Exercise set 1

Linear regression and k-nearest neighbor (KNN) prediction, one predictor

- In simple linear regression the prediction of Y for a given $X = x_0$ is

$$\hat{f}(x) = \hat{\alpha}_0 + \hat{\alpha}_1 x_0$$

where

$$\hat{\alpha}_1 = \frac{n\sum_{i=1}^n x_i y_i - [\sum_{i=1}^n y_i]\,[\sum_{i=1}^n x_i]}{n\sum_{i=1}^n x_i^2 - [\sum_{i=1}^n x_i]^2}$$

and

$$\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \bar{x}.$$

- In *KNN* with one predictor the prediction of Y for a given $X = x_0$ is the average of the k y_i's which have the smallest corresponding absolute difference $|x_i - x_0|$.

Consider the data set

	x	y
1	1.00	1.88
2	2.00	4.54
3	3.00	10.12
4	4.00	9.14
5	5.00	11.26

Task 1

a) What is the prediction for $x = 3$ if the linear regression model is used? Use a calculator (or use R as a calculator) to answer this question.

b) Use the `lm`-function in `R` to fit the same model as in a). Use the `predict`-function to compute the prediction of $x = 3$.

c) What is the prediction for $x = 3$ if KNN with $K = 3$ is used? For $K = 1$ and $K = 5$? Use a calculator (or use R as a calculator) to answer this question.

d) Consider the R-function

```
knn=function(x0,x,y,K=20)
{
  d=abs(x-x0)
  o=order(d)[1:K]
  ypred=mean(y[o])
  return(ypred)
}
```

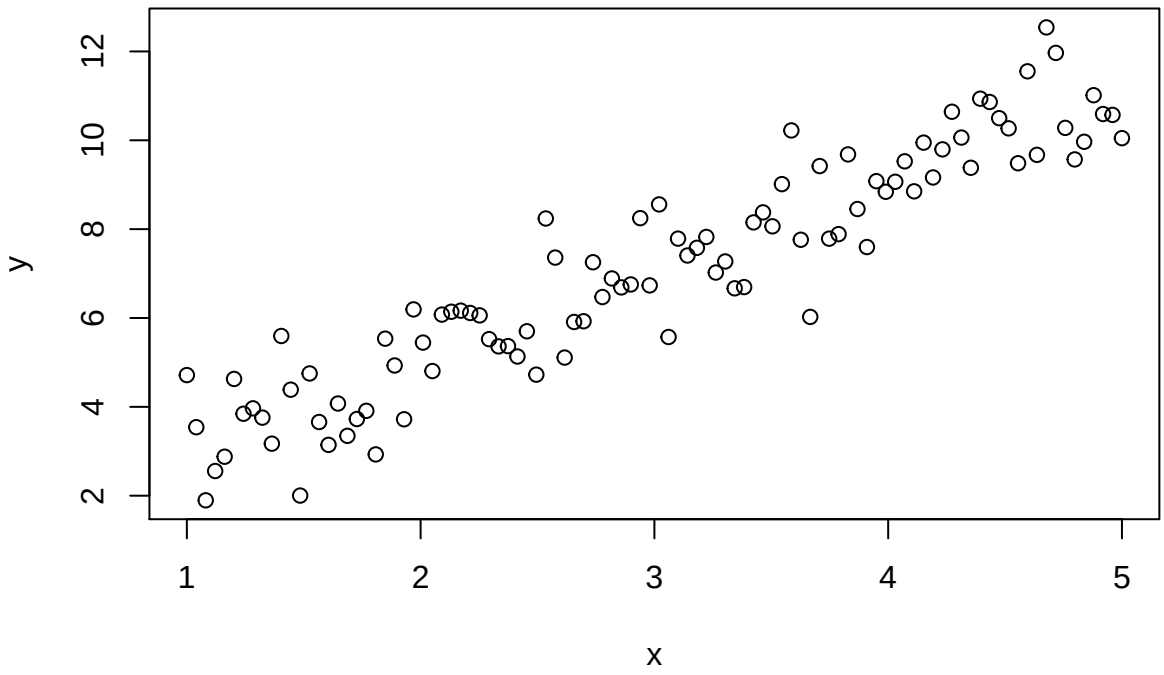
Experiment with `R` to see and explain what happens in each row of the function if $x_0 = 3$ and $K = 3$. Start by

```
K=3
x0=3
x=1:5
```

```r
y=c(1.88,4.54,10.12,9.14,11.26)
d=abs(x-x0)
d
```

```
## [1] 2 1 0 1 2
```

e) Which of the 4 methods, linear regression or KNN with K=1,2 or 3, would you use if you had seen the following plot? Why?



Test and training data

- *Training data* is data which is used to fit the model.
- *Test data* is data which is used to test the predictions of the model (and not used to fit the model).

Task 2

a) Load, summarize, and read the help to the dataset College from the R-package ISLR by

```r
library(ISLR)
## summary(College)
## help(College)
```

b) Divide the data into a 50/50 training and test data by

```r
set.seed(123)
n=nrow(College)
train_indicator=sample(1:n,size=floor(n/2))
```

```
train=College[train_indicator,]
test=College[-train_indicator,]
```

c) Fit a linear prediction model, on the training data, for number of applications, Apps, with the predictors Private and Accept by

```
m1=lm(Apps~Private+Accept,data=train)
summary(m1)
```

d) Compute training-MSE by

```
pred_train=predict(m1)
mse_train=mean((train$Apps-pred_train)^2)
mse_train
```

e) Now, redo c) and d) but by only using Accept as the predictor. Fit the model and compute the training-MSE. Is it larger or smaller?

f) Finally compute the test-MSE for the two regression models fitted in c) and e). This is how you do it for the one in c)

```
pred_test=predict(m1,newdata = test)
mse_test=mean((test$Apps-pred_test)^2)
```

g) Comment on the relationships between the MSE's between the models in trainingset and testset respectively.

h) Use KNN to predict Apps. Compare the test-MSE with the linear regression predictions.

Task 3

Do Exercise 3.10a-f in the book and, also, do the extra task

g*) Compare the predictions by splitting the data into a training and a test part, 50% in each.