

Övningsuppgift A - analys av bränsledata för bilar

Per Idenfeldt, Oliver Grahn Thuna, Daniel Berg, Gabriel Junhager

9/30/2019

Contents

1	Introduktion	1
2	Variabelselektion	1
2.1	Variabelselektion - forward och backward	3
3	Konstruktion av modell	8
3.1	MSEP	8
4	Jämförelse av amerikanska - och icke-amerikanska bilar	9
5	Referens	9

1 Introduktion

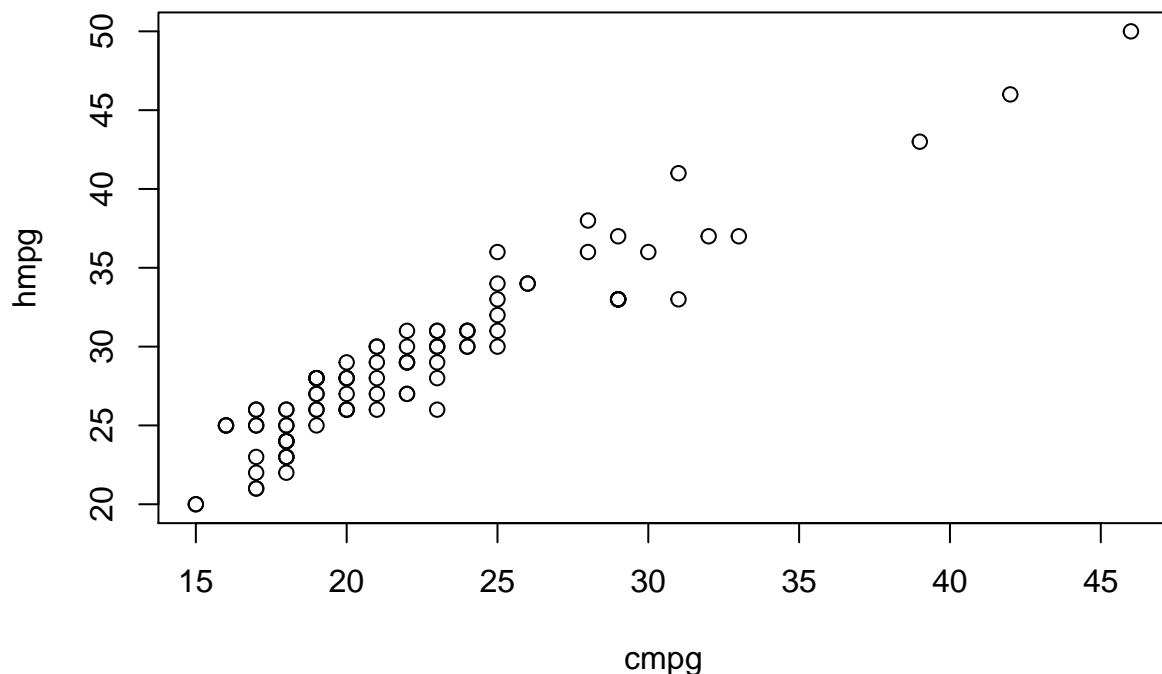
Denna rapport syftar till att utifrån de två datasetten “*Consumer Reports: The 1993 Cars - Annual Auto Issue*”, *Yonkers, NY: Consumers Union*” samt “*PACE New Car & Truck 1993 Buying Guide*” (1993), *Milwaukee, WI: Pace Publ. Inc*” bygga en modell om bilar bensinförbrukning. Detta görs genom att välja lämpliga variabler med hjälp av statistiska metoder som stegvis variabelselektion. Efter detta byggs en modell med hjälp av multipel linjär regression och dess prediktiva förmåga analyseras tillsammans med frågan om amerikanska bilar och icke-amerikanska bilar bränsleförbrukning skiljer sig på ett signifikant sätt.

2 Variabelselektion

Vi börjar med att undersöka data som är icke-kategorisk, annat data undersöks senare.

Variabler som helt klart är irrelevanta till bränsleförbrukning utesluts också automatiskt, till och exempel standard på krockkudde.

Vektorerna V7 och V8 står för hur många miles man kommer per gallon i stad respektive motorväg. Vi misstänker att vi kommer kunna kombinera dem i en variabel, hur ser de ut om vi plottar dem mot varandra?



Figur 1: Plot mellan city miles per gallon och highway miles per gallon

Vi ser en klar linjär trend. Korrelationen som visas nedan verkar också relativt hög.

```
## [1] 0.9439358
```

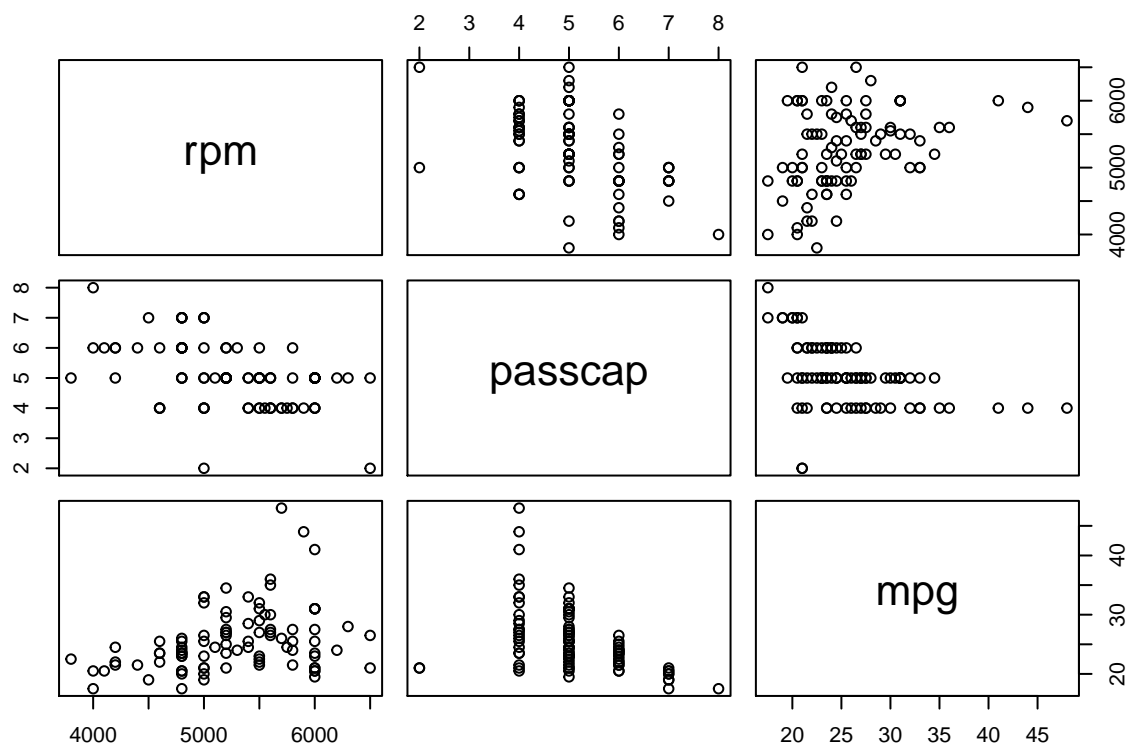
Vi kombinerar helt enkelt dessa variabler istället för att göra en modell åt varje, även fast de kan ha mindre skillnader.

Nu gör vi en korrelationsmatris utav dessa numeriska variabler. % latex table generated in R 3.6.1 by xtable 1.8-4 package % Mon Sep 30 14:02:39 2019

	minprice	midprice	maxprice	cylinders	engineize	horsepower	rpm	engineerev	fueltankcap	passcap	lencar	wheelbase	width	weight	mpg
minprice	1.00	0.97	0.91	0.62	0.65	0.80	-0.04	-0.47	0.64	0.06	0.55	0.52	0.49	0.67	-0.61
midprice	0.97	1.00	0.98	0.59	0.60	0.79	-0.00	-0.43	0.62	0.06	0.50	0.50	0.46	0.65	-0.59
maxprice	0.91	0.98	1.00	0.54	0.54	0.74	0.03	-0.37	0.58	0.05	0.44	0.47	0.41	0.61	-0.54
cylinders	0.62	0.59	0.54	1.00	0.87	0.68	-0.44	-0.74	0.67	0.40	0.68	0.68	0.78	0.80	-0.66
engineize	0.65	0.60	0.54	0.87	1.00	0.73	-0.55	-0.82	0.76	0.37	0.78	0.73	0.87	0.85	-0.68
horsepower	0.80	0.79	0.74	0.68	0.73	1.00	0.04	-0.60	0.71	0.01	0.55	0.49	0.64	0.74	-0.66
rpm	-0.04	-0.00	0.03	-0.44	-0.55	0.04	1.00	0.49	-0.33	-0.47	-0.44	-0.47	-0.54	0.34	0.34
engineerev	-0.47	-0.43	-0.37	-0.74	-0.82	-0.60	0.49	1.00	-0.61	-0.33	-0.69	-0.64	-0.78	-0.74	0.65
fueltankcap	0.64	0.62	0.58	0.67	0.76	0.71	-0.33	-0.61	1.00	0.47	0.69	0.76	0.80	0.89	-0.81
passcap	0.06	0.06	0.05	0.40	0.37	0.01	-0.47	-0.33	0.47	1.00	0.49	0.69	0.49	0.55	-0.45
lencar	0.55	0.50	0.44	0.68	0.78	0.55	-0.44	-0.69	0.69	0.49	1.00	0.82	0.82	0.81	-0.61
wheelbase	0.52	0.50	0.47	0.68	0.73	0.49	-0.47	-0.64	0.76	0.69	0.82	1.00	0.81	0.87	-0.65
width	0.49	0.46	0.41	0.78	0.87	0.64	-0.54	-0.78	0.80	0.49	0.82	0.81	1.00	0.87	-0.69
weight	0.67	0.65	0.61	0.80	0.85	0.74	-0.43	-0.74	0.89	0.55	0.81	0.87	0.87	1.00	-0.84
mpg	-0.61	-0.59	-0.54	-0.66	-0.68	-0.66	0.34	0.65	-0.81	-0.45	-0.61	-0.65	-0.69	-0.84	1.00

Figur 2: Korrelationsmatris på data som endast är numerisk och relevant

Vi säger arbiträrt att vi vill testa alla variabler som fick $|r| < 0.5$, genom att plotta dem mot mpg.



Figur 3: Plotten av de variablerna som har dålig korrelation med *mpg*

Av denna figur kan vi inte riktigt avgöra om variablerna bör vara med i modellen eller ej, så vi har kvar dem och utför ytterligare tester.

2.1 Variabelselektion - forward och backward

Eftersom vi har ett stort antal variabler att välja bland för att bygga upp vår modell utför vi både *forward selection* och *backward elimination*. De variabler vi utför detta på omfattar all förutom de kategoriska och de som saknar data i vissa punkter. Vad vi får ut är en modell som har eliminerat ett stort antal av våra förklarande variabler.

```
##
## Call:
## lm(formula = mpg ~ weight + wheelbase + fueltankcap + domestic +
##     width + enginerev + enginesize + minprice + passcap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.5669 -1.4095  0.0993  1.3611  8.6676
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.661431  12.139692   0.219  0.82701
## weight      -0.008281   0.001675  -4.945 3.92e-06 ***
## wheelbase    0.318107   0.096147   3.309  0.00139 **
## fueltankcap -0.641508   0.188855  -3.397  0.00105 **
```

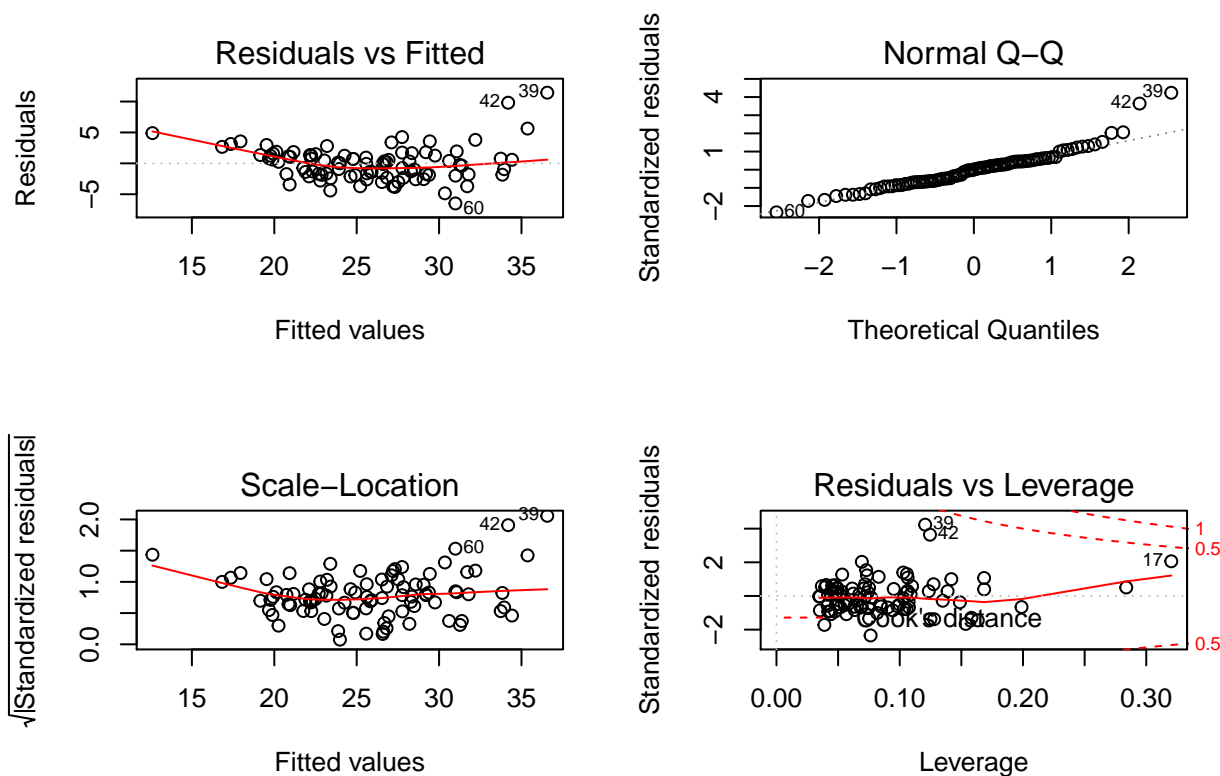
```
## domestic      -2.164667    0.700386   -3.091    0.00272 **
## width         0.319103    0.198430    1.608    0.11160
## enginerev     0.002587    0.001003    2.579    0.01169 *
## enginesize    1.814765    0.679593    2.670    0.00911 **
## minprice     -0.121207    0.053650   -2.259    0.02649 *
## passcap      -0.724709    0.419511   -1.728    0.08780 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.506 on 83 degrees of freedom
## Multiple R-squared:  0.8056, Adjusted R-squared:  0.7846
## F-statistic: 38.23 on 9 and 83 DF,  p-value: < 2.2e-16
```

VIF av den framtagna modellen:

```
##      weight  wheelbase fueltankcap  domestic      width  enginerev
##  14.300130   6.299636   5.620167   1.814558   8.239058   3.636914
##  enginesize  minprice    passcap
##    7.282365   3.226134   2.783643
```

Ett VIF-test visar oss att variabeln weight är mycket korrelerad med andra variabler i vår modell. Om man tänker rent praktiskt så är detta mycket logiskt eftersom att vikten av en bil till viss del avgörs av de variablerna som vi redan har i vår modell. Är det verkligen nödvändigt att ha med denna variabel? Vi tar bort den och betraktar hur modellen ser ut.

```
##
## Call:
## lm(formula = mpg ~ wheelbase + fueltankcap + passcap + enginerev +
##      minprice + domestic + width, data = cars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.4898 -1.8394  0.0158  1.2732 11.4361
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.850119   12.698952   1.248  0.21541
## wheelbase     0.153398    0.103860   1.477  0.14338
## fueltankcap  -1.053296    0.189060  -5.571 2.91e-07 ***
## passcap      -1.147245    0.471535  -2.433  0.01707 *
## enginerev     0.002793    0.001019   2.741  0.00746 **
## minprice     -0.152612    0.053255  -2.866  0.00524 **
## domestic     -1.175666    0.771363  -1.524  0.13119
## width         0.202137    0.201100   1.005  0.31767
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.879 on 85 degrees of freedom
## Multiple R-squared:  0.7372, Adjusted R-squared:  0.7155
## F-statistic: 34.06 on 7 and 85 DF,  p-value: < 2.2e-16
```



```
## wheelbase fueltankcap passcap enginerev minprice domestic
## 5.567336 4.265823 2.663558 2.841073 2.407469 1.666951
## width
## 6.409084
```

Vi märker att modellen får betydligt högre säkerhet i skattningarna på några av dess parametrar. Detta är typiskt för problem med multikolaritet. Det finns fortfarande en viss osäkerhet i vissa parametrar, kan detta lösas genom att även ta bort width variablen? Denna visar också relativt höga VIF-värden och har inte väldigt hög signifikansnivå.

```
##
## Call:
## lm(formula = mpg ~ wheelbase + fueltankcap + passcap + enginerev +
## minprice + domestic, data = cars)
##
## Residuals:
## Min 1Q Median 3Q Max
## -6.6268 -1.8208 -0.0243 1.4521 11.7952
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.2338745 8.6092177 2.931 0.00433 **
## wheelbase 0.1952111 0.0951716 2.051 0.04330 *
## fueltankcap -0.9542848 0.1613814 -5.913 6.61e-08 ***
## passcap -1.2217060 0.4657073 -2.623 0.01030 *
## enginerev 0.0023672 0.0009267 2.554 0.01240 *
## minprice -0.1597589 0.0527812 -3.027 0.00326 **
```

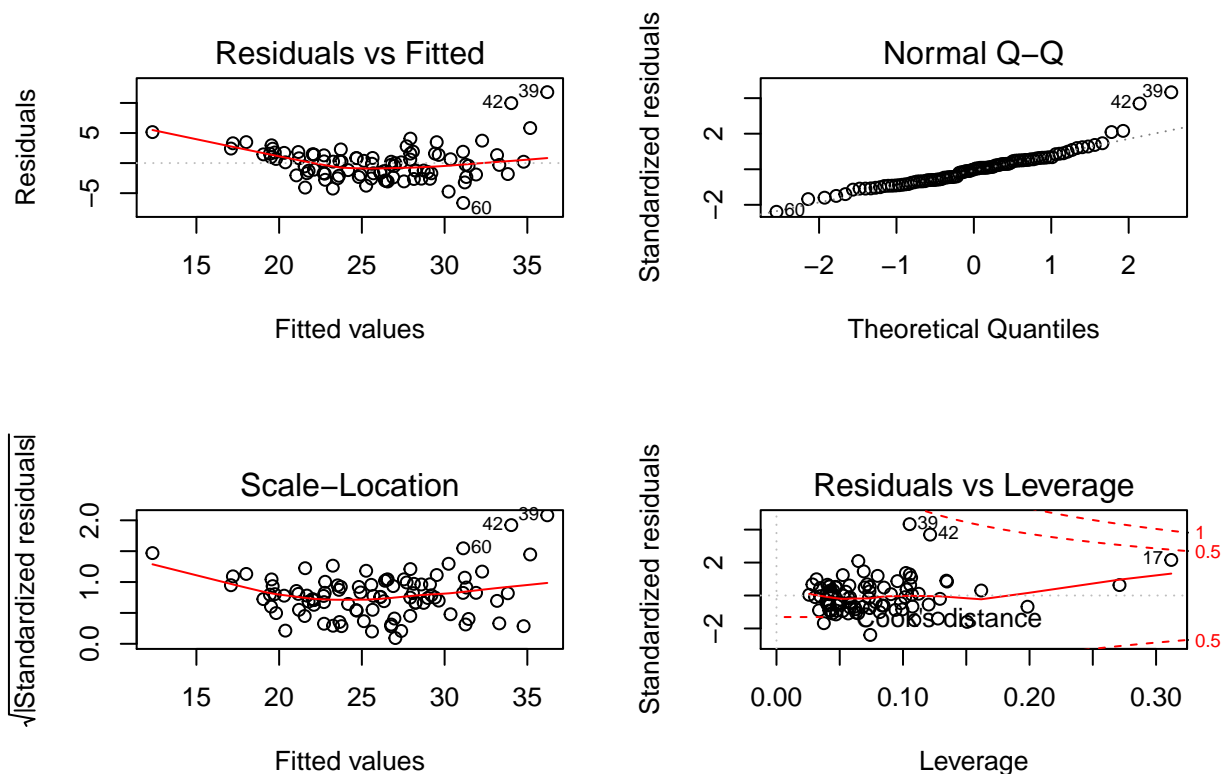
```
## domestic    -0.9068639  0.7235670  -1.253  0.21348
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.879 on 86 degrees of freedom
## Multiple R-squared:  0.7341, Adjusted R-squared:  0.7155
## F-statistic: 39.56 on 6 and 86 DF,  p-value: < 2.2e-16

##   wheelbase fueltankcap   passcap  enginerev   minprice   domestic
##   4.674244   3.107832    2.597820   2.349101   2.364555   1.466595
```

Enligt våra VIF-värden så har vi inte längre några problem med kolinearit. Modellen har relativt okej R^2 -värde, och alla lutningskoefficienter förutom den för domestic har goda t-värden. Detta tyder på att vi inte längre har lika starka multikollinearitet-problem som vi hade tidigare.

Värt att notera: Det är egentligen inte viktigt att intercept har hög säkerhet för vår modell. Detta eftersom att det är inte meningsfullt att tänka sig vad en bil med 0 i alla värden har för bränsleförbrukning. I vår modell har denna hypotetiska bil en bränsleförbrukning på 25.23, vilket är mer än vad vi förväntar oss av en bil utan säten eller bränsletank och med 0 rpm.

Vi undersöker residualer och möjliga outliers med nedanstående plottar.



```
##
## Call:
## lm(formula = mpg ~ wheelbase + fueltankcap + passcap + enginerev +
##     minprice + domestic, data = cars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -6.6268 -1.8208 -0.0243 1.4521 11.7952
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.2338745  8.6092177   2.931  0.00433 **
## wheelbase   0.1952111  0.0951716   2.051  0.04330 *
## fueltankcap -0.9542848  0.1613814  -5.913 6.61e-08 ***
## passcap     -1.2217060  0.4657073  -2.623  0.01030 *
## enginerev    0.0023672  0.0009267   2.554  0.01240 *
## minprice    -0.1597589  0.0527812  -3.027  0.00326 **
## domestic    -0.9068639  0.7235670  -1.253  0.21348
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.879 on 86 degrees of freedom
## Multiple R-squared:  0.7341, Adjusted R-squared:  0.7155
## F-statistic: 39.56 on 6 and 86 DF,  p-value: < 2.2e-16

##   wheelbase fueltankcap   passcap  enginerev   minprice   domestic
##   4.674244   3.107832    2.597820   2.349101   2.364555   1.466595
```

Observationerna 39 och 42 ligger precis innanför Cook's distance. När vi tittar på vår QQ-plot så ser vi att även här så orsakar 39 och 42 trubbel, och gör även att variansen för residylerna inte blir lika normalfördelat som det annars skulle vara.

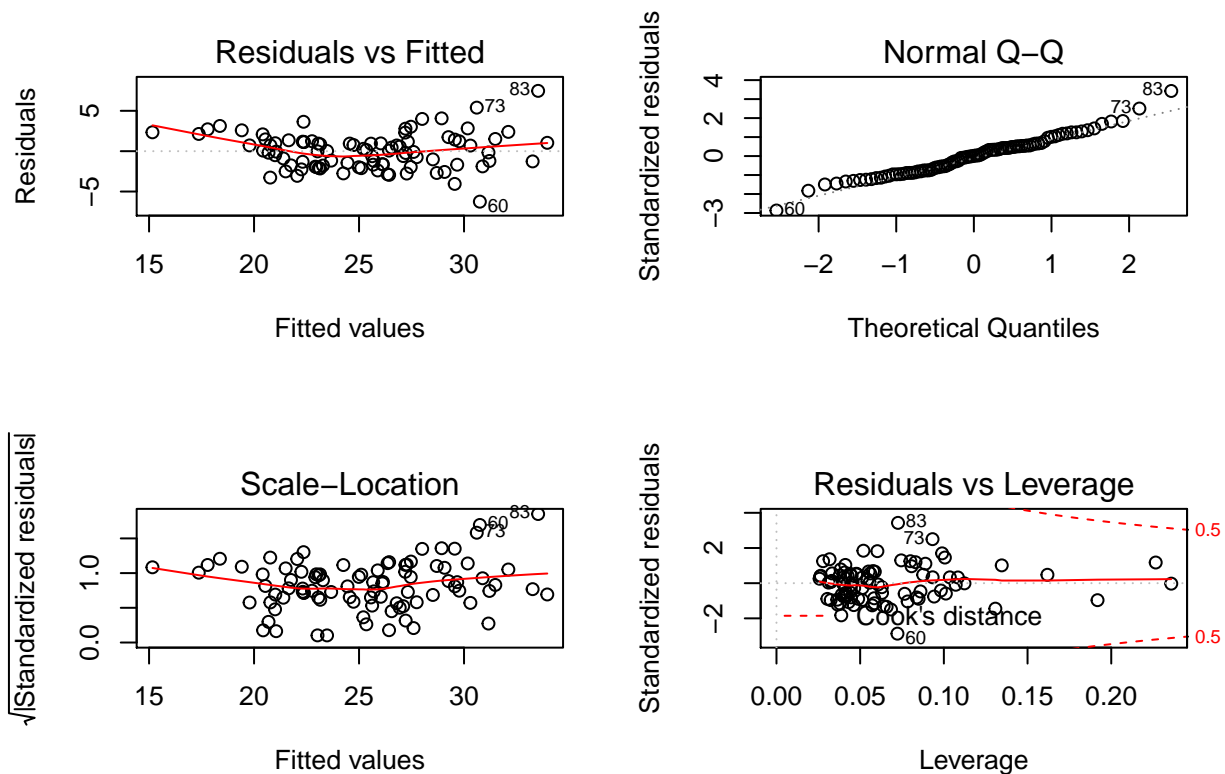
```
##   manufacturer model  type minprice midprice maxprice cmpg hmpg airbags
## 39           Geo Metro Small    6.7      8.4      10.0   46   50      0
## 42           Honda Civic Small    8.4     12.1     15.8   42   46      1
##   drivetrain cylinders enginesize horsepower  rpm  enginerev
## 39           1         3         1.0        55 5700      3755
## 42           1         4         1.5       102 5900      2650
##   manualtransmissions fueltankcap passcap  lencar wheelbase width  Uturn
## 39                   1         10.6      4     151      93    63    34
## 42                   1         11.9      4     173     103    67    36
##   rearseatroom luggagecap weight domestic
## 39           27.5         10   1695      0
## 42           28         12   2350      0
```

Ovanför ser vi att dessa observationer är båda små bilar med väldigt höga bränslekostnader, vilket kan ha att göra med dessa specifika modeller. Vi väljer att ta bort dessa outliers och ser om vår modell blir märkbart bättre.

```
##
## Call:
## lm(formula = mpg ~ fueltankcap + passcap + enginerev + minprice +
##     domestic, data = cars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.2536 -1.7167  0.0233  1.2022  7.4715
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.2930033  3.1103282  12.955 < 2e-16 ***
## fueltankcap -0.7751716  0.1206219  -6.426 7.24e-09 ***
## passcap     -0.5457420  0.2792933  -1.954  0.05399 .
```

```
## engine rev    0.0013391  0.0007269   1.842  0.06895 .
## minprice     -0.1183507  0.0390999  -3.027  0.00327 **
## domestic     -0.4921866  0.5703853  -0.863  0.39062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.263 on 85 degrees of freedom
## Multiple R-squared:  0.7625, Adjusted R-squared:  0.7485
## F-statistic: 54.58 on 5 and 85 DF,  p-value: < 2.2e-16

## fuel tank cap    pass cap    engine rev    min price    domestic
##      2.638116      1.475880      2.117328      2.044312      1.440809
```



Residylerna till vår nya modell ser genast mycket bättre ut, och vi får även aningen bättre R^2 -värde.

3 Konstruktion av modell

3.1 MSEP

```
## [1] "mpg~ width"
## [1] 9.953143

## [1] "mpg~ weight+wheelbase+fuel tank cap+pass cap+engine rev+min price+domestic"
## [1] 5.054161

## [1] "mpg~ wheelbase+fuel tank cap+pass cap+engine rev+min price+domestic"
## [1] 5.693809
```


4 Jämförelse av amerikanska - och icke-amerikanska bilar

I modell 4 kan vi notera att variabeln “domestic” har ett p -värde på 0.39062. Eftersom detta är baserat på en nollhypotes där koefficienten för domestic antas vara noll så ser vi att koefficientens värde, nämligen -0.538925 , inte alls är signifikant skilt från noll, förutsatt att alla antaganden för en multilinjär modell är uppfyllda förstås. I detta fall kan vi då av datat dra slutsatsen att bränsleförbrukningen inte påverkas märkbart av att bilen är amerikansk eller inte.

5 Referens

Formula with dynamic number of variables: <https://stackoverflow.com/questions/4951442/formula-with-dynamic-number-of-variables>