Probability and Statistics
Assignment 5

Hien Le - hien.le@student.auc.nl
Deniz Ovalioglu - deniz.ovalioglu@student.auc.nl

May 2018

a. The second line of the script transforms the table cardata in the first line into a 4-column table, with the columns being presented in the order of the indices in the vector i.e. MPG, VOL, HP, SP, WT. More concisely, it extracts 4 rows of the original table in the order 4 (MPG), 2 (VOL), 3 (HP), 5 (SP), 6 (WT).

b.

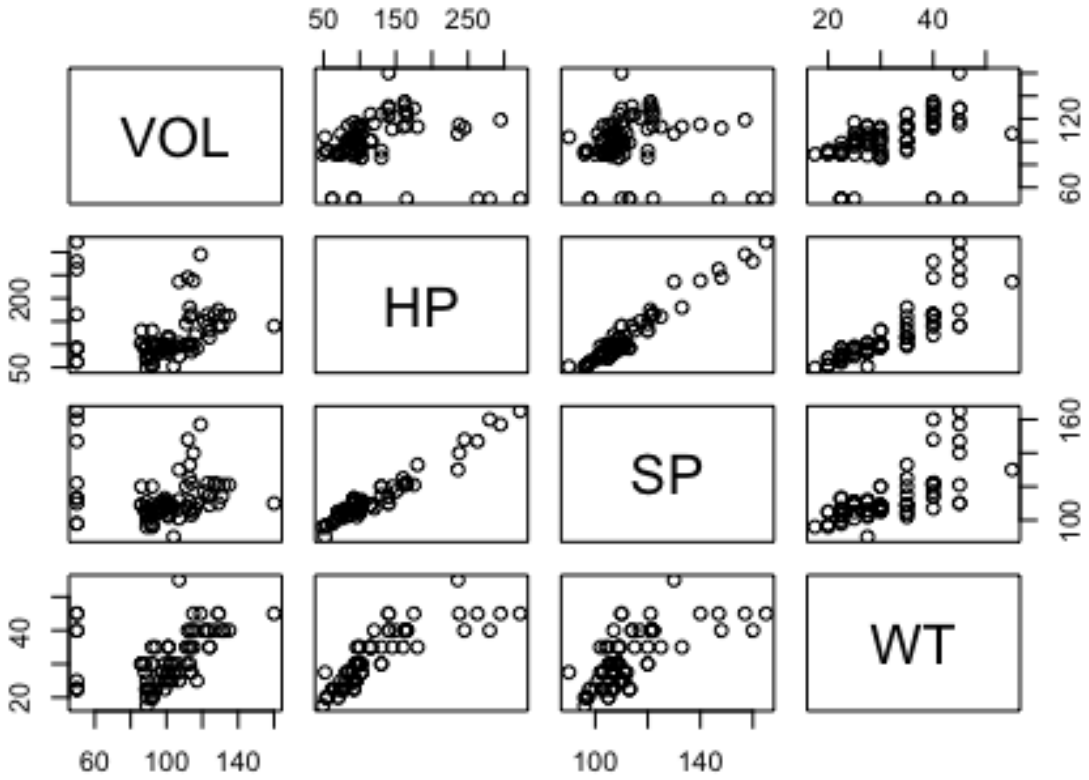

Figure 1: Scatter plot of candidate explanatory variables against each other

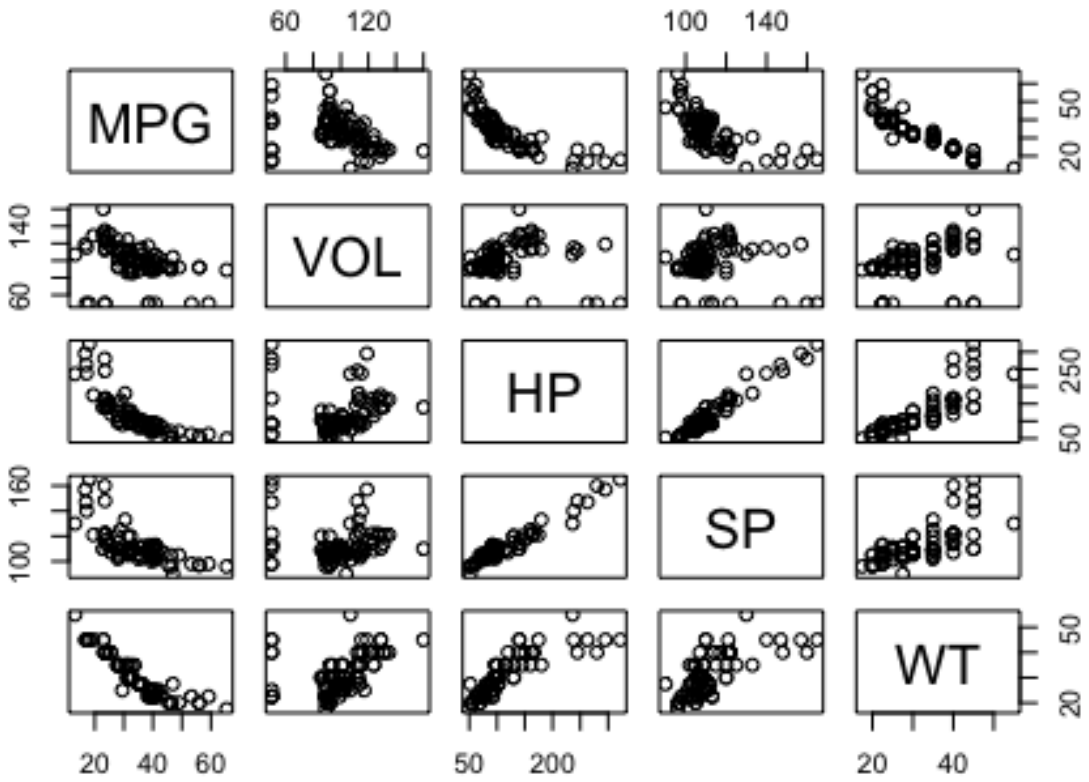

Figure 2: Scatter plot of candidate explanatory variables against MPG

WT seems to have a strong linear relationship with MPG, HP and SP seem to have a polynomial relationship with MPG.

c. (full model)

- The fitted regression equation: $MPG = -0.016VOL + 0.392HP - 1.295SP - 1.860WT + 192.438$.
- The estimated σ^2: $3.653^2 = 13.344$.
- The determination coefficient R^2:
+ Multiple R^2: 0.873
+ Adjusted R^2: 0.867
$\Rightarrow$ R^2 is close to 1 $\Rightarrow$ good fit.
- One explanatory variable is insignificant: VOL, with p-value of 0.495.

d. (step-up)

Results can be seen in Table 1:

Table 1: Step-up testing descriptive values

	fitted regression equation - $MPG =$	**estimated** σ^2	R^2 **(Adjusted)**
lm1A	$50.220 - 0.166VOL$	87.572	0.125
lm1B	$50.066 - 0.139HP$	38.118	0.619
lm1C	$88.938 - 0.491SP$	53.305	0.467
lm1D	$68.165 - 1.112WT$	18.327	0.817
lm2A	$66.855 - 0.990WT - 0.021HP$	18.105	0.819
lm2B	$75.649 - 0.997WT - 0.098SP$	17.506	0.825
lm2C	$68.876 - 1.101WT - 0.011VOL$	18.507	0.815
lm3A	$194.130 - 1.922WT - 1.320SP + 0.405HP$	13.250	0.8676

- Selected model: model 3A - MPG = $\beta_0 + \beta_1 HP + \beta_2 SP + \beta_3 WT$.
- Fitted equation: see Table 1.
- σ^2: 13.250 on 78 dfs.
- Multiple R^2: 0.873, adjusted R^2: 0.868.

See appendix for code. The step-up process could be stopped after adding HP with WT and SP (which produced a significant result in step 2B), as step 3B indicated an insignificant outcome.

e. (step-down)

- Selected model: MPG = $\beta_0 + \beta_1 HP + \beta_2 SP + \beta_3 WT$.
- Fitted regression equation: $MPG = 194.130 + 0.405HP - 1.320SP - 1.922WT$. - Looking at the results of function *step* (see Appendix), it can be concluded that the AIC score decreased significantly (by 1.5 points) after the removal of the VOL variable. Hence the model now only consists of only 3 variables, all of which are significant.
- σ^2: $3.64^2 = 13.250$ on 78 dfs.
- Multiple R^2: 0.873, adjusted R^2: 0.868 (good fit).

f.

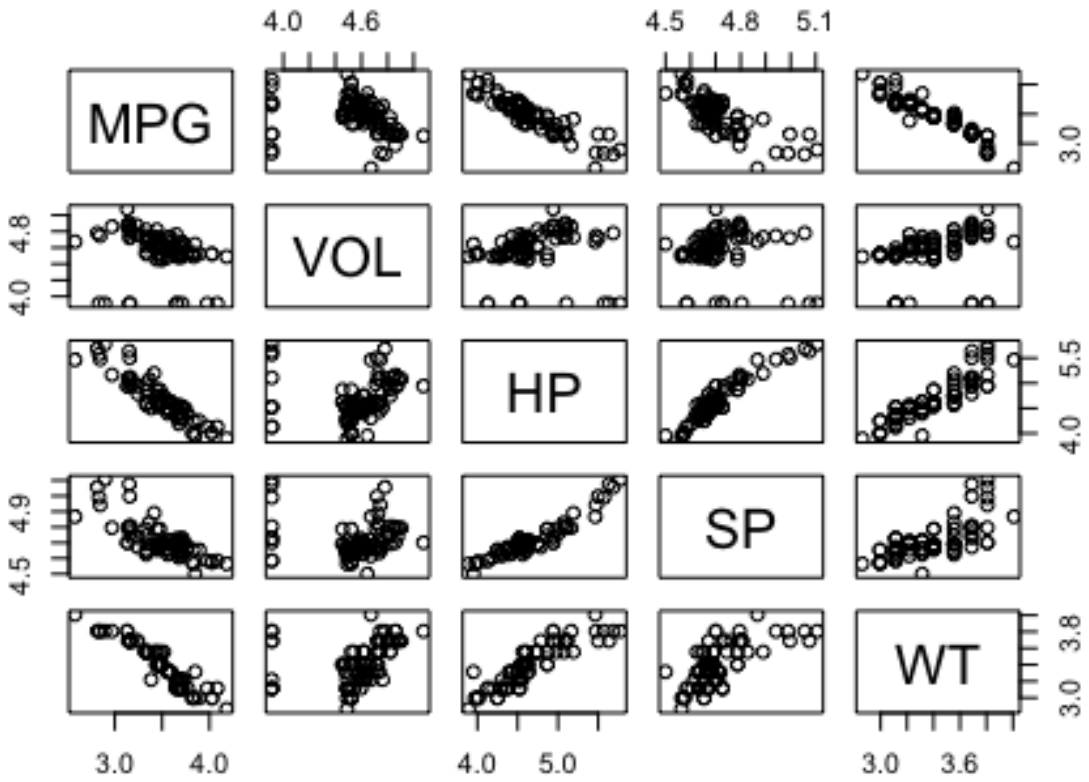

Figure 3: Scatter plot of candidate explanatory variables against MPG with a logarithmic transformation of all variables

log(HP) and log(WT) seem to have a linear relationship with log(MPG).

g.

- Selected model after step-down: $log(MPG) = \beta_0 + \beta_1 log(HP) + \beta_3 log(WT)$.
- Fitted regression equation: $log(MPG) = 7.190 - 0.268 log(HP) - 0.725 log(WT)$.
- σ^2: $0.088^2 = 0.008$ on 79 dfs.
- Multiple R^2: 0.920, adjusted R^2: 0.920 (good fit).

h. Comparison of all the models, but first need to check if they satisfy the iid-error assumption $\rightarrow$ need to make QQ plot and residuals-fitted-value plot:

The following are the plots (Q-Q plot + residual vs fitted-values plot) for each of the models in c, d, e and g:

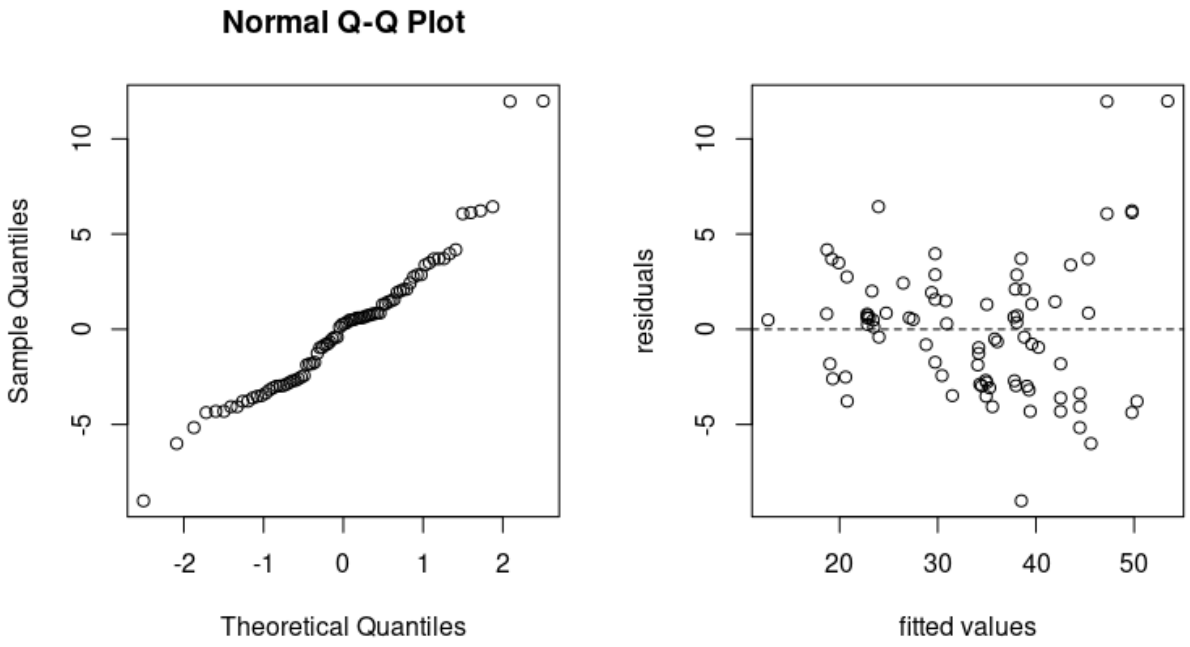

Figure 4: Plots for full model in section c

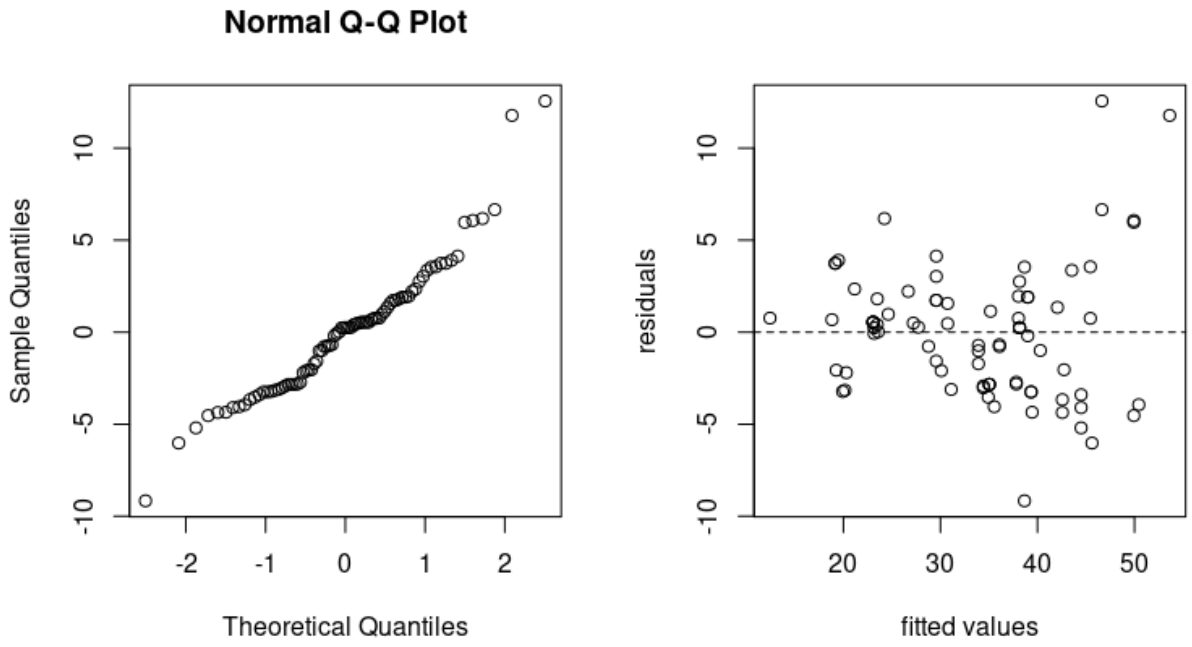

Figure 5: Plots for stepped-up model in section d

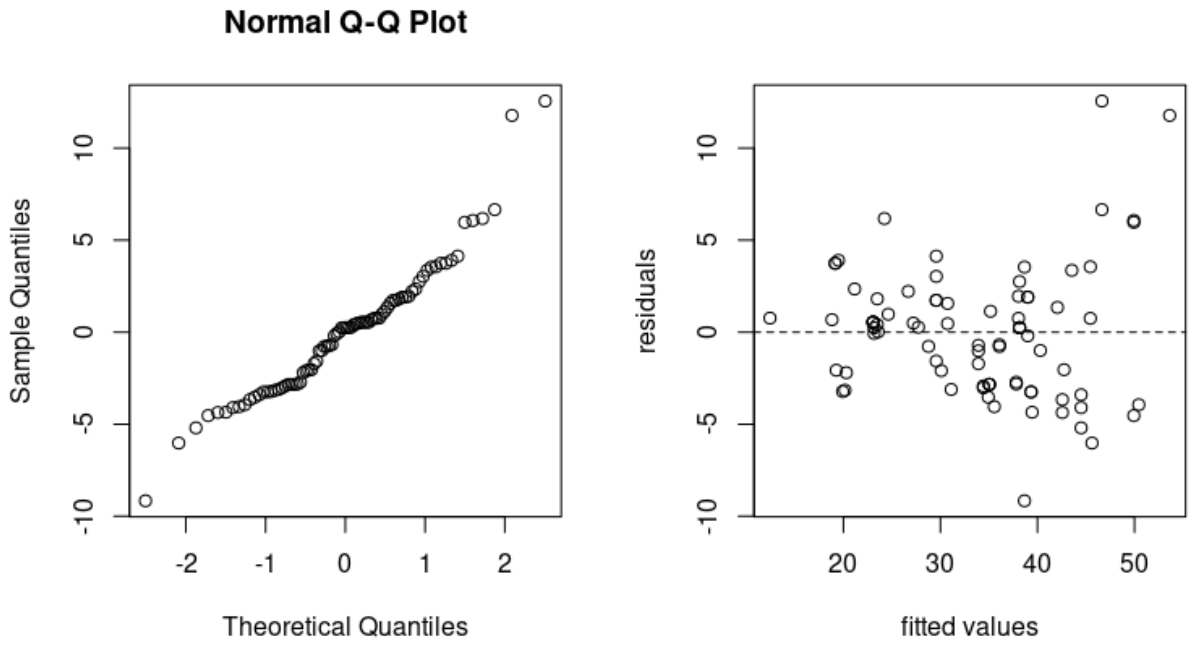

Figure 6: Plots for AIC stepped-down model in section e

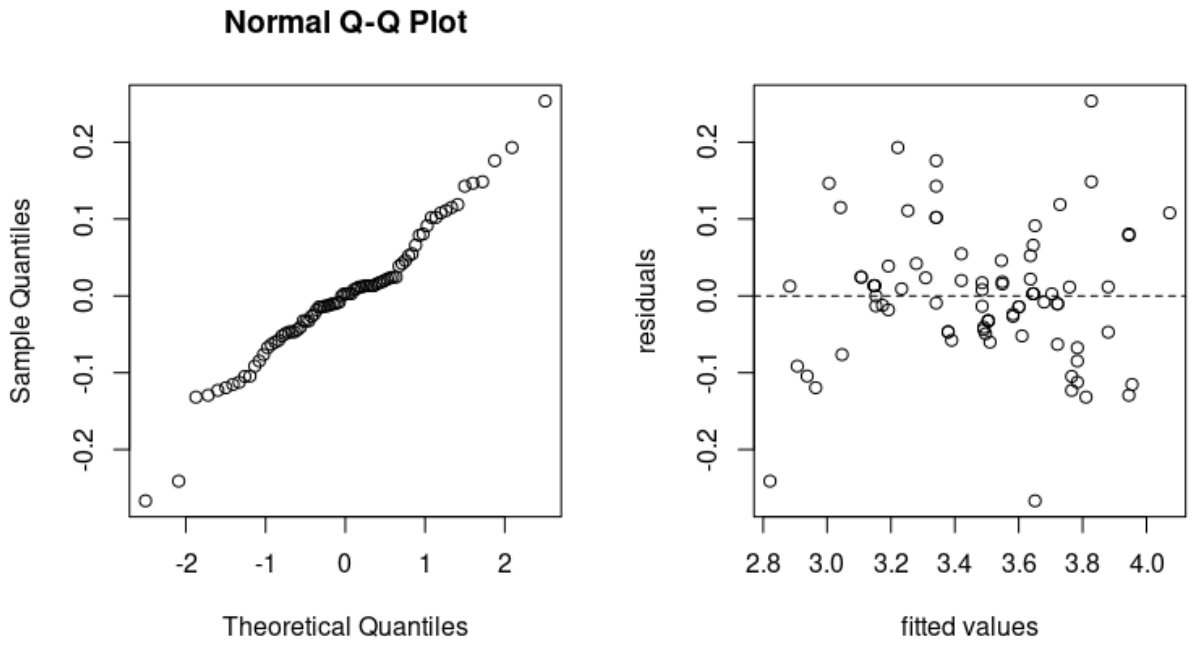

Figure 7: Plots for log model in section g

Comments:

- It's worth noting that the stepped-up model in d and the stepped-down model in e are the same. The Q-Q plot and fitted values vs. residuals plot of this model indicates that the assumption of i.i.d errors is satisfied. This model is significant with p-value $< 2.2e-16$, and all three variables are significant at 0.001 level, it also has a high R^2 value of 0.868 (good fit).

- Meanwhile, the full model in c, while satisfying the assumption of i.i.d errors and having a high R^2 value, has one insignificant variable VOL. This model also has a higher Residual standard error than the one in d and e, conveying the fact that it is not as good of a fit as the other one.

- The stepped-down log model in g produces the highest R^2 of 0.917, while being significant at 0.001 and satisfying the i.i.d errors assumption (strong linearity in Q-Q plot and no trend in the other plot). Since it has a different response from the other two models, we cannot use its σ^2 as a comparison metric.

- To conclude, we prefer the stepped-down log model in part g as its R^2 value is higher than the other models.

Appendix

```
cardata=read.table("carmpgdat_new.txt",header=TRUE)
cardata=cardata[,c(4,2,3,5,6)]

## section b:
cardata_without_MPG=cardata[,c(2,3,4,5)]
pairs(cardata)
pairs(cardata_without_MPG)
## end of section b

## section c:
carlm = lm(MPG~VOL+HP+SP+WT, data=cardata)
summary(carlm)

Call:
lm(formula = MPG ~ VOL + HP + SP + WT, data = cardata)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0108 -2.7731  0.2733  1.8362 11.9854

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.43775   23.53161   8.178 4.62e-12 ***
VOL          -0.01565    0.02283  -0.685    0.495
HP            0.39221    0.08141   4.818 7.13e-06 ***
SP           -1.29482    0.24477  -5.290 1.11e-06 ***
WT           -1.85980    0.21336  -8.717 4.22e-13 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.653 on 77 degrees of freedom
Multiple R-squared:  0.8733,Adjusted R-squared:  0.8667
F-statistic: 132.7 on 4 and 77 DF,  p-value: < 2.2e-16

## end of section c

## section d:
> lm1A=lm(MPG~VOL,data=cardata)
> summary(lm1A)

Call:
lm(formula = MPG ~ VOL, data = cardata)

Residuals:
    Min      1Q  Median      3Q     Max
-24.901  -4.624  -0.909   4.797  29.987

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.22001    4.74859  10.576  < 2e-16 ***
VOL         -0.16637    0.04691  -3.547 0.000656 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.358 on 80 degrees of freedom
Multiple R-squared:  0.1359,Adjusted R-squared:  0.1251
F-statistic: 12.58 on 1 and 80 DF,  p-value: 0.0006556

>
> lm1B=lm(MPG~HP,data=cardata)
> summary(lm1B)

Call:
lm(formula = MPG ~ HP, data = cardata)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7198 -4.1224 -0.9077  3.1009 22.1461

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.06608    1.56949   31.90   <2e-16 ***
```

```
HP          -0.13902    0.01207  -11.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.174 on 80 degrees of freedom
Multiple R-squared:  0.6239,Adjusted R-squared:  0.6192
F-statistic: 132.7 on 1 and 80 DF,  p-value: < 2.2e-16

>
> lm1C=lm(MPG~SP,data=cardata)
> summary(lm1C)

Call:
lm(formula = MPG ~ SP, data = cardata)

Residuals:
    Min      1Q  Median      3Q     Max
-12.066  -4.961  -1.015   4.257  23.564

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.93774    6.54647   13.59  < 2e-16 ***
SP          -0.49065    0.05779   -8.49 8.84e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.301 on 80 degrees of freedom
Multiple R-squared:  0.474,Adjusted R-squared:  0.4674
F-statistic: 72.08 on 1 and 80 DF,  p-value: 8.837e-13

>
> lm1D=lm(MPG~WT,data=cardata)
> summary(lm1D)

Call:
lm(formula = MPG ~ WT, data = cardata)

Residuals:
     Min       1Q   Median       3Q      Max
-10.8601  -2.2698  -1.1768   0.4899  16.6983

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 68.16545    1.86695   36.51   <2e-16 ***
WT          -1.11222    0.05842  -19.04   <2e-16 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.281 on 80 degrees of freedom
Multiple R-squared:  0.8192,Adjusted R-squared:  0.8169
F-statistic: 362.4 on 1 and 80 DF,  p-value: < 2.2e-16

>
> # we choose WT first because it had the highest R^2 value
> lm2A = lm(MPG~WT+HP, data=cardata)
> summary(lm2A) #insignificant

Call:
lm(formula = MPG ~ WT + HP, data = cardata)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7084  -2.1636  -0.9201   0.8802  16.9040

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 66.85500    2.07929  32.153  < 2e-16 ***
WT          -0.99037    0.10475  -9.455 1.26e-14 ***
HP          -0.02097    0.01500  -1.398    0.166
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.255 on 79 degrees of freedom
Multiple R-squared:  0.8235,Adjusted R-squared:  0.8191
F-statistic: 184.4 on 2 and 79 DF,  p-value: < 2.2e-16

>
> lm2B = lm(MPG~WT+SP,data=cardata)
> summary(lm2B) #significant at 0.05

Call:
lm(formula = MPG ~ WT + SP, data = cardata)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5160  -2.5085  -0.8544   0.9377  16.6276

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 75.64938    3.89181  19.438   <2e-16 ***
```

```
WT          -0.99738    0.07774 -12.830   <2e-16 ***
SP          -0.09816    0.04508  -2.177   0.0325 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.184 on 79 degrees of freedom
Multiple R-squared:  0.8294,Adjusted R-squared:  0.8251
F-statistic: 192.1 on 2 and 79 DF,  p-value: < 2.2e-16

>
> lm2C = lm(MPG~WT+VOL,data=cardata)
> summary(lm2C) # insignificant

Call:
lm(formula = MPG ~ WT + VOL, data = cardata)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7595  -2.4173  -1.0671   0.5798  16.7439

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 68.87609    2.43465  28.290   <2e-16 ***
WT          -1.10100    0.06362 -17.307   <2e-16 ***
VOL         -0.01070    0.02337  -0.458    0.648
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.302 on 79 degrees of freedom
Multiple R-squared:  0.8197,Adjusted R-squared:  0.8151
F-statistic: 179.5 on 2 and 79 DF,  p-value: < 2.2e-16

>
> # because combination of WT and SP was significant, we carry on with another step
> lm3A=lm(MPG~WT+SP+HP, data=cardata)
> summary(lm3A) #significant

Call:
lm(formula = MPG ~ WT + SP + HP, data = cardata)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1633 -2.8387  0.2464  1.7889 12.5566

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 194.12962   23.32213   8.324 2.22e-12 ***
WT           -1.92210    0.19238  -9.991 1.31e-15 ***
SP           -1.32000    0.24118  -5.473 5.19e-07 ***
HP            0.40518    0.07891   5.135 2.03e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.64 on 78 degrees of freedom
Multiple R-squared:  0.8725,Adjusted R-squared:  0.8676
F-statistic: 177.9 on 3 and 78 DF,  p-value: < 2.2e-16

>
> lm3B = lm(MPG~WT+SP+VOL,data=cardata)
> summary(lm3B) #insignificant => we stop here

Call:
lm(formula = MPG ~ WT + SP + VOL, data = cardata)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0003  -2.7013  -0.5674   1.2842  16.7766

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.18296    5.12341  15.845  < 2e-16 ***
WT          -0.91127    0.09318  -9.780 3.35e-15 ***
SP          -0.13484    0.04992  -2.701  0.00847 **
VOL         -0.04121    0.02516  -1.638  0.10554
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.14 on 78 degrees of freedom
Multiple R-squared:  0.8351,Adjusted R-squared:  0.8287
F-statistic: 131.7 on 3 and 78 DF,  p-value: < 2.2e-16

## end of section d

## section e

> car_AIC_step_down = step(carlm)
Start:  AIC=217.3
MPG ~ VOL + HP + SP + WT

       Df Sum of Sq    RSS    AIC
```

```
- VOL   1      6.27 1033.7 215.80
<none>              1027.4 217.30
- HP    1    309.67 1337.0 236.90
- SP    1    373.36 1400.7 240.72
- WT    1   1013.76 2041.2 271.59

Step:  AIC=215.8
MPG ~ HP + SP + WT

       Df Sum of Sq    RSS    AIC
<none>              1033.7 215.80
- HP    1    349.37 1383.0 237.68
- SP    1    396.97 1430.6 240.45
- WT    1   1322.87 2356.5 281.37


> summary(car_AIC_step_down)

Call:
lm(formula = MPG ~ HP + SP + WT, data = cardata)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1633 -2.8387  0.2464  1.7889 12.5566

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 194.12962   23.32213   8.324 2.22e-12 ***
HP            0.40518    0.07891   5.135 2.03e-06 ***
SP           -1.32000    0.24118  -5.473 5.19e-07 ***
WT           -1.92210    0.19238  -9.991 1.31e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.64 on 78 degrees of freedom
Multiple R-squared:  0.8725,Adjusted R-squared:  0.8676
F-statistic: 177.9 on 3 and 78 DF,  p-value: < 2.2e-16

## end of section e

## section f
pairs(log(cardata))
## end of section f

## section g
```

```
carlm_log = lm(log(MPG)~log(VOL)+log(HP)+log(SP)+log(WT),data=cardata)
carlm_log_SP = lm(log(MPG)~log(VOL)+log(HP)+log(WT),data=cardata)
carlm_log_VOL = lm(log(MPG)~log(HP)+log(WT),data=cardata)

# result of full model:
> summary(carlm_log)
Call:
lm(formula = log(MPG) ~ log(VOL) + log(HP) + log(SP) + log(WT),
    data = cardata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.258695 -0.044754  0.000879  0.043550  0.226129

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.36931    2.88108   1.864   0.0662 .
log(VOL)    -0.03163    0.04447  -0.711   0.4792
log(HP)     -0.50246    0.31861  -1.577   0.1189
log(SP)      0.51289    0.75163   0.682   0.4971
log(WT)     -0.53594    0.23468  -2.284   0.0251 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08811 on 77 degrees of freedom
Multiple R-squared:  0.9203,Adjusted R-squared:  0.9162
F-statistic: 222.4 on 4 and 77 DF,  p-value: < 2.2e-16

# result of model after removing log(SP):
> summary(carlm_log_SP)
Call:
lm(formula = log(MPG) ~ log(VOL) + log(HP) + log(WT), data = cardata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26023 -0.04594 -0.00199  0.04223  0.22767

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.33080    0.19348  37.889  < 2e-16 ***
log(VOL)    -0.04095    0.04218  -0.971    0.335
log(HP)     -0.28829    0.05461  -5.279 1.14e-06 ***
log(WT)     -0.68333    0.09144  -7.473 9.81e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.08781 on 78 degrees of freedom
Multiple R-squared:  0.9199,Adjusted R-squared:  0.9168
F-statistic: 298.4 on 3 and 78 DF,  p-value: < 2.2e-16

// result of model after removing log(VOL):
> summary(carlm_log_VOL)
Call:
lm(formula = log(MPG) ~ log(HP) + log(WT), data = cardata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26671 -0.04685  0.00274  0.03515  0.25354

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.19011    0.12817  56.099  < 2e-16 ***
log(HP)     -0.26818    0.05052  -5.309 9.86e-07 ***
log(WT)     -0.72456    0.08095  -8.950 1.22e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08778 on 79 degrees of freedom
Multiple R-squared:  0.9189,Adjusted R-squared:  0.9168
F-statistic: 447.5 on 2 and 79 DF,  p-value: < 2.2e-16

## end of section g

## section h: diagnostics

diagnose_by_plots = function(model) {
  par(mfrow=c(2,2))
  par(pty="s")
  qqnorm(residuals(model))
  plot(fitted(model),residuals(model),xlab="fitted values",ylab="residuals")
  abline(0,0,lty=2)
}

# model in c: full lm
diagnose_by_plots(carlm)

# model in d: lm2c
diagnose_by_plots(lm3A)

# model in e:
```

```
diagnose_by_plots(car_AIC_step_down)

# model in g:
diagnose_by_plots(carlm_log_VOL)

## end of section h
```