Data 603- Project

Shora Dehkordi, Ryan leeson, Guarav Kumar, Maryam Sarafraz

03/12/2019

```
taxi_data = read.csv ("./taxitrip_sample_df_final.csv")
#head (taxi_data, 4)
#tail (taxi_data, 7)

#   convert day_of_week to a numerical value
transform (taxi_data, day_of_week = as.numeric (day_of_week))

#   Filter for weekend
#   Sunday = 1
#   Saturday = 7
taxi_data$weekend = 1
taxi_data$weekend[ taxi_data$day_of_week > 1 & taxi_data$day_of_week < 6] = 0

#   convert months to a numerical value
transform (taxi_data, months = as.numeric (months))

#   Filtering for season
taxi_data$season = "Winter"   #   Winter
taxi_data$season[taxi_data$months > 2 & taxi_data$months < 6] = "Spring"   #
Spring
taxi_data$season[taxi_data$months > 5 & taxi_data$months < 9] = "Summer"   #
Summer
taxi_data$season[taxi_data$months > 8 & taxi_data$months < 12] = "Fall"   #
Fall

transform (taxi_data, hours = as.numeric (hours))

taxi_data$time_of_day = "Night"   #   Night
taxi_data$time_of_day[taxi_data$hours >= 6 & taxi_data$hours < 12] =
"Morning"   #   Morning
taxi_data$time_of_day[taxi_data$hours >= 12 & taxi_data$hours < 18] =
"Afternoon"   #   Afternoon
taxi_data$time_of_day[taxi_data$hours >= 18 & taxi_data$hours < 24] =
"Evening"   #   Evening

head (taxi_data, 4)
```

Linear model with log transformation

```
#original model
taxi_fulllm_log = lm ( log (fare) ~ factor(payment_type) + factor(company) +
avg_miles + avg_minutes + factor(time_of_day) + factor(season) +
factor(weekend) + factor(hour_type), data = taxi_data)
```

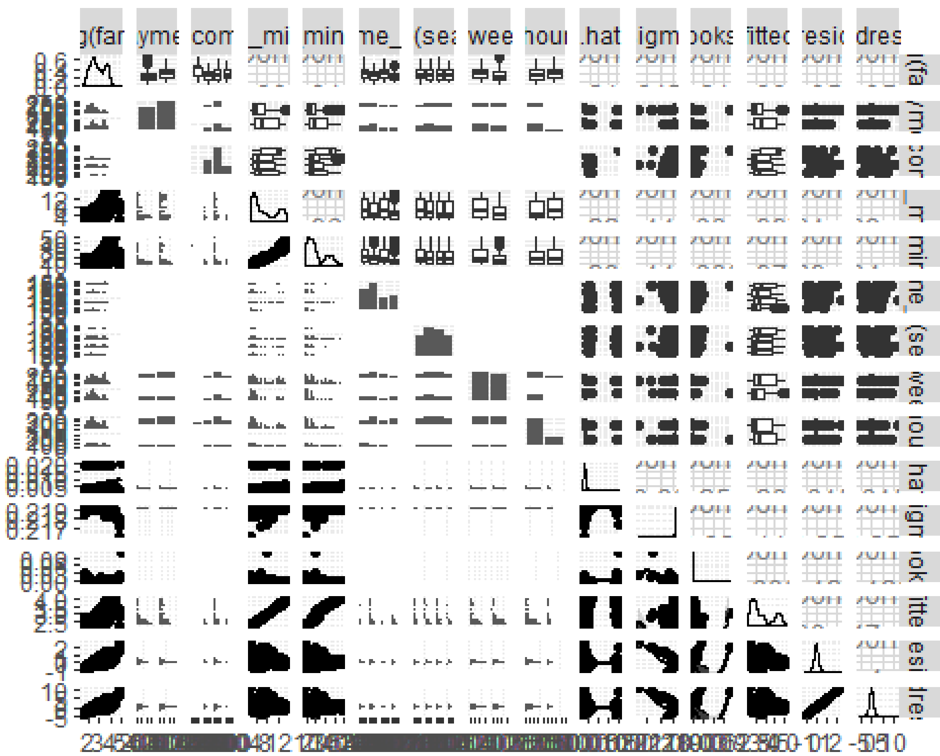
multi-colinearity

```
vif (taxi_fulllm_log)
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## factor(payment_type)   1.065684  1        1.032320
## factor(company)        1.033313  3        1.005477
## avg_miles             13.801971  1        3.715100
## avg_minutes           13.862480  1        3.723235
## factor(time_of_day)    1.221832  3        1.033956
## factor(season)         1.017771  3        1.002940
## factor(weekend)        1.081302  1        1.039857
## factor(hour_type)      1.113469  1        1.055211
```

avg_miles and avg_minutes are co-linear

```
ggpairs (taxi_fulllm_log, lower = list ( continuous = "smooth_loess", combo =
"facethist", discrete = "facetbar", na = "na"), cardinality_threshold = 25)
```



Model varaible testing

```
taxi_fulllm_log_nomin = lm ( log (fare) ~ factor(payment_type) +
factor(company) + avg_miles + factor(time_of_day) + factor(season) +
factor(weekend) + factor(hour_type), data = taxi_data)

taxi_stepw = ols_step_both_p ( taxi_fulllm_log_nomin, pent = 0.05, prem =
0.1, details = FALSE)
```

```
## Stepwise Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. factor(payment_type)
## 2. factor(company)
## 3. avg_miles
## 4. factor(time_of_day)
## 5. factor(season)
## 6. factor(weekend)
## 7. factor(hour_type)
##
## We are selecting variables based on p value...
##
## Variables Entered/Removed:
##
## - avg_miles added
## - factor(hour_type) added
## - factor(company) added
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.933       RMSE               0.223
## R-Squared               0.871       Coef. Var          7.882
## Adj. R-Squared          0.871       MSE                0.050
## Pred R-Squared          0.870       MAE                0.168
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## -----------------------------------------------------------------------
##                 Sum of
##                Squares          DF    Mean Square       F          Sig.
## -----------------------------------------------------------------------
## Regression    1671.611           5        334.322    6738.242    0.0000
## Residual       247.780        4994          0.050
## Total         1919.391        4999
## -----------------------------------------------------------------------
##
##                                           Parameter Estimates
## -------------------------------------------------------------------------
```

```
------------------------------
##                        model       Beta     Std. Error     Std. Beta        t
Sig       lower      upper
## --------------------------------------------------------------------------
------------------------------
##                  (Intercept)     2.033          0.030                  66.939
0.000     1.974      2.093
##                    avg_miles     0.127          0.001         0.931    181.806
0.000     0.126      0.128
## factor(hour_type)rush_hour     0.035          0.008         0.023      4.570
0.000     0.020      0.050
##           factor(company)101    -0.065          0.030        -0.047     -2.150
0.032    -0.124     -0.006
##           factor(company)107    -0.044          0.030        -0.035     -1.455
0.146    -0.102      0.015
##           factor(company)109    -0.052          0.030        -0.033     -1.715
0.086    -0.112      0.007
## --------------------------------------------------------------------------
------------------------------
```

Hour_type, company, and avg_miles are suggested for the model.

```
taxi_formodel = ols_step_forward_p ( taxi_fulllm_log_nomin, pent = 0.05,
details = FALSE)
```

```
## Forward Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. factor(payment_type)
## 2. factor(company)
## 3. avg_miles
## 4. factor(time_of_day)
## 5. factor(season)
## 6. factor(weekend)
## 7. factor(hour_type)
##
## We are selecting variables based on p value...
##
## Variables Entered:
##
## - avg_miles
## - factor(hour_type)
## - factor(company)
##
## No more variables to be added.
##
## Final Model Output
## ------------------
```

```
##
##                         Model Summary
## -------------------------------------------------------------
## R                        0.933       RMSE              0.223
## R-Squared                0.871       Coef. Var         7.882
## Adj. R-Squared           0.871       MSE               0.050
## Pred R-Squared           0.870       MAE               0.168
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## ------------------------------------------------------------------------
##                 Sum of
##                Squares          DF    Mean Square       F          Sig.
## ------------------------------------------------------------------------
## Regression    1671.611           5        334.322    6738.242    0.0000
## Residual       247.780        4994          0.050
## Total         1919.391        4999
## ------------------------------------------------------------------------
##
##                                           Parameter Estimates
## ---------------------------------------------------------------------------
-----------------------------
##                      model      Beta    Std. Error    Std. Beta       t
Sig      lower     upper
## ---------------------------------------------------------------------------
-----------------------------
##                (Intercept)     2.033         0.030                  66.939
0.000     1.974     2.093
##                  avg_miles     0.127         0.001        0.931    181.806
0.000     0.126     0.128
## factor(hour_type)rush_hour     0.035         0.008        0.023      4.570
0.000     0.020     0.050
##          factor(company)101    -0.065         0.030       -0.047     -2.150
0.032    -0.124    -0.006
##          factor(company)107    -0.044         0.030       -0.035     -1.455
0.146    -0.102     0.015
##          factor(company)109    -0.052         0.030       -0.033     -1.715
0.086    -0.112     0.007
## ---------------------------------------------------------------------------
-----------------------------
```

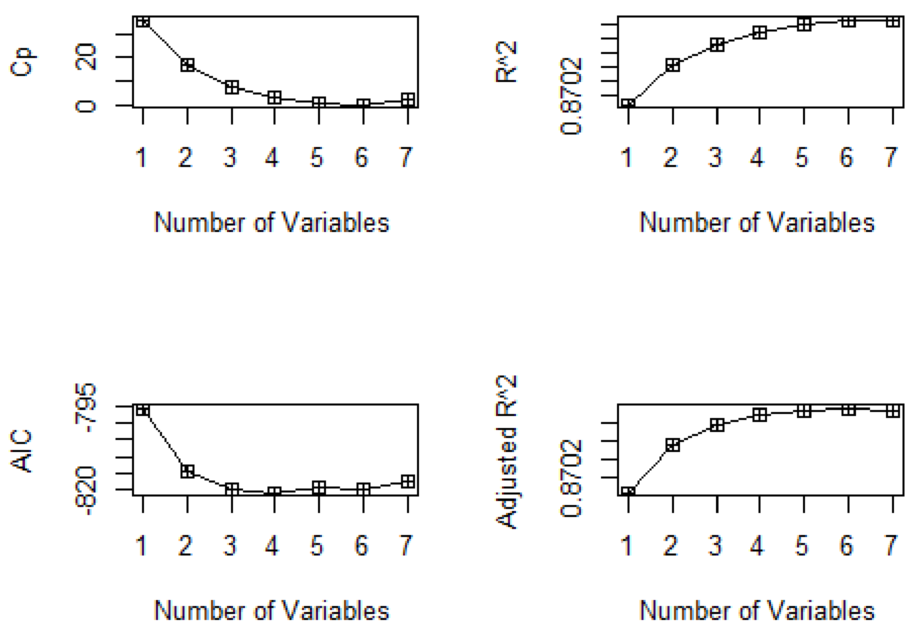
Hour_type, company and avg_miles are suggested for the model.

```
taxi_backmodel = ols_step_backward_p ( taxi_fulllm_log_nomin, prem = 0.05,
details = FALSE)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . factor(payment_type)
## 2 . factor(company)
## 3 . avg_miles
## 4 . factor(time_of_day)
## 5 . factor(season)
## 6 . factor(weekend)
## 7 . factor(hour_type)
##
## We are eliminating variables based on p value...
##
## Variables Removed:
##
## - factor(weekend)
## - factor(season)
## - factor(payment_type)
## - factor(time_of_day)
##
## No more variables satisfy the condition of p value = 0.05
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.933       RMSE              0.223
## R-Squared               0.871       Coef. Var         7.882
## Adj. R-Squared          0.871       MSE               0.050
## Pred R-Squared          0.870       MAE               0.168
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## ------------------------------------------------------------------------
##                 Sum of
##                Squares          DF    Mean Square       F          Sig.
## ------------------------------------------------------------------------
## Regression    1671.611           5        334.322    6738.242    0.0000
## Residual       247.780        4994          0.050
## Total         1919.391        4999
## ------------------------------------------------------------------------
##
##                                           Parameter Estimates
```

```
## ----------------------------------------------------------------------------
------------------------------
##                      model      Beta     Std. Error     Std. Beta        t
Sig       lower      upper
## ----------------------------------------------------------------------------
------------------------------
##                (Intercept)     2.033         0.030                   66.939
0.000     1.974      2.093
##          factor(company)101    -0.065         0.030        -0.047     -2.150
0.032    -0.124     -0.006
##          factor(company)107    -0.044         0.030        -0.035     -1.455
0.146    -0.102      0.015
##          factor(company)109    -0.052         0.030        -0.033     -1.715
0.086    -0.112      0.007
##                   avg_miles     0.127         0.001         0.931    181.806
0.000     0.126      0.128
## factor(hour_type)rush_hour     0.035         0.008         0.023      4.570
0.000     0.020      0.050
## ----------------------------------------------------------------------------
------------------------------
```

Hour_type, company and avg_miles are suggested for the model.




```r
ks_stat2 = data.frame ( c(1, 2, 3, 4, 5, 6, 7), ks$cp, ks$aic, ks$adjr,
ks$rsq)
names (ks_stat2) = c( "Predictors", "CP", "AIC", "Adjusted R^2", "R^2")
ks_stat2
```

```
##   Predictors          CP       AIC Adjusted R^2       R^2
## 1          1 36.39513391 -795.4552    0.8700412 0.8700672
## 2          2 17.23082764 -814.5276    0.8705619 0.8706137
## 3          3  7.88033499 -819.8655    0.8707775 0.8709067
## 4          4  2.87533920 -820.8756    0.8708810 0.8710876
## 5          5  0.23870034 -819.5210    0.8709233 0.8712073
## 6          6  0.08278259 -819.6825    0.8709532 0.8712630
## 7          7  2.00000000 -817.7655    0.8709295 0.8712651
```

Cp (0.08278259) suggests using the six variable model AIC (-820.8756) suggersts using the four variable model Adj.rsq (0.8709532) suggests using the six variable model

```r
best.subset = regsubsets ( log (fare) ~ factor(payment_type) +
factor(company) + avg_miles + factor(time_of_day) + factor(season) +
factor(weekend) + factor(hour_type), data = taxi_data, nv = 10)
summary ( best.subset)
```

```
## Subset selection object
## Call: regsubsets.formula(log(fare) ~ factor(payment_type) +
factor(company) +
##     avg_miles + factor(time_of_day) + factor(season) + factor(weekend) +
##     factor(hour_type), data = taxi_data, nv = 10)
## 13 Variables  (and intercept)
##                                    Forced in Forced out
## factor(payment_type)Credit Card     FALSE      FALSE
## factor(company)101                  FALSE      FALSE
## factor(company)107                  FALSE      FALSE
## factor(company)109                  FALSE      FALSE
## avg_miles                           FALSE      FALSE
## factor(time_of_day)Evening          FALSE      FALSE
## factor(time_of_day)Morning          FALSE      FALSE
## factor(time_of_day)Night            FALSE      FALSE
## factor(season)Spring                FALSE      FALSE
## factor(season)Summer                FALSE      FALSE
## factor(season)Winter                FALSE      FALSE
## factor(weekend)1                    FALSE      FALSE
## factor(hour_type)rush_hour          FALSE      FALSE
## 1 subsets of each size up to 10
## Selection Algorithm: exhaustive
##           factor(payment_type)Credit Card factor(company)101
## 1  ( 1 )  " "                             " "
## 2  ( 1 )  " "                             " "
## 3  ( 1 )  " "                             "*"
## 4  ( 1 )  " "                             "*"
## 5  ( 1 )  " "                             "*"
## 6  ( 1 )  "*"                             "*"
## 7  ( 1 )  "*"                             "*"
## 8  ( 1 )  "*"                             "*"
## 9  ( 1 )  "*"                             "*"
## 10  ( 1 ) "*"                             "*"
```

```
##           factor(company)107 factor(company)109 avg_miles
## 1  ( 1 )  " "                " "                "*"
## 2  ( 1 )  " "                " "                "*"
## 3  ( 1 )  " "                " "                "*"
## 4  ( 1 )  " "                " "                "*"
## 5  ( 1 )  " "                " "                "*"
## 6  ( 1 )  " "                " "                "*"
## 7  ( 1 )  " "                " "                "*"
## 8  ( 1 )  "*"                "*"                "*"
## 9  ( 1 )  "*"                "*"                "*"
## 10  ( 1 ) "*"                "*"                "*"
##           factor(time_of_day)Evening factor(time_of_day)Morning
## 1  ( 1 )  " "                        " "
## 2  ( 1 )  " "                        " "
## 3  ( 1 )  " "                        " "
## 4  ( 1 )  " "                        " "
## 5  ( 1 )  " "                        " "
## 6  ( 1 )  " "                        " "
## 7  ( 1 )  "*"                        " "
## 8  ( 1 )  " "                        " "
## 9  ( 1 )  "*"                        " "
## 10  ( 1 ) "*"                        "*"
##           factor(time_of_day)Night factor(season)Spring
## 1  ( 1 )  " "                      " "
## 2  ( 1 )  " "                      " "
## 3  ( 1 )  " "                      " "
## 4  ( 1 )  "*"                      " "
## 5  ( 1 )  "*"                      " "
## 6  ( 1 )  "*"                      " "
## 7  ( 1 )  "*"                      " "
## 8  ( 1 )  "*"                      " "
## 9  ( 1 )  "*"                      " "
## 10  ( 1 ) "*"                      " "
##           factor(season)Summer factor(season)Winter factor(weekend)1
## 1  ( 1 )  " "                  " "                  " "
## 2  ( 1 )  " "                  " "                  " "
## 3  ( 1 )  " "                  " "                  " "
## 4  ( 1 )  " "                  " "                  " "
## 5  ( 1 )  "*"                  " "                  " "
## 6  ( 1 )  "*"                  " "                  " "
## 7  ( 1 )  "*"                  " "                  " "
## 8  ( 1 )  "*"                  " "                  " "
## 9  ( 1 )  "*"                  " "                  " "
## 10  ( 1 ) "*"                  " "                  " "
##           factor(hour_type)rush_hour
## 1  ( 1 )  " "
## 2  ( 1 )  "*"
## 3  ( 1 )  "*"
## 4  ( 1 )  "*"
## 5  ( 1 )  "*"
```

```
## 6  ( 1 )  "*"
## 7  ( 1 )  "*"
## 8  ( 1 )  "*"
## 9  ( 1 )  "*"
## 10  ( 1 ) "*"

reg.summary = summary ( best.subset)
```

Four variables: company, avg_miles, time_of_day, hour_type Six variables: company, avg_miles, time_of_day, hour_type, payment_type, season

```
summary (taxi_fulllm_log_nomin)

##
## Call:
## lm(formula = log(fare) ~ factor(payment_type) + factor(company) +
##     avg_miles + factor(time_of_day) + factor(season) + factor(weekend) +
##     factor(hour_type), data = taxi_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12071 -0.14201  0.00246  0.13667  2.46221
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.0443151  0.0319170  64.051  < 2e-16 ***
## factor(payment_type)Credit Card  0.0095212  0.0065460   1.455 0.145869
## factor(company)101           -0.0623113  0.0302515  -2.060 0.039472 *
## factor(company)107           -0.0414008  0.0299412  -1.383 0.166807
## factor(company)109           -0.0507664  0.0304879  -1.665 0.095949 .
## avg_miles                     0.1263960  0.0007347 172.031  < 2e-16 ***
## factor(time_of_day)Evening   -0.0126504  0.0081505  -1.552 0.120701
## factor(time_of_day)Morning   -0.0071164  0.0096406  -0.738 0.460448
## factor(time_of_day)Night     -0.0260796  0.0102921  -2.534 0.011309 *
## factor(season)Spring         -0.0079156  0.0090413  -0.875 0.381350
## factor(season)Summer          0.0083958  0.0092650   0.906 0.364879
## factor(season)Winter         -0.0067953  0.0097145  -0.699 0.484272
## factor(weekend)1             -0.0018860  0.0065550  -0.288 0.773573
## factor(hour_type)rush_hour    0.0284062  0.0080691   3.520 0.000435 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2226 on 4986 degrees of freedom
## Multiple R-squared:  0.8713, Adjusted R-squared:  0.8709
## F-statistic:  2596 on 13 and 4986 DF,  p-value: < 2.2e-16
```

company, time_of_day, hour_type, avg_miles are significant

Models

```
#   four variable model
taxi_lm_red_4 = lm ( log (fare) ~ factor(company) + avg_miles +
```

```r
factor(time_of_day) + factor(hour_type), data = taxi_data)

#   three variable model
taxi_lm_red_3 = lm ( log (fare) ~ factor(company) + avg_miles +
factor(hour_type), data = taxi_data)

#   six variable model
taxi_lm_red_6 = lm ( log (fare) ~ factor(company) + avg_miles +
factor(time_of_day) + factor(hour_type) + factor(payment_type) +
factor(season), data = taxi_data)

taxi_fulllm_log = lm ( log (fare) ~ factor(payment_type) + factor(company) +
avg_miles + avg_minutes + factor(time_of_day) + factor(season) +
factor(weekend) + factor(hour_type), data = taxi_data)
```

Partial F-test

```r
#   full and 6 variables
anova (taxi_fulllm_log_nomin, taxi_lm_red_6)

## Analysis of Variance Table
##
## Model 1: log(fare) ~ factor(payment_type) + factor(company) + avg_miles +
##     factor(time_of_day) + factor(season) + factor(weekend) +
##     factor(hour_type)
## Model 2: log(fare) ~ factor(company) + avg_miles + factor(time_of_day) +
##     factor(hour_type) + factor(payment_type) + factor(season)
##   Res.Df    RSS Df  Sum of Sq      F Pr(>F)
## 1   4986 247.09
## 2   4987 247.10 -1 -0.0041025 0.0828 0.7736

#   full and 4 variables
anova (taxi_fulllm_log_nomin, taxi_lm_red_4)

## Analysis of Variance Table
##
## Model 1: log(fare) ~ factor(payment_type) + factor(company) + avg_miles +
##     factor(time_of_day) + factor(season) + factor(weekend) +
##     factor(hour_type)
## Model 2: log(fare) ~ factor(company) + avg_miles + factor(time_of_day) +
##     factor(hour_type)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   4986 247.09
## 2   4991 247.43 -5  -0.34072 1.3751 0.2303

#   full and 3 variables
anova (taxi_fulllm_log_nomin, taxi_lm_red_3)

## Analysis of Variance Table
##
## Model 1: log(fare) ~ factor(payment_type) + factor(company) + avg_miles +
##     factor(time_of_day) + factor(season) + factor(weekend) +
```

```
##     factor(hour_type)
## Model 2: log(fare) ~ factor(company) + avg_miles + factor(hour_type)
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1   4986 247.09
## 2   4994 247.78 -8  -0.68787 1.735 0.08523 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary (taxi_lm_red_4)
```

```
##
## Call:
## lm(formula = log(fare) ~ factor(company) + avg_miles + factor(time_of_day)
+
##     factor(hour_type), data = taxi_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12536 -0.14200  0.00203  0.13620  2.46892
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.0465634  0.0309846  66.051  < 2e-16 ***
## factor(company)101          -0.0634126  0.0302420  -2.097 0.036058 *
## factor(company)107          -0.0427242  0.0299361  -1.427 0.153591
## factor(company)109          -0.0519230  0.0304801  -1.704 0.088536 .
## avg_miles                    0.1266584  0.0007101 178.371  < 2e-16 ***
## factor(time_of_day)Evening -0.0125143  0.0081370  -1.538 0.124126
## factor(time_of_day)Morning -0.0068521  0.0096192  -0.712 0.476293
## factor(time_of_day)Night   -0.0262982  0.0101465  -2.592 0.009574 **
## factor(hour_type)rush_hour  0.0284734  0.0080591   3.533 0.000414 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2227 on 4991 degrees of freedom
## Multiple R-squared:  0.8711, Adjusted R-squared:  0.8709
## F-statistic:  4216 on 8 and 4991 DF,  p-value: < 2.2e-16
```

```
anova (taxi_fulllm_log_nomin, taxi_lm_red_3)
```


```
## Analysis of Variance Table
##
## Model 1: log(fare) ~ factor(payment_type) + factor(company) + avg_miles +
##     factor(time_of_day) + factor(season) + factor(weekend) +
##     factor(hour_type)
## Model 2: log(fare) ~ factor(company) + avg_miles + factor(hour_type)
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1   4986 247.09
## 2   4994 247.78 -8  -0.68787 1.735 0.08523 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-tests suggest that the three variable model is not significantly different from the full model. So, four variables can be removed from the model.

```
summary (taxi_lm_red_3)

##
## Call:
## lm(formula = log(fare) ~ factor(company) + avg_miles + factor(hour_type),
##     data = taxi_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12493 -0.14283  0.00216  0.13761  2.47551
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.0333172  0.0303755  66.939   <2e-16 ***
## factor(company)101        -0.0650179  0.0302379  -2.150   0.0316 *
## factor(company)107        -0.0435584  0.0299347  -1.455   0.1457
## factor(company)109        -0.0522789  0.0304849  -1.715   0.0864 .
## avg_miles                  0.1269968  0.0006985 181.806   <2e-16 ***
## factor(hour_type)rush_hour  0.0350409  0.0076679   4.570    5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2227 on 4994 degrees of freedom
## Multiple R-squared:  0.8709, Adjusted R-squared:  0.8708
## F-statistic:  6738 on 5 and 4994 DF,  p-value: < 2.2e-16

taxi_lm_red_2 = lm ( log (fare) ~ avg_miles + factor(hour_type), data =
taxi_data)
anova (taxi_fulllm_log_nomin, taxi_lm_red_2)

## Analysis of Variance Table
##
## Model 1: log(fare) ~ factor(payment_type) + factor(company) + avg_miles +
##     factor(time_of_day) + factor(season) + factor(weekend) +
##     factor(hour_type)
## Model 2: log(fare) ~ avg_miles + factor(hour_type)
##   Res.Df    RSS  Df Sum of Sq      F   Pr(>F)
## 1   4986 247.09
## 2   4997 248.34 -11   -1.2504 2.2937 0.008551 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Company must stay in the model because p-value < 0.05

So, the three variable model is the best reduced model.

```
print ("Adj. R2")
```

```
## [1] "Adj. R2"

summary (taxi_lm_red_3)$adj.r.sq

## [1] 0.8707775

print ("RMSE")

## [1] "RMSE"

sigma (taxi_lm_red_3)

## [1] 0.2227457
```

Interactions

```
taxi_lm_red_3_int = lm ( log (fare) ~ (factor(company) + avg_miles +
factor(hour_type)) ^2, data = taxi_data)
summary (taxi_lm_red_3_int)

##
## Call:
## lm(formula = log(fare) ~ (factor(company) + avg_miles +
factor(hour_type))^2,
##     data = taxi_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12838 -0.13917  0.00279  0.13683  2.48401
##
## Coefficients:
##                                               Estimate Std. Error
## (Intercept)                                  1.8309688  0.0779500
## factor(company)101                           0.1575354  0.0786259
## factor(company)107                           0.1462659  0.0782790
## factor(company)109                           0.1652790  0.0791152
## avg_miles                                    0.1459032  0.0067757
## factor(hour_type)rush_hour                   0.0780647  0.0766069
## factor(company)101:avg_miles                -0.0218820  0.0069033
## factor(company)107:avg_miles                -0.0167747  0.0068376
## factor(company)109:avg_miles                -0.0218591  0.0069511
## factor(company)101:factor(hour_type)rush_hour -0.0525482  0.0766263
## factor(company)107:factor(hour_type)rush_hour -0.0511805  0.0758327
## factor(company)109:factor(hour_type)rush_hour -0.0222869  0.0769850
## avg_miles:factor(hour_type)rush_hour         0.0002865  0.0016437
##                                             t value Pr(>|t|)
## (Intercept)                                  23.489  < 2e-16 ***
## factor(company)101                            2.004  0.04517 *
## factor(company)107                            1.869  0.06175 .
## factor(company)109                            2.089  0.03675 *
## avg_miles                                    21.533  < 2e-16 ***
## factor(hour_type)rush_hour                    1.019  0.30824
```

```
## factor(company)101:avg_miles                      -3.170  0.00153 **
## factor(company)107:avg_miles                      -2.453  0.01419 *
## factor(company)109:avg_miles                      -3.145  0.00167 **
## factor(company)101:factor(hour_type)rush_hour  -0.686  0.49289
## factor(company)107:factor(hour_type)rush_hour  -0.675  0.49976
## factor(company)109:factor(hour_type)rush_hour  -0.289  0.77221
## avg_miles:factor(hour_type)rush_hour             0.174  0.86163
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2224 on 4987 degrees of freedom
## Multiple R-squared:  0.8715, Adjusted R-squared:  0.8712
## F-statistic:  2819 on 12 and 4987 DF,  p-value: < 2.2e-16
```

```
anova (taxi_lm_red_3_int, taxi_lm_red_3)
```

```
## Analysis of Variance Table
##
## Model 1: log(fare) ~ (factor(company) + avg_miles + factor(hour_type))^2
## Model 2: log(fare) ~ factor(company) + avg_miles + factor(hour_type)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   4987 246.62
## 2   4994 247.78 -7   -1.1635 3.3612 0.001403 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

At least one of the interactions are significant.

The individual t-tests suggest only the interaction between avg_miles and company is significant.

```
taxi_lm_red_3_int_red = lm ( log (fare) ~ factor(company) + avg_miles +
factor(hour_type) + avg_miles*factor(company), data = taxi_data)
summary (taxi_lm_red_3_int_red)
```

```
##
## Call:
## lm(formula = log(fare) ~ factor(company) + avg_miles + factor(hour_type) +
##     avg_miles * factor(company), data = taxi_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12689 -0.13809  0.00289  0.13628  2.48544
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.843278   0.075028  24.568  < 2e-16 ***
## factor(company)101              0.143579   0.075720   1.896  0.05799 .
## factor(company)107              0.132462   0.075384   1.757  0.07895 .
## factor(company)109              0.157893   0.076166   2.073  0.03822 *
## avg_miles                       0.145565   0.006739  21.602  < 2e-16 ***
```

```
## factor(hour_type)rush_hour    0.034894   0.007657   4.557 5.31e-06 ***
## factor(company)101:avg_miles -0.021517   0.006873  -3.131  0.00175 **
## factor(company)107:avg_miles -0.016405   0.006807  -2.410  0.01598 *
## factor(company)109:avg_miles -0.021425   0.006921  -3.096  0.00197 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2224 on 4991 degrees of freedom
## Multiple R-squared:  0.8714, Adjusted R-squared:  0.8712
## F-statistic:  4229 on 8 and 4991 DF,  p-value: < 2.2e-16
```

```
anova (taxi_lm_red_3_int_red, taxi_lm_red_3_int)
```

```
## Analysis of Variance Table
##
## Model 1: log(fare) ~ factor(company) + avg_miles + factor(hour_type) +
##     avg_miles * factor(company)
## Model 2: log(fare) ~ (factor(company) + avg_miles + factor(hour_type))^2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   4991 246.75
## 2   4987 246.62  4   0.13814 0.6983  0.593
```

The partial F-test suggests the interactions for company and hour_type, and hour_type and avg_miles are insignificant.

```
anova (taxi_lm_red_3_int_red, taxi_lm_red_3)
```

```
## Analysis of Variance Table
##
## Model 1: log(fare) ~ factor(company) + avg_miles + factor(hour_type) +
##     avg_miles * factor(company)
## Model 2: log(fare) ~ factor(company) + avg_miles + factor(hour_type)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4991 246.75
## 2   4994 247.78 -3   -1.0254 6.9133 0.0001216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The partial F-test suggests the interaction between avg_miles and company is significant.

```
summary (taxi_lm_red_3_int_red)
```

```
##
## Call:
## lm(formula = log(fare) ~ factor(company) + avg_miles + factor(hour_type) +
##     avg_miles * factor(company), data = taxi_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12689 -0.13809  0.00289  0.13628  2.48544
##
```

```
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.843278   0.075028  24.568  < 2e-16 ***
## factor(company)101            0.143579   0.075720   1.896  0.05799 .
## factor(company)107            0.132462   0.075384   1.757  0.07895 .
## factor(company)109            0.157893   0.076166   2.073  0.03822 *
## avg_miles                     0.145565   0.006739  21.602  < 2e-16 ***
## factor(hour_type)rush_hour    0.034894   0.007657   4.557 5.31e-06 ***
## factor(company)101:avg_miles -0.021517   0.006873  -3.131  0.00175 **
## factor(company)107:avg_miles -0.016405   0.006807  -2.410  0.01598 *
## factor(company)109:avg_miles -0.021425   0.006921  -3.096  0.00197 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2224 on 4991 degrees of freedom
## Multiple R-squared:  0.8714, Adjusted R-squared:  0.8712
## F-statistic:  4229 on 8 and 4991 DF,  p-value: < 2.2e-16
```

```
print ("Adj. R2")
```

```
## [1] "Adj. R2"
```

```
summary (taxi_lm_red_3_int_red)$adj.r.sq
```

```
## [1] 0.8712349
```

```
print ("RMSE")
```

```
## [1] "RMSE"
```

```
sigma (taxi_lm_red_3_int_red)
```

```
## [1] 0.2223511
```

Higher Orders

```
taxi_lm_red_3_int_red_12o = lm ( log (fare) ~ factor(company) + poly
(avg_miles, 12, raw = TRUE) + factor(hour_type) + avg_miles*factor(company),
data = taxi_data)
summary (taxi_lm_red_3_int_red_12o)
```

```
##
## Call:
## lm(formula = log(fare) ~ factor(company) + poly(avg_miles, 12,
##     raw = TRUE) + factor(hour_type) + avg_miles * factor(company),
##     data = taxi_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23914 -0.11432 -0.00574  0.11377  2.60994
##
## Coefficients: (1 not defined because of singularities)
##                                       Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                         -1.533e+01  3.993e+00  -3.839 0.000125
## factor(company)101                   2.446e-01  6.926e-02   3.532 0.000415
## factor(company)107                   2.368e-01  6.896e-02   3.434 0.000599
## factor(company)109                   2.499e-01  6.966e-02   3.587 0.000337
## poly(avg_miles, 12, raw = TRUE)1     4.525e+01  1.020e+01   4.437 9.34e-06
## poly(avg_miles, 12, raw = TRUE)2    -5.128e+01  1.112e+01  -4.610 4.13e-06
## poly(avg_miles, 12, raw = TRUE)3     3.286e+01  6.870e+00   4.783 1.77e-06
## poly(avg_miles, 12, raw = TRUE)4    -1.325e+01  2.686e+00  -4.932 8.42e-07
## poly(avg_miles, 12, raw = TRUE)5     3.559e+00  7.033e-01   5.060 4.34e-07
## poly(avg_miles, 12, raw = TRUE)6    -6.576e-01  1.271e-01  -5.175 2.37e-07
## poly(avg_miles, 12, raw = TRUE)7     8.471e-02  1.605e-02   5.278 1.36e-07
## poly(avg_miles, 12, raw = TRUE)8    -7.585e-03  1.412e-03  -5.374 8.07e-08
## poly(avg_miles, 12, raw = TRUE)9     4.626e-04  8.471e-05   5.461 4.97e-08
## poly(avg_miles, 12, raw = TRUE)10 -1.831e-05  3.305e-06  -5.540 3.19e-08
## poly(avg_miles, 12, raw = TRUE)11  4.236e-07  7.550e-08   5.610 2.13e-08
## poly(avg_miles, 12, raw = TRUE)12 -4.346e-09  7.662e-10  -5.673 1.48e-08
## factor(hour_type)rush_hour           4.109e-02  7.008e-03   5.864 4.81e-09
## avg_miles                                   NA         NA      NA       NA
## factor(company)101:avg_miles        -2.936e-02  6.296e-03  -4.664 3.19e-06
## factor(company)107:avg_miles        -2.487e-02  6.237e-03  -3.987 6.78e-05
## factor(company)109:avg_miles        -2.839e-02  6.336e-03  -4.480 7.62e-06
##
## (Intercept)                         ***
## factor(company)101                   ***
## factor(company)107                   ***
## factor(company)109                   ***
## poly(avg_miles, 12, raw = TRUE)1    ***
## poly(avg_miles, 12, raw = TRUE)2    ***
## poly(avg_miles, 12, raw = TRUE)3    ***
## poly(avg_miles, 12, raw = TRUE)4    ***
## poly(avg_miles, 12, raw = TRUE)5    ***
## poly(avg_miles, 12, raw = TRUE)6    ***
## poly(avg_miles, 12, raw = TRUE)7    ***
## poly(avg_miles, 12, raw = TRUE)8    ***
## poly(avg_miles, 12, raw = TRUE)9    ***
## poly(avg_miles, 12, raw = TRUE)10 ***
## poly(avg_miles, 12, raw = TRUE)11 ***
## poly(avg_miles, 12, raw = TRUE)12 ***
## factor(hour_type)rush_hour           ***
## avg_miles
## factor(company)101:avg_miles        ***
## factor(company)107:avg_miles        ***
## factor(company)109:avg_miles        ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2029 on 4980 degrees of freedom
## Multiple R-squared:  0.8932, Adjusted R-squared:  0.8928
## F-statistic:  2192 on 19 and 4980 DF,  p-value: < 2.2e-16
```

```
taxi_lm_red_3_int_red_8o = lm ( log (fare) ~ factor(company) + poly
(avg_miles, 8, raw = TRUE) + factor(hour_type) + avg_miles*factor(company),
data = taxi_data)
summary (taxi_lm_red_3_int_red_8o)

##
## Call:
## lm(formula = log(fare) ~ factor(company) + poly(avg_miles, 8,
##     raw = TRUE) + factor(hour_type) + avg_miles * factor(company),
##     data = taxi_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23493 -0.11576 -0.00942  0.11663  2.62346
##
## Coefficients: (1 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      3.012e+00  3.554e-01   8.474  < 2e-16 ***
## factor(company)101               2.463e-01  6.950e-02   3.544 0.000398 ***
## factor(company)107               2.389e-01  6.920e-02   3.453 0.000559 ***
## factor(company)109               2.523e-01  6.989e-02   3.610 0.000310 ***
## poly(avg_miles, 8, raw = TRUE)1 -2.606e+00  5.666e-01  -4.599 4.35e-06 ***
## poly(avg_miles, 8, raw = TRUE)2  1.983e+00  3.669e-01   5.405 6.79e-08 ***
## poly(avg_miles, 8, raw = TRUE)3 -6.964e-01  1.239e-01  -5.619 2.03e-08 ***
## poly(avg_miles, 8, raw = TRUE)4  1.387e-01  2.406e-02   5.764 8.69e-09 ***
## poly(avg_miles, 8, raw = TRUE)5 -1.644e-02  2.776e-03  -5.922 3.40e-09 ***
## poly(avg_miles, 8, raw = TRUE)6  1.145e-03  1.877e-04   6.103 1.12e-09 ***
## poly(avg_miles, 8, raw = TRUE)7 -4.320e-05  6.853e-06  -6.303 3.17e-10 ***
## poly(avg_miles, 8, raw = TRUE)8  6.792e-07  1.043e-07   6.511 8.20e-11 ***
## factor(hour_type)rush_hour       4.292e-02  7.021e-03   6.114 1.05e-09 ***
## avg_miles                               NA         NA      NA       NA
## factor(company)101:avg_miles    -3.003e-02  6.317e-03  -4.753 2.06e-06 ***
## factor(company)107:avg_miles    -2.563e-02  6.256e-03  -4.096 4.27e-05 ***
## factor(company)109:avg_miles    -2.903e-02  6.357e-03  -4.567 5.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2036 on 4984 degrees of freedom
## Multiple R-squared:  0.8924, Adjusted R-squared:  0.892
## F-statistic:  2755 on 15 and 4984 DF,  p-value: < 2.2e-16

taxi_lm_red_3_int_red_7o = lm ( log (fare) ~ factor(company) + poly
(avg_miles, 7, raw = TRUE) + factor(hour_type) + avg_miles*factor(company),
data = taxi_data)
summary (taxi_lm_red_3_int_red_7o)

##
## Call:
## lm(formula = log(fare) ~ factor(company) + poly(avg_miles, 7,
##     raw = TRUE) + factor(hour_type) + avg_miles * factor(company),
```

```
##     data = taxi_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26132 -0.11499 -0.00873  0.12154  2.60750
##
## Coefficients: (1 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.137e+00  2.093e-01   5.435 5.73e-08 ***
## factor(company)101               2.589e-01  6.976e-02   3.712 0.000208 ***
## factor(company)107               2.522e-01  6.945e-02   3.631 0.000285 ***
## factor(company)109               2.669e-01  7.015e-02   3.805 0.000143 ***
## poly(avg_miles, 7, raw = TRUE)1  6.165e-01  2.770e-01   2.225 0.026093 *
## poly(avg_miles, 7, raw = TRUE)2 -2.014e-01  1.492e-01  -1.350 0.177026
## poly(avg_miles, 7, raw = TRUE)3  6.684e-02  4.042e-02   1.653 0.098300 .
## poly(avg_miles, 7, raw = TRUE)4 -1.301e-02  6.026e-03  -2.159 0.030862 *
## poly(avg_miles, 7, raw = TRUE)5  1.343e-03  4.995e-04   2.689 0.007201 **
## poly(avg_miles, 7, raw = TRUE)6 -6.844e-05  2.155e-05  -3.176 0.001501 **
## poly(avg_miles, 7, raw = TRUE)7  1.357e-06  3.766e-07   3.603 0.000318 ***
## factor(hour_type)rush_hour       4.367e-02  7.049e-03   6.195 6.31e-10 ***
## avg_miles                               NA         NA      NA       NA
## factor(company)101:avg_miles    -3.182e-02  6.338e-03  -5.020 5.33e-07 ***
## factor(company)107:avg_miles    -2.751e-02  6.276e-03  -4.383 1.19e-05 ***
## factor(company)109:avg_miles    -3.091e-02  6.377e-03  -4.847 1.29e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2044 on 4985 degrees of freedom
## Multiple R-squared:  0.8914, Adjusted R-squared:  0.8911
## F-statistic:  2924 on 14 and 4985 DF,  p-value: < 2.2e-16
```

```
taxi_lm_red_3_int_red_2o = lm ( log (fare) ~ factor(company) + poly
(avg_miles, 2, raw = TRUE) + factor(hour_type) + avg_miles*factor(company),
data = taxi_data)
summary (taxi_lm_red_3_int_red_2o)
```

```
##
## Call:
## lm(formula = log(fare) ~ factor(company) + poly(avg_miles, 2,
##     raw = TRUE) + factor(hour_type) + avg_miles * factor(company),
##     data = taxi_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2222 -0.1224 -0.0137  0.1304  2.5659
##
## Coefficients: (1 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.5173916  0.0716373  21.182  < 2e-16 ***
## factor(company)101               0.2378765  0.0712525   3.338 0.000848 ***
```

```
## factor(company)107               0.2296931  0.0709425   3.238 0.001213 **
## factor(company)109               0.2429038  0.0716529   3.390 0.000704 ***
## poly(avg_miles, 2, raw = TRUE)1  0.2452783  0.0074251  33.034  < 2e-16 ***
## poly(avg_miles, 2, raw = TRUE)2 -0.0056156  0.0002183 -25.720  < 2e-16 ***
## factor(hour_type)rush_hour       0.0394301  0.0071978   5.478 4.51e-08 ***
## avg_miles                               NA         NA      NA       NA
## factor(company)101:avg_miles    -0.0315514  0.0064708  -4.876 1.12e-06 ***
## factor(company)107:avg_miles    -0.0270061  0.0064101  -4.213 2.56e-05 ***
## factor(company)109:avg_miles    -0.0303784  0.0065129  -4.664 3.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.209 on 4990 degrees of freedom
## Multiple R-squared:  0.8865, Adjusted R-squared:  0.8863
## F-statistic:  4330 on 9 and 4990 DF,  p-value: < 2.2e-16
```

```
taxi_lm_red_3_int_red_4o = lm ( log (fare) ~ factor(company) + poly
(avg_miles, 4, raw = TRUE) + factor(hour_type) + avg_miles*factor(company),
data = taxi_data)
summary (taxi_lm_red_3_int_red_4o)
```

```
##
## Call:
## lm(formula = log(fare) ~ factor(company) + poly(avg_miles, 4,
##     raw = TRUE) + factor(hour_type) + avg_miles * factor(company),
##     data = taxi_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23919 -0.11706 -0.01297  0.12316  2.60906
##
## Coefficients: (1 not defined because of singularities)
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.102e+00  7.864e-02  14.018  < 2e-16 ***
## factor(company)101              2.600e-01  7.012e-02   3.708 0.000211 ***
## factor(company)107              2.521e-01  6.982e-02   3.611 0.000307 ***
## factor(company)109              2.680e-01  7.052e-02   3.801 0.000146 ***
## poly(avg_miles, 4, raw = TRUE)1  5.289e-01  2.647e-02  19.983  < 2e-16 ***
## poly(avg_miles, 4, raw = TRUE)2 -6.579e-02  5.950e-03 -11.057  < 2e-16 ***
## poly(avg_miles, 4, raw = TRUE)3  4.803e-03  5.352e-04   8.974  < 2e-16 ***
## poly(avg_miles, 4, raw = TRUE)4 -1.282e-04  1.634e-05  -7.849 5.12e-15 ***
## factor(hour_type)rush_hour       4.275e-02  7.087e-03   6.033 1.73e-09 ***
## avg_miles                               NA         NA      NA       NA
## factor(company)101:avg_miles    -3.304e-02  6.367e-03  -5.190 2.19e-07 ***
## factor(company)107:avg_miles    -2.830e-02  6.307e-03  -4.487 7.38e-06 ***
## factor(company)109:avg_miles    -3.201e-02  6.409e-03  -4.995 6.08e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2056 on 4988 degrees of freedom
```

```
## Multiple R-squared:  0.8902, Adjusted R-squared:  0.8899
## F-statistic:  3676 on 11 and 4988 DF,  p-value: < 2.2e-16
```

```
anova (taxi_lm_red_3_int_red_8o, taxi_lm_red_3_int_red_12o)
```

```
## Analysis of Variance Table
##
## Model 1: log(fare) ~ factor(company) + poly(avg_miles, 8, raw = TRUE) +
##     factor(hour_type) + avg_miles * factor(company)
## Model 2: log(fare) ~ factor(company) + poly(avg_miles, 12, raw = TRUE) +
##     factor(hour_type) + avg_miles * factor(company)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4984 206.59
## 2   4980 204.98  4     1.617 9.8214 6.52e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print ("Adj. R2")
```

```
## [1] "Adj. R2"
```

```
summary (taxi_lm_red_3_int_red)$adj.r.sq
```

```
## [1] 0.8712349
```

```
print ("RMSE")
```

```
## [1] "RMSE"
```

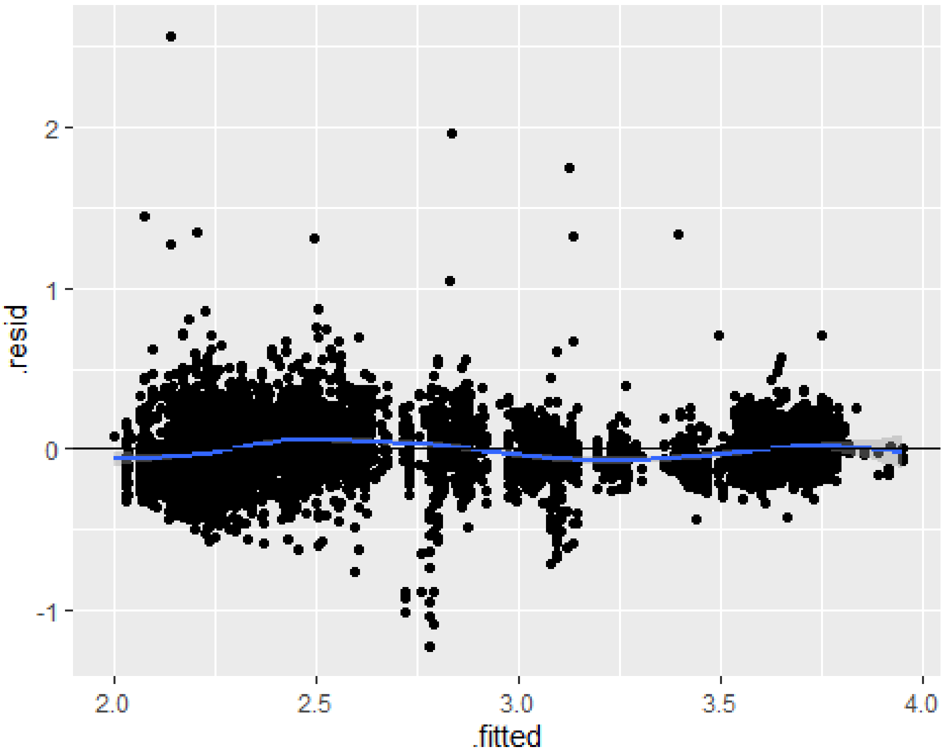
```
sigma (taxi_lm_red_3_int_red)
```

```
## [1] 0.2223511
```

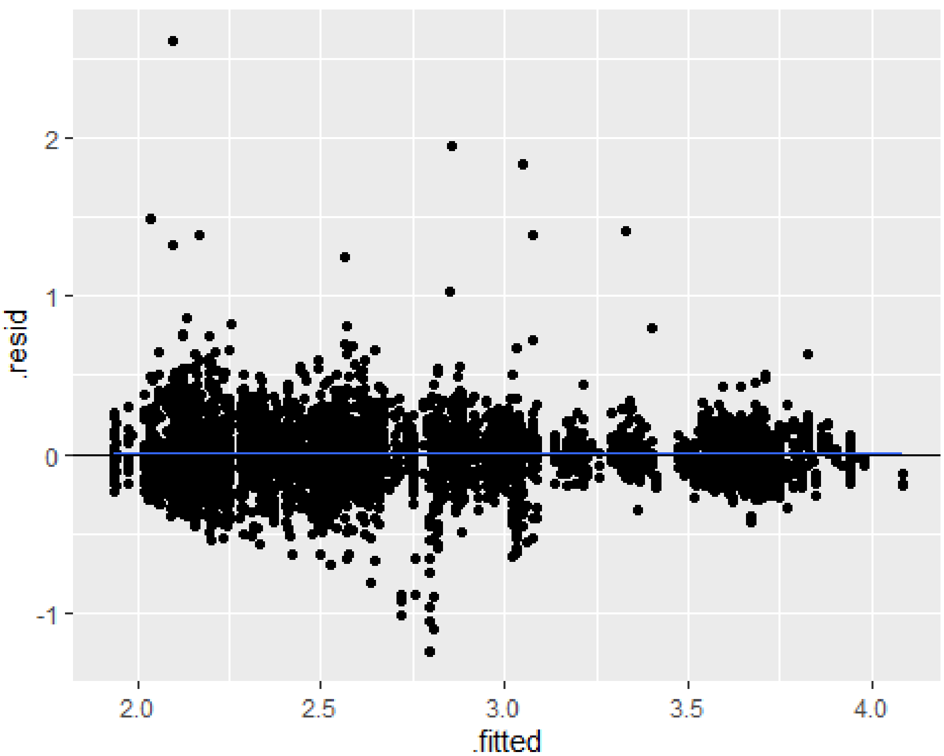
Test of assumptions

```
#   2nd order model
ggplot (taxi_lm_red_3_int_red_2o, aes ( x = .fitted, y = .resid)) +
  geom_point () + geom_smooth () +
  geom_hline (yintercept = 0)
```

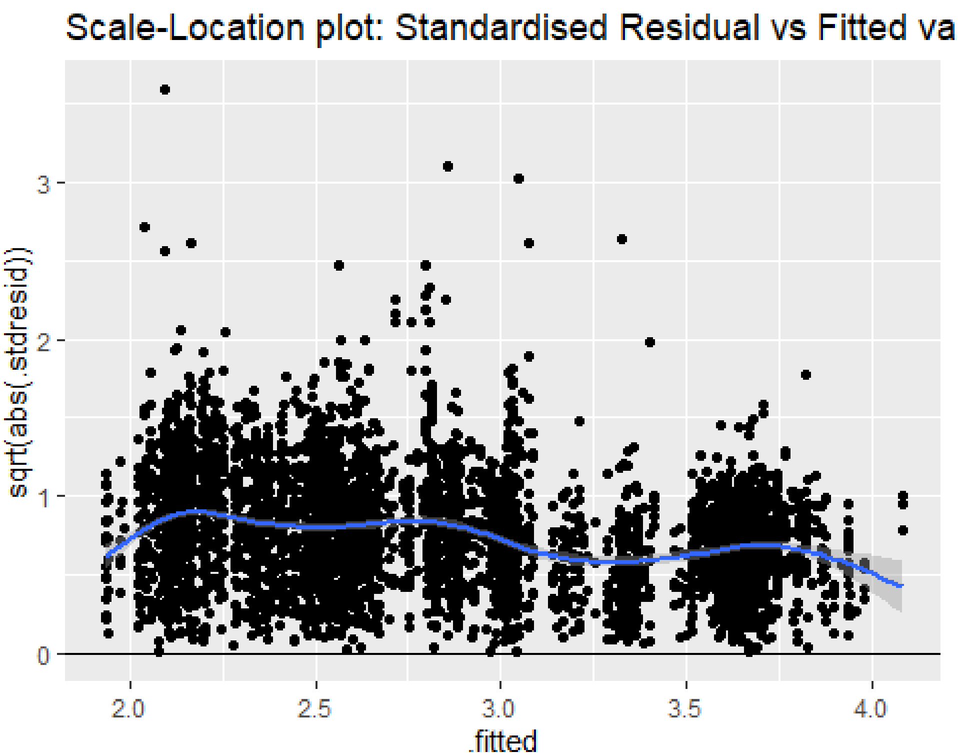
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



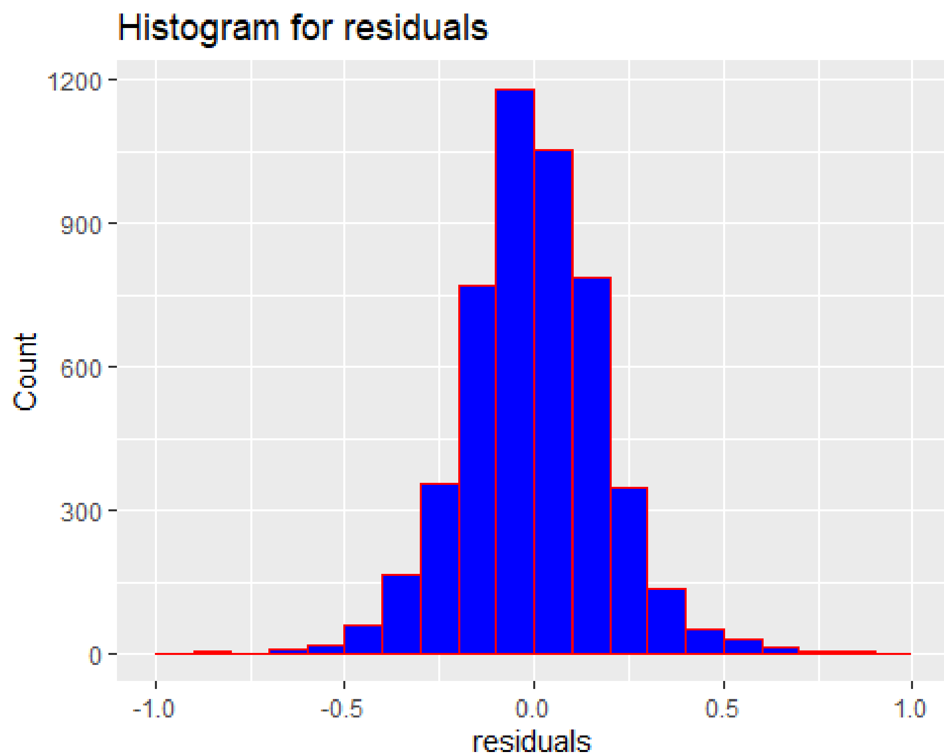

```
#   12th order model
ggplot (taxi_lm_red_3_int_red_12o, aes ( x = .fitted, y = .resid)) +
  geom_point () + geom_smooth () +
  geom_hline (yintercept = 0)

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



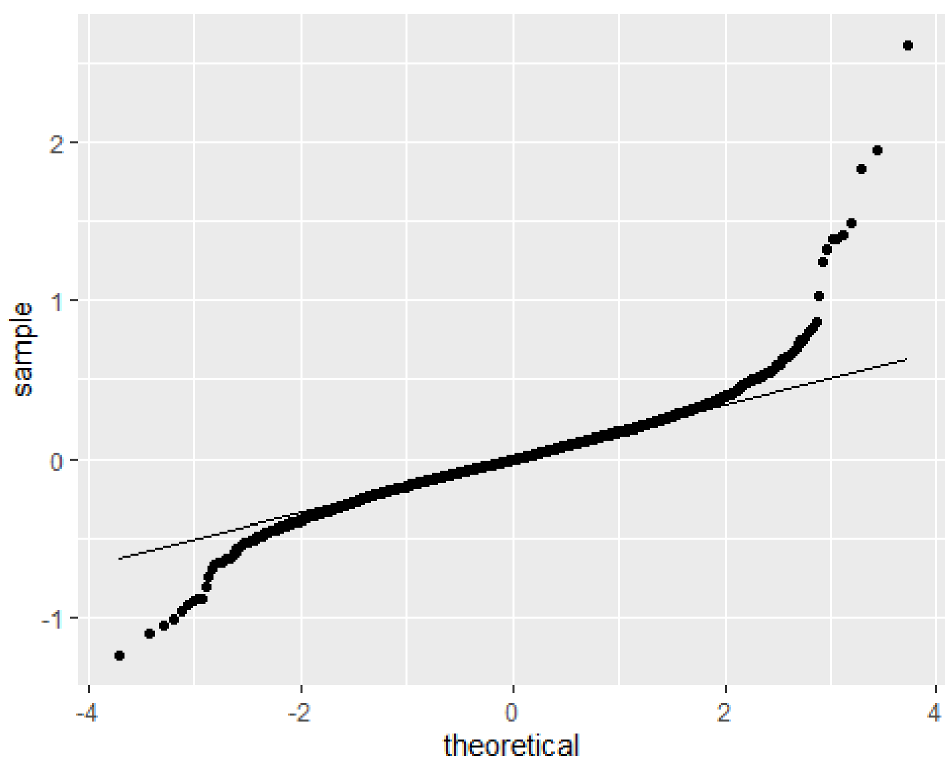

```
ggplot (taxi_lm_red_3_int_red_12o, aes ( x = .fitted, y = sqrt ( abs
(.stdresid)))) +
  geom_point () + geom_smooth () +
  geom_hline (yintercept = 0) +
  ggtitle ("Scale-Location plot: Standardised Residual vs Fitted values")

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
#   BP test
bptest (taxi_lm_red_3_int_red_12o)

##
##  studentized Breusch-Pagan test
##
## data:  taxi_lm_red_3_int_red_12o
## BP = 140.39, df = 19, p-value < 2.2e-16

#   H0 : heteroscedasticity is not present

par ( mfrow = c(1,2))
ggplot ( data = taxi_data, aes ( residuals (taxi_lm_red_3_int_red_12o))) +
  geom_histogram (breaks = seq (-1, 1, by = 0.1), col = "red", fill = "blue")
+
  labs ( title = "Histogram for residuals") +
  labs ( x = "residuals", y = "Count")
```



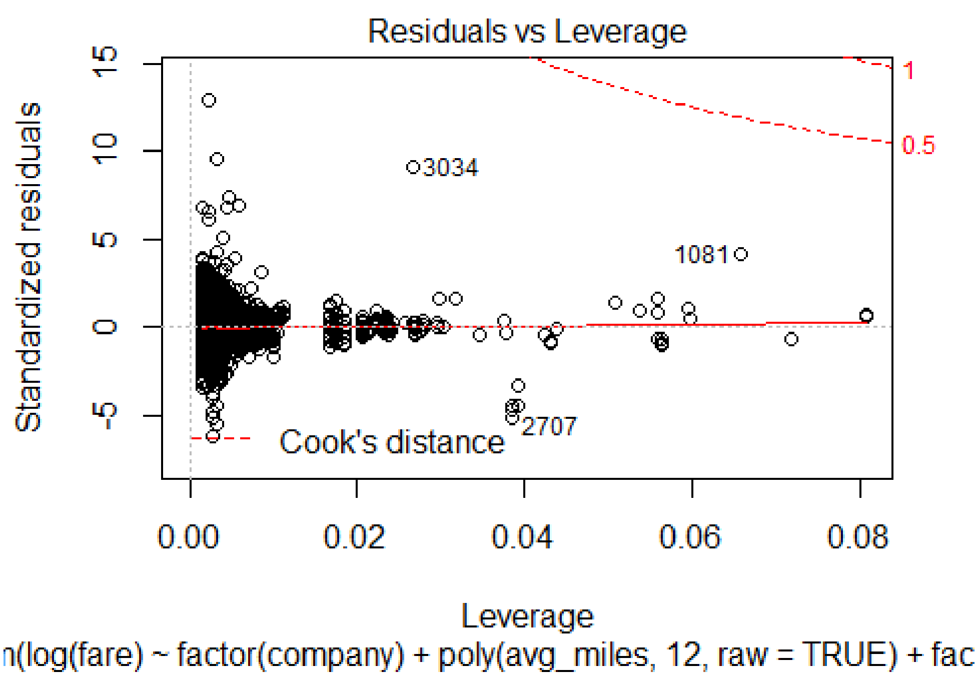

```
ggplot (taxi_data, aes ( sample = taxi_lm_red_3_int_red_12o$residuals)) +
  stat_qq () +
  stat_qq_line ()
```



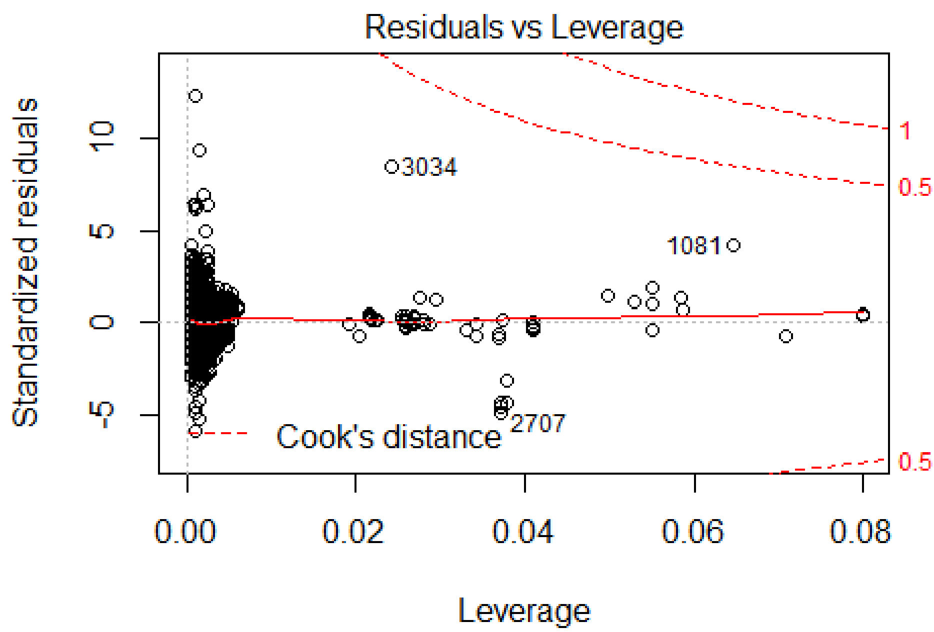
```r
shapiro.test ( residuals (taxi_lm_red_3_int_red_12o))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(taxi_lm_red_3_int_red_12o)
## W = 0.93038, p-value < 2.2e-16

#   H0 : model is normal
```

Test for outliers

```r
# order 12 cook's distance for 12th order model
plot (taxi_lm_red_3_int_red_12o, which = 5)
```

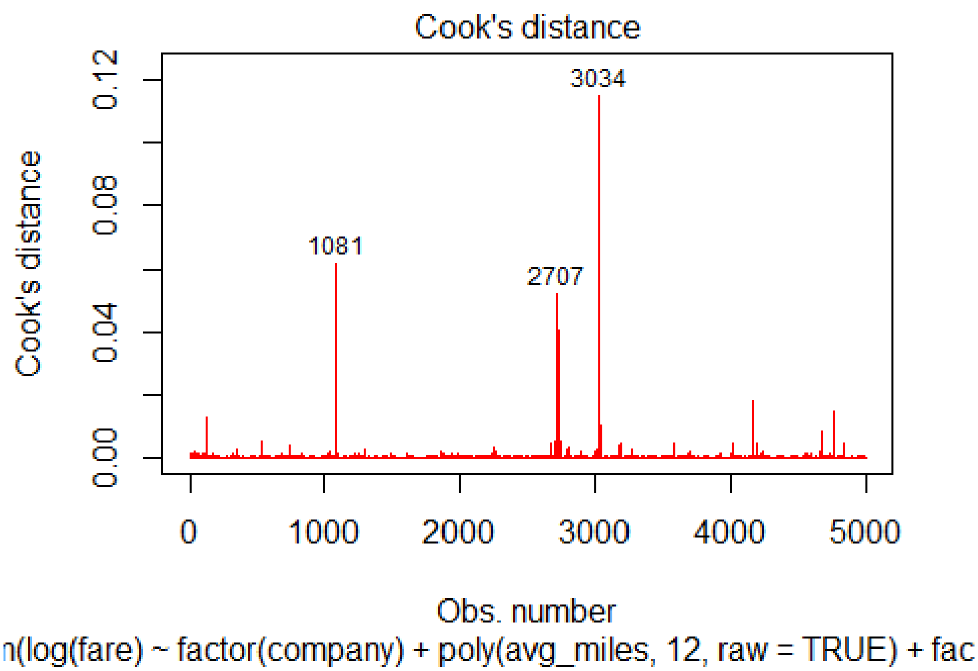



```r
plot (taxi_lm_red_3_int_red_2o, which = 5)
```



n(log(fare) ~ factor(company) + poly(avg_miles, 2, raw = TRUE) + factc

```
taxi_data[cooks.distance (taxi_lm_red_3_int_red_12o) > 0.5,]

##  [1] X                    pickup_area          dropoff_area
##  [4] trip_miles           trip_seconds         fare
##  [7] trip_start_timestamp tips                 tolls
## [10] trip_total           payment_type         company
## [13] extras               pickup_dropoff       avg_miles
## [16] avg_minutes          hours                months
## [19] day_of_week          hour_type            tip_pct
## [22] tip_type             pickup_dropoff_dummy weekend
## [25] season               time_of_day
## <0 rows> (or 0-length row.names)

plot (taxi_lm_red_3_int_red_12o, pch = 10, col = "red", which = c(4))
```




```
lev = hatvalues (taxi_lm_red_3_int_red_12o)
p = length ( coef (taxi_lm_red_3_int_red_12o))
n = nrow (taxi_data)
outlier = lev[lev > (2*p/n)]
print (outlier)

##           1           2           3           4           5           6
## 0.016676409 0.017452349 0.018464550 0.017452266 0.016675352 0.018464320
##           7           8           9          10          11          12
## 0.016675555 0.017556643 0.018464566 0.016675804 0.017556612 0.017452336
##          13          14          15          16          17          18
## 0.017452336 0.016675805 0.016675805 0.018464561 0.018348104 0.016675805
##          19          20          21          22          23          24
## 0.016675805 0.016675805 0.018464561 0.017556612 0.016675805 0.017556612
##          25          26          27          28          29          30
## 0.016675805 0.016675805 0.018464561 0.017452336 0.017452336 0.017556612
##          31          32          33          34          35          36
## 0.018464561 0.016675805 0.016675805 0.016675805 0.017452336 0.018464561
##          37          38          39          40          41          42
## 0.016675805 0.016675805 0.017556612 0.016675805 0.016675805 0.016675805
##          43          44          45          46          47          48
## 0.016675805 0.016675805 0.017556612 0.018464561 0.018464561 0.016675805
##          49          50         263         288         300         401
## 0.017556612 0.017452336 0.027882558 0.027882558 0.027882558 0.030311639
##         408         419         436         437         528         609
## 0.029538562 0.030311639 0.029538562 0.029538562 0.050813131 0.027292326
##         613         615        1081        1293        1525        1555
```

```
## 0.026568880 0.026568880 0.065819290 0.080778190 0.023836965 0.026965312
##        1562        1571        1578        1586        1599        1851
## 0.026965312 0.026965312 0.026965312 0.026965312 0.027691955 0.009980171
##        1852        1853        1854        1855        1857        1858
## 0.009980171 0.009980171 0.009476617 0.009980171 0.009476617 0.009980171
##        1861        1862        1863        1865        1867        1869
## 0.009476617 0.043073803 0.009476617 0.010026624 0.043073803 0.009476617
##        1870        1872        1873        1875        1876        1877
## 0.009476617 0.009980171 0.009980171 0.009476617 0.009980171 0.010501116
##        1878        1879        1880        1881        1882        1885
## 0.009980171 0.010026624 0.009476617 0.009980171 0.010026624 0.009980171
##        1888        1889        1891        1893        1895        1898
## 0.009476617 0.009980171 0.009476617 0.009980171 0.009476617 0.009980171
##        1899        1954        2117        2122        2142        2143
## 0.010026624 0.034659120 0.023515103 0.024307470 0.023515103 0.023515103
##        2149        2268        2314        2577        2651        2652
## 0.023515103 0.080737400 0.037875639 0.042468872 0.008483282 0.011169110
##        2653        2654        2655        2656        2657        2658
## 0.008483282 0.010711035 0.010711035 0.010711035 0.008483282 0.011169110
##        2659        2660        2661        2662        2663        2664
## 0.008483282 0.008483282 0.008483282 0.010655647 0.010168802 0.008483282
##        2665        2666        2667        2668        2669        2670
## 0.043831834 0.008483282 0.008483282 0.008954655 0.008483282 0.008483282
##        2671        2672        2673        2674        2675        2676
## 0.010168802 0.011169110 0.010655647 0.010168802 0.008954655 0.008483282
##        2677        2678        2679        2680        2681        2682
## 0.010655647 0.010168802 0.011169110 0.008483282 0.010168802 0.008954655
##        2683        2684        2685        2686        2687        2688
## 0.008954655 0.008954655 0.008483282 0.008483282 0.010655647 0.008954655
##        2689        2690        2691        2692        2693        2694
## 0.008954655 0.008483282 0.008483282 0.011169110 0.008954655 0.008483282
##        2695        2696        2697        2698        2699        2700
## 0.010168802 0.008483282 0.010168802 0.008483282 0.010711035 0.008483282
##        2707        2711        2722        2725        2733        2751
## 0.038497645 0.038497645 0.038497645 0.039233726 0.039233726 0.020746701
##        2752        2753        2754        2755        2756        2757
## 0.021468253 0.020746701 0.020746701 0.022386787 0.020746701 0.023304309
##        2758        2759        2760        2761        2762        2763
## 0.022386787 0.020746701 0.023304309 0.020746701 0.022386787 0.022386787
##        2764        2765        2766        2767        2768        2769
## 0.023304309 0.023304309 0.020746701 0.022386787 0.024072665 0.023304309
##        2770        2771        2772        2773        2774        2775
## 0.022386787 0.020746701 0.021468253 0.020746701 0.021468253 0.020746701
##        2776        2777        2778        2779        2780        2781
## 0.023304309 0.024072665 0.020746701 0.020746701 0.023180784 0.022386787
##        2782        2783        2784        2785        2786        2787
## 0.020746701 0.020746701 0.024072665 0.056483668 0.022386787 0.023304309
##        2788        2789        2790        2791        2792        2793
## 0.056483668 0.021468253 0.020746701 0.022386787 0.020746701 0.022386787
##        2794        2795        2796        2797        2798        2799
```

```
## 0.022386787 0.020746701 0.022386787 0.020746701 0.020746701 0.021468253
##        2800        2901        3034        3182        3269        3301
## 0.056483668 0.071917744 0.026686464 0.059468496 0.053841627 0.023538707
##        3305        3306        3320        3361        3445        3578
## 0.023538707 0.024324170 0.023538707 0.037673532 0.059697392 0.031762239
##        3724        3727        3739        4012        4573        4652
## 0.008588367 0.025674476 0.008588367 0.029783553 0.055820084 0.055899106
##        4673
## 0.055899106
```

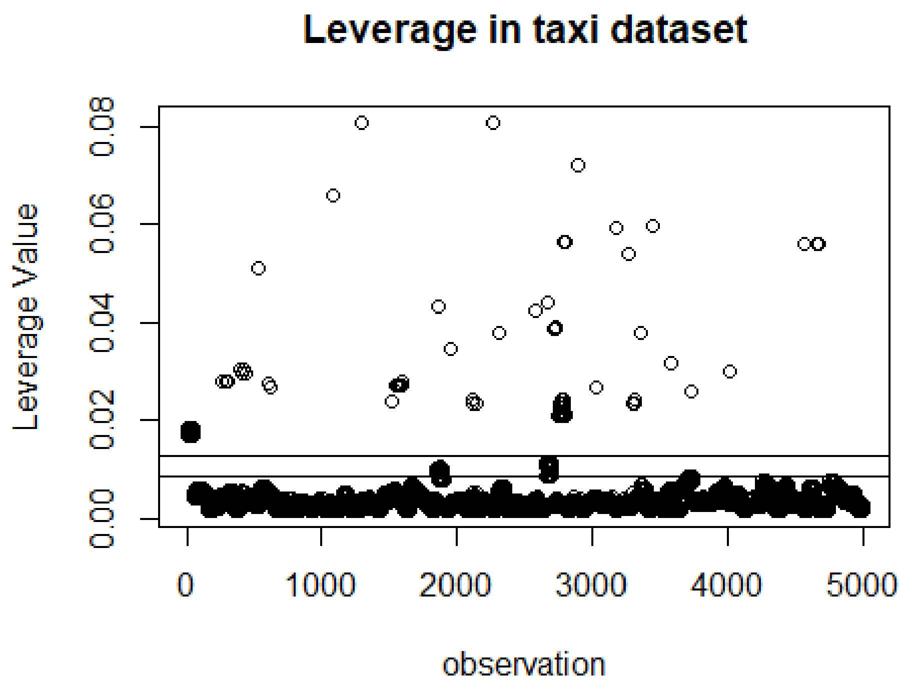
```r
lev = hatvalues (taxi_lm_red_3_int_red_12o)
p = length ( coef (taxi_lm_red_3_int_red_12o))
n = nrow (taxi_data)
outlier = lev[lev > (3*p/n)]
print (outlier)
```

```
##          1          2          3          4          5          6
## 0.01667641 0.01745235 0.01846455 0.01745227 0.01667535 0.01846432
##          7          8          9         10         11         12
## 0.01667555 0.01755664 0.01846457 0.01667580 0.01755661 0.01745234
##         13         14         15         16         17         18
## 0.01745234 0.01667581 0.01667581 0.01846456 0.01834810 0.01667581
##         19         20         21         22         23         24
## 0.01667581 0.01667581 0.01846456 0.01755661 0.01667581 0.01755661
##         25         26         27         28         29         30
## 0.01667581 0.01667581 0.01846456 0.01745234 0.01745234 0.01755661
##         31         32         33         34         35         36
## 0.01846456 0.01667581 0.01667581 0.01667581 0.01745234 0.01846456
##         37         38         39         40         41         42
## 0.01667581 0.01667581 0.01755661 0.01667581 0.01667581 0.01667581
##         43         44         45         46         47         48
## 0.01667581 0.01667581 0.01755661 0.01846456 0.01846456 0.01667581
##         49         50        263        288        300        401
## 0.01755661 0.01745234 0.02788256 0.02788256 0.02788256 0.03031164
##        408        419        436        437        528        609
## 0.02953856 0.03031164 0.02953856 0.02953856 0.05081313 0.02729233
##        613        615       1081       1293       1525       1555
## 0.02656888 0.02656888 0.06581929 0.08077819 0.02383696 0.02696531
##       1562       1571       1578       1586       1599       1862
## 0.02696531 0.02696531 0.02696531 0.02696531 0.02769196 0.04307380
##       1867       1954       2117       2122       2142       2143
## 0.04307380 0.03465912 0.02351510 0.02430747 0.02351510 0.02351510
##       2149       2268       2314       2577       2665       2707
## 0.02351510 0.08073740 0.03787564 0.04246887 0.04383183 0.03849765
##       2711       2722       2725       2733       2751       2752
## 0.03849765 0.03849765 0.03923373 0.03923373 0.02074670 0.02146825
##       2753       2754       2755       2756       2757       2758
## 0.02074670 0.02074670 0.02238679 0.02074670 0.02330431 0.02238679
##       2759       2760       2761       2762       2763       2764
## 0.02074670 0.02330431 0.02074670 0.02238679 0.02238679 0.02330431
```

```
##       2765       2766       2767       2768       2769       2770 
## 0.02330431 0.02074670 0.02238679 0.02407267 0.02330431 0.02238679 
##       2771       2772       2773       2774       2775       2776 
## 0.02074670 0.02146825 0.02074670 0.02146825 0.02074670 0.02330431 
##       2777       2778       2779       2780       2781       2782 
## 0.02407267 0.02074670 0.02074670 0.02318078 0.02238679 0.02074670 
##       2783       2784       2785       2786       2787       2788 
## 0.02074670 0.02407267 0.05648367 0.02238679 0.02330431 0.05648367 
##       2789       2790       2791       2792       2793       2794 
## 0.02146825 0.02074670 0.02238679 0.02074670 0.02238679 0.02238679 
##       2795       2796       2797       2798       2799       2800 
## 0.02074670 0.02238679 0.02074670 0.02074670 0.02146825 0.05648367 
##       2901       3034       3182       3269       3301       3305 
## 0.07191774 0.02668646 0.05946850 0.05384163 0.02353871 0.02353871 
##       3306       3320       3361       3445       3578       3727 
## 0.02432417 0.02353871 0.03767353 0.05969739 0.03176224 0.02567448 
##       4012       4573       4652       4673 
## 0.02978355 0.05582008 0.05589911 0.05589911
```

(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,263,288,300,401,408,419,436,437,528,609,613,615,1081,1293,1525,1555,1562,1571,1578,1586,1599,1862,1867,1954,2117,2122,2142,2143,2149,2268,2314,2577,2665,2707,2711,2722,2725,2733,2751,2752,2753,2754,2755,2756,2757,2758,2759,2760,2761,2762,2763,2764,2765,2766,2767,2768,2769 2770,2771,2772,2773,2774,2775,2776,2777,2778,2779,2780,2781,2782,2783,2784,2785,2786,2787,2788,2789,2790,2791,2792,2793,2794,2795,2796,2797,2798,2799,2800,2901,3034,3182,3269,3301,3305,3306,3320,3361,3445,3578,3727,4012,4573,4652,4673)

```
plot (rownames (taxi_data), lev, main = "Leverage in taxi dataset", xlab = 
"observation", ylab = "Leverage Value")
abline (h = 2*p/n, lty = 1)
abline (h = 3*p/n, lty = 1)
```



New dataset

```
taxi_data2 = taxi_data[-
c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,
29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,263,288,300
,401,408,419,436,437,528,609,613,615,1081,1293,1525,1555,1562,1571,1578,1586,
1599,1862,1867,1954,2117,2122,2142,2143,2149,2268,2314,2577,2665,2707,2711,27
22,2725,2733,2751,2752,2753,2754,2755,2756,2757,2758,2759,2760,2761,2762,2763
,2764,2765,2766,2767,2768,27692770,2771,2772,2773,2774,2775,2776,2777,2778,27
79,2780,2781,2782,2783,2784,2785,2786,2787,2788,2789,2790,2791,2792,2793,2794
,2795,2796,2797,2798,2799,2800,2901,3034,3182,3269,3301,3305,3306,3320,3361,3
445,3578,3727,4012,4573,4652,4673), ]

nrow (taxi_data2)

## [1] 4848

taxi2_fulllm_log = lm ( log (fare) ~ factor(payment_type) + factor(company) +
avg_miles + avg_minutes + factor(time_of_day) + factor(season) +
factor(weekend) + factor(hour_type), data = taxi_data2)

vif (taxi2_fulllm_log)

##                           GVIF Df GVIF^(1/(2*Df))
## factor(payment_type)  1.066114  1        1.032528
## factor(company)       1.017828  2        1.004428
## avg_miles            13.262871  1        3.641822
## avg_minutes          13.289386  1        3.645461
```

```
## factor(time_of_day)   1.226729  3        1.034645
## factor(season)        1.016830  3        1.002786
## factor(weekend)       1.082269  1        1.040321
## factor(hour_type)     1.117561  1        1.057147
```

avg_minutes should be removed

Model varaible testing

```
taxi2_fulllm_log_nomin = lm ( log (fare) ~ factor(payment_type) +
factor(company) + avg_miles + factor(time_of_day) + factor(season) +
factor(weekend) + factor(hour_type), data = taxi_data2)

taxi_stepw = ols_step_both_p ( taxi2_fulllm_log_nomin, pent = 0.05, prem =
0.1, details = FALSE)
```

```
## Stepwise Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. factor(payment_type)
## 2. factor(company)
## 3. avg_miles
## 4. factor(time_of_day)
## 5. factor(season)
## 6. factor(weekend)
## 7. factor(hour_type)
##
## We are selecting variables based on p value...
##
## Variables Entered/Removed:
##
## - avg_miles added
## - factor(hour_type) added
## - factor(company) added
## - factor(time_of_day) added
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.932       RMSE                0.220
## R-Squared               0.868       Coef. Var           7.821
## Adj. R-Squared          0.868       MSE                 0.049
## Pred R-Squared          0.867       MAE                 0.168
## -------------------------------------------------------------
```

```
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## ------------------------------------------------------------------------
##                 Sum of
##                Squares           DF    Mean Square       F          Sig.
## ------------------------------------------------------------------------
## Regression    1543.092            7        220.442    4535.485    0.0000
## Residual       235.242         4840          0.049
## Total         1778.334         4847
## ------------------------------------------------------------------------
##
##                                           Parameter Estimates
## --------------------------------------------------------------------------
------------------------------
##                       model      Beta    Std. Error    Std. Beta       t
Sig      lower     upper
## --------------------------------------------------------------------------
------------------------------
##                (Intercept)     1.991         0.010                 198.298
0.000     1.971     2.011
##                  avg_miles     0.126         0.001        0.926    173.606
0.000     0.124     0.127
## factor(hour_type)rush_hour     0.028         0.008        0.019      3.398
0.001     0.012     0.043
##         factor(company)107     0.022         0.007        0.018      2.976
0.003     0.008     0.037
##         factor(company)109     0.013         0.010        0.009      1.401
0.161    -0.005     0.032
## factor(time_of_day)Evening    -0.013         0.008       -0.010     -1.597
0.110    -0.029     0.003
## factor(time_of_day)Morning    -0.008         0.010       -0.005     -0.813
0.416    -0.027     0.011
##   factor(time_of_day)Night    -0.029         0.010       -0.019     -2.868
0.004    -0.049    -0.009
## --------------------------------------------------------------------------
------------------------------
```

Stepwise regression suggests a model including avg_miles, hour_type, company, and time_of_day.

```
taxi_formodel = ols_step_forward_p ( taxi2_fulllm_log_nomin, pent = 0.05,
details = FALSE)
```

```
## Forward Selection Method
## ---------------------------
##
## Candidate Terms:
```

```
##
## 1. factor(payment_type)
## 2. factor(company)
## 3. avg_miles
## 4. factor(time_of_day)
## 5. factor(season)
## 6. factor(weekend)
## 7. factor(hour_type)
##
## We are selecting variables based on p value...
##
## Variables Entered:
##
## - avg_miles
## - factor(hour_type)
## - factor(company)
## - factor(time_of_day)
##
## No more variables to be added.
##
## Final Model Output
## ------------------
##
##                         Model Summary
## -------------------------------------------------------------
## R                       0.932       RMSE                0.220
## R-Squared               0.868       Coef. Var           7.821
## Adj. R-Squared          0.868       MSE                 0.049
## Pred R-Squared          0.867       MAE                 0.168
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## ------------------------------------------------------------------------
##                 Sum of
##                Squares          DF    Mean Square       F          Sig.
## ------------------------------------------------------------------------
## Regression    1543.092           7        220.442    4535.485    0.0000
## Residual       235.242        4840          0.049
## Total         1778.334        4847
## ------------------------------------------------------------------------
##
##                                           Parameter Estimates
## ---------------------------------------------------------------------------
------------------------------
##                      model      Beta    Std. Error    Std. Beta       t
Sig      lower     upper
## ---------------------------------------------------------------------------
```

```
------------------------------
##                 (Intercept)     1.991         0.010                    198.298
0.000     1.971     2.011
##                   avg_miles     0.126         0.001         0.926    173.606
0.000     0.124     0.127
## factor(hour_type)rush_hour     0.028         0.008         0.019      3.398
0.001     0.012     0.043
##          factor(company)107     0.022         0.007         0.018      2.976
0.003     0.008     0.037
##          factor(company)109     0.013         0.010         0.009      1.401
0.161    -0.005     0.032
## factor(time_of_day)Evening    -0.013         0.008        -0.010     -1.597
0.110    -0.029     0.003
## factor(time_of_day)Morning    -0.008         0.010        -0.005     -0.813
0.416    -0.027     0.011
##   factor(time_of_day)Night    -0.029         0.010        -0.019     -2.868
0.004    -0.049    -0.009
## ----------------------------------------------------------------------------
------------------------------
```

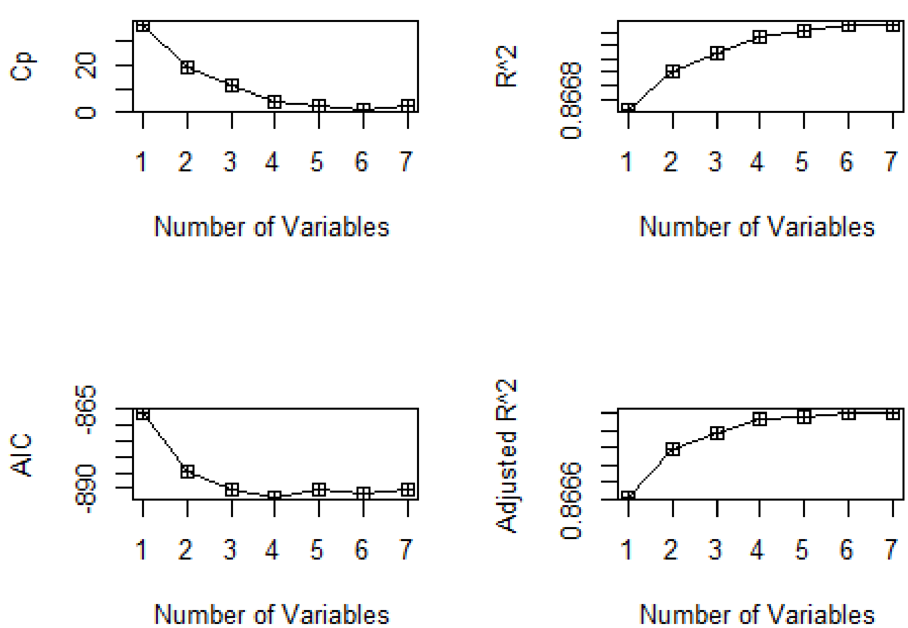
Forward regression suggests a model including avg_miles, hour_type, company, and time_of_day.

```
taxi_backmodel = ols_step_backward_p ( taxi2_fulllm_log_nomin, prem = 0.05,
details = FALSE)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . factor(payment_type)
## 2 . factor(company)
## 3 . avg_miles
## 4 . factor(time_of_day)
## 5 . factor(season)
## 6 . factor(weekend)
## 7 . factor(hour_type)
##
## We are eliminating variables based on p value...
##
## Variables Removed:
##
## - factor(weekend)
## - factor(season)
## - factor(payment_type)
##
## No more variables satisfy the condition of p value = 0.05
##
##
```

```
## Final Model Output
## ------------------
##
##                         Model Summary
## -------------------------------------------------------------
## R                        0.932       RMSE               0.220
## R-Squared                0.868       Coef. Var          7.821
## Adj. R-Squared           0.868       MSE                0.049
## Pred R-Squared           0.867       MAE                0.168
## -------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                 ANOVA
## ------------------------------------------------------------------------
##                 Sum of
##                Squares          DF    Mean Square       F          Sig.
## ------------------------------------------------------------------------
## Regression    1543.092           7        220.442    4535.485    0.0000
## Residual       235.242        4840          0.049
## Total         1778.334        4847
## ------------------------------------------------------------------------
##
##                                          Parameter Estimates
## --------------------------------------------------------------------------
------------------------------
##                      model      Beta    Std. Error    Std. Beta       t
Sig      lower     upper
## --------------------------------------------------------------------------
------------------------------
##                (Intercept)     1.991         0.010                198.298
0.000     1.971     2.011
##         factor(company)107     0.022         0.007        0.018      2.976
0.003     0.008     0.037
##         factor(company)109     0.013         0.010        0.009      1.401
0.161    -0.005     0.032
##                  avg_miles     0.126         0.001        0.926    173.606
0.000     0.124     0.127
## factor(time_of_day)Evening    -0.013         0.008       -0.010     -1.597
0.110    -0.029     0.003
## factor(time_of_day)Morning    -0.008         0.010       -0.005     -0.813
0.416    -0.027     0.011
##   factor(time_of_day)Night    -0.029         0.010       -0.019     -2.868
0.004    -0.049    -0.009
## factor(hour_type)rush_hour     0.028         0.008        0.019      3.398
0.001     0.012     0.043
## --------------------------------------------------------------------------
------------------------------
```

Backward regression suggests a model including avg_miles, hour_type, company, and time_of_day.




```
ks_stat2 = data.frame ( c(1, 2, 3, 4, 5, 6, 7), ks$cp, ks$aic, ks$adjr,
ks$rsq)
names (ks_stat2) = c( "Predictors", "CP", "AIC", "Adjusted R^2", "R^2")
ks_stat2

##   Predictors        CP       AIC Adjusted R^2       R^2
## 1          1 37.687687 -865.6793    0.8666240 0.8666515
## 2          2 18.729594 -884.5376    0.8671692 0.8672240
## 3          3 11.106158 -890.1453    0.8673775 0.8674869
## 4          4  4.660409 -892.5930    0.8675263 0.8677176
## 5          5  2.696495 -890.5629    0.8675526 0.8678259
## 6          6  1.373066 -891.8939    0.8676162 0.8679167
## 7          7  3.000000 -890.2679    0.8675991 0.8679269
```

Cp suggests using the six variable model AIC suggersts using the four variable model Adj.rsq suggests using the six variable model

```
best.subset = regsubsets ( log (fare) ~ factor(payment_type) +
factor(company) + avg_miles + factor(time_of_day) + factor(season) +
factor(weekend) + factor(hour_type), data = taxi_data2, nv = 10)
summary ( best.subset)

## Subset selection object
## Call: regsubsets.formula(log(fare) ~ factor(payment_type) +
```

```
factor(company) +
##     avg_miles + factor(time_of_day) + factor(season) + factor(weekend) +
##     factor(hour_type), data = taxi_data2, nv = 10)
## 12 Variables  (and intercept)
##                                   Forced in Forced out
## factor(payment_type)Credit Card     FALSE      FALSE
## factor(company)107                  FALSE      FALSE
## factor(company)109                  FALSE      FALSE
## avg_miles                           FALSE      FALSE
## factor(time_of_day)Evening          FALSE      FALSE
## factor(time_of_day)Morning          FALSE      FALSE
## factor(time_of_day)Night            FALSE      FALSE
## factor(season)Spring                FALSE      FALSE
## factor(season)Summer                FALSE      FALSE
## factor(season)Winter                FALSE      FALSE
## factor(weekend)1                    FALSE      FALSE
## factor(hour_type)rush_hour          FALSE      FALSE
## 1 subsets of each size up to 10
## Selection Algorithm: exhaustive
##           factor(payment_type)Credit Card factor(company)107
## 1  ( 1 )  " "                             " "
## 2  ( 1 )  " "                             " "
## 3  ( 1 )  " "                             "*"
## 4  ( 1 )  " "                             "*"
## 5  ( 1 )  " "                             "*"
## 6  ( 1 )  "*"                             "*"
## 7  ( 1 )  "*"                             "*"
## 8  ( 1 )  "*"                             "*"
## 9  ( 1 )  "*"                             "*"
## 10  ( 1 ) "*"                             "*"
##           factor(company)109 avg_miles factor(time_of_day)Evening
## 1  ( 1 )  " "                "*"       " "
## 2  ( 1 )  " "                "*"       " "
## 3  ( 1 )  " "                "*"       " "
## 4  ( 1 )  " "                "*"       " "
## 5  ( 1 )  " "                "*"       " "
## 6  ( 1 )  " "                "*"       " "
## 7  ( 1 )  " "                "*"       "*"
## 8  ( 1 )  "*"                "*"       "*"
## 9  ( 1 )  "*"                "*"       "*"
## 10  ( 1 ) "*"                "*"       "*"
##           factor(time_of_day)Morning factor(time_of_day)Night
## 1  ( 1 )  " "                        " "
## 2  ( 1 )  " "                        " "
## 3  ( 1 )  " "                        " "
## 4  ( 1 )  " "                        "*"
## 5  ( 1 )  " "                        "*"
## 6  ( 1 )  " "                        "*"
## 7  ( 1 )  " "                        "*"
## 8  ( 1 )  " "                        "*"
```

```
## 9  ( 1 )  "*"                        "*"
## 10  ( 1 ) "*"                        "*"
##           factor(season)Spring factor(season)Summer factor(season)Winter
## 1  ( 1 )  " "                  " "                  " "
## 2  ( 1 )  " "                  " "                  " "
## 3  ( 1 )  " "                  " "                  " "
## 4  ( 1 )  " "                  " "                  " "
## 5  ( 1 )  " "                  "*"                  " "
## 6  ( 1 )  " "                  "*"                  " "
## 7  ( 1 )  " "                  "*"                  " "
## 8  ( 1 )  " "                  "*"                  " "
## 9  ( 1 )  " "                  "*"                  " "
## 10  ( 1 ) " "                  "*"                  " "
##           factor(weekend)1 factor(hour_type)rush_hour
## 1  ( 1 )  " "              " "
## 2  ( 1 )  " "              "*"
## 3  ( 1 )  " "              "*"
## 4  ( 1 )  " "              "*"
## 5  ( 1 )  " "              "*"
## 6  ( 1 )  " "              "*"
## 7  ( 1 )  " "              "*"
## 8  ( 1 )  " "              "*"
## 9  ( 1 )  " "              "*"
## 10  ( 1 ) "*"              "*"
```

```
reg.summary = summary ( best.subset)
```

```
summary (taxi2_fulllm_log_nomin)
```

```
##
## Call:
## lm(formula = log(fare) ~ factor(payment_type) + factor(company) +
##     avg_miles + factor(time_of_day) + factor(season) + factor(weekend) +
##     factor(hour_type), data = taxi_data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12023 -0.14077  0.00088  0.13719  2.45631
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.9895970  0.0125187 158.930  < 2e-16 ***
## factor(payment_type)Credit Card  0.0118244  0.0065766   1.798 0.072249 .
## factor(company)107              0.0223905  0.0074328   3.012 0.002605 **
## factor(company)109              0.0133864  0.0095095   1.408 0.159287
## avg_miles                       0.1254528  0.0007502 167.218  < 2e-16 ***
## factor(time_of_day)Evening     -0.0132796  0.0082240  -1.615 0.106434
## factor(time_of_day)Morning     -0.0082096  0.0096953  -0.847 0.397173
## factor(time_of_day)Night       -0.0283325  0.0103219  -2.745 0.006075 **
## factor(season)Spring           -0.0069575  0.0090874  -0.766 0.443940
```

```
## factor(season)Summer             0.0080966  0.0093190   0.869 0.384981
## factor(season)Winter            -0.0060359  0.0097500  -0.619 0.535901
## factor(weekend)1                -0.0040270  0.0065930  -0.611 0.541367
## factor(hour_type)rush_hour       0.0275061  0.0081171   3.389 0.000708 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2204 on 4835 degrees of freedom
## Multiple R-squared:  0.8679, Adjusted R-squared:  0.8676
## F-statistic:  2648 on 12 and 4835 DF,  p-value: < 2.2e-16
```

Season, weekend, payment_type are insignificant.

Models

```
taxi2_lm_red_4 = lm ( log (fare) ~ factor(company) + avg_miles +
factor(time_of_day) + factor(hour_type), data = taxi_data2)
taxi2_lm_red_3 = lm ( log (fare) ~ factor(company) + avg_miles +
factor(hour_type), data = taxi_data2)

taxi2_lm_red_6 = lm ( log (fare) ~ factor(company) + avg_miles +
factor(time_of_day) + factor(hour_type) + factor(payment_type) +
factor(season), data = taxi_data2)

taxi2_fulllm_log = lm ( log (fare) ~ factor(payment_type) + factor(company) +
avg_miles + avg_minutes + factor(time_of_day) + factor(season) +
factor(weekend) + factor(hour_type), data = taxi_data2)

nrow (taxi_data2)

## [1] 4848
```

Partial F-test

```
#   full and 6 variables
anova (taxi2_fulllm_log_nomin, taxi2_lm_red_6)

## Analysis of Variance Table
##
## Model 1: log(fare) ~ factor(payment_type) + factor(company) + avg_miles +
##     factor(time_of_day) + factor(season) + factor(weekend) +
##     factor(hour_type)
## Model 2: log(fare) ~ factor(company) + avg_miles + factor(time_of_day) +
##     factor(hour_type) + factor(payment_type) + factor(season)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   4835 234.87
## 2   4836 234.89 -1 -0.018122 0.3731 0.5414

#   full and 4 variables
anova (taxi2_fulllm_log_nomin, taxi2_lm_red_4)

## Analysis of Variance Table
##
```

```
## Model 1: log(fare) ~ factor(payment_type) + factor(company) + avg_miles +
##     factor(time_of_day) + factor(season) + factor(weekend) +
##     factor(hour_type)
## Model 2: log(fare) ~ factor(company) + avg_miles + factor(time_of_day) +
##     factor(hour_type)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   4835 234.87
## 2   4840 235.24 -5  -0.37212 1.5321 0.1762
```

```
#   full and 3 variables
anova (taxi2_fulllm_log_nomin, taxi2_lm_red_3)
```

```
## Analysis of Variance Table
##
## Model 1: log(fare) ~ factor(payment_type) + factor(company) + avg_miles +
##     factor(time_of_day) + factor(season) + factor(weekend) +
##     factor(hour_type)
## Model 2: log(fare) ~ factor(company) + avg_miles + factor(hour_type)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1   4835 234.87
## 2   4843 235.65 -8  -0.78239 2.0133 0.04112 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary (taxi2_lm_red_4)
```

```
##
## Call:
## lm(formula = log(fare) ~ factor(company) + avg_miles + factor(time_of_day)
+
##     factor(hour_type), data = taxi_data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12774 -0.14176  0.00164  0.13741  2.46241
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.9909060  0.0100400 198.298  < 2e-16 ***
## factor(company)107          0.0220582  0.0074133   2.976 0.002939 **
## factor(company)109          0.0133249  0.0095109   1.401 0.161275
## avg_miles                   0.1258093  0.0007247 173.606  < 2e-16 ***
## factor(time_of_day)Evening -0.0131126  0.0082130  -1.597 0.110427
## factor(time_of_day)Morning -0.0078676  0.0096737  -0.813 0.416089
## factor(time_of_day)Night   -0.0291805  0.0101762  -2.868 0.004155 **
## factor(hour_type)rush_hour  0.0275516  0.0081073   3.398 0.000683 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2205 on 4840 degrees of freedom
```

```
## Multiple R-squared:  0.8677, Adjusted R-squared:  0.8675 
## F-statistic:  4535 on 7 and 4840 DF,  p-value: < 2.2e-16
```

```r
anova (taxi2_lm_red_4, taxi2_lm_red_3)
```

```
## Analysis of Variance Table
## 
## Model 1: log(fare) ~ factor(company) + avg_miles + factor(time_of_day) + 
##     factor(hour_type)
## Model 2: log(fare) ~ factor(company) + avg_miles + factor(hour_type)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
## 1   4840 235.24                              
## 2   4843 235.65 -3  -0.41027 2.8137 0.03783 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

time_of_day is significant according to the above results, so, the four variable model is selected.

```r
print ("Adj. R2")
```

```
## [1] "Adj. R2"
```

```r
summary (taxi2_lm_red_4)$adj.r.sq
```

```
## [1] 0.8675263
```

```r
print ("RMSE")
```

```
## [1] "RMSE"
```

```r
sigma (taxi2_lm_red_4)
```

```
## [1] 0.2204626
```

Interactions

```r
taxi2_lm_red_4_int = lm ( log (fare) ~ (factor(company) + avg_miles +
factor(hour_type) + factor(time_of_day)) ^2, data = taxi_data2)
summary (taxi2_lm_red_4_int)
```

```
## 
## Call:
## lm(formula = log(fare) ~ (factor(company) + avg_miles + factor(hour_type) 
+ 
##     factor(time_of_day))^2, data = taxi_data2)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.12453 -0.13768  0.00405  0.13555  2.47132 
## 
## Coefficients: (1 not defined because of singularities)
##                                                         Estimate
```

```
## (Intercept)                                          2.004e+00
## factor(company)107                                   5.524e-03
## factor(company)109                                   1.485e-02
## avg_miles                                            1.235e-01
## factor(hour_type)rush_hour                          -1.090e-02
## factor(time_of_day)Evening                           4.776e-03
## factor(time_of_day)Morning                           1.568e-02
## factor(time_of_day)Night                            -4.497e-02
## factor(company)107:avg_miles                         5.003e-03
## factor(company)109:avg_miles                         8.222e-05
## factor(company)107:factor(hour_type)rush_hour       -9.454e-05
## factor(company)109:factor(hour_type)rush_hour        2.700e-02
## factor(company)107:factor(time_of_day)Evening       -1.386e-02
## factor(company)109:factor(time_of_day)Evening       -1.048e-02
## factor(company)107:factor(time_of_day)Morning       -3.347e-02
## factor(company)109:factor(time_of_day)Morning        2.120e-04
## factor(company)107:factor(time_of_day)Night         -2.277e-02
## factor(company)109:factor(time_of_day)Night         -9.368e-03
## avg_miles:factor(hour_type)rush_hour                 1.821e-03
## avg_miles:factor(time_of_day)Evening                -3.763e-03
## avg_miles:factor(time_of_day)Morning                -2.492e-04
## avg_miles:factor(time_of_day)Night                   4.619e-03
## factor(hour_type)rush_hour:factor(time_of_day)Evening  7.188e-02
## factor(hour_type)rush_hour:factor(time_of_day)Morning -2.256e-02
## factor(hour_type)rush_hour:factor(time_of_day)Night          NA
##                                                     Std. Error t value
## (Intercept)                                          1.903e-02 105.316
## factor(company)107                                   2.006e-02   0.275
## factor(company)109                                   2.636e-02   0.563
## avg_miles                                            1.862e-03  66.347
## factor(hour_type)rush_hour                           2.239e-02  -0.487
## factor(time_of_day)Evening                           2.050e-02   0.233
## factor(time_of_day)Morning                           2.482e-02   0.632
## factor(time_of_day)Night                             2.359e-02  -1.906
## factor(company)107:avg_miles                         1.709e-03   2.928
## factor(company)109:avg_miles                         2.144e-03   0.038
## factor(company)107:factor(hour_type)rush_hour        1.925e-02  -0.005
## factor(company)109:factor(hour_type)rush_hour        2.390e-02   1.130
## factor(company)107:factor(time_of_day)Evening        1.926e-02  -0.719
## factor(company)109:factor(time_of_day)Evening        2.438e-02  -0.430
## factor(company)107:factor(time_of_day)Morning        2.320e-02  -1.443
## factor(company)109:factor(time_of_day)Morning        2.900e-02   0.007
## factor(company)107:factor(time_of_day)Night          2.305e-02  -0.988
## factor(company)109:factor(time_of_day)Night          3.169e-02  -0.296
## avg_miles:factor(hour_type)rush_hour                 1.750e-03   1.040
## avg_miles:factor(time_of_day)Evening                 1.782e-03  -2.111
## avg_miles:factor(time_of_day)Morning                 2.079e-03  -0.120
## avg_miles:factor(time_of_day)Night                   2.515e-03   1.837
## factor(hour_type)rush_hour:factor(time_of_day)Evening  1.861e-02   3.861
## factor(hour_type)rush_hour:factor(time_of_day)Morning  2.093e-02  -1.078
```

```
## factor(hour_type)rush_hour:factor(time_of_day)Night           NA      NA
##                                                         Pr(>|t|)
## (Intercept)                                              < 2e-16 ***
## factor(company)107                                      0.783031
## factor(company)109                                      0.573125
## avg_miles                                                < 2e-16 ***
## factor(hour_type)rush_hour                              0.626591
## factor(time_of_day)Evening                              0.815778
## factor(time_of_day)Morning                              0.527558
## factor(time_of_day)Night                                0.056724 .
## factor(company)107:avg_miles                            0.003425 **
## factor(company)109:avg_miles                            0.969409
## factor(company)107:factor(hour_type)rush_hour           0.996082
## factor(company)109:factor(hour_type)rush_hour           0.258648
## factor(company)107:factor(time_of_day)Evening           0.471984
## factor(company)109:factor(time_of_day)Evening           0.667231
## factor(company)107:factor(time_of_day)Morning           0.149204
## factor(company)109:factor(time_of_day)Morning           0.994167
## factor(company)107:factor(time_of_day)Night             0.323317
## factor(company)109:factor(time_of_day)Night             0.767548
## avg_miles:factor(hour_type)rush_hour                    0.298220
## avg_miles:factor(time_of_day)Evening                    0.034797 *
## avg_miles:factor(time_of_day)Morning                    0.904585
## avg_miles:factor(time_of_day)Night                      0.066321 .
## factor(hour_type)rush_hour:factor(time_of_day)Evening 0.000114 ***
## factor(hour_type)rush_hour:factor(time_of_day)Morning 0.281164
## factor(hour_type)rush_hour:factor(time_of_day)Night          NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2196 on 4824 degrees of freedom
## Multiple R-squared:  0.8692, Adjusted R-squared:  0.8686
## F-statistic:  1394 on 23 and 4824 DF,  p-value: < 2.2e-16
```

From the individual t-tests, company*hour_type appears to be insignificant.

```
anova (taxi2_lm_red_4_int, taxi2_lm_red_4)
```

```
## Analysis of Variance Table
##
## Model 1: log(fare) ~ (factor(company) + avg_miles + factor(hour_type) +
##     factor(time_of_day))^2
## Model 2: log(fare) ~ factor(company) + avg_miles + factor(time_of_day) +
##     factor(hour_type)
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
## 1   4824 232.56
## 2   4840 235.24 -16   -2.6828 3.4781 3.043e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The f-test suggests that the interactions are significant

```r
#   Interaction model without compnay*time_of_day
taxi2_lm_red_4_int_red = lm ( log (fare) ~ factor(company) + avg_miles +
factor(hour_type) + factor(time_of_day) + avg_miles*factor(company) +
factor(hour_type)*factor(time_of_day) + avg_miles*factor(time_of_day), data =
taxi_data2)


#   Partial F-test
anova (taxi2_lm_red_4_int_red, taxi2_lm_red_4_int)

## Analysis of Variance Table
##
## Model 1: log(fare) ~ factor(company) + avg_miles + factor(hour_type) +
##     factor(time_of_day) + avg_miles * factor(company) + factor(hour_type)
*
##     factor(time_of_day) + avg_miles * factor(time_of_day)
## Model 2: log(fare) ~ (factor(company) + avg_miles + factor(hour_type) +
##     factor(time_of_day))^2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   4833 232.89
## 2   4824 232.56  9   0.32733 0.7544  0.659
```

The partial F-test indicates company*time_of_day is an insignificant interaction (F= 0.7544, df= 9, 4824, p-value = 0.659)

```r
#   Interaction model without hour_type*time_of_day
taxi2_lm_red_4_int_red_2 = lm ( log (fare) ~ factor(company) + avg_miles +
factor(hour_type) + factor(time_of_day) + avg_miles*factor(company) +
avg_miles*factor(time_of_day), data = taxi_data2)

#   Partial F-Test
anova (taxi2_lm_red_4_int_red_2, taxi2_lm_red_4_int_red)

## Analysis of Variance Table
##
## Model 1: log(fare) ~ factor(company) + avg_miles + factor(hour_type) +
##     factor(time_of_day) + avg_miles * factor(company) + avg_miles *
##     factor(time_of_day)
## Model 2: log(fare) ~ factor(company) + avg_miles + factor(hour_type) +
##     factor(time_of_day) + avg_miles * factor(company) + factor(hour_type)
*
##     factor(time_of_day) + avg_miles * factor(time_of_day)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4835 233.94
## 2   4833 232.89  2    1.0484 10.878 1.933e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The partial F-test indicates hour_type*time_of_day must be kept (F= 10.878, df= 2, 4833, p-value < 0.05)

```
#   Interaction model without avg_miles*time_of_day
taxi2_lm_red_4_int_red_3 = lm ( log (fare) ~ factor(company) + avg_miles +
factor(hour_type) + factor(time_of_day) + avg_miles*factor(company) +
factor(hour_type)*factor(time_of_day), data = taxi_data2)


#   Partial F-test
anova (taxi2_lm_red_4_int_red_3, taxi2_lm_red_4_int_red)

## Analysis of Variance Table
##
## Model 1: log(fare) ~ factor(company) + avg_miles + factor(hour_type) +
##     factor(time_of_day) + avg_miles * factor(company) + factor(hour_type)
*
##     factor(time_of_day)
## Model 2: log(fare) ~ factor(company) + avg_miles + factor(hour_type) +
##     factor(time_of_day) + avg_miles * factor(company) + factor(hour_type)
*
##     factor(time_of_day) + avg_miles * factor(time_of_day)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   4836 233.52
## 2   4833 232.89  3   0.63494 4.3922 0.004303 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The partial F-test indicates avg_miles*time_of_day must be kept (F= 4.3922, df= 3, 4833, p-value= 0.004303 < 0.05)

```
#   Interaction model without avg_mlies*company
taxi2_lm_red_4_int_red_4 = lm ( log (fare) ~ factor(company) + avg_miles +
factor(hour_type) + factor(time_of_day) +
factor(hour_type)*factor(time_of_day) + avg_miles*factor(time_of_day), data =
taxi_data2)

#   Partial F-test
anova (taxi2_lm_red_4_int_red_4, taxi2_lm_red_4_int_red)

## Analysis of Variance Table
##
## Model 1: log(fare) ~ factor(company) + avg_miles + factor(hour_type) +
##     factor(time_of_day) + factor(hour_type) * factor(time_of_day) +
##     avg_miles * factor(time_of_day)
## Model 2: log(fare) ~ factor(company) + avg_miles + factor(hour_type) +
##     factor(time_of_day) + avg_miles * factor(company) + factor(hour_type)
*
##     factor(time_of_day) + avg_miles * factor(time_of_day)
##   Res.Df    RSS Df Sum of Sq     F   Pr(>F)
## 1   4835 233.50
## 2   4833 232.89  2   0.61727 6.405 0.001667 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The partial F-test indicates avg_miles*company must be kept (F= 6.405, df= 2, 4833, p-value = 0.001667 < 0.05)

```
taxi2_lm_red_4_int_red_2o = lm ( log (fare) ~ factor(company) + poly
(avg_miles, 2, raw = TRUE) + factor(hour_type) + factor(time_of_day) +
avg_miles*factor(company) + factor(hour_type)*factor(time_of_day) +
avg_miles*factor(time_of_day), data = taxi_data2)
summary (taxi2_lm_red_4_int_red_2o)

##
## Call:
## lm(formula = log(fare) ~ factor(company) + poly(avg_miles, 2,
##     raw = TRUE) + factor(hour_type) + factor(time_of_day) + avg_miles *
##     factor(company) + factor(hour_type) * factor(time_of_day) +
##     avg_miles * factor(time_of_day), data = taxi_data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21669 -0.12196 -0.01271  0.12728  2.55163
##
## Coefficients: (2 not defined because of singularities)
##                                                         Estimate
## (Intercept)                                            1.7707983
## factor(company)107                                    -0.0096445
## factor(company)109                                     0.0032356
## poly(avg_miles, 2, raw = TRUE)1                        0.2172737
## poly(avg_miles, 2, raw = TRUE)2                       -0.0058373
## factor(hour_type)rush_hour                             0.0115151
## factor(time_of_day)Evening                            -0.0095488
## factor(time_of_day)Morning                            -0.0038840
## factor(time_of_day)Night                              -0.0453562
## avg_miles                                                     NA
## factor(company)107:avg_miles                           0.0047487
## factor(company)109:avg_miles                           0.0015099
## factor(hour_type)rush_hour:factor(time_of_day)Evening  0.0727421
## factor(hour_type)rush_hour:factor(time_of_day)Morning -0.0239461
## factor(hour_type)rush_hour:factor(time_of_day)Night           NA
## factor(time_of_day)Evening:avg_miles                  -0.0043307
## factor(time_of_day)Morning:avg_miles                  -0.0011791
## factor(time_of_day)Night:avg_miles                     0.0010516
##                                                       Std. Error t value
## (Intercept)                                            0.0168012 105.397
## factor(company)107                                     0.0120889  -0.798
## factor(company)109                                     0.0158999   0.203
## poly(avg_miles, 2, raw = TRUE)1                        0.0039331  55.242
## poly(avg_miles, 2, raw = TRUE)2                        0.0002245 -25.998
## factor(hour_type)rush_hour                             0.0115699   0.995
```

```
## factor(time_of_day)Evening                                    0.0144856  -0.659
## factor(time_of_day)Morning                                    0.0177275  -0.219
## factor(time_of_day)Night                                      0.0163608  -2.772
## avg_miles                                                            NA      NA
## factor(company)107:avg_miles                                  0.0015708   3.023
## factor(company)109:avg_miles                                  0.0019688   0.767
## factor(hour_type)rush_hour:factor(time_of_day)Evening  0.0173367   4.196
## factor(hour_type)rush_hour:factor(time_of_day)Morning  0.0195823  -1.223
## factor(hour_type)rush_hour:factor(time_of_day)Night           NA      NA
## factor(time_of_day)Evening:avg_miles                          0.0016341  -2.650
## factor(time_of_day)Morning:avg_miles                          0.0019356  -0.609
## factor(time_of_day)Night:avg_miles                            0.0022699   0.463
##                                                              Pr(>|t|)
## (Intercept)                                                   < 2e-16 ***
## factor(company)107                                            0.42503
## factor(company)109                                            0.83876
## poly(avg_miles, 2, raw = TRUE)1                               < 2e-16 ***
## poly(avg_miles, 2, raw = TRUE)2                               < 2e-16 ***
## factor(hour_type)rush_hour                                    0.31966
## factor(time_of_day)Evening                                    0.50980
## factor(time_of_day)Morning                                    0.82659
## factor(time_of_day)Night                                      0.00559 **
## avg_miles                                                          NA
## factor(company)107:avg_miles                                  0.00252 **
## factor(company)109:avg_miles                                  0.44316
## factor(hour_type)rush_hour:factor(time_of_day)Evening 2.77e-05 ***
## factor(hour_type)rush_hour:factor(time_of_day)Morning  0.22145
## factor(hour_type)rush_hour:factor(time_of_day)Night         NA
## factor(time_of_day)Evening:avg_miles                          0.00807 **
## factor(time_of_day)Morning:avg_miles                          0.54243
## factor(time_of_day)Night:avg_miles                            0.64319
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2056 on 4832 degrees of freedom
## Multiple R-squared:  0.8851, Adjusted R-squared:  0.8848
## F-statistic:  2482 on 15 and 4832 DF,  p-value: < 2.2e-16
```

```
taxi2_lm_red_4_int_red_9o = lm ( log (fare) ~ factor(company) + poly
(avg_miles, 9, raw = TRUE) + factor(hour_type) + factor(time_of_day) +
avg_miles*factor(company) + factor(hour_type)*factor(time_of_day) +
avg_miles*factor(time_of_day), data = taxi_data2)
summary (taxi2_lm_red_4_int_red_9o)
```

```
##
## Call:
## lm(formula = log(fare) ~ factor(company) + poly(avg_miles, 9,
##     raw = TRUE) + factor(hour_type) + factor(time_of_day) + avg_miles *
##     factor(company) + factor(hour_type) * factor(time_of_day) +
##     avg_miles * factor(time_of_day), data = taxi_data2)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23388 -0.11380 -0.00598  0.11740  2.59517
##
## Coefficients: (2 not defined because of singularities)
##                                                          Estimate
## (Intercept)                                             1.844e+00
## factor(company)107                                     -1.093e-02
## factor(company)109                                      2.850e-03
## poly(avg_miles, 9, raw = TRUE)1                         3.336e-01
## poly(avg_miles, 9, raw = TRUE)2                        -4.561e-01
## poly(avg_miles, 9, raw = TRUE)3                         3.611e-01
## poly(avg_miles, 9, raw = TRUE)4                        -1.306e-01
## poly(avg_miles, 9, raw = TRUE)5                         2.582e-02
## poly(avg_miles, 9, raw = TRUE)6                        -2.984e-03
## poly(avg_miles, 9, raw = TRUE)7                         2.013e-04
## poly(avg_miles, 9, raw = TRUE)8                        -7.342e-06
## poly(avg_miles, 9, raw = TRUE)9                         1.118e-07
## factor(hour_type)rush_hour                              1.038e-02
## factor(time_of_day)Evening                             -2.315e-02
## factor(time_of_day)Morning                             -6.448e-03
## factor(time_of_day)Night                               -6.461e-02
## avg_miles                                                      NA
## factor(company)107:avg_miles                            4.918e-03
## factor(company)109:avg_miles                            1.329e-03
## factor(hour_type)rush_hour:factor(time_of_day)Evening  7.359e-02
## factor(hour_type)rush_hour:factor(time_of_day)Morning -2.100e-02
## factor(hour_type)rush_hour:factor(time_of_day)Night           NA
## factor(time_of_day)Evening:avg_miles                   -2.715e-03
## factor(time_of_day)Morning:avg_miles                   -5.278e-04
## factor(time_of_day)Night:avg_miles                      2.255e-03
##                                                        Std. Error t value
## (Intercept)                                             8.961e-01   2.057
## factor(company)107                                      1.179e-02  -0.927
## factor(company)109                                      1.550e-02   0.184
## poly(avg_miles, 9, raw = TRUE)1                         1.643e+00   0.203
## poly(avg_miles, 9, raw = TRUE)2                         1.233e+00  -0.370
## poly(avg_miles, 9, raw = TRUE)3                         4.985e-01   0.724
## poly(avg_miles, 9, raw = TRUE)4                         1.203e-01  -1.085
## poly(avg_miles, 9, raw = TRUE)5                         1.810e-02   1.426
## poly(avg_miles, 9, raw = TRUE)6                         1.710e-03  -1.745
## poly(avg_miles, 9, raw = TRUE)7                         9.847e-05   2.044
## poly(avg_miles, 9, raw = TRUE)8                         3.155e-06  -2.327
## poly(avg_miles, 9, raw = TRUE)9                         4.308e-08   2.595
## factor(hour_type)rush_hour                              1.128e-02   0.920
## factor(time_of_day)Evening                              1.418e-02  -1.632
## factor(time_of_day)Morning                              1.729e-02  -0.373
## factor(time_of_day)Night                                1.613e-02  -4.007
## avg_miles                                                      NA      NA
```

```
## factor(company)107:avg_miles                          1.535e-03   3.205
## factor(company)109:avg_miles                          1.920e-03   0.693
## factor(hour_type)rush_hour:factor(time_of_day)Evening 1.693e-02   4.347
## factor(hour_type)rush_hour:factor(time_of_day)Morning 1.908e-02  -1.100
## factor(hour_type)rush_hour:factor(time_of_day)Night          NA      NA
## factor(time_of_day)Evening:avg_miles                  1.608e-03  -1.688
## factor(time_of_day)Morning:avg_miles                  1.890e-03  -0.279
## factor(time_of_day)Night:avg_miles                    2.238e-03   1.007
##                                                      Pr(>|t|)
## (Intercept)                                           0.03972 *
## factor(company)107                                    0.35385
## factor(company)109                                    0.85412
## poly(avg_miles, 9, raw = TRUE)1                       0.83912
## poly(avg_miles, 9, raw = TRUE)2                       0.71148
## poly(avg_miles, 9, raw = TRUE)3                       0.46881
## poly(avg_miles, 9, raw = TRUE)4                       0.27776
## poly(avg_miles, 9, raw = TRUE)5                       0.15388
## poly(avg_miles, 9, raw = TRUE)6                       0.08110 .
## poly(avg_miles, 9, raw = TRUE)7                       0.04100 *
## poly(avg_miles, 9, raw = TRUE)8                       0.02000 *
## poly(avg_miles, 9, raw = TRUE)9                       0.00948 **
## factor(hour_type)rush_hour                            0.35775
## factor(time_of_day)Evening                            0.10270
## factor(time_of_day)Morning                            0.70919
## factor(time_of_day)Night                             6.25e-05 ***
## avg_miles                                                  NA
## factor(company)107:avg_miles                          0.00136 **
## factor(company)109:avg_miles                          0.48863
## factor(hour_type)rush_hour:factor(time_of_day)Evening 1.41e-05 ***
## factor(hour_type)rush_hour:factor(time_of_day)Morning  0.27133
## factor(hour_type)rush_hour:factor(time_of_day)Night         NA
## factor(time_of_day)Evening:avg_miles                  0.09151 .
## factor(time_of_day)Morning:avg_miles                  0.78003
## factor(time_of_day)Night:avg_miles                    0.31377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2003 on 4825 degrees of freedom
## Multiple R-squared:  0.8911, Adjusted R-squared:  0.8906
## F-statistic:  1795 on 22 and 4825 DF,  p-value: < 2.2e-16
```

```
taxi2_lm_red_4_int_red_15o = lm ( log (fare) ~ factor(company) + poly
(avg_miles, 15, raw = TRUE) + factor(hour_type) + factor(time_of_day) +
avg_miles*factor(company) + factor(hour_type)*factor(time_of_day) +
avg_miles*factor(time_of_day), data = taxi_data2)
summary (taxi2_lm_red_4_int_red_15o)
```

```
##
## Call:
## lm(formula = log(fare) ~ factor(company) + poly(avg_miles, 15,
```

```
##     raw = TRUE) + factor(hour_type) + factor(time_of_day) + avg_miles *
##     factor(company) + factor(hour_type) * factor(time_of_day) +
##     avg_miles * factor(time_of_day), data = taxi_data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23468 -0.11554 -0.00498  0.11478  2.59910
##
## Coefficients: (3 not defined because of singularities)
##                                                           Estimate
## (Intercept)                                              3.683e+01
## factor(company)107                                      -1.139e-02
## factor(company)109                                       2.518e-03
## poly(avg_miles, 15, raw = TRUE)1                        -1.048e+02
## poly(avg_miles, 15, raw = TRUE)2                         1.382e+02
## poly(avg_miles, 15, raw = TRUE)3                        -1.059e+02
## poly(avg_miles, 15, raw = TRUE)4                         5.273e+01
## poly(avg_miles, 15, raw = TRUE)5                        -1.801e+01
## poly(avg_miles, 15, raw = TRUE)6                         4.354e+00
## poly(avg_miles, 15, raw = TRUE)7                        -7.561e-01
## poly(avg_miles, 15, raw = TRUE)8                         9.468e-02
## poly(avg_miles, 15, raw = TRUE)9                        -8.469e-03
## poly(avg_miles, 15, raw = TRUE)10                        5.258e-04
## poly(avg_miles, 15, raw = TRUE)11                       -2.116e-05
## poly(avg_miles, 15, raw = TRUE)12                        4.524e-07
## poly(avg_miles, 15, raw = TRUE)13                               NA
## poly(avg_miles, 15, raw = TRUE)14                       -2.233e-10
## poly(avg_miles, 15, raw = TRUE)15                        3.492e-12
## factor(hour_type)rush_hour                               1.133e-02
## factor(time_of_day)Evening                              -2.348e-02
## factor(time_of_day)Morning                              -6.182e-03
## factor(time_of_day)Night                                -6.486e-02
## avg_miles                                                       NA
## factor(company)107:avg_miles                             4.983e-03
## factor(company)109:avg_miles                             1.322e-03
## factor(hour_type)rush_hour:factor(time_of_day)Evening  7.227e-02
## factor(hour_type)rush_hour:factor(time_of_day)Morning -2.481e-02
## factor(hour_type)rush_hour:factor(time_of_day)Night           NA
## factor(time_of_day)Evening:avg_miles                    -2.320e-03
## factor(time_of_day)Morning:avg_miles                    -4.727e-04
## factor(time_of_day)Night:avg_miles                       2.646e-03
##                                                         Std. Error t value
## (Intercept)                                              2.361e+01   1.560
## factor(company)107                                       1.179e-02  -0.966
## factor(company)109                                       1.549e-02   0.163
## poly(avg_miles, 15, raw = TRUE)1                         6.708e+01  -1.562
## poly(avg_miles, 15, raw = TRUE)2                         8.362e+01   1.652
## poly(avg_miles, 15, raw = TRUE)3                         6.068e+01  -1.745
## poly(avg_miles, 15, raw = TRUE)4                         2.867e+01   1.839
## poly(avg_miles, 15, raw = TRUE)5                         9.343e+00  -1.928
```

```
## poly(avg_miles, 15, raw = TRUE)6                          2.167e+00   2.009
## poly(avg_miles, 15, raw = TRUE)7                          3.637e-01  -2.079
## poly(avg_miles, 15, raw = TRUE)8                          4.430e-02   2.137
## poly(avg_miles, 15, raw = TRUE)9                          3.878e-03  -2.184
## poly(avg_miles, 15, raw = TRUE)10                         2.369e-04   2.219
## poly(avg_miles, 15, raw = TRUE)11                         9.433e-06  -2.244
## poly(avg_miles, 15, raw = TRUE)12                         2.004e-07   2.258
## poly(avg_miles, 15, raw = TRUE)13                                NA      NA
## poly(avg_miles, 15, raw = TRUE)14                         9.879e-11  -2.260
## poly(avg_miles, 15, raw = TRUE)15                         1.552e-12   2.250
## factor(hour_type)rush_hour                                1.127e-02   1.005
## factor(time_of_day)Evening                                1.416e-02  -1.658
## factor(time_of_day)Morning                                1.726e-02  -0.358
## factor(time_of_day)Night                                  1.613e-02  -4.022
## avg_miles                                                        NA      NA
## factor(company)107:avg_miles                              1.534e-03   3.248
## factor(company)109:avg_miles                              1.918e-03   0.689
## factor(hour_type)rush_hour:factor(time_of_day)Evening  1.690e-02   4.276
## factor(hour_type)rush_hour:factor(time_of_day)Morning  1.906e-02  -1.301
## factor(hour_type)rush_hour:factor(time_of_day)Night           NA      NA
## factor(time_of_day)Evening:avg_miles                      1.611e-03  -1.440
## factor(time_of_day)Morning:avg_miles                      1.889e-03  -0.250
## factor(time_of_day)Night:avg_miles                        2.240e-03   1.182
##                                                         Pr(>|t|)
## (Intercept)                                              0.11889
## factor(company)107                                       0.33391
## factor(company)109                                       0.87091
## poly(avg_miles, 15, raw = TRUE)1                         0.11844
## poly(avg_miles, 15, raw = TRUE)2                         0.09857 .
## poly(avg_miles, 15, raw = TRUE)3                         0.08101 .
## poly(avg_miles, 15, raw = TRUE)4                         0.06598 .
## poly(avg_miles, 15, raw = TRUE)5                         0.05390 .
## poly(avg_miles, 15, raw = TRUE)6                         0.04461 *
## poly(avg_miles, 15, raw = TRUE)7                         0.03767 *
## poly(avg_miles, 15, raw = TRUE)8                         0.03261 *
## poly(avg_miles, 15, raw = TRUE)9                         0.02900 *
## poly(avg_miles, 15, raw = TRUE)10                        0.02651 *
## poly(avg_miles, 15, raw = TRUE)11                        0.02490 *
## poly(avg_miles, 15, raw = TRUE)12                        0.02399 *
## poly(avg_miles, 15, raw = TRUE)13                             NA
## poly(avg_miles, 15, raw = TRUE)14                        0.02385 *
## poly(avg_miles, 15, raw = TRUE)15                        0.02450 *
## factor(hour_type)rush_hour                               0.31496
## factor(time_of_day)Evening                               0.09741 .
## factor(time_of_day)Morning                               0.72021
## factor(time_of_day)Night                                5.85e-05 ***
## avg_miles                                                     NA
## factor(company)107:avg_miles                             0.00117 **
## factor(company)109:avg_miles                             0.49075
## factor(hour_type)rush_hour:factor(time_of_day)Evening 1.94e-05 ***
```

```
## factor(hour_type)rush_hour:factor(time_of_day)Morning  0.19322
## factor(hour_type)rush_hour:factor(time_of_day)Night         NA
## factor(time_of_day)Evening:avg_miles                 0.15001
## factor(time_of_day)Morning:avg_miles                 0.80242
## factor(time_of_day)Night:avg_miles                   0.23746
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1999 on 4820 degrees of freedom
## Multiple R-squared:  0.8917, Adjusted R-squared:  0.8911
## F-statistic:  1469 on 27 and 4820 DF,  p-value: < 2.2e-16
```

```r
print ("Adj. R2")
```

```
## [1] "Adj. R2"
```

```r
summary (taxi2_lm_red_4_int_red_9o)$adj.r.sq
```

```
## [1] 0.8906006
```

```r
print ("RMSE")
```

```
## [1] "RMSE"
```

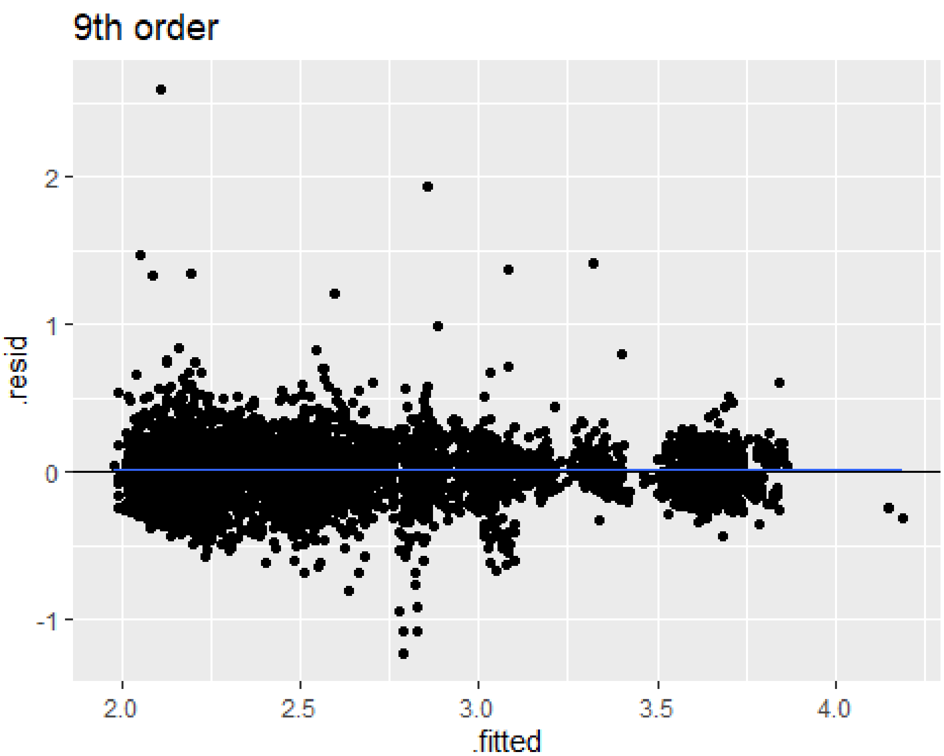
```r
sigma (taxi2_lm_red_4_int_red_9o)
```

```
## [1] 0.2003446
```

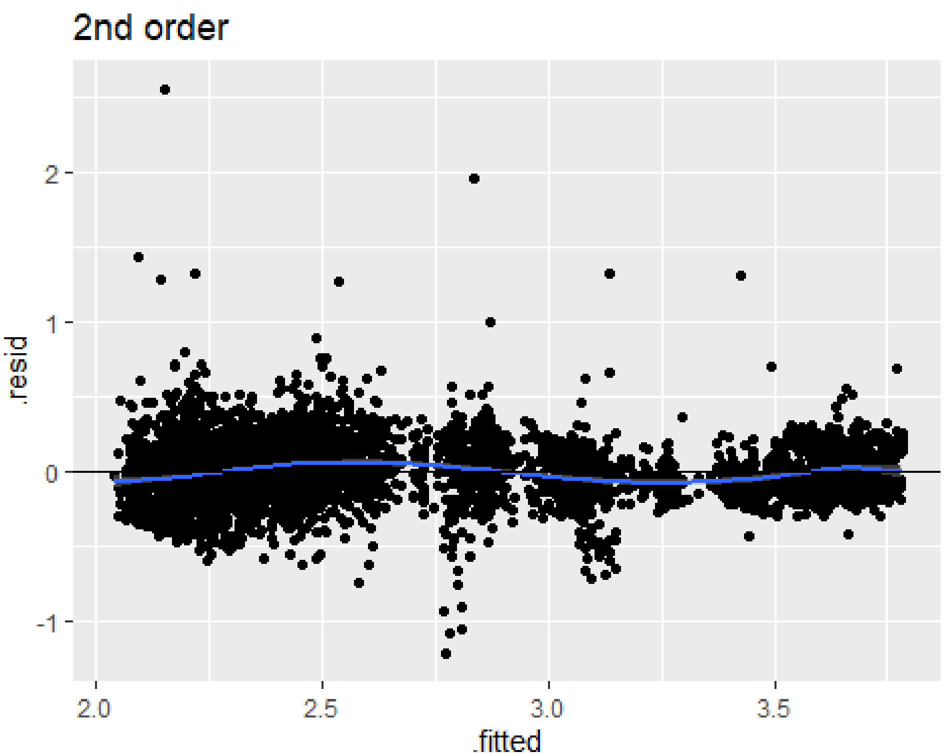
Test of assumptions

```r
ggplot (taxi2_lm_red_4_int_red_9o, aes ( x = .fitted, y = .resid)) +
  geom_point () + geom_smooth () +
  geom_hline (yintercept = 0) +
  ggtitle ("9th order")
```

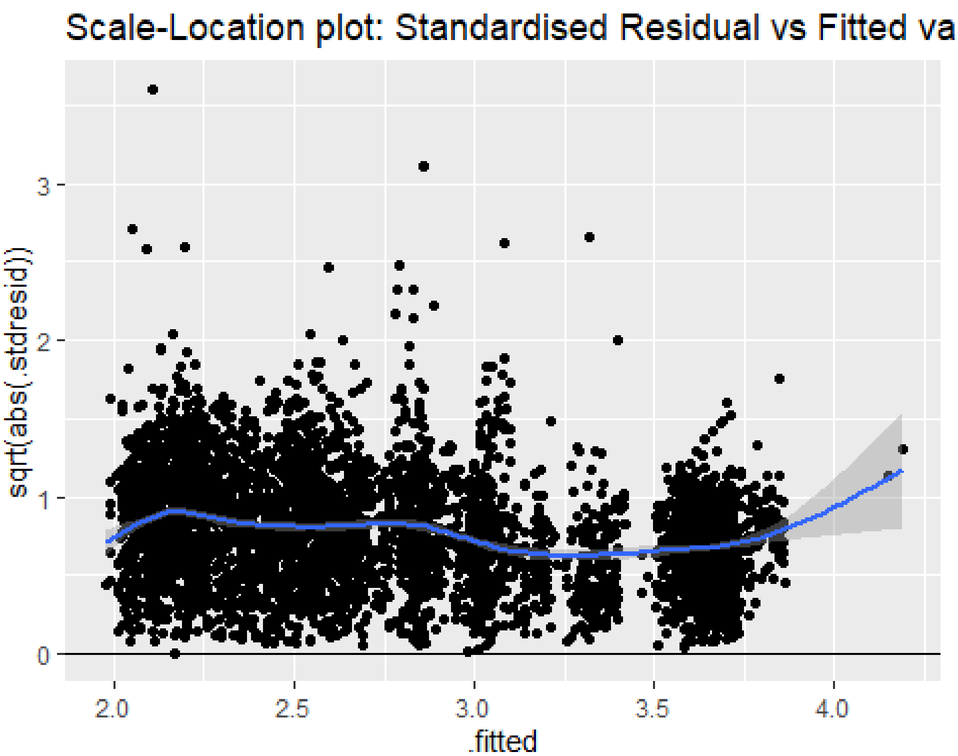
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



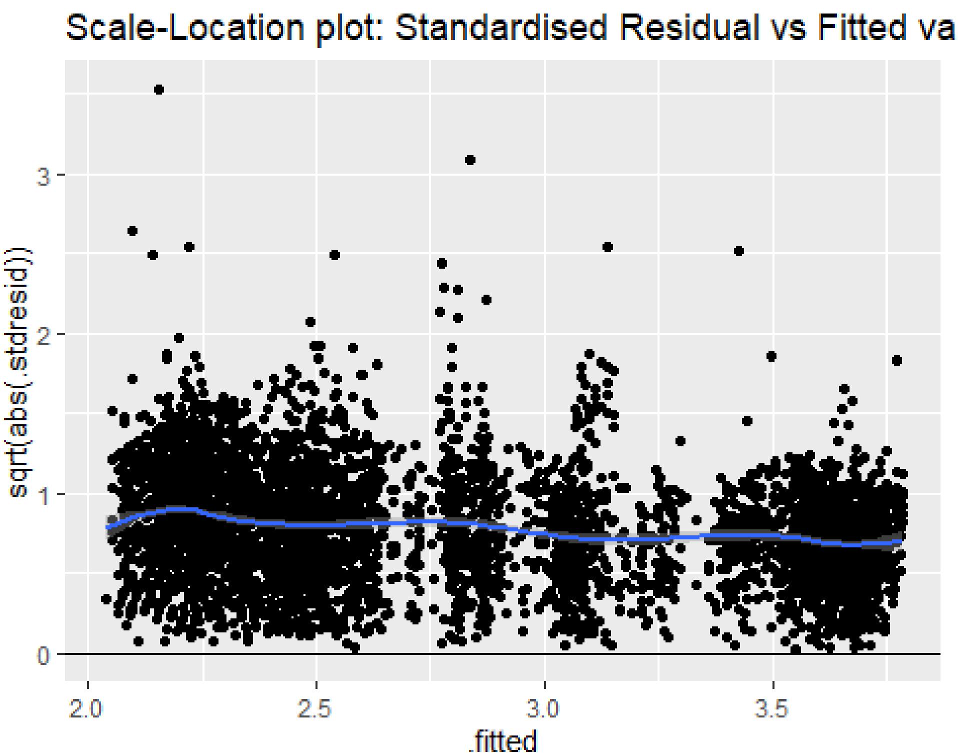

```
ggplot (taxi2_lm_red_4_int_red_2o, aes ( x = .fitted, y = .resid)) +
  geom_point () + geom_smooth () +
  geom_hline (yintercept = 0) +
  ggtitle ("2nd order")

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



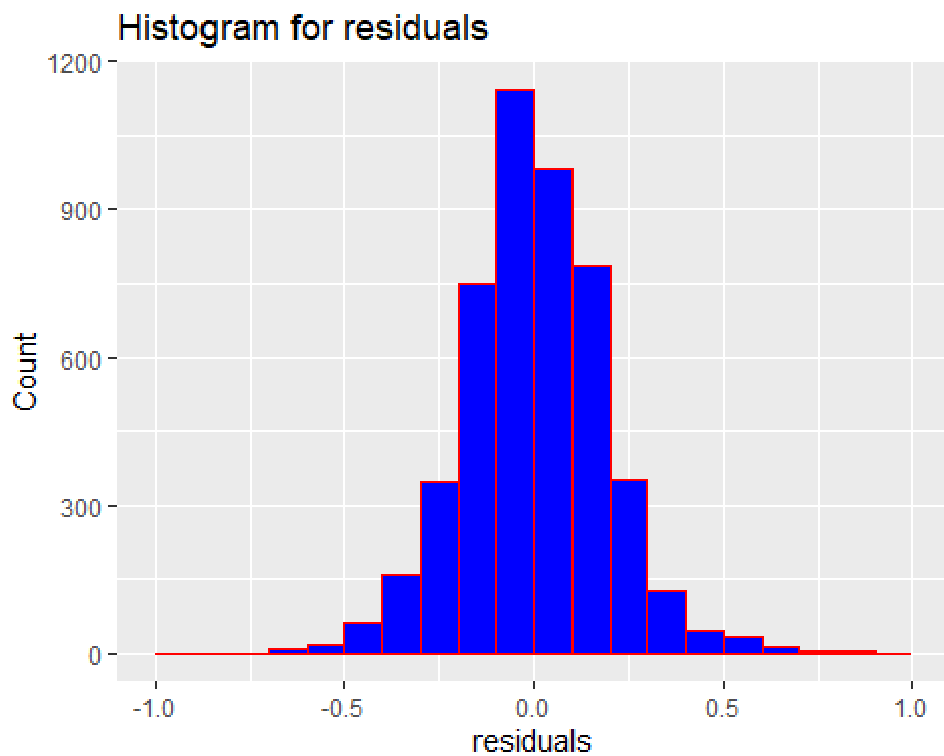

```
ggplot (taxi2_lm_red_4_int_red_9o, aes ( x = .fitted, y = sqrt ( abs
(.stdresid)))) +
  geom_point () + geom_smooth () +
  geom_hline (yintercept = 0) +
  ggtitle ("Scale-Location plot: Standardised Residual vs Fitted values, 9th
order")

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



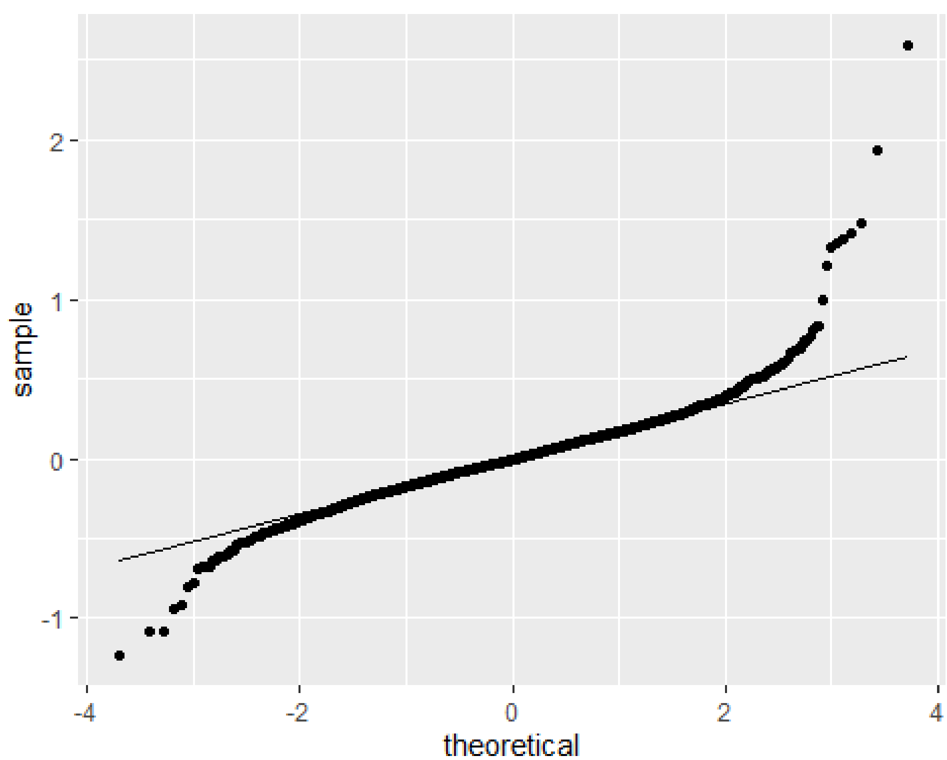

```
ggplot (taxi2_lm_red_4_int_red_2o, aes ( x = .fitted, y = sqrt ( abs
(.stdresid)))) +
  geom_point () + geom_smooth () +
  geom_hline (yintercept = 0) +
  ggtitle ("Scale-Location plot: Standardised Residual vs Fitted values, 2th
order")

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



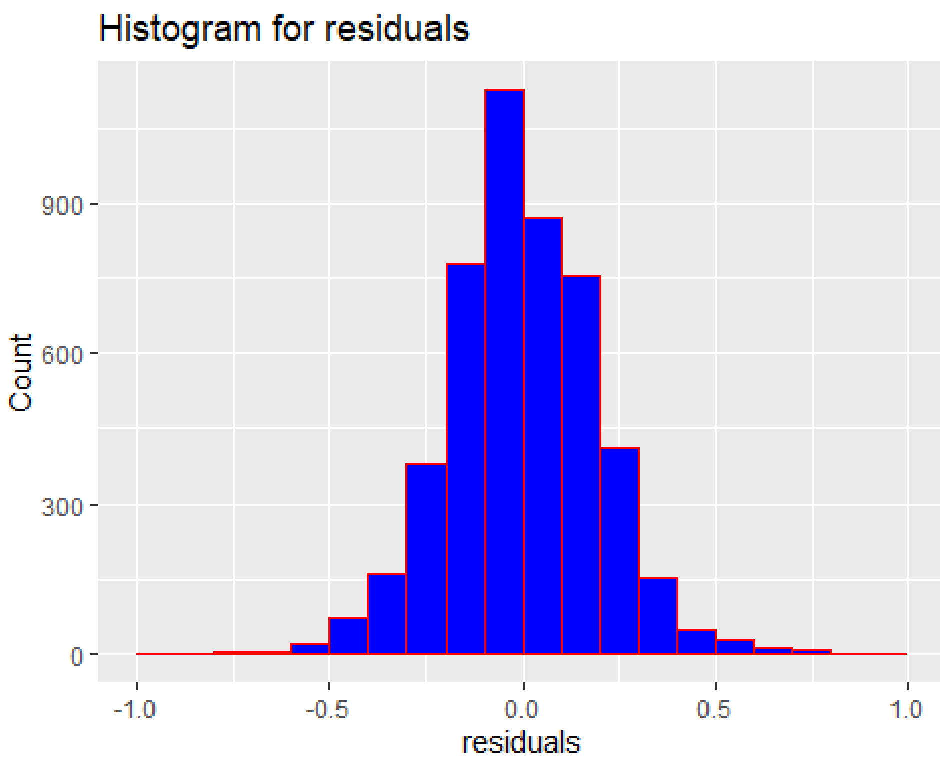

```
#   Bp test, 9th order model
bptest (taxi2_lm_red_4_int_red_9o)

##
##  studentized Breusch-Pagan test
##
## data:  taxi2_lm_red_4_int_red_9o
## BP = 76.953, df = 22, p-value = 5.103e-08

#   H0 : heteroscedasticity is not present

#   Bp test, 2nd order model
bptest (taxi2_lm_red_4_int_red_2o)

##
##  studentized Breusch-Pagan test
##
## data:  taxi2_lm_red_4_int_red_2o
## BP = 62.779, df = 15, p-value = 8.343e-08

ggplot ( data = taxi_data2, aes ( residuals (taxi2_lm_red_4_int_red_9o))) +
  geom_histogram (breaks = seq (-1, 1, by = 0.1), col = "red", fill = "blue")
+
  labs ( title = "Histogram for residuals") +
  labs ( x = "residuals", y = "Count")
```



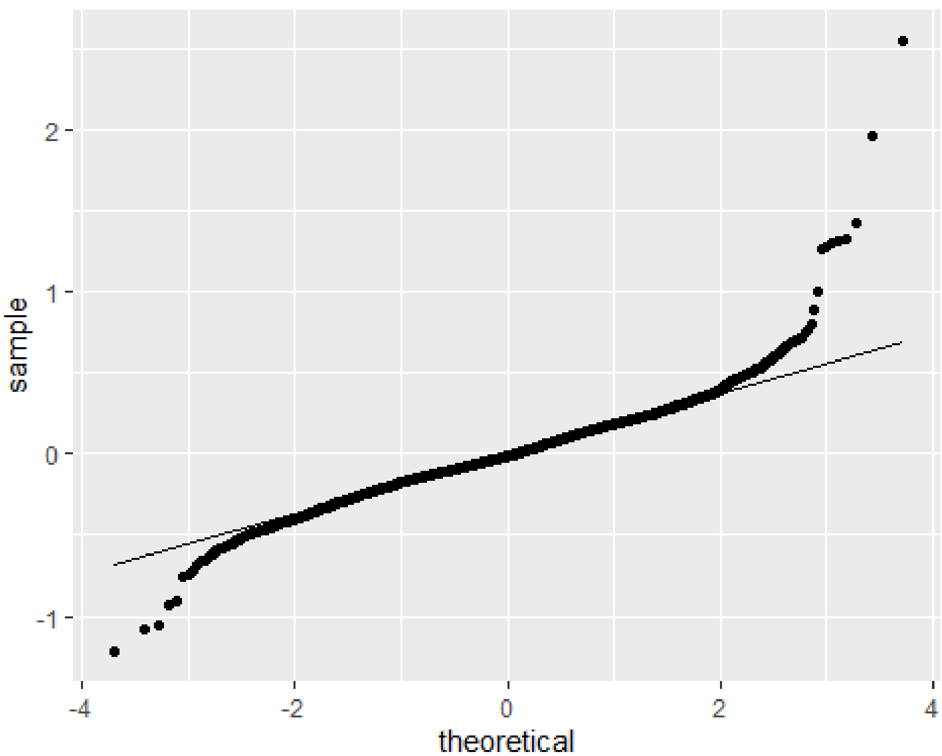

```r
ggplot (taxi_data2, aes ( sample = taxi2_lm_red_4_int_red_9o$residuals)) +
  stat_qq () +
  stat_qq_line ()
```



```
shapiro.test ( residuals (taxi2_lm_red_4_int_red_9o))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(taxi2_lm_red_4_int_red_9o)
## W = 0.93923, p-value < 2.2e-16

#   H0 : model is normal

ggplot ( data = taxi_data2, aes ( residuals (taxi2_lm_red_4_int_red_2o))) +
  geom_histogram (breaks = seq (-1, 1, by = 0.1), col = "red", fill = "blue")
+
  labs ( title = "Histogram for residuals") +
  labs ( x ="residuals", y = "Count")
```



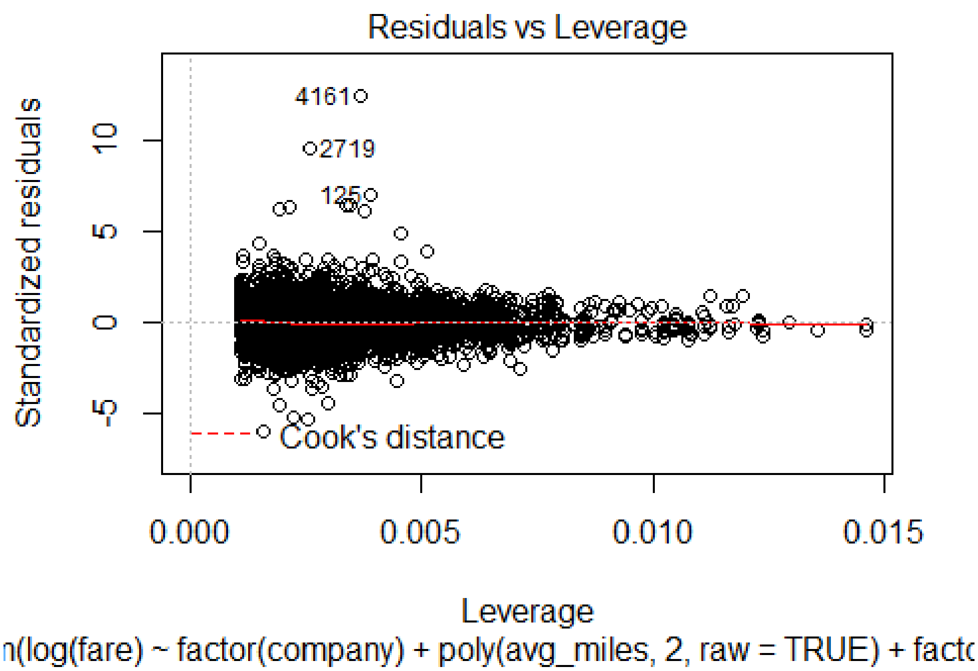

```
ggplot (taxi_data2, aes ( sample = taxi2_lm_red_4_int_red_2o$residuals)) +
  stat_qq () +
  stat_qq_line ()
```



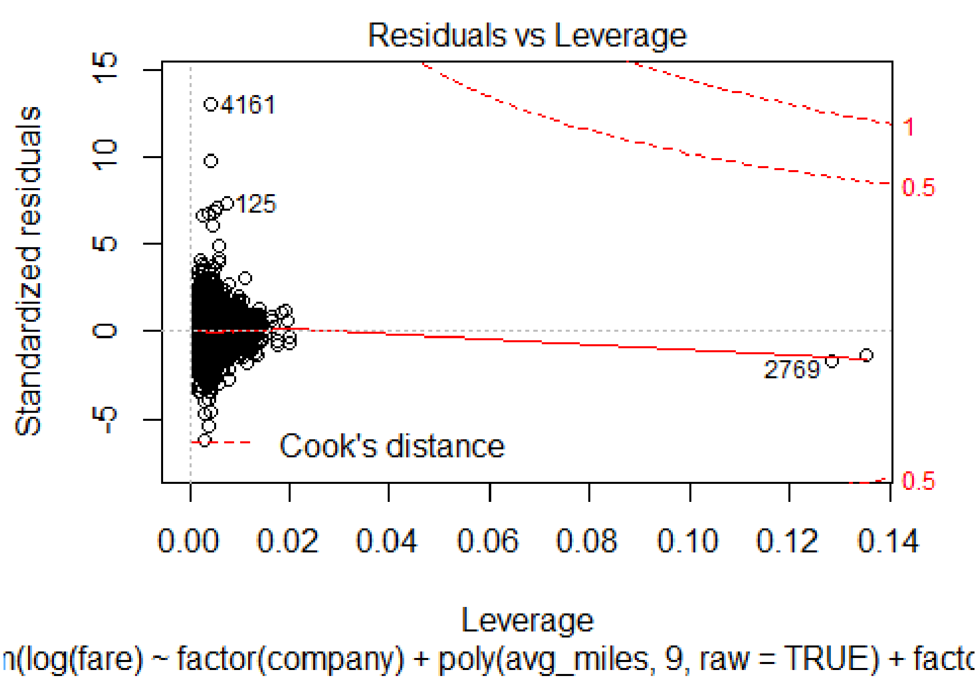

```
shapiro.test ( residuals (taxi2_lm_red_4_int_red_2o))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(taxi2_lm_red_4_int_red_2o)
## W = 0.94872, p-value < 2.2e-16

#   H0 : model is normal
```

Outliers

```
plot (taxi2_lm_red_4_int_red_2o, which = 5)
```




```r
plot (taxi2_lm_red_4_int_red_9o, which = 5)
```



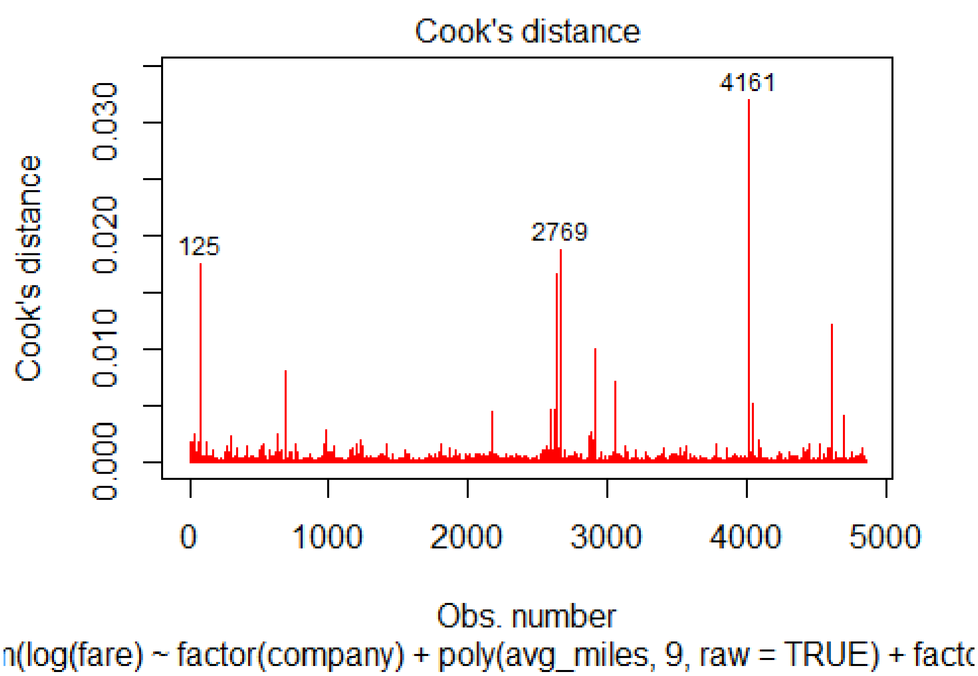

```r
taxi_data[cooks.distance (taxi2_lm_red_4_int_red_9o) > 0.5,]
```

```
##  [1] X                    pickup_area          dropoff_area
##  [4] trip_miles           trip_seconds         fare
##  [7] trip_start_timestamp tips                 tolls
## [10] trip_total           payment_type         company
## [13] extras               pickup_dropoff       avg_miles
## [16] avg_minutes          hours                months
## [19] day_of_week          hour_type            tip_pct
## [22] tip_type             pickup_dropoff_dummy weekend
## [25] season               time_of_day
## <0 rows> (or 0-length row.names)
```

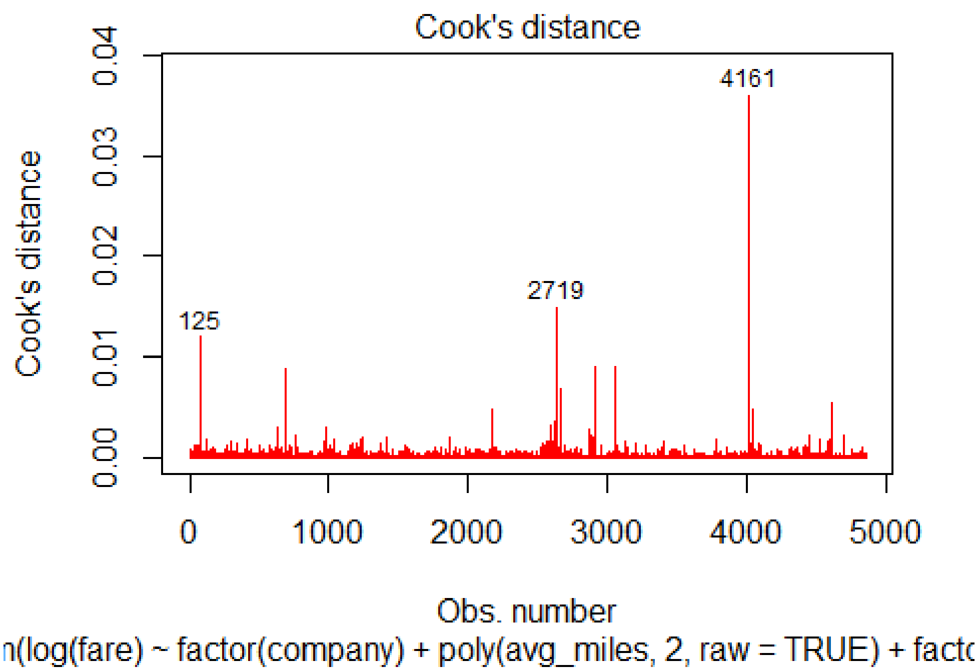
```
taxi_data[cooks.distance (taxi2_lm_red_4_int_red_2o) > 0.5,]
```

```
##  [1] X                    pickup_area          dropoff_area
##  [4] trip_miles           trip_seconds         fare
##  [7] trip_start_timestamp tips                 tolls
## [10] trip_total           payment_type         company
## [13] extras               pickup_dropoff       avg_miles
## [16] avg_minutes          hours                months
## [19] day_of_week          hour_type            tip_pct
## [22] tip_type             pickup_dropoff_dummy weekend
## [25] season               time_of_day
## <0 rows> (or 0-length row.names)
```

```
plot (taxi2_lm_red_4_int_red_9o, pch = 10, col = "red", which = c(4))
```



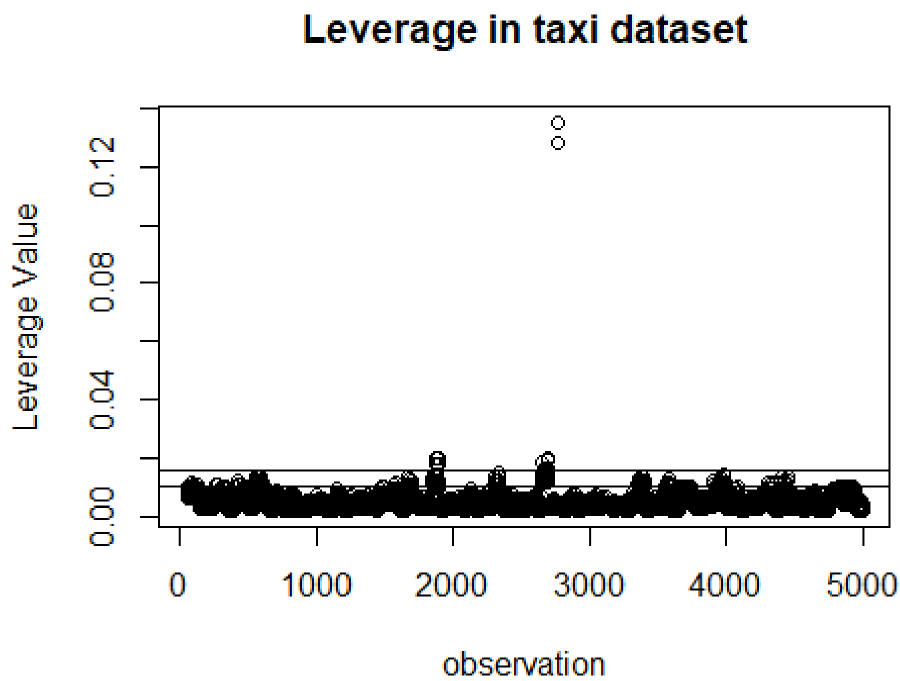

```
plot (taxi2_lm_red_4_int_red_2o, pch = 10, col = "red", which = c(4))
```




```
lev = hatvalues (taxi2_lm_red_4_int_red_9o)
p = length ( coef (taxi2_lm_red_4_int_red_9o))
n = nrow (taxi_data2)
outlier = lev[lev > (3*p/n)]
print (outlier)
```

```
##       1881       1883       1887       1898       1900       2655
## 0.02013312 0.01745808 0.01745808 0.02013312 0.01745808 0.01856337
##       2656       2677       2686       2692       2693       2769
## 0.01621518 0.01647276 0.01952012 0.01913764 0.01670323 0.12835472
##       2770
## 0.13520686
```

```
plot (rownames (taxi_data2), lev, main = "Leverage in taxi dataset", xlab =
"observation", ylab = "Leverage Value")
abline (h = 2*p/n, lty = 1)
abline (h = 3*p/n, lty = 1)
```


Leverage in taxi dataset
Leverage Value
0.00
0.04
0.08
0.12
0
1000
2000
3000
4000
5000
observation