Statistical Learning (5454) - Assignment 1

Matthias Hochholzer, Lukas Pirnbacher, Anne Valder

Due: 2024-03-25

Exercise 1

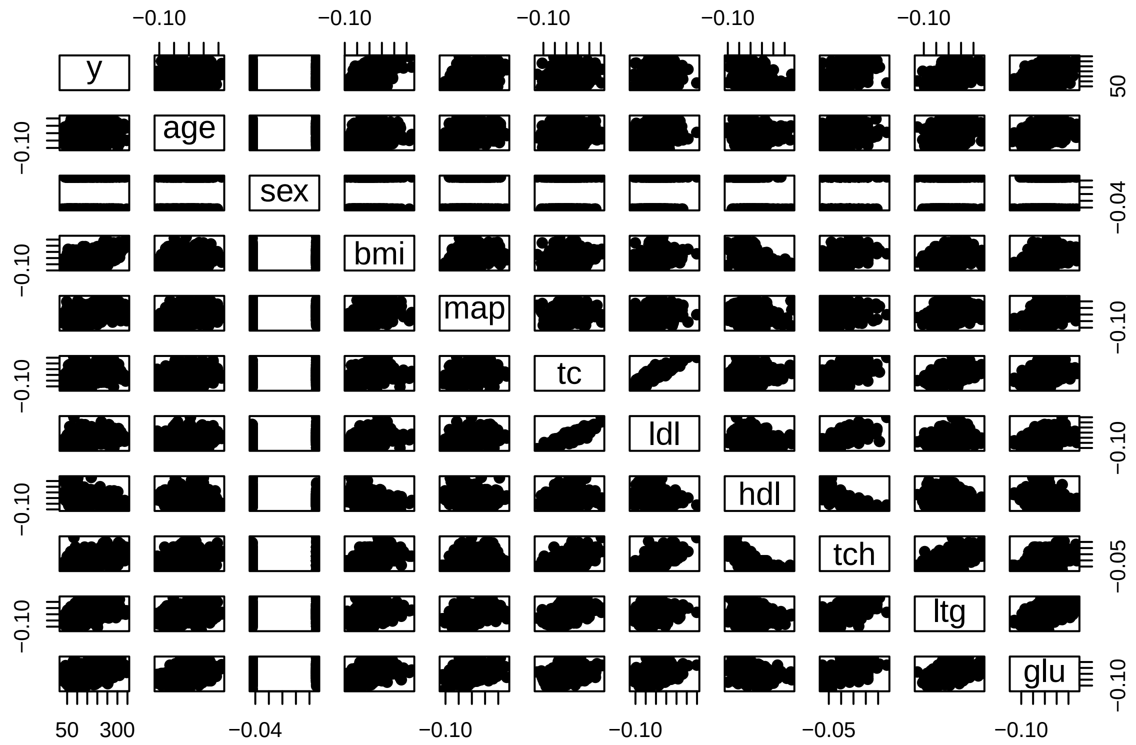
```
##        y              age                 sex                bmi
##  Min.   : 25.0   Min.   :-0.107226   Min.   :-0.04464   Min.   :-0.090275
##  1st Qu.: 87.0   1st Qu.:-0.037299   1st Qu.:-0.04464   1st Qu.:-0.034229
##  Median :140.5   Median : 0.005383   Median :-0.04464   Median :-0.007284
##  Mean   :152.1   Mean   : 0.000000   Mean   : 0.00000   Mean   : 0.000000
##  3rd Qu.:211.5   3rd Qu.: 0.038076   3rd Qu.: 0.05068   3rd Qu.: 0.031248
##  Max.   :346.0   Max.   : 0.110727   Max.   : 0.05068   Max.   : 0.170555
##       map                  tc                 ldl
##  Min.   :-0.112400   Min.   :-0.126781   Min.   :-0.115613
##  1st Qu.:-0.036656   1st Qu.:-0.034248   1st Qu.:-0.030358
##  Median :-0.005671   Median :-0.004321   Median :-0.003819
##  Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.000000
##  3rd Qu.: 0.035644   3rd Qu.: 0.028358   3rd Qu.: 0.029844
##  Max.   : 0.132044   Max.   : 0.153914   Max.   : 0.198788
##       hdl                 tch                 ltg
##  Min.   :-0.102307   Min.   :-0.076395   Min.   :-0.126097
##  1st Qu.:-0.035117   1st Qu.:-0.039493   1st Qu.:-0.033249
##  Median :-0.006584   Median :-0.002592   Median :-0.001948
##  Mean   : 0.000000   Mean   : 0.000000   Mean   : 0.000000
##  3rd Qu.: 0.029312   3rd Qu.: 0.034309   3rd Qu.: 0.032433
##  Max.   : 0.181179   Max.   : 0.185234   Max.   : 0.133599
##       glu
##  Min.   :-0.137767
##  1st Qu.:-0.033179
##  Median :-0.001078
##  Mean   : 0.000000
##  3rd Qu.: 0.027917
##  Max.   : 0.135612
```

- Explanation: Random selection

Table 1: Correlation Matrix

	y	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu
y	1.00	0.19	0.04	0.59	0.44	0.21	0.17	-0.39	0.43	0.57	0.38
age	0.19	1.00	0.17	0.19	0.34	0.26	0.22	-0.08	0.20	0.27	0.30
sex	0.04	0.17	1.00	0.09	0.24	0.04	0.14	-0.38	0.33	0.15	0.21
bmi	0.59	0.19	0.09	1.00	0.40	0.25	0.26	-0.37	0.41	0.45	0.39
map	0.44	0.34	0.24	0.40	1.00	0.24	0.19	-0.18	0.26	0.39	0.39
tc	0.21	0.26	0.04	0.25	0.24	1.00	0.90	0.05	0.54	0.52	0.33

	y	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu
ldl	0.17	0.22	0.14	0.26	0.19	0.90	1.00	-0.20	0.66	0.32	0.29
hdl	-0.39	-0.08	-0.38	-0.37	-0.18	0.05	-0.20	1.00	-0.74	-0.40	-0.27
tch	0.43	0.20	0.33	0.41	0.26	0.54	0.66	-0.74	1.00	0.62	0.42
ltg	0.57	0.27	0.15	0.45	0.39	0.52	0.32	-0.40	0.62	1.00	0.46
glu	0.38	0.30	0.21	0.39	0.39	0.33	0.29	-0.27	0.42	0.46	1.00



- Explanation standardized - Interpret correlation

```
##
## Call:
## lm(formula = y ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -154.436  -37.748   -1.375   37.421  153.466
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  152.706      2.711  56.319  < 2e-16 ***
## age            9.856     62.721   0.157 0.875213
## sex         -240.347     64.936  -3.701 0.000245 ***
## bmi          499.266     70.415   7.090 6.35e-12 ***
## map          354.976     70.187   5.058 6.55e-07 ***
## tc          -861.163    436.264  -1.974 0.049095 *
## ldl          541.190    354.923   1.525 0.128119
## hdl          116.045    221.425   0.524 0.600518
## tch          166.516    166.601   0.999 0.318178
## ltg          773.896    179.728   4.306 2.11e-05 ***
## glu           63.631     68.817   0.925 0.355729
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.18 on 389 degrees of freedom
## Multiple R-squared:  0.5258, Adjusted R-squared:  0.5136
## F-statistic: 43.13 on 10 and 389 DF,  p-value: < 2.2e-16
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   152.71       2.71   56.32   <2e-16 ***
## age             9.86      62.72    0.16     0.88
## sex          -240.35      64.94   -3.70   <2e-16 ***
## bmi           499.27      70.41    7.09   <2e-16 ***
## map           354.98      70.19    5.06   <2e-16 ***
## tc           -861.16     436.26   -1.97     0.05 *
## ldl           541.19     354.92    1.52     0.13
## hdl           116.05     221.42    0.52     0.60
## tch           166.52     166.60    1.00     0.32
## ltg           773.90     179.73    4.31   <2e-16 ***
## glu            63.63      68.82    0.92     0.36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 2854.869
```

```
## [1] 2945.384
```

```
##
## Call:
## lm(formula = y ~ sex + bmi + map + tc + ltg, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -154.487  -39.583   -2.167   36.677  143.460
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  152.676      2.744  55.634  < 2e-16 ***
## sex         -143.624     60.008  -2.393  0.01716 *
## bmi          580.467     67.332   8.621  < 2e-16 ***
## map          344.751     68.041   5.067 6.23e-07 ***
## tc          -218.311     67.313  -3.243  0.00128 **
## ltg          657.293     75.344   8.724  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.85 on 394 degrees of freedom
## Multiple R-squared:  0.5077, Adjusted R-squared:  0.5014
## F-statistic: 81.26 on 5 and 394 DF,  p-value: < 2.2e-16
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   152.68       2.74   55.63   <2e-16 ***
## sex          -143.62      60.01   -2.39     0.02 *
## bmi           580.47      67.33    8.62   <2e-16 ***
## map           344.75      68.04    5.07   <2e-16 ***
## tc           -218.31      67.31   -3.24   <2e-16 ***
## ltg           657.29      75.34    8.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 2963.644
```

```
## [1] 3022.301
```

```
## Analysis of Variance Table
##
## Model 1: y ~ age + sex + bmi + map + tc + ldl + hdl + tch + ltg + glu
## Model 2: y ~ sex + bmi + map + tc + ltg
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1    389 1141947
## 2    394 1185458 -5    -43510 2.9643 0.01221 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Start:  AIC=3204.71
## y ~ age + sex + bmi + map + tc + ldl + hdl + tch + ltg + glu
##
##        Df Sum of Sq     RSS    AIC
## - age   1        72 1142020 3202.7
## - hdl   1       806 1142754 3203.0
## - glu   1      2510 1144457 3203.6
## - tch   1      2933 1144880 3203.7
## <none>              1141947 3204.7
## - ldl   1      6825 1148773 3205.1
## - tc    1     11438 1153386 3206.7
## - sex   1     40216 1182164 3216.6
## - ltg   1     54429 1196377 3221.3
## - map   1     75090 1217038 3228.2
## - bmi   1    147581 1289529 3251.3
##
## Step:  AIC=3202.74
## y ~ sex + bmi + map + tc + ldl + hdl + tch + ltg + glu
##
##        Df Sum of Sq     RSS    AIC
## - hdl   1       824 1142844 3201.0
## - glu   1      2656 1144676 3201.7
## - tch   1      2916 1144936 3201.8
## <none>              1142020 3202.7
## - ldl   1      6890 1148910 3203.1
## - tc    1     11478 1153497 3204.7
## - sex   1     40274 1182294 3214.6
## - ltg   1     54900 1196920 3219.5
## - map   1     79224 1221244 3227.6
## - bmi   1    147570 1289590 3249.3
##
## Step:  AIC=3201.03
## y ~ sex + bmi + map + tc + ldl + tch + ltg + glu
##
##        Df Sum of Sq     RSS    AIC
## - tch   1      2185 1145029 3199.8
## - glu   1      2705 1145549 3200.0
## <none>              1142844 3201.0
## - ldl   1      8808 1151653 3202.1
## - tc    1     27555 1170400 3208.6
## - sex   1     40811 1183656 3213.1
## - map   1     78720 1221564 3225.7
## - ltg   1     92523 1235368 3230.2
## - bmi   1    147071 1289915 3247.4
##
## Step:  AIC=3199.79
## y ~ sex + bmi + map + tc + ldl + ltg + glu
##
##        Df Sum of Sq     RSS    AIC
## - glu   1      3071 1148100 3198.9
## <none>              1145029 3199.8
## - ldl   1     36551 1181580 3210.4
## - sex   1     39159 1184188 3211.2
## - tc    1     61374 1206403 3218.7
## - map   1     76944 1221973 3223.8
## - bmi   1    146794 1291823 3246.0
## - ltg   1    239636 1384665 3273.8
##
## Step:  AIC=3198.86
## y ~ sex + bmi + map + tc + ldl + ltg
##
##        Df Sum of Sq     RSS    AIC
## <none>              1148100 3198.9
## - sex   1     37042 1185142 3209.6
## - ldl   1     37358 1185458 3209.7
## - tc    1     61253 1209352 3217.7
## - map   1     84790 1232890 3225.4
## - bmi   1    158343 1306443 3248.5
## - ltg   1    262231 1410331 3279.1
```

```
##
## Call:
## lm(formula = y ~ sex + bmi + map + tc + ldl + ltg, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -157.214  -38.027   -2.143   36.163  149.530
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  152.723      2.704  56.477  < 2e-16 ***
## sex         -225.947     63.453  -3.561 0.000415 ***
## bmi          509.713     69.234   7.362 1.07e-12 ***
## map          362.152     67.222   5.387 1.23e-07 ***
## tc          -775.933    169.455  -4.579 6.28e-06 ***
## ldl          554.531    155.071   3.576 0.000392 ***
## ltg          805.250     84.993   9.474  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 54.05 on 393 degrees of freedom
## Multiple R-squared:  0.5232, Adjusted R-squared:  0.5159
## F-statistic: 71.88 on 6 and 393 DF,  p-value: < 2.2e-16
```

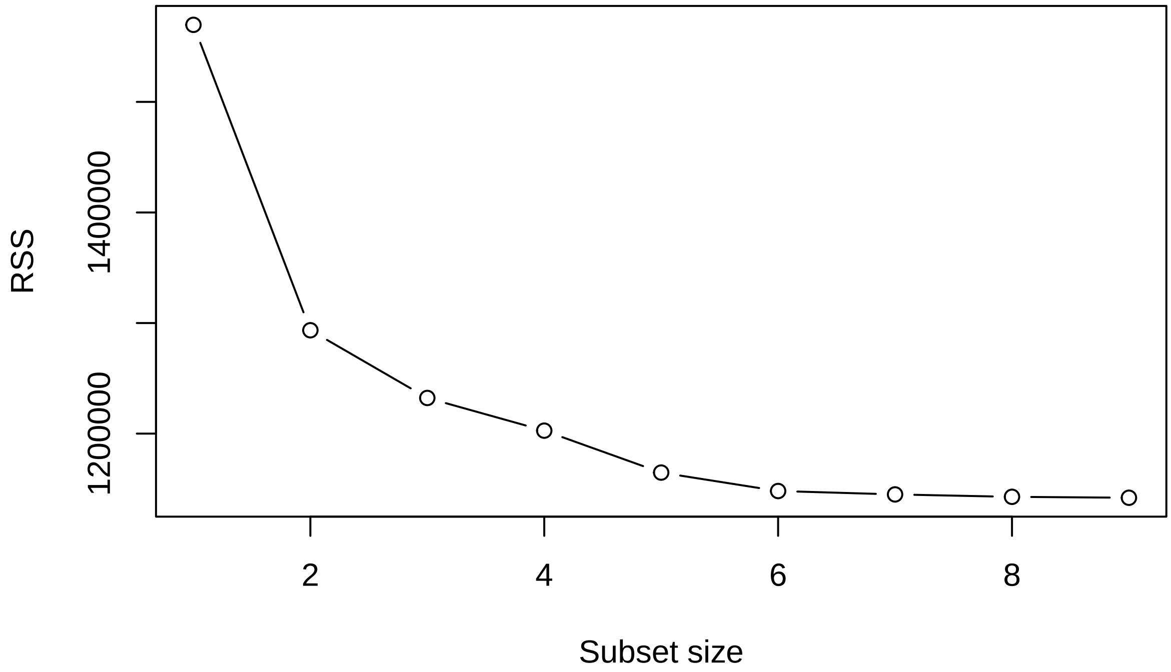
```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   152.72       2.70   56.48 < 2.2e-16 ***
## sex          -225.95      63.45   -3.56 < 2.2e-16 ***
## bmi           509.71      69.23    7.36 < 2.2e-16 ***
## map           362.15      67.22    5.39 < 2.2e-16 ***
## tc           -775.93     169.46   -4.58 < 2.2e-16 ***
## ldl           554.53     155.07    3.58 < 2.2e-16 ***
## ltg           805.25      84.99    9.47 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 2870.25
```

```
## [1] 2966.798
```

```
## Analysis of Variance Table
##
## Model 1: y ~ age + sex + bmi + map + tc + ldl + hdl + tch + ltg + glu
## Model 2: y ~ sex + bmi + map + tc + ldl + ltg
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    389 1141947
## 2    393 1148100 -4   -6152.4 0.5239 0.7182
```

```
## Subset selection object
## Call: regsubsets.formula(y ~ ., data = train, nvmax = 9, really.big = TRUE)
## 10 Variables  (and intercept)
##     Forced in Forced out
## age     FALSE      FALSE
## sex     FALSE      FALSE
## bmi     FALSE      FALSE
## map     FALSE      FALSE
## tc      FALSE      FALSE
## ldl     FALSE      FALSE
## hdl     FALSE      FALSE
## tch     FALSE      FALSE
## ltg     FALSE      FALSE
## glu     FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: exhaustive
##          age sex bmi map tc  ldl hdl tch ltg glu
## 1  ( 1 ) " " " " "*" " " " " " " " " " " " " " "
## 2  ( 1 ) " " " " "*" " " " " " " " " " " "*" " "
## 3  ( 1 ) " " " " "*" "*" " " " " " " " " "*" " "
## 4  ( 1 ) " " " " "*" "*" "*" " " " " " " "*" " "
## 5  ( 1 ) " " "*" "*" "*" " " " " "*" " " "*" " "
## 6  ( 1 ) " " "*" "*" "*" "*" "*" " " " " "*" " "
## 7  ( 1 ) " " "*" "*" "*" "*" "*" " " " " "*" "*"
## 8  ( 1 ) " " "*" "*" "*" "*" "*" " " "*" "*" "*"
## 9  ( 1 ) " " "*" "*" "*" "*" "*" "*" "*" "*" "*"
```



```
##   Adj.R2 BIC AIC
## 1      7   6   5
```

```
##
## Call:
## lm(formula = select_model(5, lm_subset, "Y"), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -148.699  -38.009   -0.413   36.673  148.969
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   152.69       2.72  56.131  < 2e-16 ***
## sex          -233.35      63.90  -3.652 0.000295 ***
## bmi           506.03      68.90   7.344 1.20e-12 ***
## map           358.97      67.63   5.308 1.86e-07 ***
## hdl          -289.95      68.92  -4.207 3.21e-05 ***
## ltg           467.58      68.89   6.787 4.22e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.37 on 394 degrees of freedom
## Multiple R-squared:  0.5163, Adjusted R-squared:  0.5101
## F-statistic:  84.1 on 5 and 394 DF,  p-value: < 2.2e-16
```

```
## [1] 2911.967
## [1] 2956.157
```

```
## Analysis of Variance Table
##
## Model 1: y ~ age + sex + bmi + map + tc + ldl + hdl + tch + ltg + glu
## Model 2: y ~ sex + bmi + map + hdl + ltg
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1    389 1141947
## 2    394 1164787 -5    -22839 1.556 0.1716
```

```
## No id variables; using all as measure variables
## No id variables; using all as measure variables
## No id variables; using all as measure variables
## No id variables; using all as measure variables
```

Table 2: Results all models

	full	small	stepwise	subset
X.Intercept.	152.71	152.68	152.72	152.69
age	9.86	NA	NA	NA
sex	-240.35	-143.62	-225.95	-233.36
bmi	499.27	580.47	509.71	506.03
map	354.98	344.75	362.15	358.97
tc	-861.16	-218.31	-775.93	NA
ldl	541.19	NA	554.53	NA
hdl	116.05	NA	NA	-289.96
tch	166.52	NA	NA	NA
ltg	773.90	657.29	805.25	467.58
glu	63.63	NA	NA	NA
MSE in sample	2854.87	2963.64	2870.25	2911.97
MSE out of sample	2945.38	3022.30	2966.80	2956.16

Exercise 2

Exercise 3

Exercise 4

Exercise 5