MA 58900 Applied Statistics

Auto MPG Analysis Using R

Xinyang Yang

04/23/2018

Dataset Introduction

This dataset is from UCI (Univ. of California at Irvine) Machine Learning Repository (Auto MPG Data Set). The dataset 9 variables and 398 observations. The variables are:

1. mpg: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
9. car name: string (unique for each instance)

Several preprocesses are implemented: delete missing value and delete useless information (car name):

```
> car=read.table("C:/Users/yangx/Documents/MA598_2018/project/car_nohead.txt")
> names(car)=c("mpg","cylinders","displacement","horsepower","weight","acceleration","modelyear","origin")
> head(car)
  mpg cylinders displacement horsepower weight acceleration modelyear origin
1  18         8          307        130   3504         12.0        70      1
2  15         8          350        165   3693         11.5        70      1
3  18         8          318        150   3436         11.0        70      1
4  16         8          304        150   3433         12.0        70      1
5  17         8          302        140   3449         10.5        70      1
6  15         8          429        198   4341         10.0        70      1
> dim(car)
[1] 394   8
```

After preprocessing, there are 7 variables and 394 observations kept.

Objective

In this project, there are three parts. Induvial Variable Analysis which analysis induvial variable using chi-squared test is the first part. Second part is One-Way ANOVA which analysis data by groups. Last part is to find best model to predict auto MPG.

Induvial Variable Analysis

In this dataset, there three variables (model year, cylinders and origin.) can be easily divide to groups. First, all observations can be divided by model year to 13 groups from year 1970 to year 1982.

```
> m_group = table(modelyear)
> m_group
modelyear
70 71 72 73 74 75 76 77 78 79 80 81 82
29 28 28 40 27 30 34 28 36 29 27 28 30
> chisq.test(m_group)

	Chi-squared test for given probabilities

data:  m_group
X-squared = 6.1624, df = 12, p-value = 0.9077
```

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982
Number	29	28	28	40	27	30	34	28	36	29	27	28	30

Table 1.

From table, he numbers of cars manufactured in different years in this dataset are likely equal. In chi-squared test, since the p-value is 0.9077 is greater than 0.05, so fail to reject the null hypothesis, there is no difference between groups of model year.

```
> c_group = table(cylinders)
> c_group
cylinders
  3   4   5   6   8 
  4 200   3  84 103 
> chisq.test(c_group)

	Chi-squared test for given probabilities

data:  c_group
X-squared = 338.11, df = 4, p-value < 2.2e-16

> 
> o_group = table(origin)
> o_group
origin
  1   2   3 
```

```
247  68  79
> chisq.test(o_group)

	Chi-squared test for given probabilities

data:  o_group
X-squared = 153.26, df = 2, p-value < 2.2e-16
```

After same test for cylinders and origin groups, the p-values are all less than 0.05, so reject the null hypothesis, there are difference between groups of cylinders or origin.

One-Way ANOVA

From induvial variable analysis, the best way to divide the dataset is dividing it by make year because there is smaller difference between groups than dividing by other variables such as cylinders and origin.

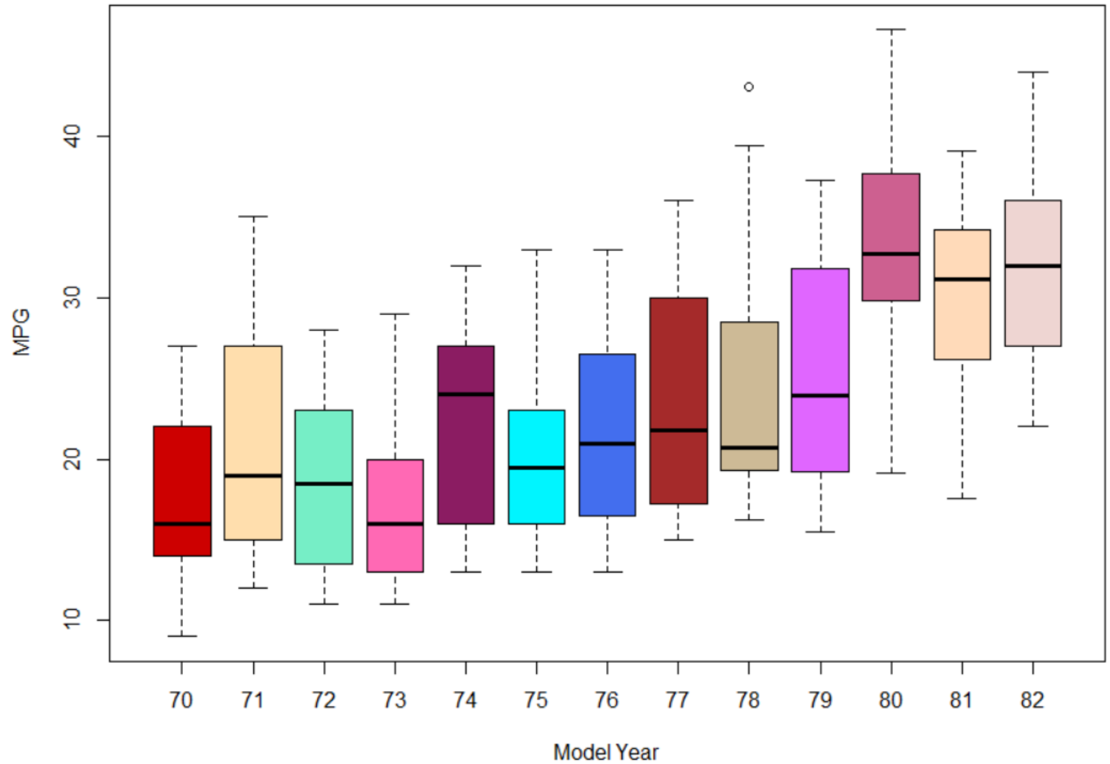
```
> group1 = mpg[modelyear=="70"]
> group2 = mpg[modelyear=="71"]
> group3 = mpg[modelyear=="72"]
> group4 = mpg[modelyear=="73"]
> group5 = mpg[modelyear=="74"]
> group6 = mpg[modelyear=="75"]
>
> group7 = mpg[modelyear=="76"]
> group8 = mpg[modelyear=="77"]
> group9 = mpg[modelyear=="78"]
> group10 = mpg[modelyear=="79"]
> group11 = mpg[modelyear=="80"]
> group12 = mpg[modelyear=="81"]
> group13 = mpg[modelyear=="82"]
>
> treatment=c(rep(70,length(group1)),rep(71,length(group2)),rep(72,length(group3)),
+
rep(73,length(group4)),rep(74,length(group5)),rep(75,length(group6)),rep(76,length(g
roup7)),rep(77,length(group8)),rep(78,length(group9)),
+
rep(79,length(group10)),rep(80,length(group11)),rep(81,length(group12)),rep(82,lengt
h(group13)) )
> treatmentfactor=factor(treatment)
> y=c(group1, group2, group3,group4, group5, group6,group7, group8,
group9,group10, group11, group12,group13)
> g=lm(y~treatmentfactor)
```

```
> summary(g)

Call:
lm(formula = y ~ treatmentfactor)

Residuals:
    Min      1Q  Median      3Q     Max
-14.704  -4.714  -1.000   4.268  19.039

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        17.6897     1.1095  15.944  < 2e-16 ***
treatmentfactor71   3.5603     1.5830   2.249 0.025073 *
treatmentfactor72   1.0246     1.5830   0.647 0.517837
treatmentfactor73  -0.5897     1.4572  -0.405 0.685954
treatmentfactor74   5.0140     1.5978   3.138 0.001833 **
treatmentfactor75   2.5770     1.5559   1.656 0.098485 .
treatmentfactor76   3.8839     1.5102   2.572 0.010498 *
treatmentfactor77   5.6853     1.5830   3.592 0.000372 ***
treatmentfactor78   6.3715     1.4908   4.274 2.43e-05 ***
treatmentfactor79   7.4034     1.5690   4.719 3.34e-06 ***
treatmentfactor80  16.1140     1.5978  10.085  < 2e-16 ***
treatmentfactor81  12.4961     1.5830   7.894 3.15e-14 ***
treatmentfactor82  14.3103     1.5559   9.198  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.975 on 381 degrees of freedom
Multiple R-squared:  0.4292,    Adjusted R-squared:  0.4112
F-statistic: 23.88 on 12 and 381 DF,  p-value: < 2.2e-16
```

From result in R, the mean of MPG of each group can be calculate:

$\mu_1 = \beta_0 = 17.69$
$\mu_2 = \beta_0 + \beta_1 = 21.25$
$\mu_3 = \beta_0 + \beta_2 = 18.71$
$\mu_4 = \beta_0 + \beta_3 = 17.10$
$\mu_5 = \beta_0 + \beta_4 = 22.70$
$\mu_6 = \beta_0 + \beta_5 = 20.27$
$\mu_7 = \beta_0 + \beta_6 = 21.57$
$\mu_8 = \beta_0 + \beta_7 = 23.37$

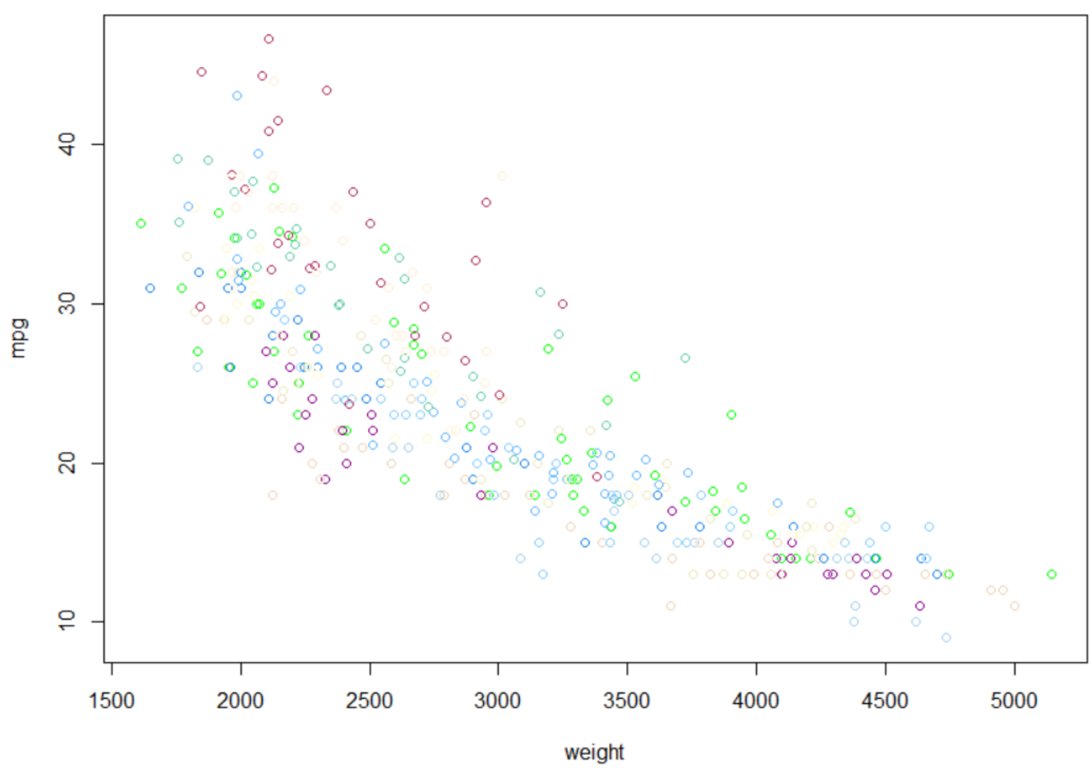
$\mu_9 = \beta_0 + \beta_8 = 24.06$
$\mu_{10} = \beta_0 + \beta_9 = 25.09$
$\mu_{11} = \beta_0 + \beta_{10} = 33.80$
$\mu_{12} = \beta_0 + \beta_{11} = 30.18$
$\mu_{13} = \beta_0 + \beta_{12} = 32.00$



From experiment between different variables (explained in next part), mpg and weight have most significant linear relationship:

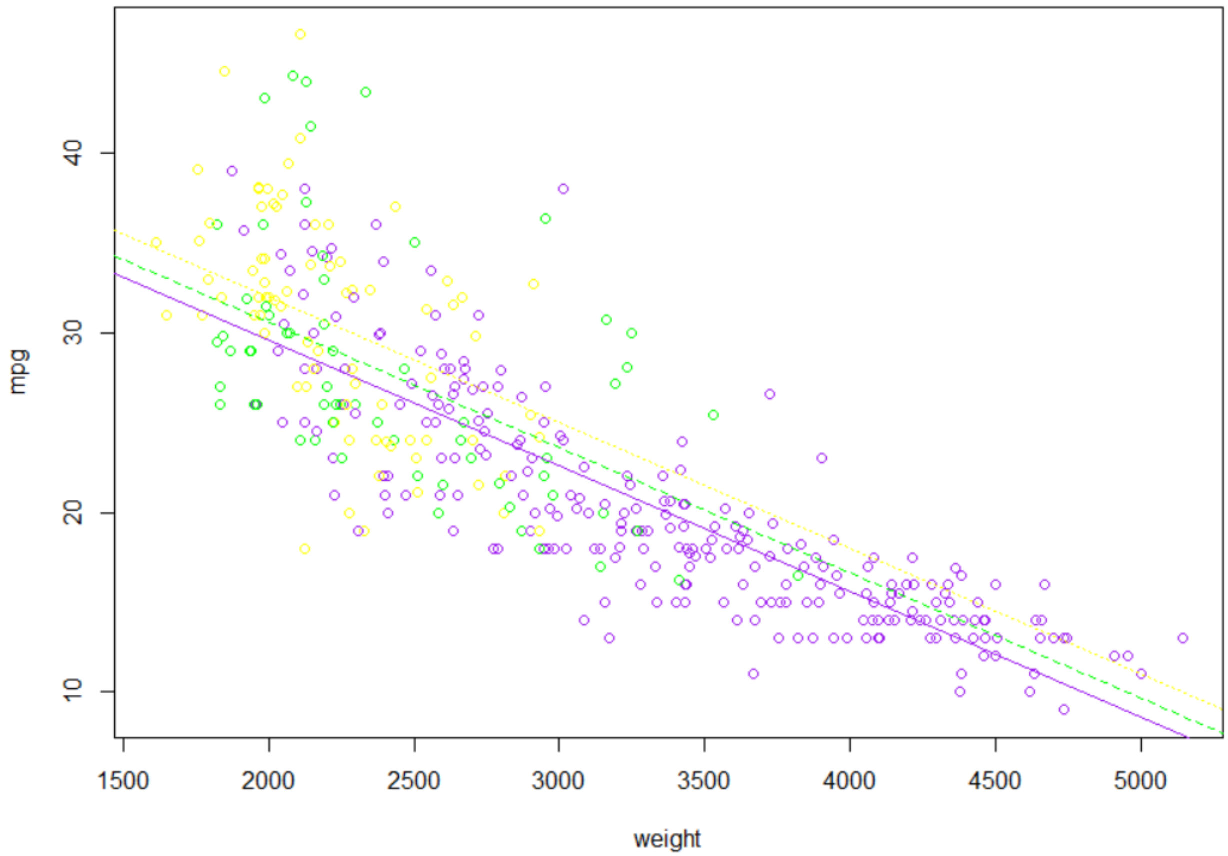
```
> m_fac = factor(modelyear)
> g = lm(mpg~weight+m_fac)
> summary(g)

Call:
lm(formula = mpg ~ weight + m_fac)

Residuals:
    Min      1Q  Median      3Q     Max
-10.2407  -2.0525  -0.0081   1.9892  13.4745

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.0940138  0.8971067  43.578  < 2e-16 ***
```

```
weight     -0.0063462  0.0002017 -31.471  < 2e-16 ***
m_fac71     1.1655209  0.8381202   1.391   0.1651
m_fac72     0.1673958  0.8351028   0.200   0.8412
m_fac73    -0.2962591  0.7683815  -0.386   0.7000
m_fac74     1.8735314  0.8483752   2.208   0.0278 *
m_fac75     1.3332036  0.8213328   1.623   0.1054
m_fac76     2.0177299  0.7985130   2.527   0.0119 *
m_fac77     3.3027599  0.8380850   3.941 9.66e-05 ***
m_fac78     3.1286360  0.7927852   3.946 9.45e-05 ***
m_fac79     5.3888638  0.8297774   6.494 2.61e-10 ***
m_fac80    10.2044807  0.8631589  11.822  < 2e-16 ***
m_fac81     7.1486740  0.8517783   8.393 9.49e-16 ***
m_fac82     8.3536507  0.8419331   9.922  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.15 on 380 degrees of freedom
Multiple R-squared:  0.8417,    Adjusted R-squared:  0.8363
F-statistic: 155.5 on 13 and 380 DF,  p-value: < 2.2e-16

> plot(mpg~weight,col=sample(color,13)[m_fac])
```



Because of too many groups for model years, it is too hard to draw the regression line. On the other hand, from plot above, mpg has little relation with weight depends on model year.

```
> o_fac = factor(origin)
> g2 = lm(mpg~weight+o_fac)
> summary(g2)

Call:
lm(formula = mpg ~ weight + o_fac)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.1301  -2.7638  -0.3161   2.4092  15.4958 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 43.574202   1.104181  39.463  < 2e-16 ***
weight      -0.006988   0.000318 -21.975  < 2e-16 ***
o_fac2       1.034813   0.655948   1.578 0.115472    
o_fac3       2.399267   0.661382   3.628 0.000324 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.273 on 390 degrees of freedom
Multiple R-squared:  0.7011,	Adjusted R-squared:  0.6988 
F-statistic: 304.9 on 3 and 390 DF,  p-value: < 2.2e-16

> plot(mpg~weight,col=c("purple","green","yellow")[o_fac])
> abline(43.574202 , -0.006988, lty=1, col="purple")
> abline(43.574202+1.034813, -0.006988, lty=2, col="green")
> abline(43.574202+2.399267, -0.006988, lty=3, col="yellow")
```



Because there are only three origins, it is easier to analysis origin group than model year group. From plot above, origin has very small impact on relation between mpg and weight.

Best Model to Predict MPG

- Model 1: full model

Due to number of levels in cylinders and origin is too small, they are not considered as variable for linear regression model.

```
> x1=displacement
> x2=as.numeric(horsepower)
> x3=weight
> x4=acceleration
> x5=modelyear
> y=mpg
> model1=lm(y~x1+x2+x3+x4+x5)
> summary(model1)

Call:
```

```
lm(formula = y ~ x1 + x2 + x3 + x4 + x5)

Residuals:
   Min     1Q Median     3Q    Max 
-8.728 -2.353 -0.055  1.929 14.362 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.676e+01  4.135e+00  -4.053 6.11e-05 ***
x1           3.593e-03  5.300e-03   0.678   0.4982    
x2           1.230e-02  6.749e-03   1.822   0.0692 .  
x3          -6.720e-03  5.962e-04 -11.272  < 2e-16 ***
x4           7.733e-02  7.824e-02   0.988   0.3236    
x5           7.590e-01  5.072e-02  14.966  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.413 on 388 degrees of freedom
Multiple R-squared:  0.8103,    Adjusted R-squared:  0.8078 
F-statistic: 331.4 on 5 and 388 DF,  p-value: < 2.2e-16
```

- Model 2: backward with AIC

```
> step(model1,direction = "backward")
Start:  AIC=973.39
y ~ x1 + x2 + x3 + x4 + x5

       Df Sum of Sq    RSS     AIC
- x1    1      5.36 4526.1  971.86
- x4    1     11.38 4532.1  972.38
<none>              4520.7  973.39
- x2    1     38.68 4559.4  974.75
- x3    1   1480.35 6001.1 1083.00
- x5    1   2609.84 7130.6 1150.94

Step:  AIC=971.86
y ~ x2 + x3 + x4 + x5

       Df Sum of Sq     RSS     AIC
- x4    1       6.9  4533.0  970.45
<none>               4526.1  971.86
```

```
- x2    1      36.5  4562.6  973.02
- x5    1    2653.1  7179.1 1151.62
- x3    1    7285.2 11811.3 1347.78

Step:  AIC=970.45
y ~ x2 + x3 + x5

       Df Sum of Sq     RSS     AIC
<none>              4533.0  970.45
- x2    1      39.3  4572.3  971.86
- x5    1    2795.2  7328.1 1157.71
- x3    1    8102.7 12635.7 1372.36

Call:
lm(formula = y ~ x2 + x3 + x5)

Coefficients:
(Intercept)          x2          x3          x5
 -15.840677    0.012310   -0.006411    0.759825
> model2=lm(y~x2+x3+x5)
> summary(model2)

Call:
lm(formula = y ~ x2 + x3 + x5)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0684 -2.3287 -0.0853  1.8987 14.4186

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.584e+01  4.030e+00  -3.931   0.0001 ***
x2           1.231e-02  6.692e-03   1.840   0.0666 .
x3          -6.411e-03  2.428e-04 -26.403   <2e-16 ***
x5           7.598e-01  4.900e-02  15.508   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.409 on 390 degrees of freedom
Multiple R-squared:  0.8098,    Adjusted R-squared:  0.8083
F-statistic: 553.3 on 3 and 390 DF,  p-value: < 2.2e-16
```

- Model 3: selected with Cp

```
> dt=data.frame(x1,x2,x3,x4,x5,y)
> sub = regsubsets(y~.,dt)
> rs = summary(sub)
> rs
Subset selection object
Call: regsubsets.formula(y ~ ., dt)
5 Variables  (and intercept)
   Forced in Forced out
x1     FALSE      FALSE
x2     FALSE      FALSE
x3     FALSE      FALSE
x4     FALSE      FALSE
x5     FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
         x1  x2  x3  x4  x5
1  ( 1 ) " " " " "*" " " " "
2  ( 1 ) " " " " "*" " " "*"
3  ( 1 ) " " "*" "*" " " "*"
4  ( 1 ) " " "*" "*" "*" "*"
5  ( 1 ) "*" "*" "*" "*" "*"
> rs$cp
[1] 241.889173   4.424437   3.048303   4.459632   6.000000
> rs$which[which.min(rs$cp),]
(Intercept)          x1          x2          x3          x4          x5
       TRUE       FALSE        TRUE        TRUE       FALSE        TRUE
> model3=lm(y~x3+x5)
> summary(models)
Error in summary(models) : object 'models' not found
> summary(model3)

Call:
lm(formula = y ~ x3 + x5)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8340 -2.2752 -0.1465  2.0275 14.3595

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.451e+01  3.976e+00  -3.649 0.000299 ***
x3          -6.626e-03  2.134e-04 -31.055  < 2e-16 ***
x5           7.591e-01  4.914e-02  15.447  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.42 on 391 degrees of freedom
Multiple R-squared:  0.8081,    Adjusted R-squared:  0.8071
F-statistic: 823.3 on 2 and 391 DF,  p-value: < 2.2e-16
```

Comparing models:

```
> anova(model2,model1)
Analysis of Variance Table

Model 1: y ~ x2 + x3 + x5
Model 2: y ~ x1 + x2 + x3 + x4 + x5
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    390 4533.0
2    388 4520.7  2    12.214 0.5242 0.5925
> anova(model3,model1)
Analysis of Variance Table

Model 1: y ~ x3 + x5
Model 2: y ~ x1 + x2 + x3 + x4 + x5
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    391 4572.3
2    388 4520.7  3    51.551 1.4748 0.2209
> anova(model3,model2)
Analysis of Variance Table

Model 1: y ~ x3 + x5
Model 2: y ~ x2 + x3 + x5
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    391 4572.3
2    390 4533.0  1    39.337 3.3844 0.06658 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	MODEL1	MODEL2	MODEL3
Variables	X1,X2,X3,X4,X5	X2,X3,X5	X3, X5
R2	0.8103	0.8098	0.8081
Adjusted R2	0.8078	0.8083	0.8071

Table 2

Conclusion:

From analysis of variance (ANOVA), all p value is greater than 0.05, there is no difference between three models. And R^2 values are also close. So model 3 is chosen as the best model.

$$y = -6.626e-03*x3 + 7.591e-01*x5$$

where y represents mpg, x3 represents weight, x5 represents model year.

Reference:

https://archive.ics.uci.edu/ml/datasets/auto+mpg

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition.