Car price model Linear regression

Predict the price of the car model with Linear regression

Load the data

```
load("rda/carprice.rda")
str(carprice)
```

```
## 'data.frame':    205 obs. of  26 variables:
##  $ car_ID          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ symboling       : int  3 3 1 2 2 2 1 1 1 0 ...
##  $ CarName         : chr  "alfa-romero giulia" "alfa-romero stelvio" "alfa-romero
##  $ fueltype        : chr  "gas" "gas" "gas" "gas" ...
##  $ aspiration      : chr  "std" "std" "std" "std" ...
##  $ doornumber      : chr  "two" "two" "two" "four" ...
##  $ carbody         : chr  "convertible" "convertible" "hatchback" "sedan" ...
##  $ drivewheel      : chr  "rwd" "rwd" "rwd" "fwd" ...
##  $ enginelocation  : chr  "front" "front" "front" "front" ...
##  $ wheelbase       : num  88.6 88.6 94.5 99.8 99.4 ...
##  $ carlength       : num  169 169 171 177 177 ...
##  $ carwidth        : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...
##  $ carheight       : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...
##  $ curbweight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 3053 ...
##  $ enginetype      : chr  "dohc" "dohc" "ohcv" "ohc" ...
##  $ cylindernumber  : chr  "four" "four" "six" "four" ...
##  $ enginesize      : int  130 130 152 109 136 136 136 136 131 131 ...
##  $ fuelsystem      : chr  "mpfi" "mpfi" "mpfi" "mpfi" ...
##  $ boreratio       : num  3.47 3.47 2.68 3.19 3.19 3.19 3.19 3.19 3.13 3.13 ...
##  $ stroke          : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 3.4 ...
##  $ compressionratio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...
##  $ horsepower      : int  111 111 154 102 115 110 110 110 140 160 ...
##  $ peakrpm         : int  5000 5000 5000 5500 5500 5500 5500 5500 5500 5500 ...
##  $ citympg         : int  21 21 19 24 18 19 19 19 17 16 ...
##  $ highwaympg      : int  27 27 26 30 22 25 25 25 20 22 ...
##  $ price           : num  13495 16500 16500 13950 17450 ...
```

Convert categorical variables to factors

```
carprice$symboling <-as.factor(carprice$symboling)
carprice$cylindernumber<-as.factor(carprice$cylindernumber)
carprice$enginetype <- as.factor(carprice$enginetype)
```

```
carprice$fuelsystem<-as.factor(carprice$fuelsystem)
carprice$fueltype<-as.factor(carprice$fueltype)
carprice$aspiration<-as.factor(carprice$aspiration)
carprice$doornumber<-as.factor(carprice$doornumber)
carprice$carbody <-as.factor(carprice$carbody)
carprice$drivewheel<-as.factor(carprice$drivewheel)
carprice$enginelocation <- as.factor(carprice$enginelocation)
```

Working on variable "carprice$CarName"

```
summary(as.factor(carprice$CarName))
```

```
##                      peugeot 504                   toyota corolla
##                                6                                6
##                   toyota corona                       subaru dl
##                                6                                4
##                     honda civic                       mazda 626
##                                3                                3
##                   mitsubishi g4             mitsubishi mirage g4
##                                3                                3
##             mitsubishi outlander                  toyota mark ii
##                                3                                3
##                       audi 100ls                        bmw 320i
##                                2                                2
##                           bmw x3                     honda accord
##                                2                                2
##                  honda civic cvcc                     isuzu D-Max
##                                2                                2
##                        mazda glc                  mazda glc deluxe
##                                2                                2
##                       mazda rx-4                     mazda rx-7 gs
##                                2                                2
##                    nissan clipper                     nissan latio
##                                2                                2
##                     nissan rogue                     peugeot 604sl

##                                2                                2
##                 plymouth fury iii                  porsche cayenne
##                                2                                2
##                          saab 99e                         saab 99gle
##                                2                                2
##                          saab 99le                          subaru
##                                2                                2
##               toyota corolla 1200           toyota corolla liftback
##                                2                                2
##                   toyota starlet                 volkswagen dasher
##                                2                                2
##                      volvo 144ea                   volvo 145e (sw)
##                                2                                2
```

```
##                             2                             2
##                   volvo 244dl                   volvo 264gl
##                             2                             2
##              alfa-romero giulia        alfa-romero Quadrifoglio
##                             1                             1
##             alfa-romero stelvio                   audi 100 ls
##                             1                             1
##                      audi 4000                     audi 5000
##                             1                             1
##             audi 5000s (diesel)                      audi fox
##                             1                             1
##                         bmw x1                         bmw x4
##                             1                             1
##                         bmw x5                         bmw z4
##                             1                             1
##                 buick century        buick century luxus (sw)
##                             1                             1
##           buick century special        buick electra 225 custom
##                             1                             1
##         buick opel isuzu deluxe buick regal sport coupe (turbo)
##                             1                             1
##                 buick skyhawk                   buick skylark
##                             1                             1
##               chevrolet impala           chevrolet monte carlo
##                             1                             1
##             chevrolet vega 2300             dodge challenger se
##                             1                             1
##                dodge colt (sw)              dodge colt hardtop
##                             1                             1
##            dodge coronet custom       dodge coronet custom (sw)
##                             1                             1
##                    dodge d200              dodge dart custom
##                             1                             1
##              dodge monaco (sw)                 dodge rampage
##                             1                             1
##              honda accord cvcc                honda accord lx
##                             1                             1
##             honda civic (auto)                honda civic 1300
##                             1                             1

##             honda civic 1500 gl                   honda prelude
##                             1                             1
##             isuzu D-Max V-Cross                      isuzu MU-X
##                             1                             1
##                      jaguar xf                       jaguar xj
##                             1                             1
##                      jaguar xk                maxda glc deluxe
##                             1                             1
##                      maxda rx3                     mazda glc 4
##                             1                             1
##                mazda glc custom              mazda glc custom l
##                             1                             1
##                  mazda rx2 coupe                     mercury cougar
```

```
##                 mazda rx2 coupe                 mercury cougar
##                               1                              1
##              mitsubishi lancer              mitsubishi mirage
##                               1                              1
##             mitsubishi montero              mitsubishi pajero
##                               1                              1
##                    nissan dayz                    nissan fuga
##                               1                              1
##                    nissan gt-r                    nissan juke
##                               1                              1
##                   nissan kicks                    nissan leaf
##                               1                              1
##                    nissan note                        (Other)
##                               1                             48
```

There are multiple levels in CarName. Reduce the variables by taking only the carCompany

```
carprice$carCompany <-gsub("\\ .*", "", carprice$CarName)
str(carprice$carCompany)
```

```
##  chr [1:205] "alfa-romero" "alfa-romero" "alfa-romero" "audi" "audi" ...
```

```
carprice$carCompany <- as.factor(carprice$carCompany)
summary(carprice$carCompany)
```

```
## alfa-romero        audi         bmw       buick   chevrolet       dodge
##           3           7           8           8           3           9
##       honda       isuzu      jaguar       maxda       mazda     mercury
##          13           4           3           2          15           1
##  mitsubishi      nissan      Nissan     peugeot    plymouth    porcshce
##          13          17           1          11           7           1
##     porsche     renault        saab      subaru      toyota     toyouta
##           4           2           6          12          31           1
##   vokswagen  volkswagen       volvo          vw
##           1           9          11           2
```

```
levels(carprice$carCompany)
```

```
##  [1] "alfa-romero" "audi"        "bmw"         "buick"       "chevrolet"
##  [6] "dodge"       "honda"       "isuzu"       "jaguar"      "maxda"
## [11] "mazda"       "mercury"     "mitsubishi"  "nissan"      "Nissan"
## [16] "peugeot"     "plymouth"    "porcshce"    "porsche"     "renault"
## [21] "saab"        "subaru"      "toyota"      "toyouta"     "vokswagen"
```

```
## [26] "volkswagen"  "volvo"       "vw"
```

```
levels(carprice$carCompany)[10] <- "mazda"
levels(carprice$carCompany)[14] <- "nissan"
levels(carprice$carCompany)[16] <- "porsche"
levels(carprice$carCompany)[21] <- "toyota"
levels(carprice$carCompany)[21] <- "volkswagen"
levels(carprice$carCompany)[23] <- "volkswagen"
levels(carprice$carCompany)
```

```
##  [1] "alfa-romero" "audi"        "bmw"         "buick"       "chevrolet"
##  [6] "dodge"       "honda"       "isuzu"       "jaguar"      "mazda"
## [11] "mercury"     "mitsubishi"  "nissan"      "peugeot"     "plymouth"
## [16] "porsche"     "renault"     "saab"        "subaru"      "toyota"
## [21] "volkswagen"  "volvo"
```

Check for missing values

```
sum(is.na(carprice))
```

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

Check for duplicated data

```
which(duplicated(carprice))
```

```
## integer(0)
```

Create the dummy variables

```
# For carCompany
dummy_1 <- data.frame(model.matrix( ~carCompany, data = carprice))
dummy_1<-dummy_1[,-1]

# For carbody
dummy_2 <- data.frame(model.matrix( ~carbody, data = carprice))
dummy_2<-dummy_2[,-1]

# Drivewheel
dummy_3 <- data.frame(model.matrix( ~drivewheel, data = carprice))
```

```
dummy_3<-dummy_3[,-1]

#Engine type
dummy_4 <- data.frame(model.matrix( ~enginetype, data = carprice))
dummy_4<-dummy_4[,-1]

#cylindernumber
dummy_5 <- data.frame(model.matrix( ~cylindernumber, data = carprice))
dummy_5<-dummy_5[,-1]

# Fuelsystem
dummy_6 <- data.frame(model.matrix( ~fuelsystem, data = carprice))
dummy_6<-dummy_6[,-1]

# Symboling
dummy_7 <- data.frame(model.matrix( ~symboling, data = carprice))
dummy_7<-dummy_7[,-1]
```

Variable having 2 levels are replaced to 0&1 and converted to numeric

```
# for fueltype
levels(carprice$fueltype)<-c(1,0)
# assigning 1 to diesel and 0 to gas
carprice$fueltype<- as.numeric(levels(carprice$fueltype))[carprice$fueltype]

# for aspiration
levels(carprice$aspiration)<-c(1,0)
# Assigning 1 to "std" and 0 to "turbo"
carprice$aspiration <- as.numeric(levels(carprice$aspiration))[carprice$aspiration]

# For doornumber
levels(carprice$doornumber)<-c(1,0)
# Assigning 1 if the number of doors is 4, and 0 if the number of doors is 2.
carprice$doornumber<- as.numeric(levels(carprice$doornumber))[carprice$doornumber]

# Enginelocation
levels(carprice$enginelocation)<-c(1,0)
# Assigning 1 if the engine is front and 0 if in rear
carprice$enginelocation<- as.numeric(levels(carprice$enginelocation))[carprice$engine
```

Combine the dummy variables and the numeric columns of carprice dataset

```
carprice_1 <- cbind(carprice[ , c(1,4:6,9:14,17,19:26)], dummy_1,dummy_2,dummy_3,dumm
```

Modeling

```
# Divide you data in 70:30 and create test and train datasets

set.seed(100)
indices= sample(1:nrow(carprice_1), 0.7*nrow(carprice_1))

train=carprice_1[indices,]
test = carprice_1[-indices,]



model_1 <-lm(price~.,data=train[,-1])
summary(model_1)
```

```
##
## Call:
## lm(formula = price ~ ., data = train[, -1])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2951.16  -708.57    10.27   719.33  3038.24
##
## Coefficients: (9 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -3.733e+04  1.585e+04  -2.355 0.020866 *
## fueltype             5.414e+03  7.339e+03   0.738 0.462780
## aspiration          -3.318e+03  9.167e+02  -3.619 0.000507 ***
## doornumber           1.790e+02  5.336e+02   0.336 0.738084
## enginelocation      -7.450e+03  4.544e+03  -1.640 0.104854
## wheelbase           -2.704e+01  1.171e+02  -0.231 0.817972
## carlength            3.767e+01  6.080e+01   0.620 0.537241
## carwidth             6.603e+02  2.448e+02   2.697 0.008464 **
## carheight            2.965e+01  1.605e+02   0.185 0.853912
## curbweight           3.691e+00  1.713e+00   2.155 0.034097 *
## enginesize           1.066e+02  2.753e+01   3.873 0.000214 ***
## boreratio           -1.073e+03  1.954e+03  -0.549 0.584162
## stroke              -2.310e+03  1.112e+03  -2.077 0.040938 *
## compressionratio    -3.947e+02  5.434e+02  -0.726 0.469719
## horsepower          -2.296e+01  2.534e+01  -0.906 0.367388
## peakrpm              1.760e+00  8.673e-01   2.029 0.045691 *
## citympg              6.859e+01  1.389e+02   0.494 0.622821
## highwaympg           1.768e+01  1.256e+02   0.141 0.888342
## carCompanyaudi      -6.043e+02  2.216e+03  -0.273 0.785736
## carCompanybmw        6.609e+03  2.636e+03   2.507 0.014126 *
## carCompanybuick      4.135e+03  2.512e+03   1.646 0.103463
## carCompanychevrolet -2.068e+02  5.060e+03  -0.041 0.967495
## carCompanydodge     -5.515e+03  2.187e+03  -2.522 0.013575 *
## carCompanyhonda     -2.104e+03  2.141e+03  -0.983 0.328660
```

```
## carCompanyisuzu      -1.886e+03  2.426e+03  -0.777 0.439099
## carCompanyjaguar      1.907e+03  2.850e+03   0.669 0.505306
## carCompanymazda      -4.602e+03  1.798e+03  -2.560 0.012277 *
## carCompanymercury    -3.024e+03  2.924e+03  -1.034 0.303982
## carCompanymitsubishi -6.093e+03  2.031e+03  -3.001 0.003558 **
## carCompanynissan     -3.928e+03  1.827e+03  -2.150 0.034457 *
## carCompanypeugeot    -5.380e+03  2.519e+03  -2.136 0.035628 *
## carCompanyplymouth   -4.957e+03  2.041e+03  -2.429 0.017302 *
## carCompanyporsche     5.574e+03  5.545e+03   1.005 0.317734
## carCompanyrenault    -5.789e+03  2.258e+03  -2.563 0.012170 *
## carCompanysaab       -3.090e+03  2.227e+03  -1.387 0.169029
## carCompanysubaru     -6.656e+03  2.138e+03  -3.113 0.002538 **
## carCompanytoyota     -3.537e+03  1.668e+03  -2.120 0.036988 *
## carCompanyvolkswagen -3.698e+03  1.851e+03  -1.998 0.049040 *
## carCompanyvolvo      -1.313e+03  2.514e+03  -0.522 0.602875
## carbodyhardtop       -2.198e+03  1.320e+03  -1.665 0.099739 .
## carbodyhatchback     -2.622e+03  1.250e+03  -2.097 0.039043 *
## carbodysedan         -2.756e+03  1.336e+03  -2.063 0.042208 *
## carbodywagon         -3.273e+03  1.486e+03  -2.202 0.030454 *
## drivewheelfwd         3.192e+02  1.048e+03   0.305 0.761389
## drivewheelrwd        -1.881e+03  1.416e+03  -1.328 0.187654
## enginetypedohcv              NA         NA      NA       NA
## enginetypel                  NA         NA      NA       NA
## enginetypeohc        -1.214e+03  1.354e+03  -0.897 0.372375
## enginetypeohcf               NA         NA      NA       NA
## enginetypeohcv       -7.004e+02  1.324e+03  -0.529 0.598286
## enginetyperotor       1.064e+04  5.069e+03   2.098 0.038904 *
## cylindernumberfive   -1.294e+03  3.376e+03  -0.383 0.702564
## cylindernumberfour    4.432e+02  3.959e+03   0.112 0.911142
## cylindernumbersix     4.432e+02  2.928e+03   0.151 0.880056
## cylindernumberthree          NA         NA      NA       NA
## cylindernumbertwelve         NA         NA      NA       NA
## cylindernumbertwo            NA         NA      NA       NA
## fuelsystem2bbl        1.793e+03  1.334e+03   1.344 0.182552
## fuelsystem4bbl               NA         NA      NA       NA
## fuelsystemidi                NA         NA      NA       NA
## fuelsystemmfi        -2.391e+02  2.473e+03  -0.097 0.923213
## fuelsystemmpfi        1.624e+03  1.420e+03   1.143 0.256217
## fuelsystemspdi        7.241e+02  1.692e+03   0.428 0.669868
## fuelsystemspfi               NA         NA      NA       NA
## symboling.1          -9.376e+02  1.811e+03  -0.518 0.606109
## symboling0           -9.746e+02  2.043e+03  -0.477 0.634615
## symboling1           -2.648e+02  2.127e+03  -0.124 0.901252
## symboling2           -5.110e+02  2.237e+03  -0.228 0.819847
## symboling3           -1.195e+03  2.214e+03  -0.540 0.590981
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1436 on 83 degrees of freedom
## Multiple R-squared:  0.9819, Adjusted R-squared:  0.9691
## F-statistic: 76.48 on 59 and 83 DF,  p-value: < 2.2e-16
```

Applying stepwise approach with step <- stepAIC(model_1, direction="both")

```
step
```

```
##
## Call:
## lm(formula = price ~ aspiration + enginelocation + carlength +
##     carwidth + curbweight + enginesize + stroke + peakrpm + citympg +
##     carCompanybmw + carCompanybuick + carCompanydodge + carCompanyhonda +
##     carCompanyjaguar + carCompanymazda + carCompanymercury +
##     carCompanymitsubishi + carCompanynissan + carCompanypeugeot +
##     carCompanyplymouth + carCompanyporsche + carCompanyrenault +
##     carCompanysaab + carCompanysubaru + carCompanytoyota + carCompanyvolkswagen +
##     carbodyhardtop + carbodyhatchback + carbodysedan + carbodywagon +
##     drivewheelrwd + enginetypeohc + enginetyperotor + cylindernumberfive +
##     fuelsystem2bbl + fuelsystemmpfi + symboling.1 + symboling0 +
##     symboling3, data = train[, -1])
##
## Coefficients:
##          (Intercept)            aspiration        enginelocation
##            -33403.736             -2925.410             -9358.366
##              carlength              carwidth             curbweight
##                 42.692               582.711                3.160
##           enginesize                stroke               peakrpm
##                 88.826             -2616.977                1.245
##              citympg         carCompanybmw       carCompanybuick
##                 64.759              7982.278             5641.473
##       carCompanydodge       carCompanyhonda      carCompanyjaguar
##             -4340.035             -1370.777              4514.841
##       carCompanymazda     carCompanymercury  carCompanymitsubishi
##             -3463.008             -2615.729             -4795.527
##      carCompanynissan     carCompanypeugeot    carCompanyplymouth
##             -2946.466             -3776.672             -3497.563
##     carCompanyporsche     carCompanyrenault        carCompanysaab
##              3716.283             -4410.374             -2823.918
##      carCompanysubaru      carCompanytoyota  carCompanyvolkswagen
##             -6542.488             -2505.288             -2840.989
##        carbodyhardtop      carbodyhatchback          carbodysedan
##             -2762.398             -3026.601             -3061.322
##          carbodywagon         drivewheelrwd          enginetypeohc
##             -3442.526             -2765.527             -1232.796
##       enginetyperotor    cylindernumberfive        fuelsystem2bbl
##              8717.263             -1071.502              1006.425
##        fuelsystemmpfi           symboling.1            symboling0
##               741.423              -851.155              -660.364
##           symboling3
##              -927.210
```

Variables with high VIF and is insignificant are removed one by one

Removing carlength

Remove citympg

Remove fuelsystemmpfi

Remove carcompanyporsche

Remove fuelsystem2bbl

Remove symboling0

Remove symboling.1

Remove carcompanymercury

Remove symboling3

Remove carbodyhardtop

Remove carbodyhatchback

Remove carbodysedan

Remove carbodywagon

All variables are significant now. Variable curbweight and enginesize have high VIFs. curbweight is very less significant as compared to enginesize

Remove curbweight

Remove cylindernumberfive

Remove peakrpm (higher VIF, higher p-value as compared to othe variables in model)

Remove carCompanysaab

Remove carcompanyhonda

Remove carCompany renault

Remove drivewheelrwd

Remove carCompanyvolkswagen

Remove carCompanydodge

Remove carCompanyplymouth

Remove carCompanynissan

Remove enginetypeohc

Remove carCompanymitsubishi

Remove carCompanytoyota

Remove carcompanyPeugeot

Adjusted R-squared = 0.9437. carwidth, enginesize have high VIF. Lets see what happens when these are removed.

Removing carwidth

Adjusted R-squared decreased from 0.9437 to 0.9224, a decline of 2 percent.

Removing enginesize

Adjusted R-squared decreases from 0.9437 to 0.8926. These are fairly large decreases.

Remove variables which are comparatively less significant

Remove aspiration

Remove carcompanysubaru

Remove enginetyperotor

Remove stroke

Remove carcompany jaguar

```
model_35 <- lm(formula = price ~  enginelocation +
                 carwidth + enginesize +
                 carCompanybmw + carCompanybuick, data = train[, -1])
summary(model_35)
```

```
##
## Call:
```

```
## lm(formula = price ~ enginelocation + carwidth + enginesize +
##     carCompanybmw + carCompanybuick, data = train[, -1])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5207.4 -1499.5  -322.4  1258.5  6821.0
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -69563.79    9448.61  -7.362 1.52e-11 ***
## enginelocation  -17519.74    1630.39 -10.746  < 2e-16 ***
## carwidth          1342.90     161.70   8.305 8.58e-14 ***
## enginesize          86.62       9.03   9.592  < 2e-16 ***
## carCompanybmw     8415.83    1472.49   5.715 6.55e-08 ***
## carCompanybuick   5820.80    1189.75   4.892 2.75e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2478 on 137 degrees of freedom
## Multiple R-squared:  0.9113, Adjusted R-squared:  0.908
## F-statistic: 281.4 on 5 and 137 DF,  p-value: < 2.2e-16
```

```
vif(model_35)
```

```
##  enginelocation        carwidth      enginesize   carCompanybmw
##        1.271785        2.889275        3.293096        1.037373
## carCompanybuick
##        1.535081
```

Now there are 5 variables in the model.

Test the model on test dataset

```
Predict_1 <- predict(model_35,test[,-c(1,20)])
```

Add a new column "test_predict" into the test dataset

```
test$test_price <- Predict_1
```

Calculate the test R2

```
cor(test$price,test$test_price)
```

```
## [1] 0.9267725
```

```
cor(test$price,test$test_price)^2
```

```
## [1] 0.8589072
```