

Introduction to Machine Learning - Exercise 1

Mikko Ahro

Problem 1

Task a

Read p1.csv into dataframe and drop columns “id”, “SMILES”, “InChIKey”

```
p1data <- read.csv("data/p1.csv", header=TRUE, sep=",")  
p1data <- subset(p1data, select=-c(id, SMILES, InChIKey))
```

Task b

Summary statistics for variables “pSat_PA”, “NumOfConf” and “ChemPot_kJmol”:

pSat_Pa	NumOfConf	ChemPot_kJmol
Min. : 0.0000	Min. : 2.00	Min. :-3.160
1st Qu.: 0.0000	1st Qu.: 73.25	1st Qu.: 9.723
Median : 0.0001	Median : 172.50	Median :12.781
Mean : 2.9620	Mean : 223.50	Mean :12.434
3rd Qu.: 0.0023	3rd Qu.: 324.25	3rd Qu.:15.659
Max. :562.8970	Max. :1058.00	Max. :28.096

Task c

Mean and standard deviation of column ‘ChemPot_kJmol’ are:

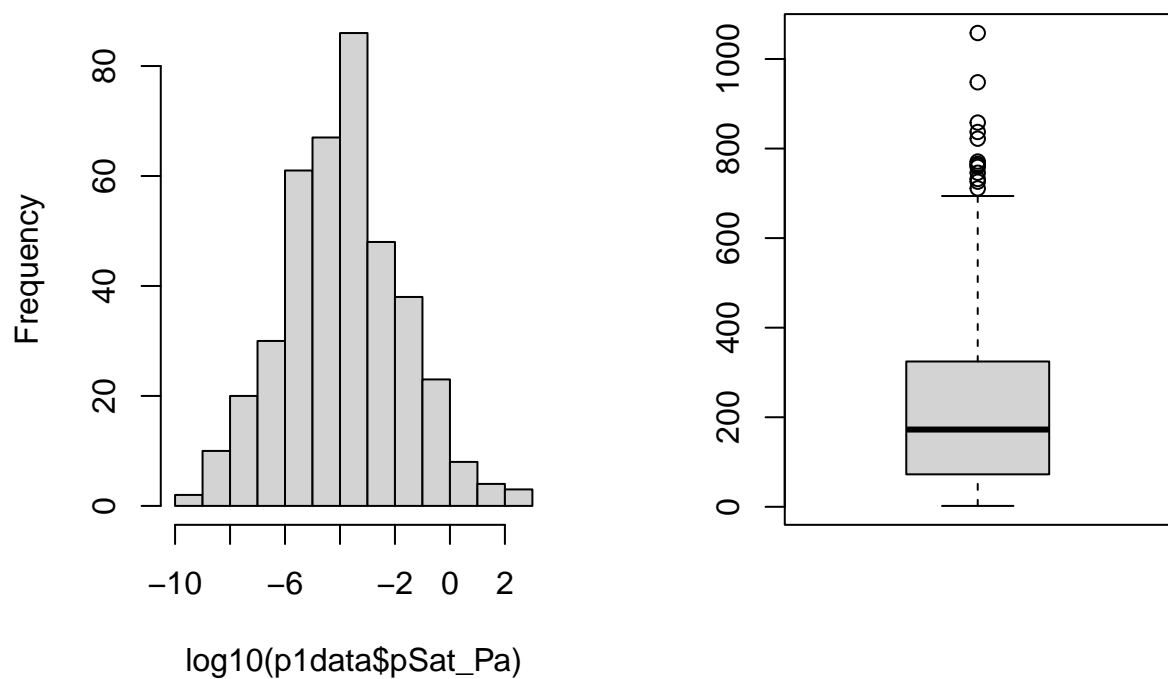
[1] “Mean: 12.4344270896”

[1] “Standard deviation: 4.77887217784492”

Task d

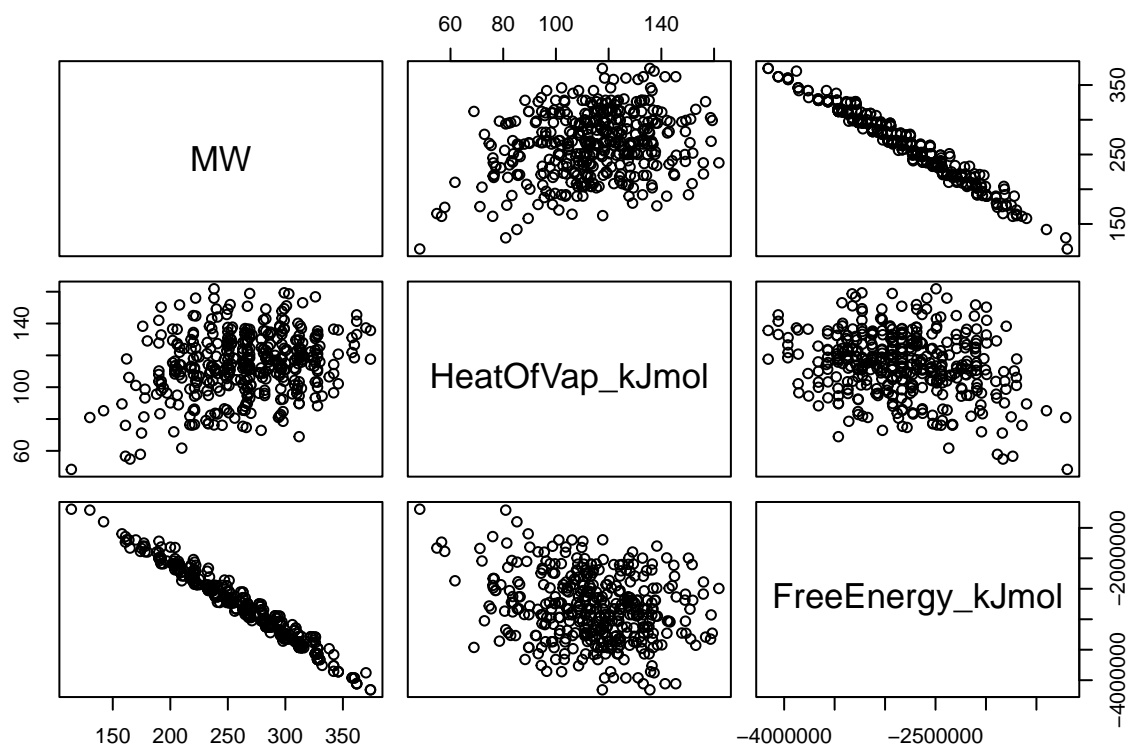
```
par(mfrow=c(1,2))  
hist(log10(p1data$pSat_Pa))  
boxplot(p1data$NumOfConf)
```

Histogram of $\log_{10}(p1data\$pSat_I)$



Task e

```
scatter_subset <- subset(p1data, select=c(MW, HeatOfVap_kJmol, FreeEnergy_kJmol))
pairs(scatter_subset)
```



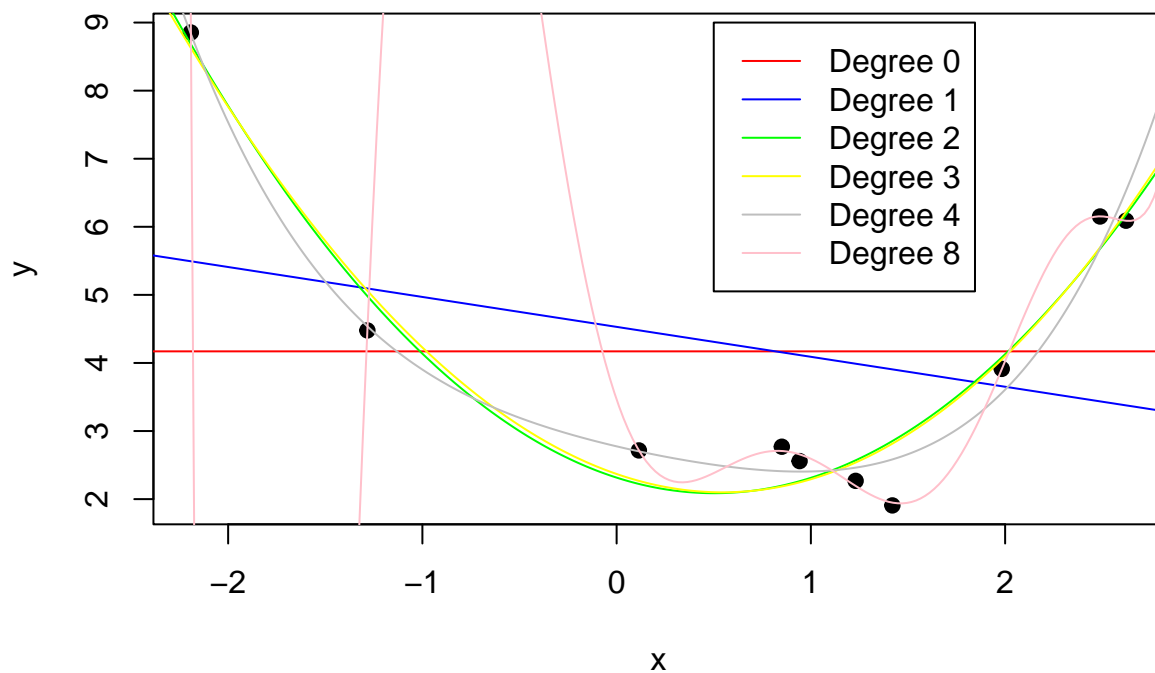
Problem 2

Task a

Degree	Train	Validation	Test	TestTRVA	CV
0	4.5122613	4.512261	4.512261e+00	4.587239	10.7991438
1	4.0885351	3.494124	5.206372e+00	4.786172	7.4522264
2	0.2185859	7.021118	1.424954e+01	14.791603	0.1596893
3	0.2168190	7.154893	1.383458e+01	14.096042	0.1590119
4	0.1187955	8.776121	1.968113e+01	15.009734	0.1589143
5	0.0965322	7.221166	2.975686e+01	20.134323	0.1590315
6	0.0075741	6.050151	1.564333e+02	12.060633	0.1592266
7	0.0049994	11.394430	1.104038e+03	15.628661	0.1611671
8	0.0020825	407.157118	1.561695e+05	10.979288	0.1602785

Task b

Fitted Polynomial Curves



Task c

```
## Loading required package: ggplot2
## Loading required package: lattice
##
## Attaching package: 'lattice'
## The following object is masked from 'package:boot':
##
##   melanoma
## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##      margin

## Loading required package: Matrix

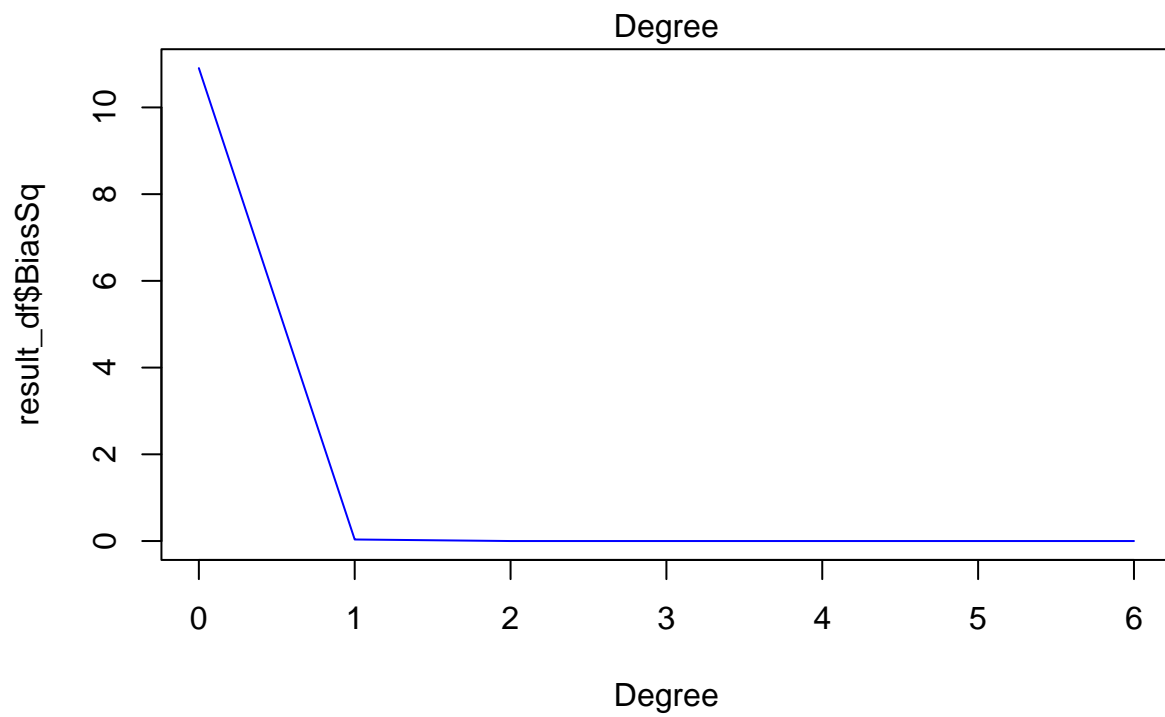
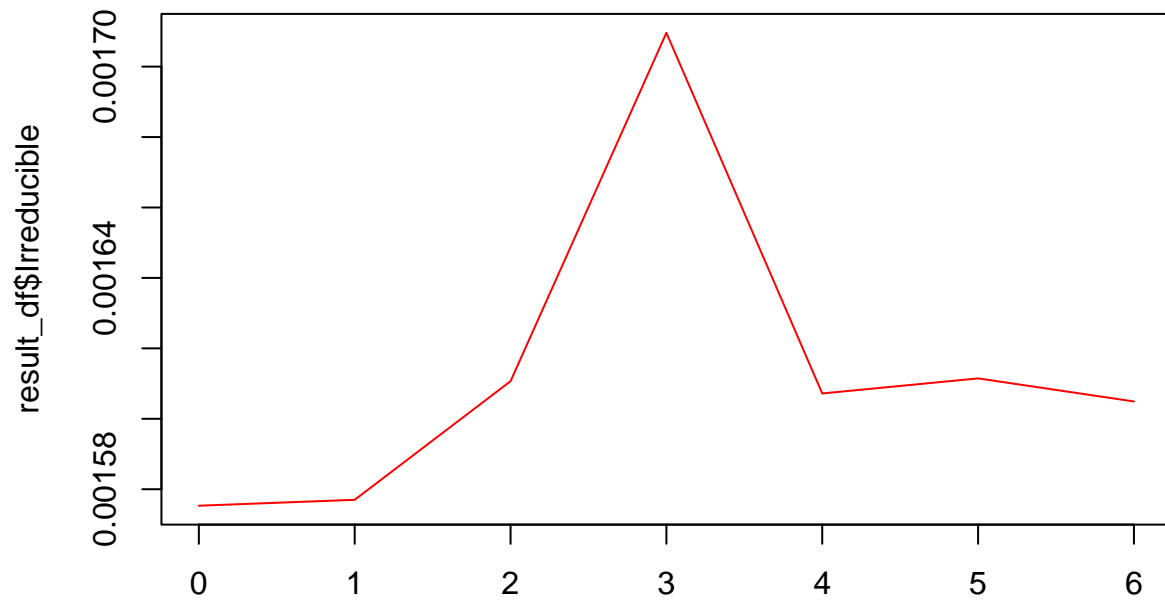
## Loaded glmnet 4.1-8

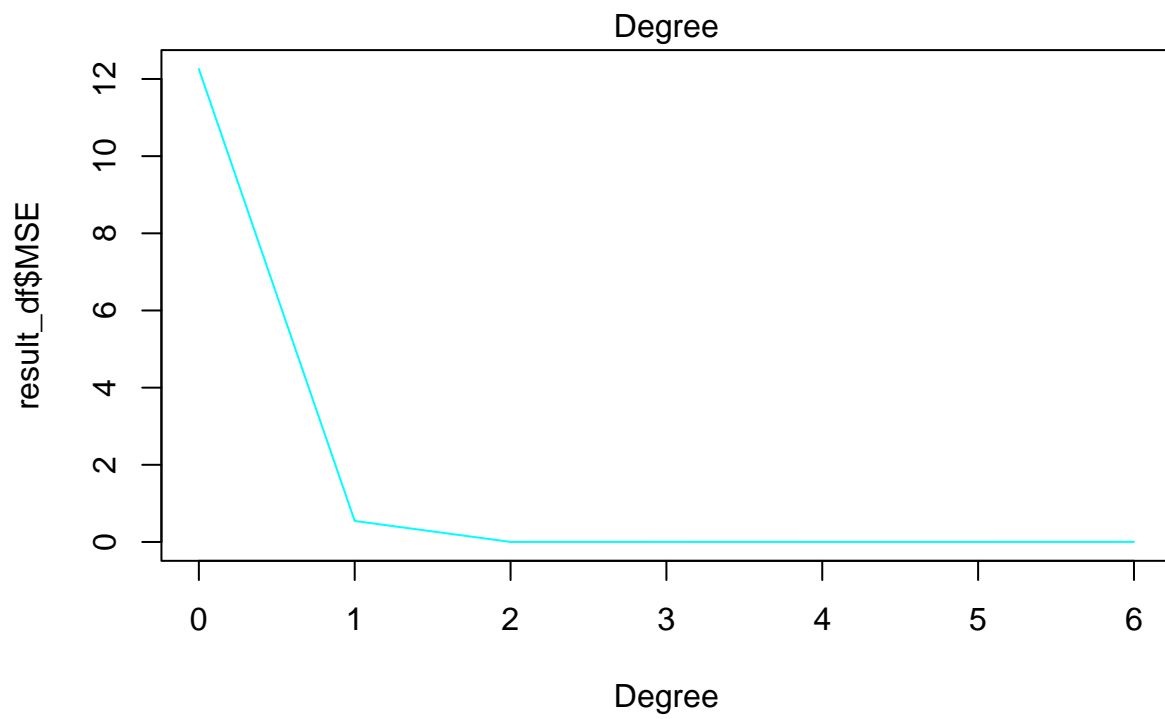
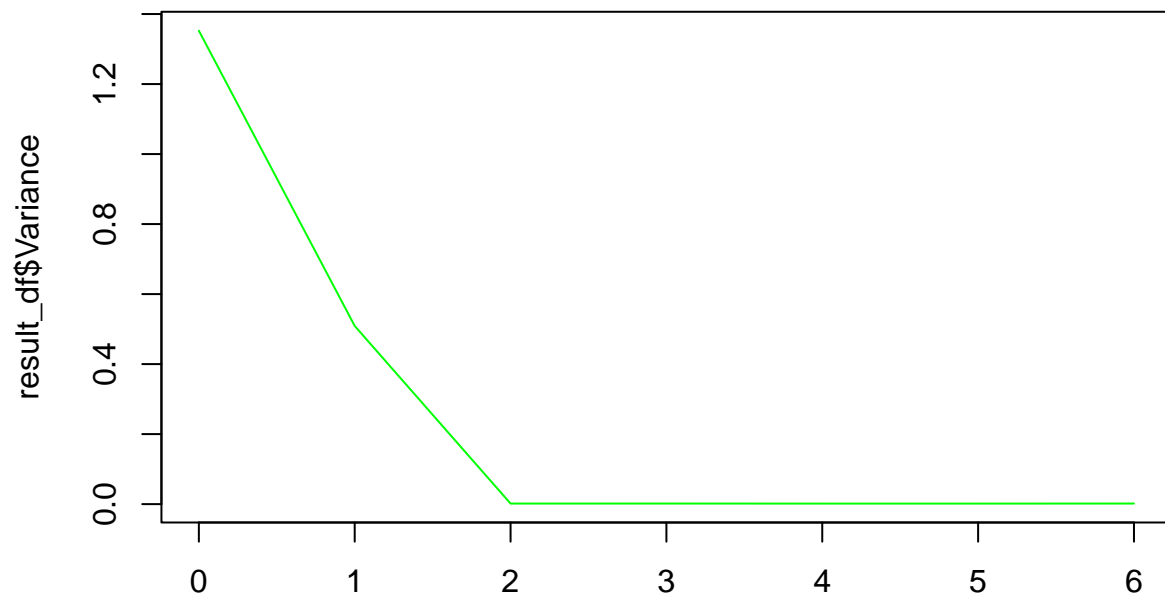
Next_Tmax ~ station + Present_Tmax + Present_Tmin + LDAPS_RHmin + LDAPS_RHmax
+ LDAPS_Tmax_lapse + LDAPS_Tmin_lapse + LDAPS_WS + LDAPS_LH + LDAPS_CC1 +
LDAPS_CC2 + LDAPS_CC3 + LDAPS_CC4 + LDAPS_PPT1 + LDAPS_PPT2 + LDAPS_PPT3 +
LDAPS_PPT4 + lat + lon + DEM + Slope + Solar.radiation
```

Problem 3

Task a

Task b





Problem 4

Not done

Problem 5

Task a

Task b

Task c

Problem 6

Task a

Task b

Task c

Problem 7

Task a

Task b