Exercise Set 1

This report contains my answers for exercise set 1.

Problem 1

Task a

Drop columns `id` and `partlybad`:

```
dt[, c("id", "partlybad") := NULL]
```

Task b

Summary of columns `T84.mean`, `UV_A.mean` and `CS.mean`:

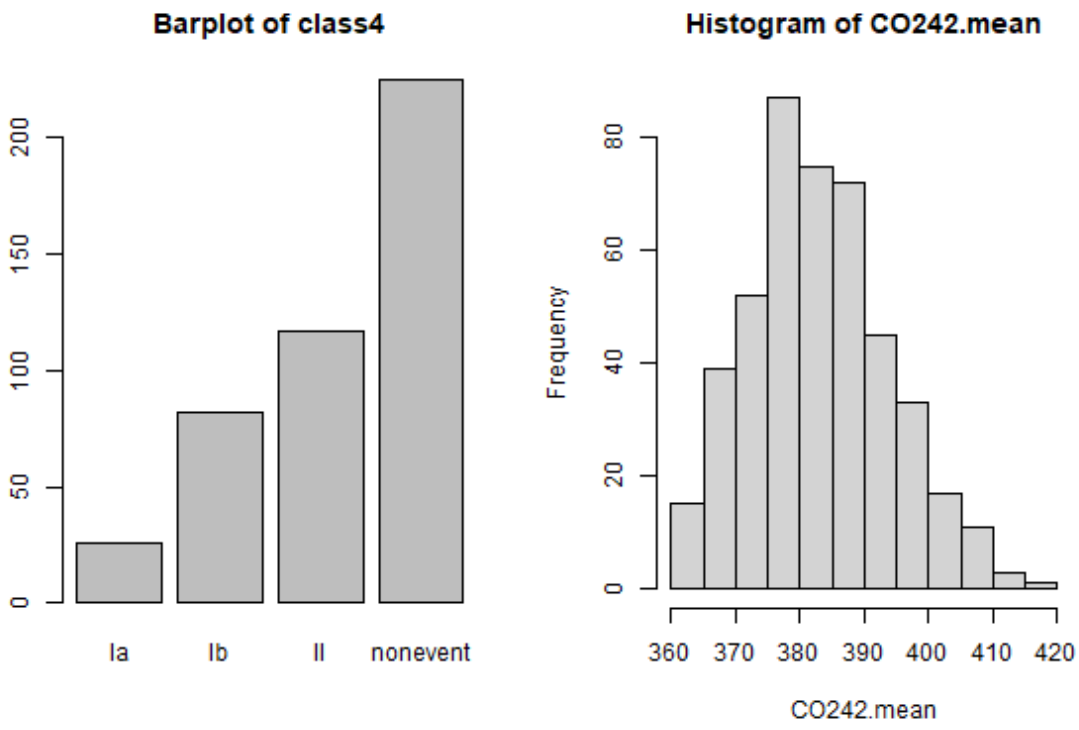
```
   T84.mean          UV_A.mean         CS.mean
 Min.   :-23.6277   Min.   : 0.439   Min.   :0.0003425
 1st Qu.: -0.7332   1st Qu.: 4.305   1st Qu.:0.0014362
 Median :  7.6616   Median :11.507   Median :0.0025485
 Mean   :  6.4396   Mean   :10.797   Mean   :0.0030832
 3rd Qu.: 14.1183   3rd Qu.:16.573   3rd Qu.:0.0041187
 Max.   : 27.0018   Max.   :22.500   Max.   :0.0137057
```

Task c

List of `t84.mean` mean and standard deviation:

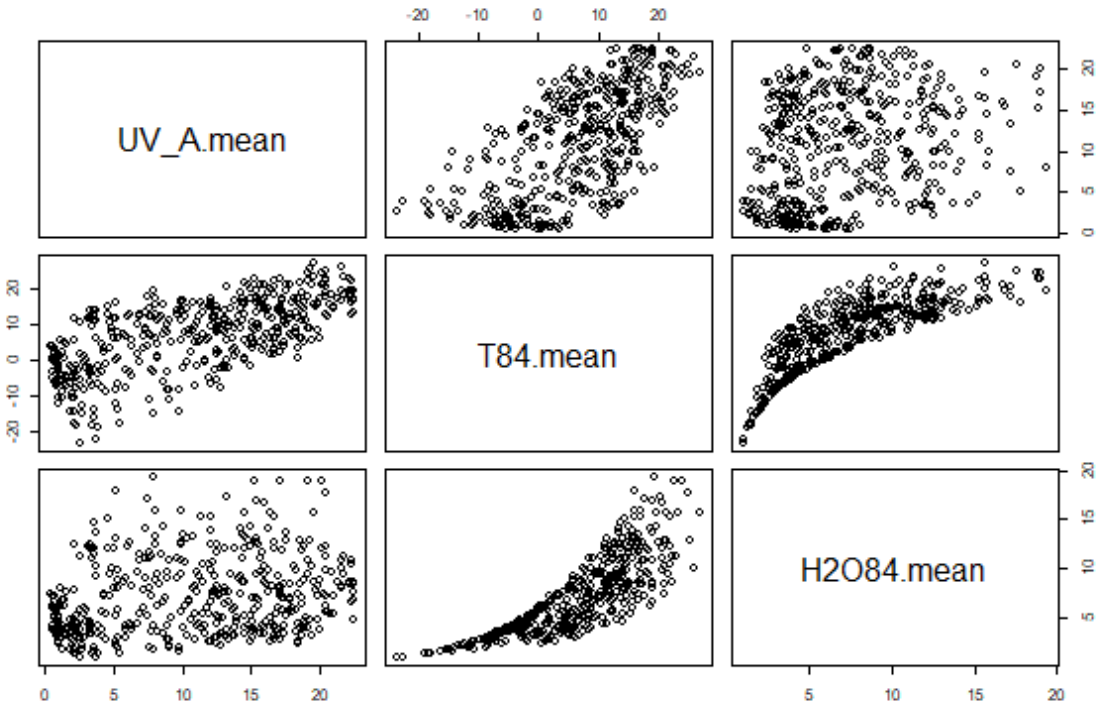
```
$t84_mean
[1] 6.439594

$t84_sd
[1] 9.827282
```

Task d



Task e

Scatterplot matrix of `UV_A.mean`,`T84.mean` and `H2O84.mean`:



Task f

A dummy model was created by first determining the most common class in column `class4` which was `nonevent`. Then the probability for an event was calculated by `p_event = nrows events/total nrows`. The `event` frequencies in the data summed up to 225 which ìs the same value as the frequency of `nonevent` values which means that `p_event=0.5`. These dummy predictions produced a score of `0.36269` on the Kaggle leaderboard.

Problem 2

Task a

MSE table for synthetic data:

Degree	Train	Validation	Test	TestTRVA	CV
0	18.4595847	32.3423634	2.210066e+01	21.6202212	28.6033242
1	4.0885351	7.1278436	8.876307e+00	9.9349426	6.7792021
2	0.2185859	0.2937319	2.458467e-01	0.2140161	0.3158100

Degree	Train	Validation	Test	TestTRVA	CV
3	0.2168190	0.2834474	2.900799e-01	0.2751106	0.3357836
4	0.1187955	0.6247259	9.690780e-01	0.2239929	0.4358991
5	0.0965322	0.5734800	4.894836e+00	1.0390890	0.4227832
6	0.0075741	3.4167870	2.132971e+02	0.8814708	0.3874940
7	0.0049994	6.8629927	1.261988e+03	0.2717186	0.4426611
8	0.0020825	401.6517804	1.542669e+05	11.2235541	2.4842370

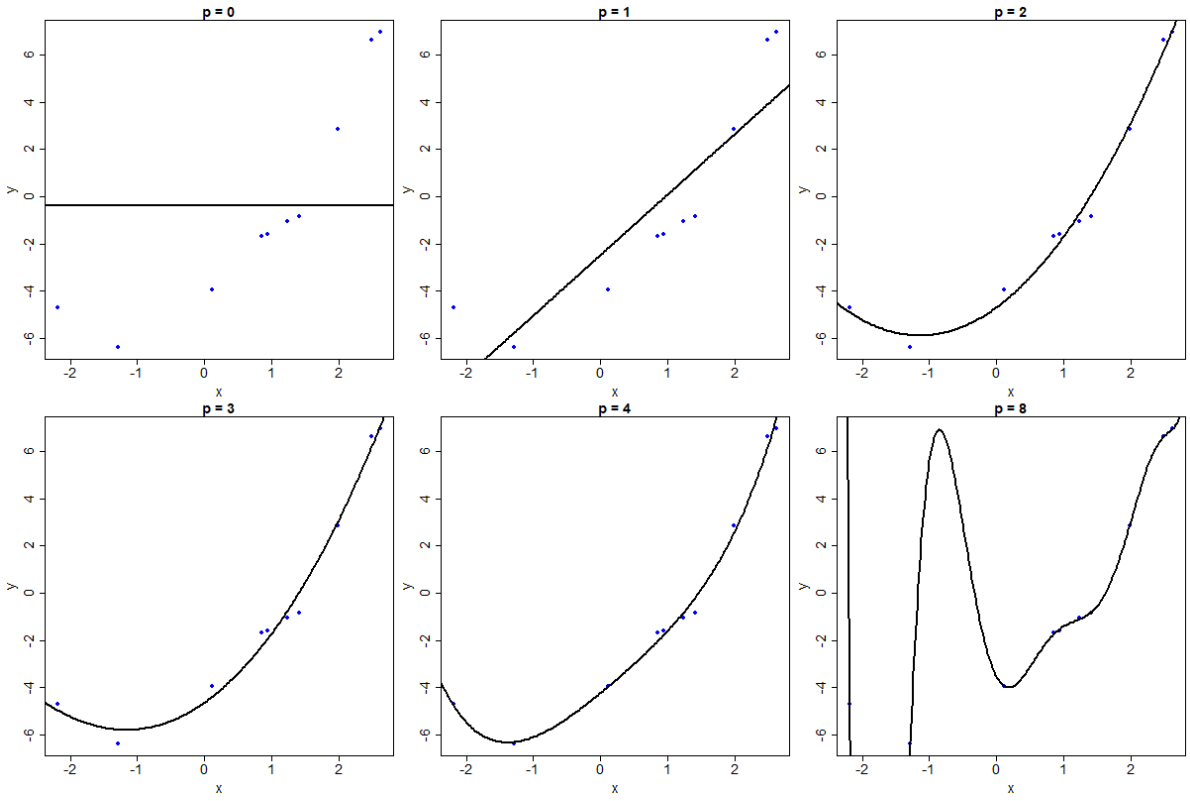
To choose the polynomial order when given a combined training and validation set but no test set I would look at the smallest error value in the table and find out its value in th `Degree` column. I would only be looking at the `Validation` and `CV` columns as they are the ones available for this task. The resulting table:

Metric	Degree	Value
Validation	3	0.2834474
CV	2	0.3158100

As the table shows I should train the model on the training set with `Degree` set to `3` and test it on the validation set.

Task b

Plots of synthetic training set points x and y and fitted polynomials with different values of p:



Task c

Table of RMSE from fitting different regressors on the real data:

Regressor	Train	Test	CV
Dummy	3.0892748	2.997846	3.086232
OLS	1.3783781	1.483817	1.444134
RF	0.5760683	1.451084	1.380194
SVR	1.4809268	1.643663	1.502207
GBM	1.4700739	1.670234	1.422283

Answers:

1. The Random Forest (RF) is the best regressor because it produces the smallest values for all cases as shown below:

Metric	Regressor	Value
Train	RF	0.5760683
Test	RF	1.4510838
CV	RF	1.3801935

2. Excluding the *Dummy* model, the values of *Train* are smaller than those of *Test*. The same is true for *CV*. The *RF Train* value is substantially lower at *0.5760683* compared to a *Test* value of *1.4510838*.

3.The regressors could be improved by excluding certain predictors from the data. Further analysis should be done to find out which ones.

Problem 3

Task a

1. Training error: The MSE keeps decreasing as flexibility increases.
2. Test error: A U-shaped curve. Error decreases and then increases again after reaching a minimum point.
3. (squared) bias: In general the more flexibility a model has, the less bias there is because the model is able to follow the patterns in the data more closely.
4. Variance: In general the variance grows as flexibility increases because with large flexibility even a small change to the data can result in a big change in $\hat{f}$.

5.irreducible (or Bayes) error: The irreducible error curve/line shows the threshold of the lowest achievable MSE for any method and therefore it is constant.

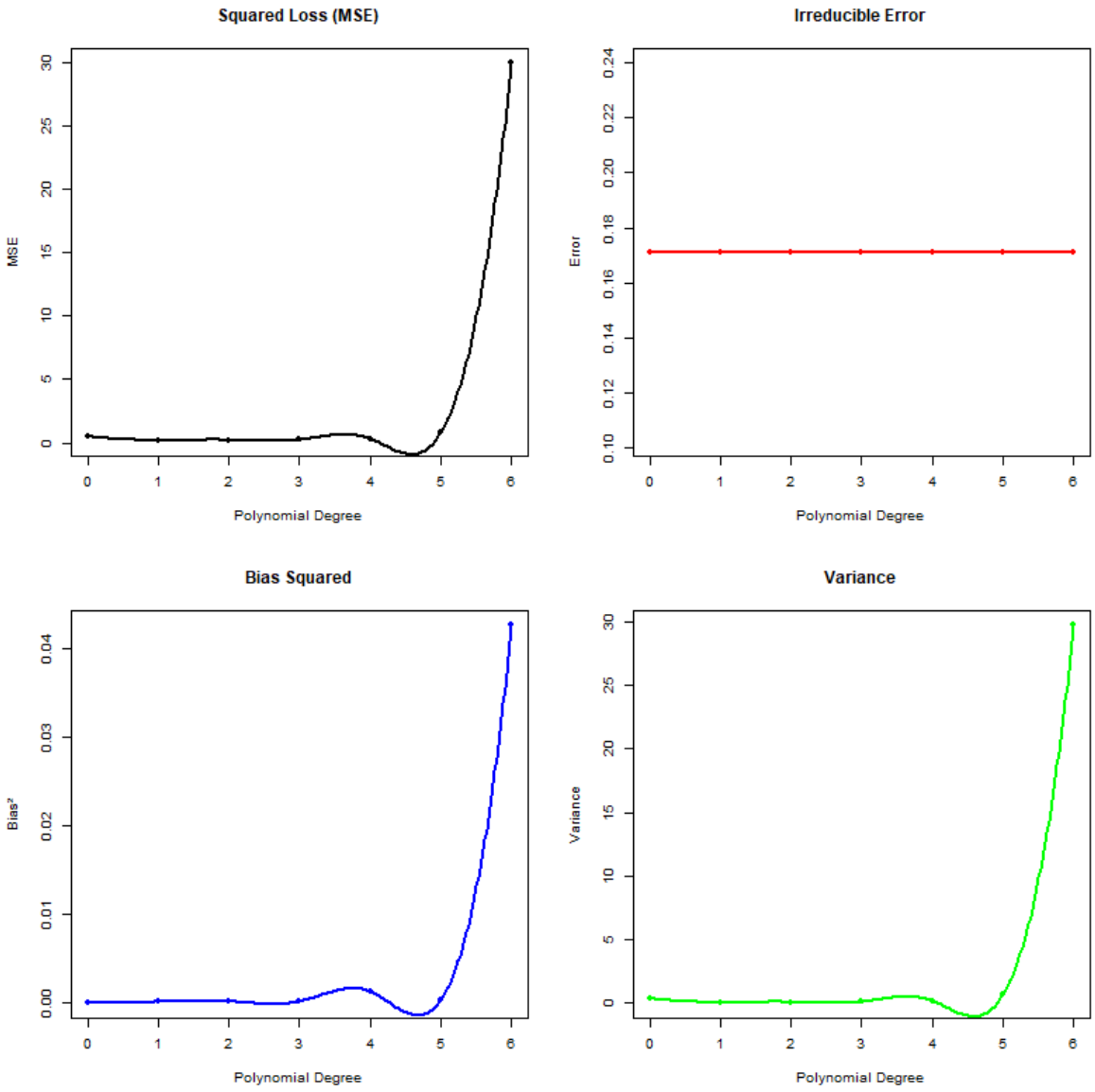
Task b

(i) Table of terms for each degree:

Degree	BiasSq	Variance	Irreducible	Total	MSE
0	0.0000000	0.3191161	0.1713681	0.4904842	0.4904842
1	0.0000613	0.0185219	0.1713681	0.1899513	0.1899513
2	0.0001024	0.0416969	0.1713681	0.2131674	0.2131674
3	0.0001094	0.0545275	0.1713681	0.2260050	0.2260050
4	0.0012264	0.1265380	0.1713681	0.2991326	0.2991326
5	0.0002717	0.6114442	0.1713681	0.7830841	0.7830841
6	0.0425999	29.7639654	0.1713681	29.9779335	29.9779335

Degree	BiasSq	Variance	Irreducible	Total	MSE

(ii) Plots:



(iii) The Irreducible error is constant which makes sense, however the other curves look really similar to each other. The variance increases a lot when the degree=6. The Total and MSE are equal.

Problem 5

Task a

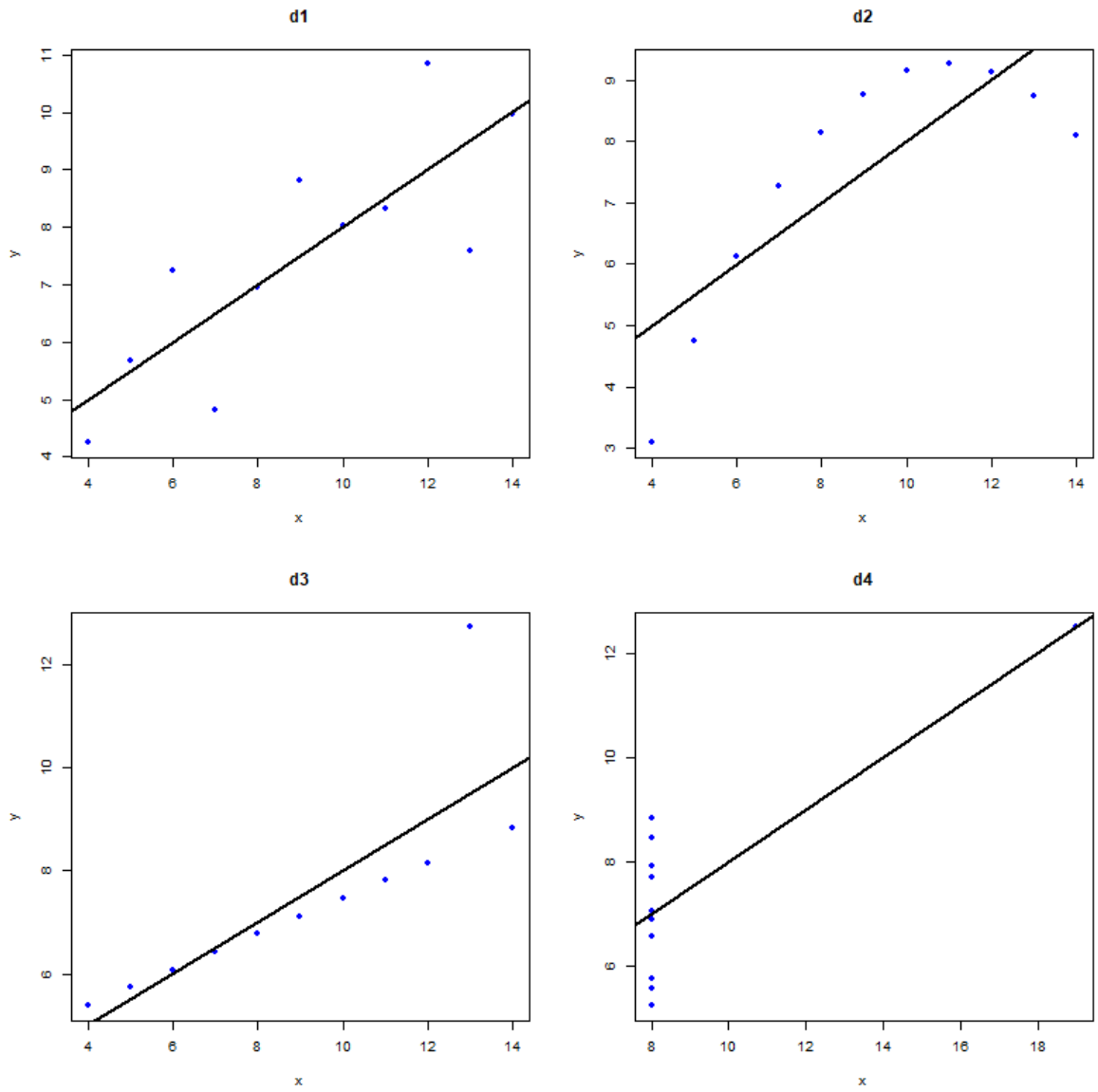
Summary data of fitted models:

intercept	slope_term	std_error_int	std_error_slope	p_value	r_squared	data_name
3.000091	0.5000909	1.124747	0.1179055	0.0021696	0.6665425	d1
3.000909	0.5000000	1.125302	0.1179637	0.0021788	0.6662420	d2
3.002454	0.4997273	1.124481	0.1178777	0.0021763	0.6663240	d3
3.001727	0.4999091	1.123921	0.1178189	0.0021646	0.6667073	d4

The slope terms suggest that when x increases 1 unit, y will increase around 0.5 units in each case.

Task b

None of the models fit the data very well. The line in d1 looks like it is close to the mean of the observations though.

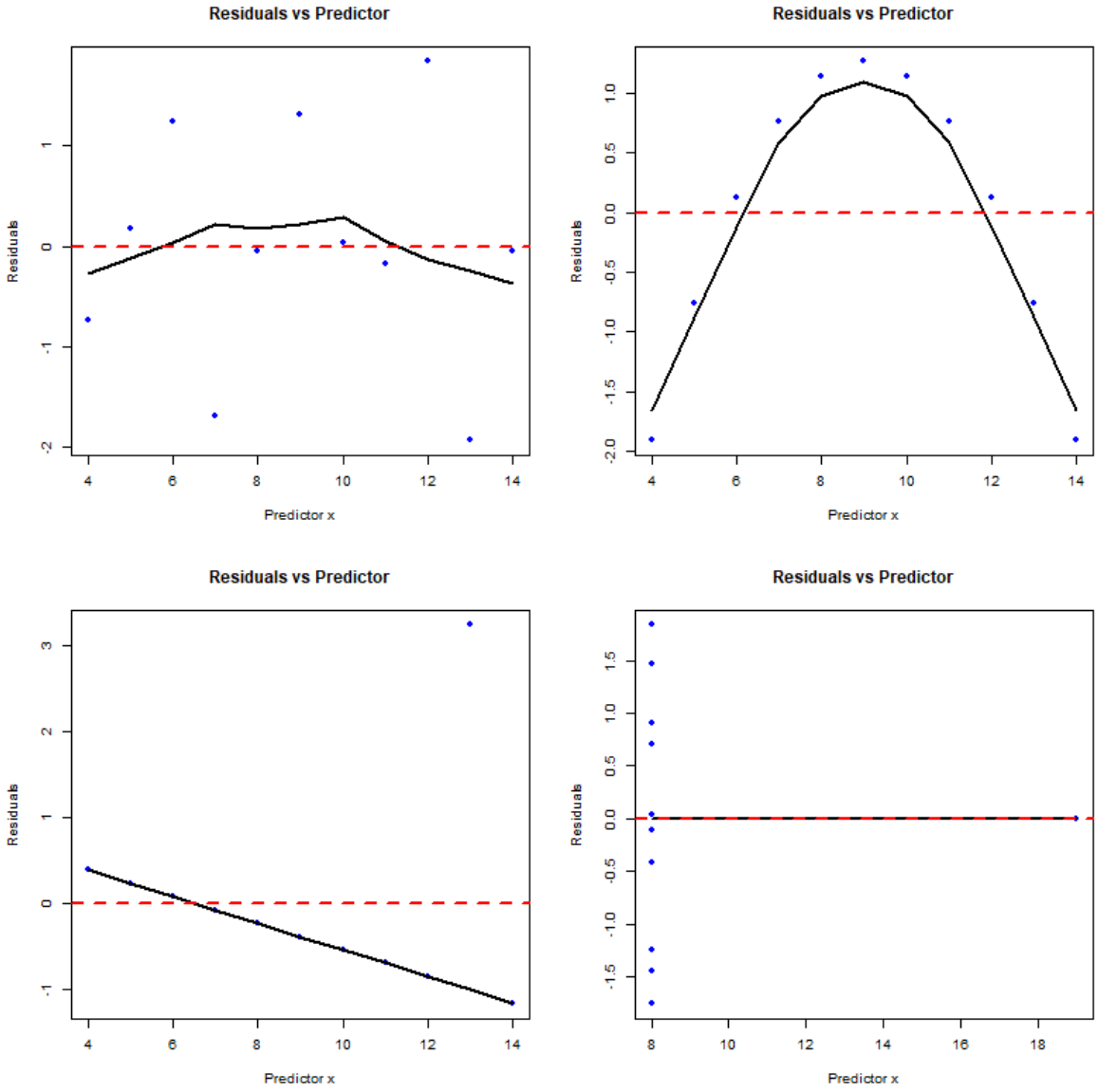


Task c

Non-linearity of the response-predictor relationships applies to d1 and d2. D3 and d4 are linear but each has one point that is outside the normal range. For d3 this looks like an outlier whereas for d4 it could be a high leverage point.

Plotting residuals is a way to identify the linearity/non-linearity of the data. Below are plots of the residuals vs predictor x for each data set. D1 implies linearity because there is little

indication of a pattern in the residuals. D2 however has a U-shape which is a strong indication of non-linearity. The decreasing line in D3 may be due to the outlier in the data. The values of x in d4 are mostly constant and so not much can be said about linearity.



Problem 6

Task a

```
[1] "Bootstrap standard errors:"
```

```
[1] 1.5441767 0.1636514
```

```
[1] "Original values:"
```

```
(Intercept)           x
   3.000909    0.500000
```

After using bootstrap with `R=5000` re-samples, the `std. error` of the result shows that the intercept `t1` is clearly more unstable than the slope `t2`. This means that for the intercept `t1`, the std. error varies `1.5441767` across re-samples when the slope `t2` only varies `0.1636514`. This is also confirmed by the biases which are `t1=0.131211434` and `t2=-0.001837802`. A smaller bias indicates that the original value is a good estimate.

Task b

The bootstrap algorithm compares the results of a statistic using n re-samples of the original data. In the above exercise it re-samples from `d2.csv` and fits a linear model for each re-sample. It then measures how accurate the original statistic (intercept and slope in this case) was based on the results it gets for the re-sampled data. So the bootstrap is computing the `standard deviation` of the statistic.

Task c

The probability of an observation to end up in the bootstrap sample is always constant because of replacement. So for a population of n sampling n data points the probability of the jth observation to not be in the sample is $P = (n-1/n)^n$. So if $n = 10000$ then $P = \left(\frac{9999}{10000}\right)^{10000} \approx 0.368$.