Statistical Learning (5454) - Assignment 2

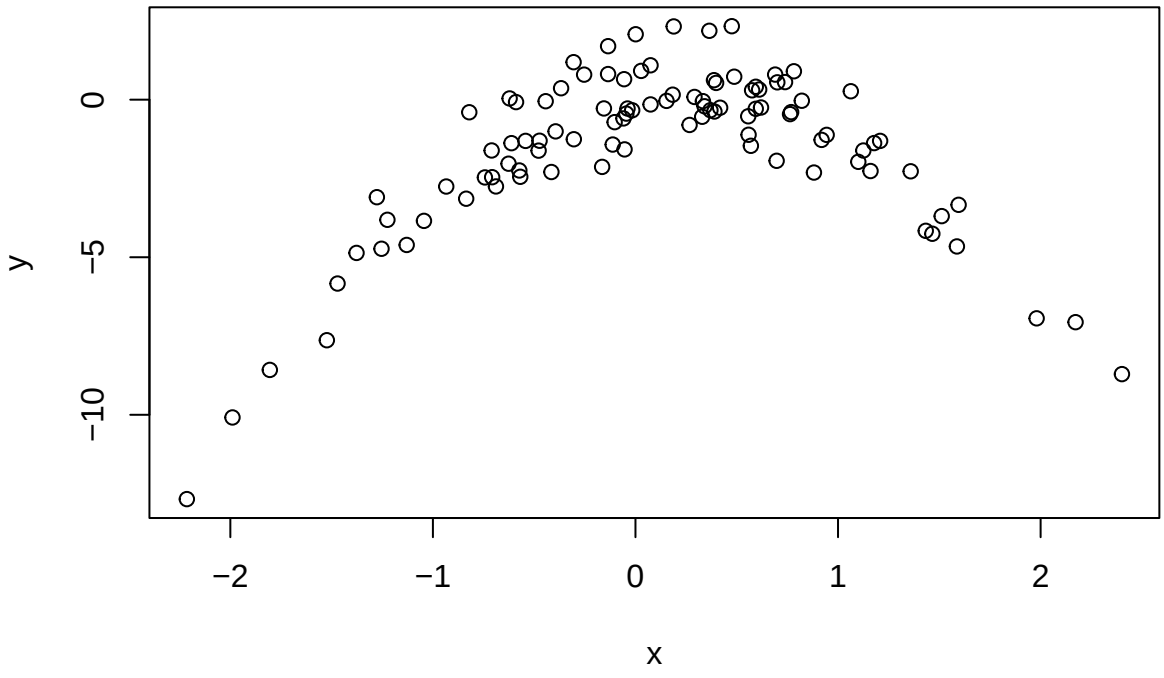
Matthias Hochholzer, Lukas Pirnbacher, Anne Valder

Due: 2024-04-22

Exercise 1

After generating the simulated data we fit four models using least squares, we calculate the AIC values and determine the in-sample error estimates by drawing suitable test data from the data generating process using twice the negative log-likelihood as loss function. Plotting the data we see a clearly non-linear relationship. In line with this we observe that the most preferable model in terms of AIC value is the second model. In terms of negative log-likelihood the fourth model appears to be the best. However, closely followed by model 2. The linear model (model 1) performs the worst out of all models measured by AIC and the negative log-likelihood.

```
#(a)
set.seed(1)
x <- rnorm(100)
y <- x - 2*x^2 + rnorm(100)
plot(x,y)
```



```r
#(b)
lm1 <- lm(y ~ x) # alternative: lm1 <- glm(y ~ x, data, family = gaussian())
lm2 <- lm(y ~ x +I(x^2))
lm3 <- lm(y ~ x + I(x^2) + I(x^3))
lm4 <- lm(y ~ x + I(x^2) + I(x^3) + I(x^4))
```

```
## [1] 478.8804 280.1670 282.0886 282.2963
```

```
## [[1]]
## 'log Lik.' 526.7693 (df=3)
##
## [[2]]
## 'log Lik.' 279.7447 (df=4)
##
## [[3]]
## 'log Lik.' 522.8529 (df=4)
##
## [[4]]
## 'log Lik.' 278.7059 (df=6)
```

Exercise 2

In exercise 2, we perform leave-one-out cross-validation (LOOCV), k-fold cross-validation (kCV) and empirical bootstrapping based on the mean squared error loss that result from fitting the four models using least squares. The simulated data and specifications are the same as in task 1. The results of task (b) across the models are shown in table 1, incidcated by "seed1" in the last column.

(c) Next, we repeat (b) using another random seed (indicated by "seed2" in table 1). For LOOCV we obtain exactly the same results for all four models regardless of the different seed. This is because in general the randomness lies in the data generation.This consistency is expected because LOOCV is deterministic in this context, not relying on random sampling. Each observation is used once as a test set while the rest are used for training, and this process is repeated for each observation in the dataset. Therefore, changing the seed does not affect LOOCV results. For kCV we observe slight differences in the errors between seed 1 and seed 2 across the models. This variation can be attributed to the random partitioning of the data into k folds. Each seed leads to a different random split, which can result in slight variations in the training and validation sets used in each fold, thus affecting the error estimates. Similar to kCV, the bootstrap error shows variations between the two seeds. Bootstrap resampling involves drawing samples with replacement from the original data set to create "new" data sets. The randomness introduced by the seed affects which observations are selected in each resample, leading to slight differences in the bootstrap error estimates between seed 1 and seed 2.

Table 1: Comparison of LOOCV, kCV, and Bootstrap Errors

	LOOCV	kCV	Bootstrap	Model	Seed
1	7.2882	6.1856	6.2931	Model 1	Seed1
5	7.2882	6.4165	6.2764	Model 1	Seed2
2	0.9374	0.9230	0.8702	Model 2	Seed1
6	0.9374	0.9012	0.8704	Model 2	Seed2
3	0.9566	0.9316	0.8554	Model 3	Seed1
7	0.9566	0.8854	0.8557	Model 3	Seed2
4	0.9539	0.9247	0.8393	Model 4	Seed1
8	0.9539	0.8962	0.8452	Model 4	Seed2

(d) The MSEs in table 1 suggest that according to LOOCV the second model is the best. With kCV the third model is the best and for the empirical bootstrap error it is the fourth model. Also, we see that the empirical bootstrap error decreases further with model complexity, reflecting potential overfitting problems of this method. Remembering the plotted data in the beginning these results are in line with our expectations since higher order regression equations fit much better to the data than the linear case.

(e) Last, we have a look at the statistical significance of the coefficient estimates that result from fitting each of the models using least squares. The results here align with our previous conclusions, as the quadratic term has the lowest p-value. \begin{table}

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.625427	0.2619366	-6.205420	0.0000000
x	0.692497	0.2909418	2.380191	0.0192385

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.056715	0.1176555	0.482043	0.6308613
x	1.017161	0.1079827	9.419666	0.0000000
I(x^2)	-2.118921	0.0847657	-24.997388	0.0000000

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0615072	0.1195037	0.5146883	0.6079538
x	0.9752803	0.1872815	5.2075636	0.0000011
I(x^2)	-2.1237910	0.0870025	-24.4106856	0.0000000
I(x^3)	0.0176386	0.0642904	0.2743580	0.7843990

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1567030	0.1394619	1.1236253	0.2640034
x	1.0308256	0.1913365	5.3874999	0.0000005
I(x^2)	-2.4098982	0.2348551	-10.2612148	0.0000000
I(x^3)	-0.0091329	0.0672288	-0.1358481	0.8922288
I(x^4)	0.0697854	0.0532401	1.3107691	0.1930956

\end{table}

Exercise 3

Exercise 4

Exercise 5