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

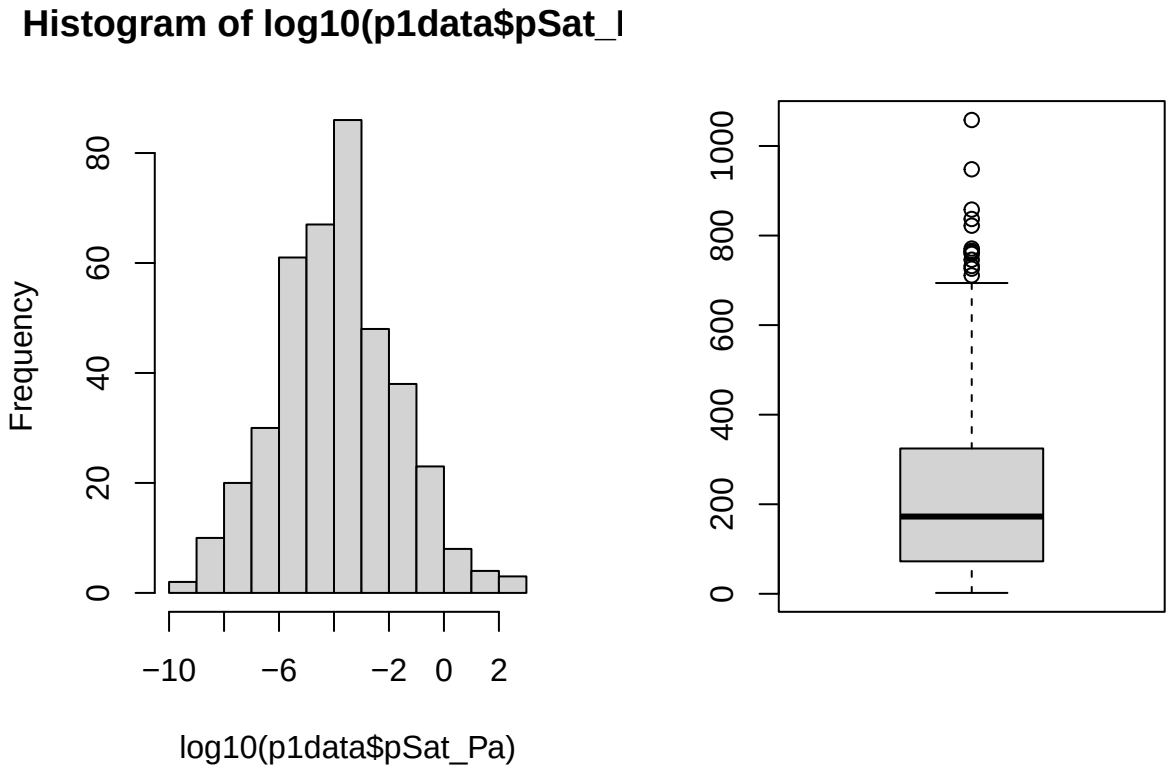
Mean and standard deviation of column ‘ChemPot_kJmol’ are:

[1] “Mean: 12.4344270896”

[1] “Standard deviation: 4.77887217784492”

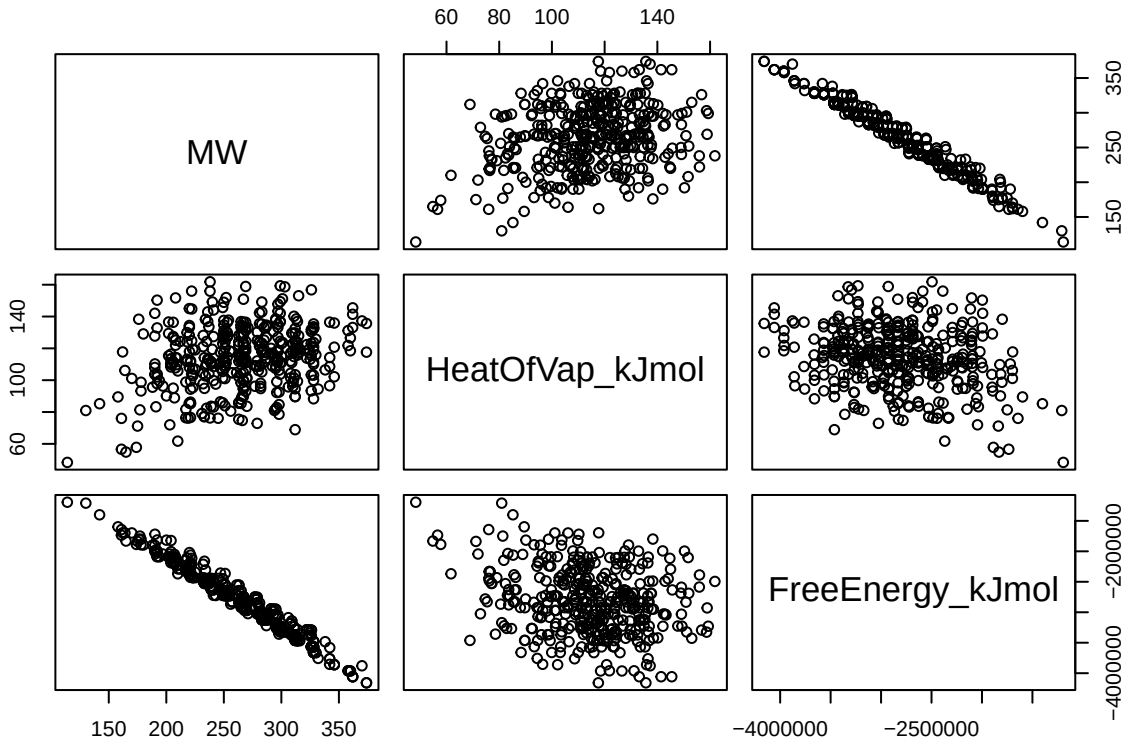
Task d

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



Task e

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

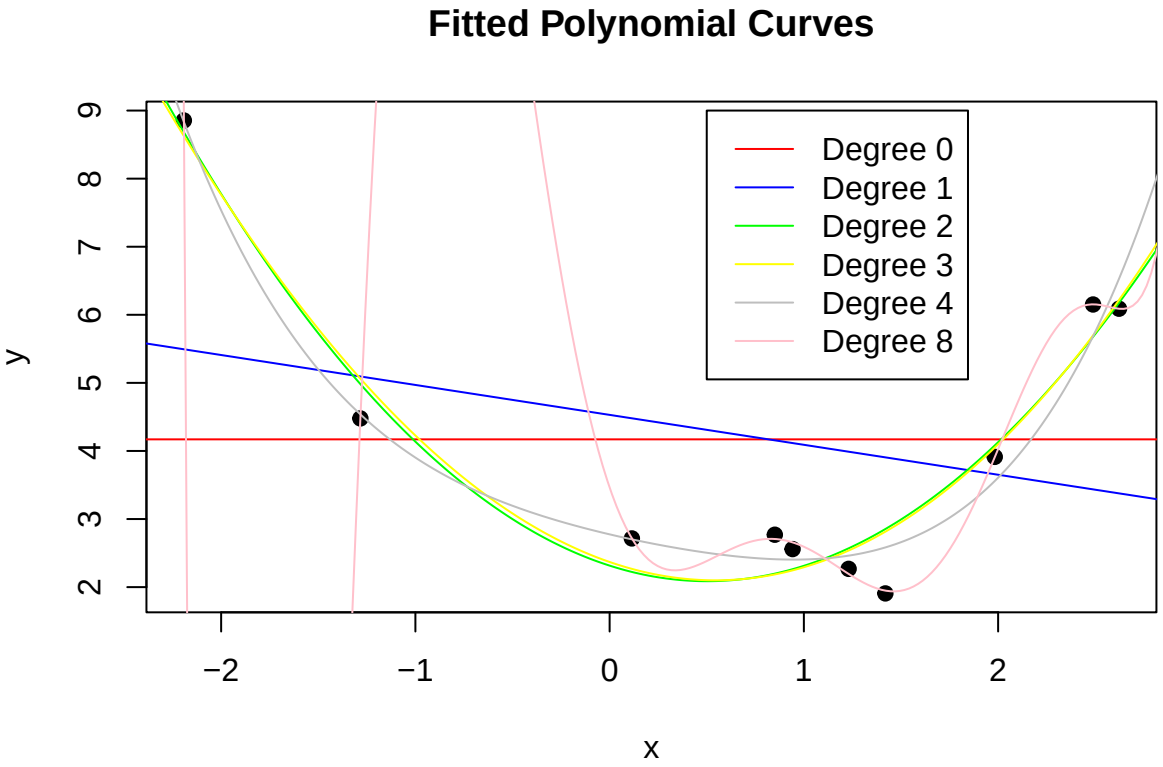


Problem 2

Task a

Degree	Train	Validation	Test	TestTRVA	CV
0	4.5122613	4.512261	4.512261e+00	4.587239	10.7949568
1	4.0885351	3.494124	5.206372e+00	4.786172	7.5235639
2	0.2185859	7.021118	1.424954e+01	14.791603	0.1589217
3	0.2168190	7.154893	1.383458e+01	14.096042	0.1588743
4	0.1187955	8.776121	1.968113e+01	15.009734	0.1598923
5	0.0965322	7.221166	2.975686e+01	20.134323	0.1591377
6	0.0075741	6.050151	1.564333e+02	12.060633	0.1604102
7	0.0049994	11.394430	1.104038e+03	15.628661	0.1594904
8	0.0020825	407.157118	1.561695e+05	10.979288	0.1608696

Task b



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.
```

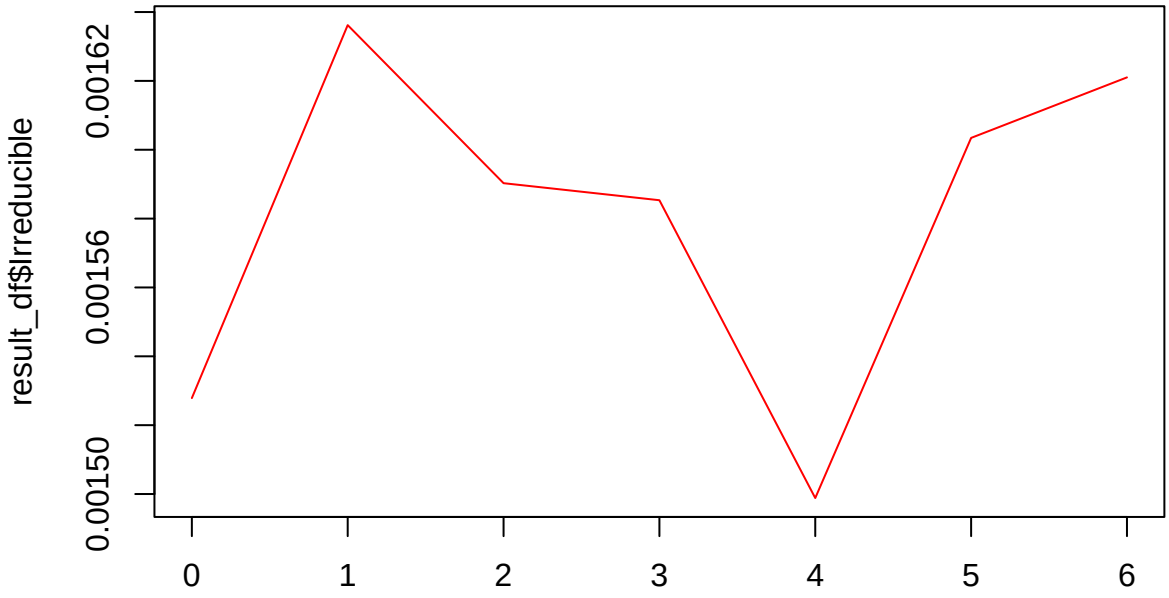
```
##
## Attaching package: 'randomForest'

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

## Loading required package: Matrix

## Loaded glmnet 4.1-8
```

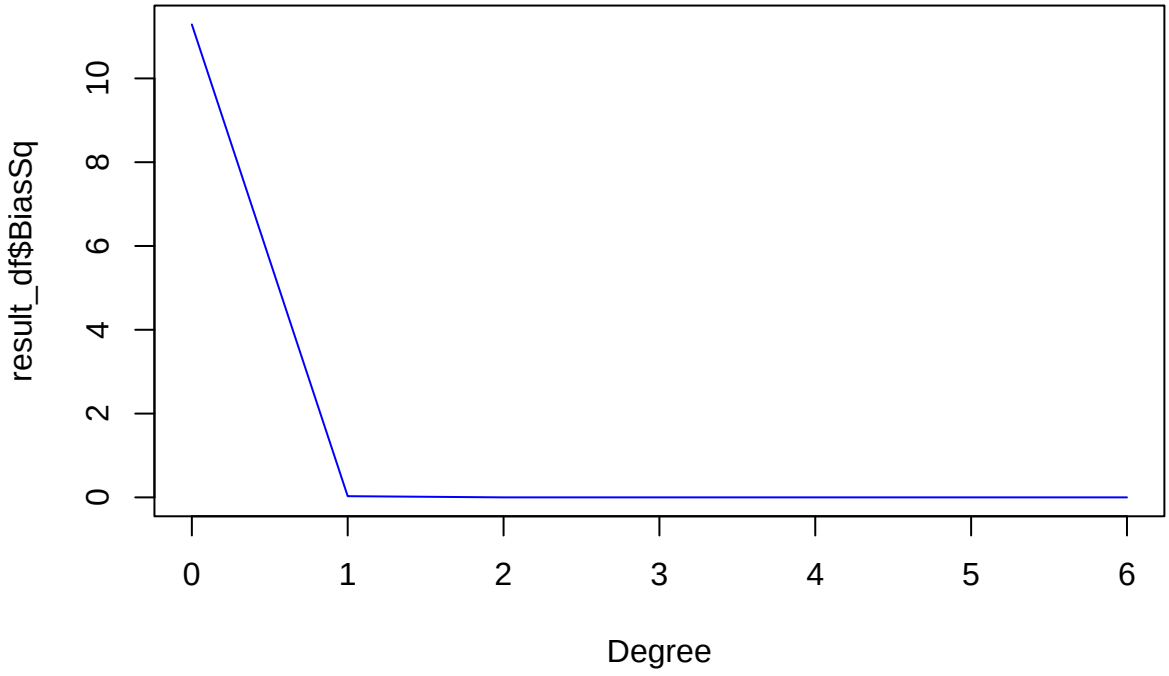
Next_Tmax ~ station + Present_Tmax + Present_Tmin + LDAPS_RHmin + LDAPS_RHmax + LDAPS_Tmax_lapse + LDAPS_Tmin_lapse + LDAPS_WS + LDAPS_LH + LDAPS_CC1 + LDAPS_CC2 + LDAPS_CC3 + LDAPS_CC4 + LDAPS_PPT1 + LDAPS_PPT2 + LDAPS_PPT3 + LDAPS_PPT4 + lat + lon + DEM + Slope + Solar.radiation

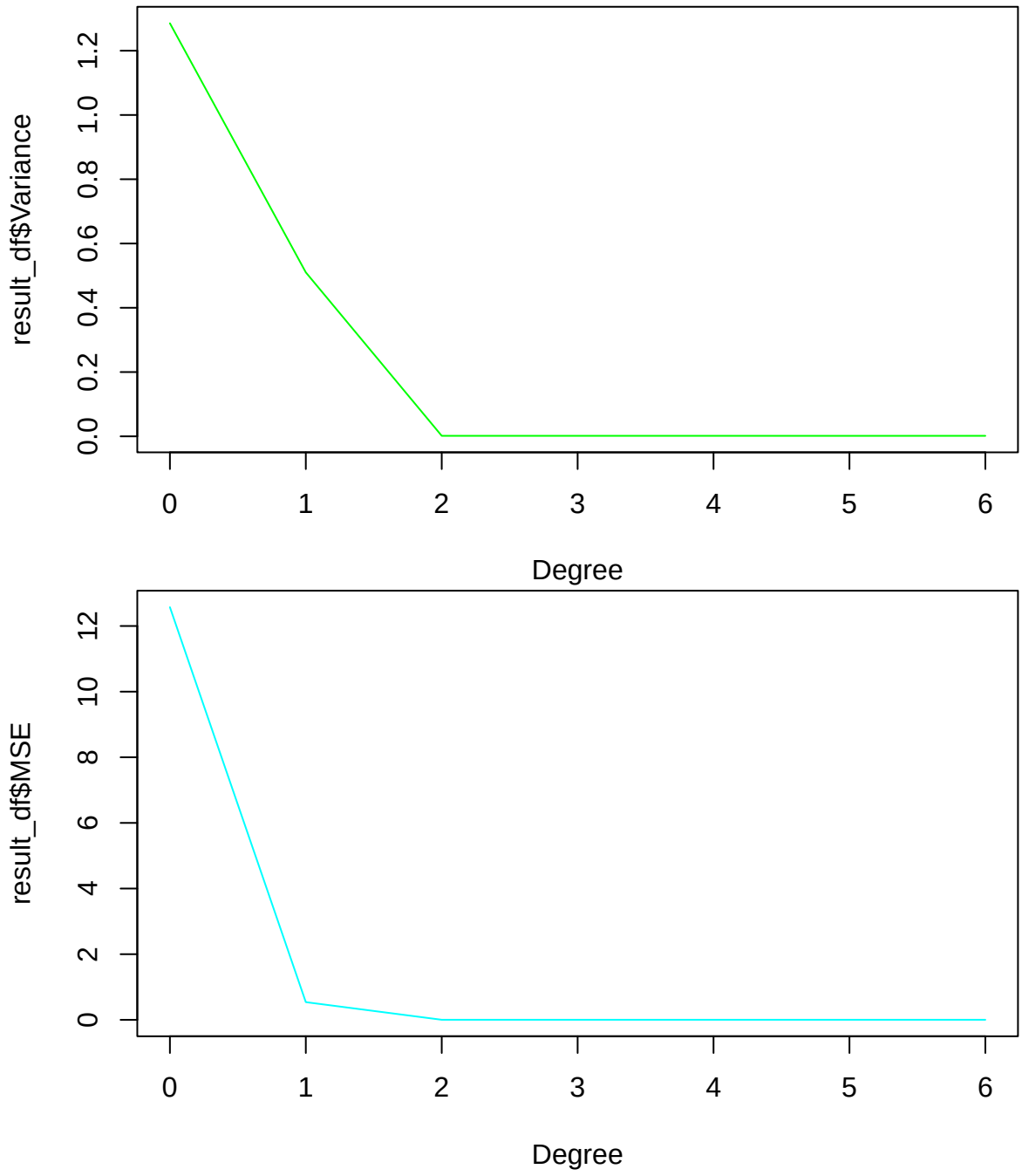
Problem 3

Task a

Task b







Problem 4

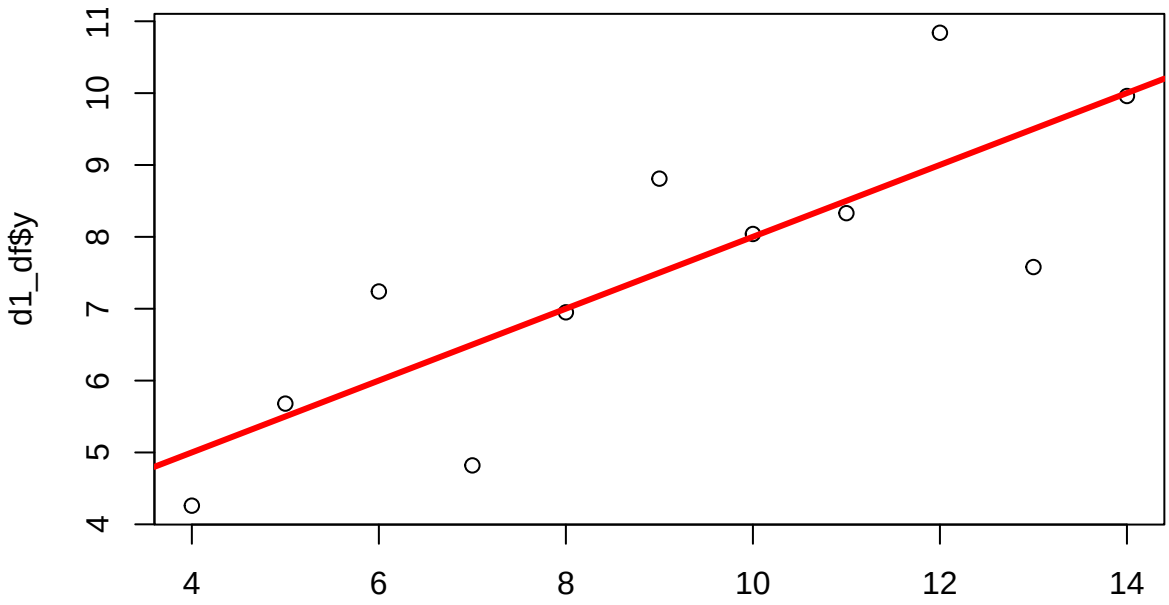
Not done

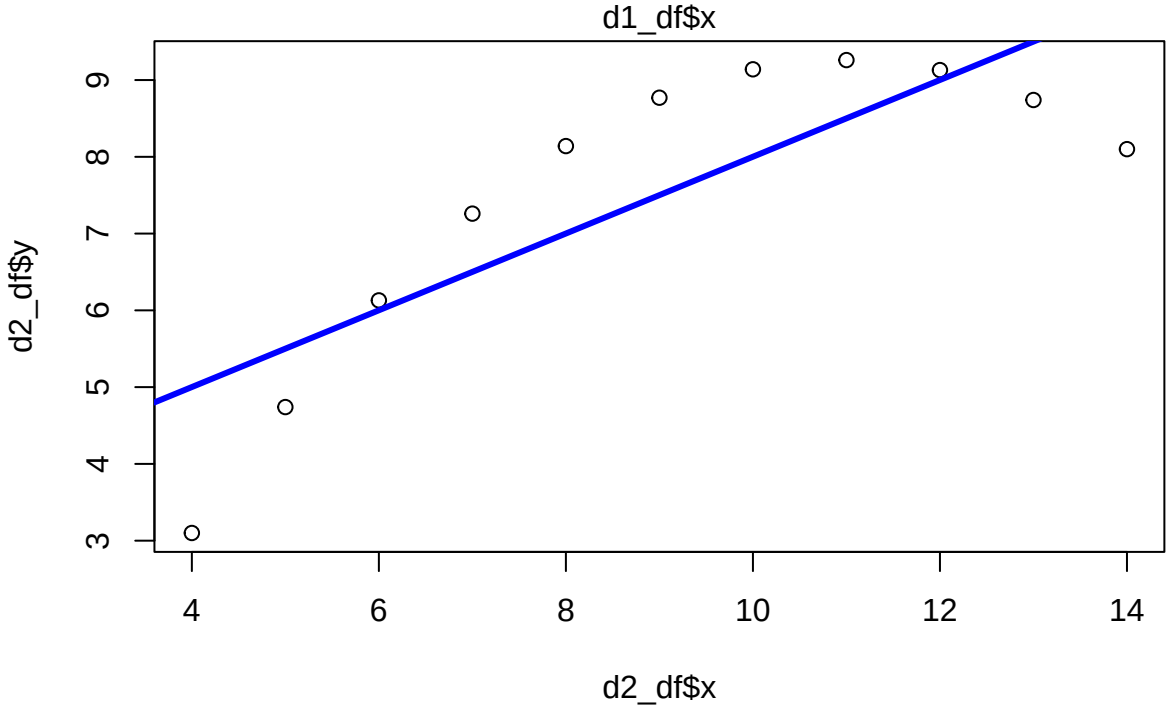
Problem 5

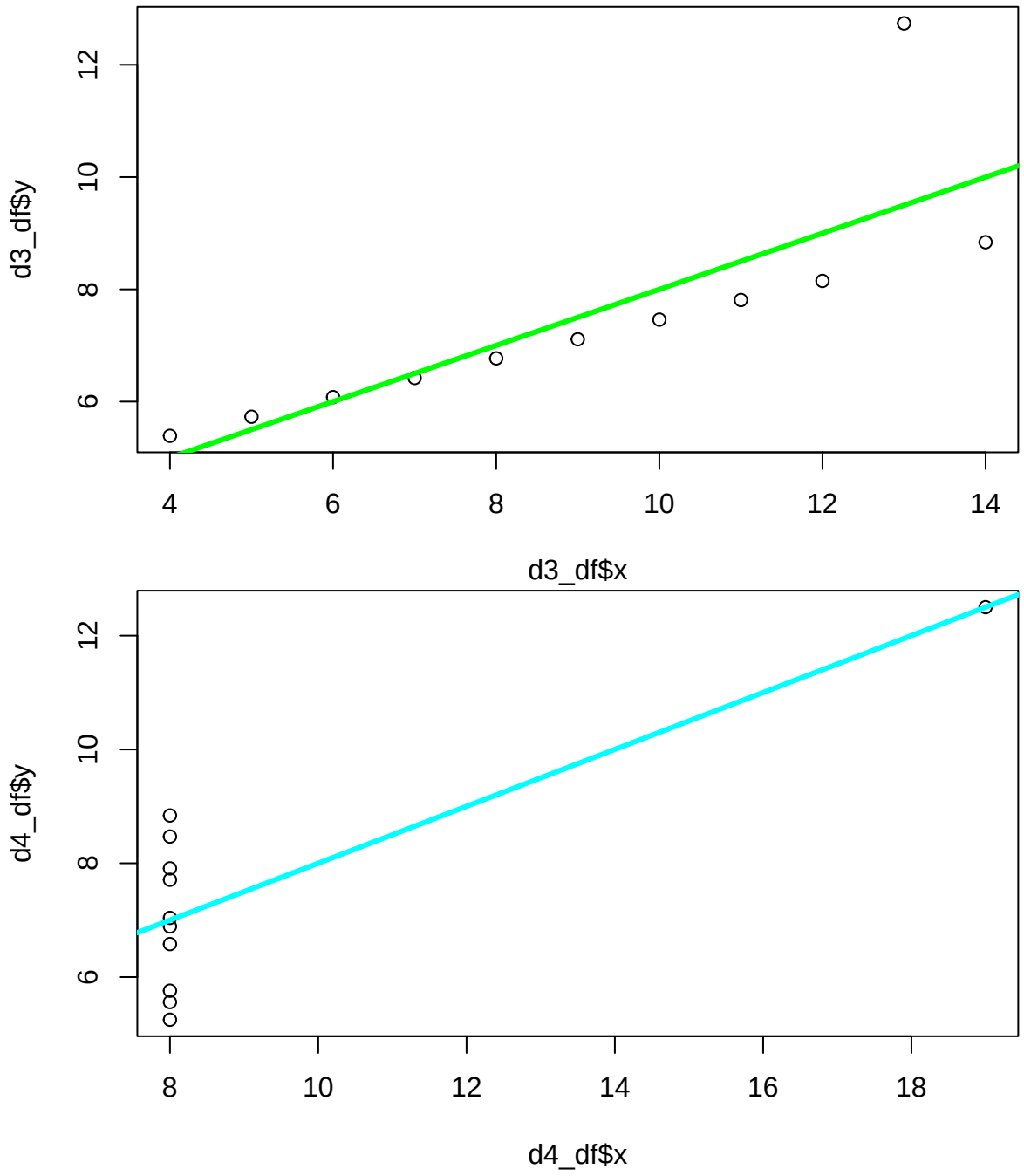
Task a

Data	Intercept	Int_SE	Int_p_value	Slope	Slope_SE	Slope_p_value	R_Squared
d1	3.000091	1.124747	0.0257341	0.5000909	0.1179055	0.0021696	0.6665425
d2	3.000909	1.125302	0.0257589	0.5000000	0.1179637	0.0021788	0.6662420
d3	3.002454	1.124481	0.0256191	0.4997273	0.1178777	0.0021763	0.6663240
d4	3.001727	1.123921	0.0255904	0.4999091	0.1178189	0.0021646	0.6667073

Task b



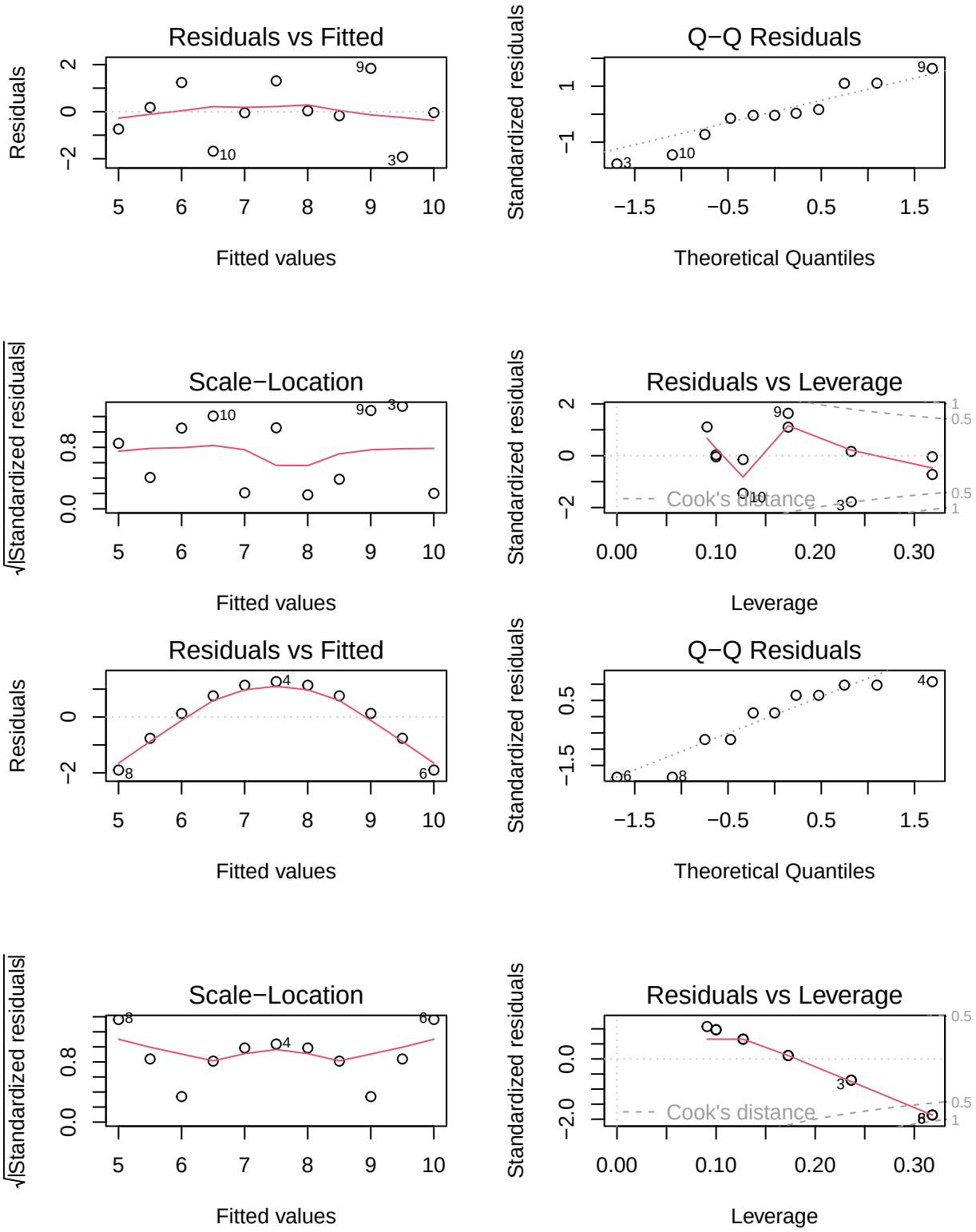


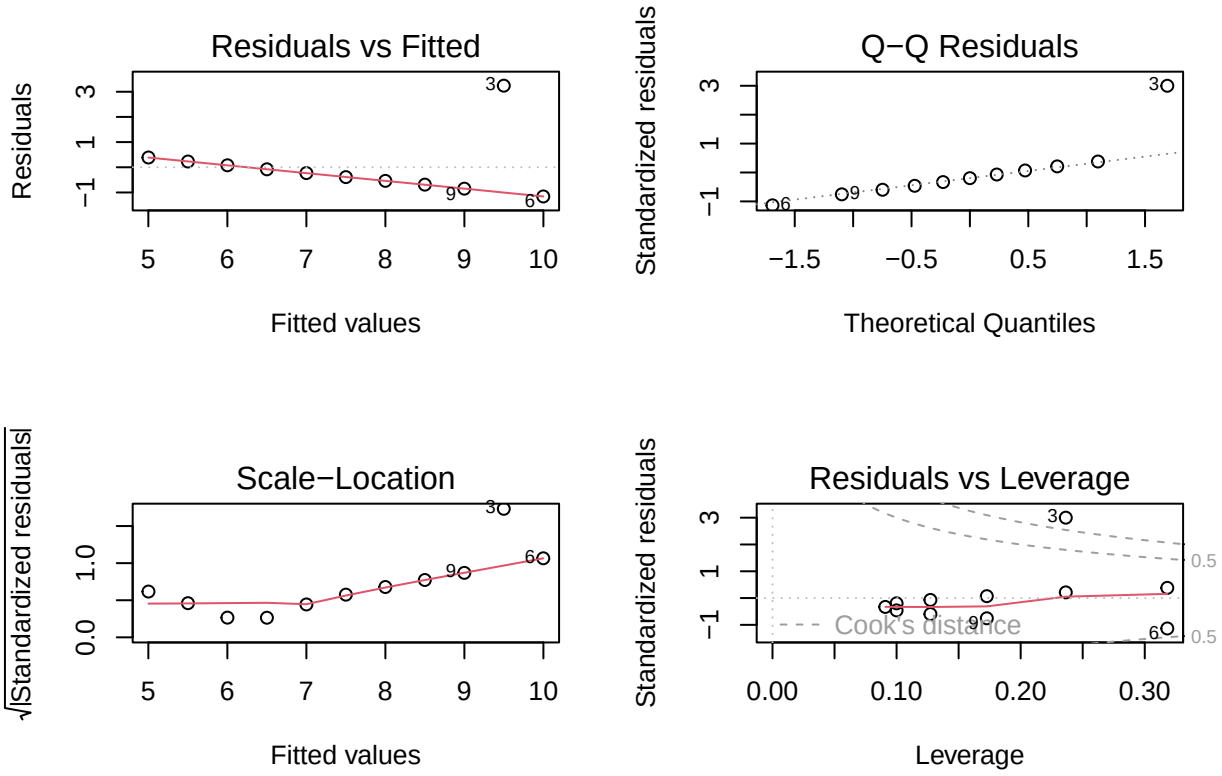


Task c

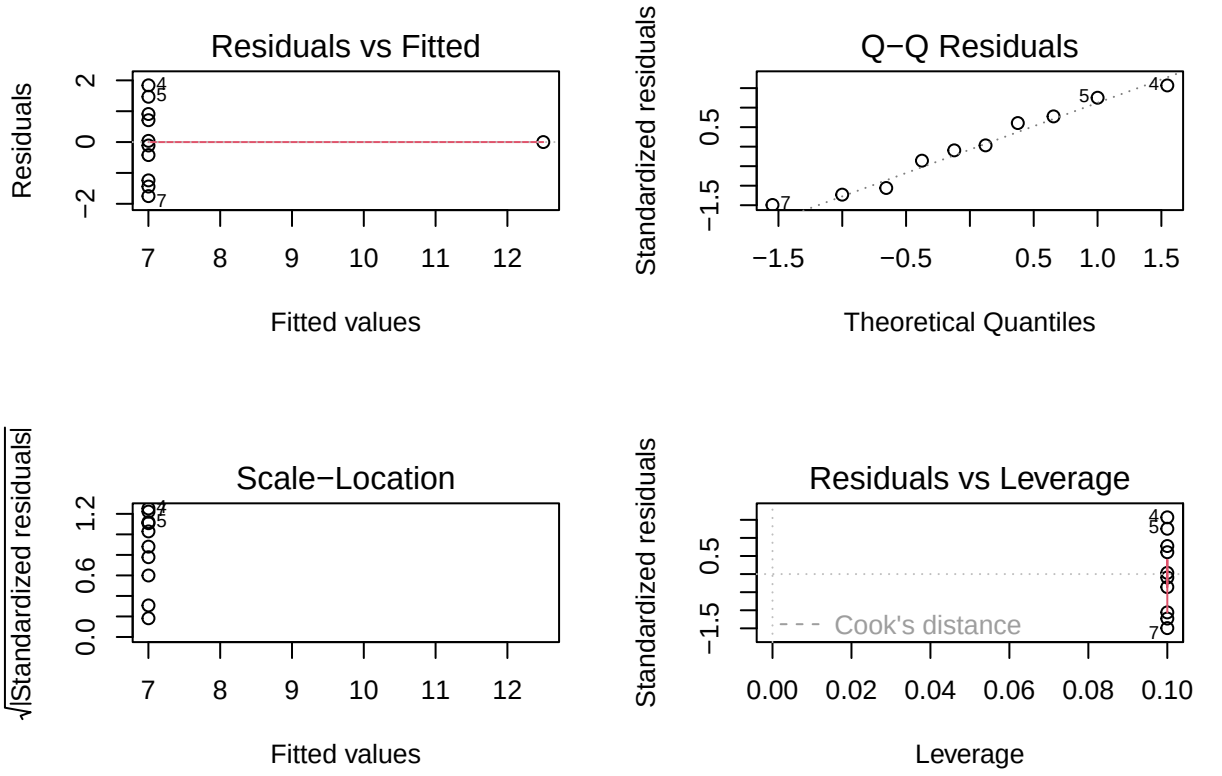
Potential problems with regression models:

1. Non-linearity of the response predictor relationship
2. Correlation of error terms
3. Non-constant variance of error terms
4. Outliers
5. High-leverage points
6. Collinearity


Residuals vs Fitted
Residuals
Fitted values
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
Residuals vs Fitted
Residuals
Fitted values
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage




```
## Warning: not plotting observations with leverage one:
##   8
```



Problem 6

Task a

Task b

Task c

Problem 7

Task a

Task b