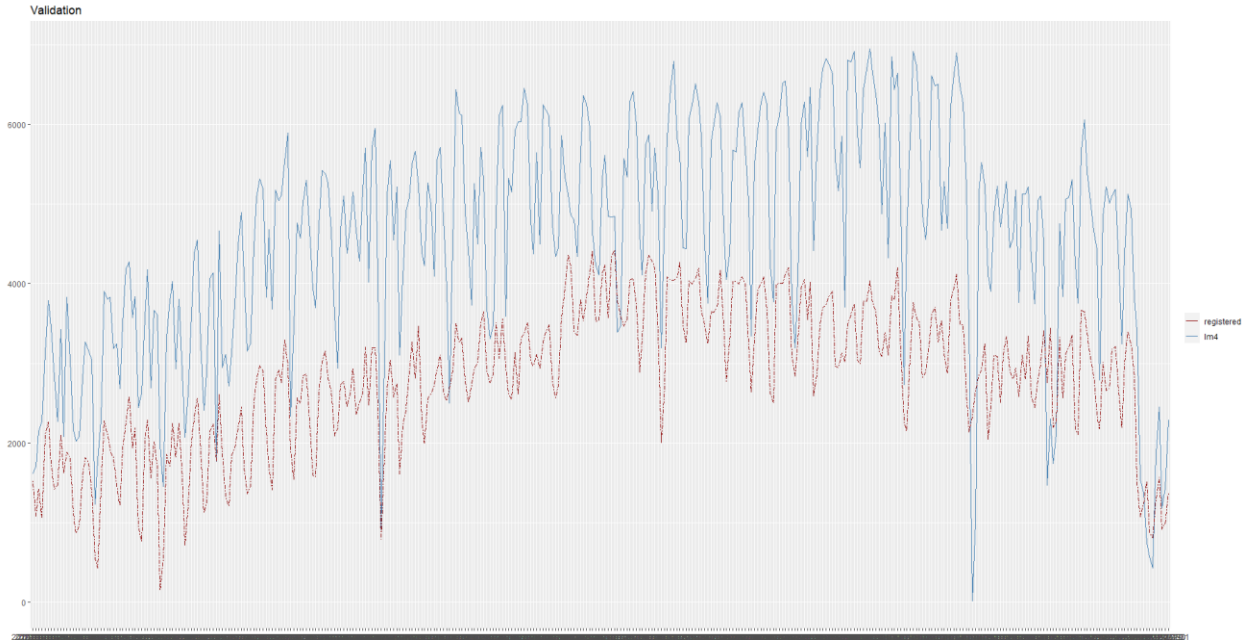
```
# relative mean squared error of all models
```


Validation
3000
2000
1000
0
dteday
casual
predict(lm1)

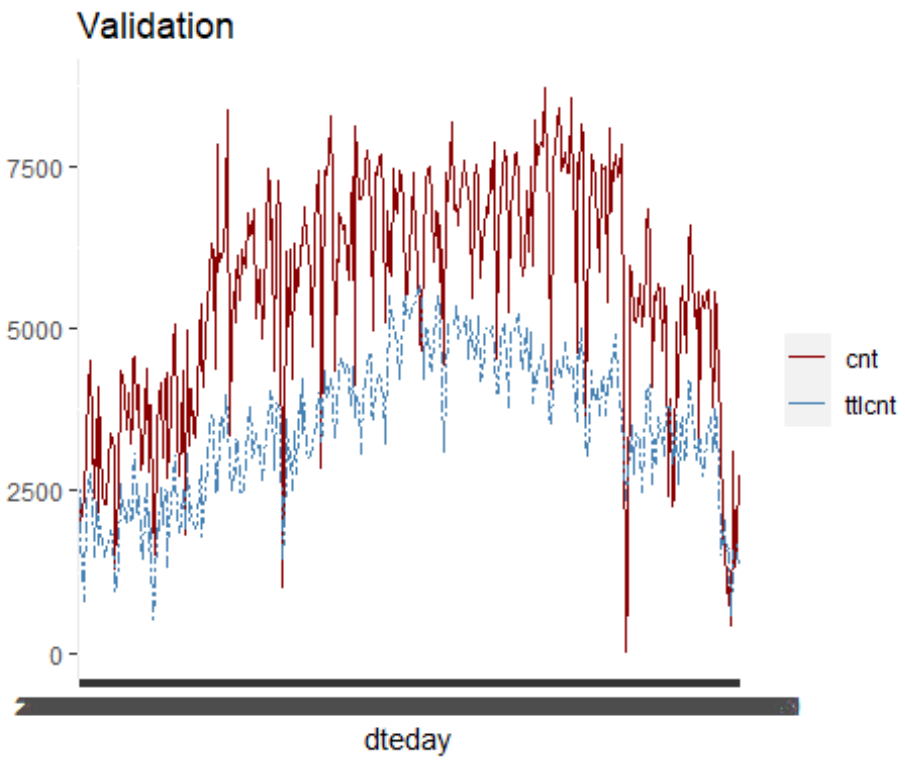

Validation
3000
2000
1000
0
casual
predict(lm1)

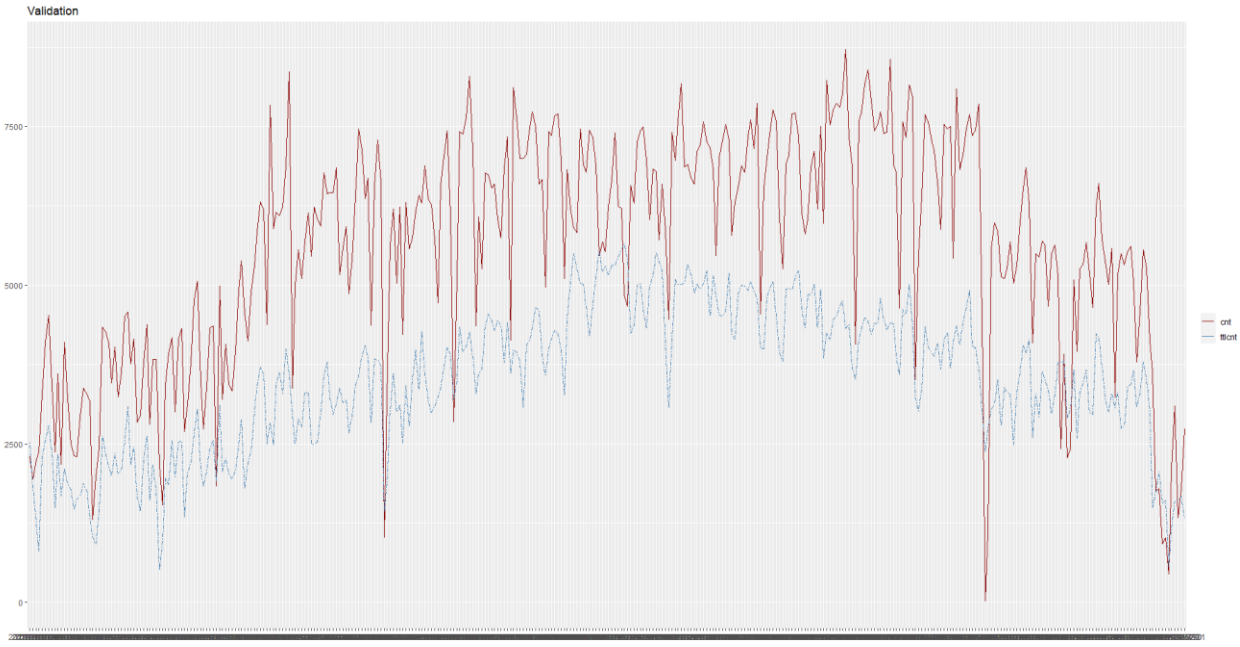
```
# different between of actual casual and predicted data in validation
```



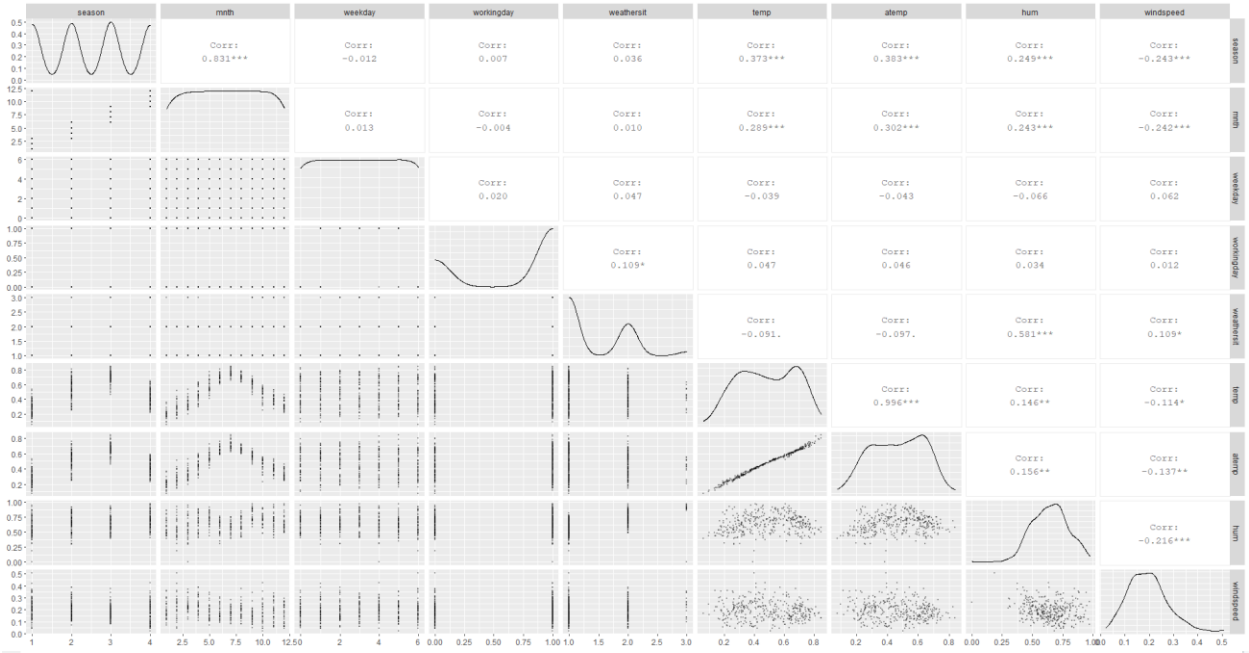



```
# different between of actual registered and predicted data in validation
```

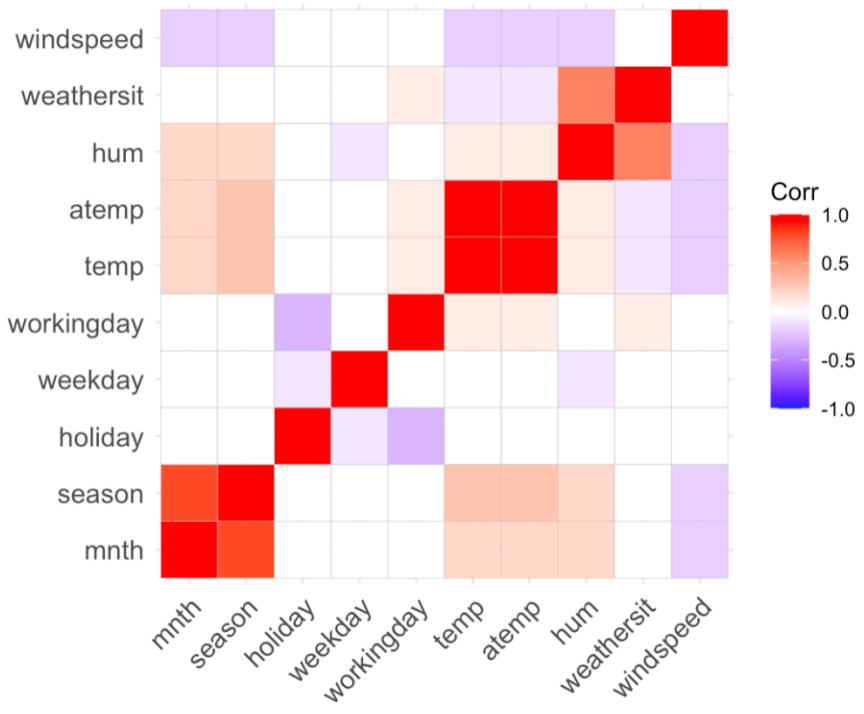





```
# different between of actual total count and predicted data in validation
```




```r
M<-cor(test)
corr <- round(cor(test), 1)
ggcorrplot(corr)
```



 # the scatter matrix plots suggest that temp correlates to atemp strongly and seasons strongly corelates with months.

Those infer that atemp and temp could not appear in model Simultaneously. Similarly, season and month could not appear in model Simultaneously.

that’s why we choose linear model 1 to predict casual.

Since linear model 4 has the smallest relative mean square error, we choose this model to predict registered.