Mandatory Assignment 1

STK2100

Håkon Berggren Olsen

Spring 2020

Problem 1.

The dataset, nuclear, contains data regarding the construction of 32 Light-Water Reactors built between 1967 and 1971. The scope of the dataset is to predict the cost of construction for further span given the dataset within this time interval.

The dataset has 32 rows and 11 columns, and has a mix of continous- and categorical variabels. Contionous data such as regarding the cost, construction and net capacity of the a given nuclear powerplant, and categorical such as cooling tower present or absent, if there exists an LWR (Light-Water Reactor) plant at the same site and such.

The description of all the variables are shown in a table below.

Name	**Continous Variables**
cost:	The capital cost of construction in millions of dollars adjusted to 1976 base.
date:	The date on which the construction permit was issued. The data are measured in years since January 1 1990 to the nearest month.
t1:	The time between application for and issue of the construction permit.
t2:	The time between issue of operating license and construction permit.
cap:	The net capacity of the power plant (MWe).
cum.n:	The cumulative number of power plants constructed by each architect-engineer.
Name	**Categorical Variables**
pr:	A binary variable where 1 indicates the prior existence of a LWR plant at the same site.
ne:	A binary variable where 1 indicates that the plant was constructed in the north-east region of the U.S.A.
ct:	A binary variable where 1 indicates the use of a cooling tower in the plant.
bw:	A binary variable where 1 indicates that the nuclear steam supply system was manufactured by Babcock-Wilcox.
pt:	A binary variable where 1 indicates those plants with partial turnkey guarantees.

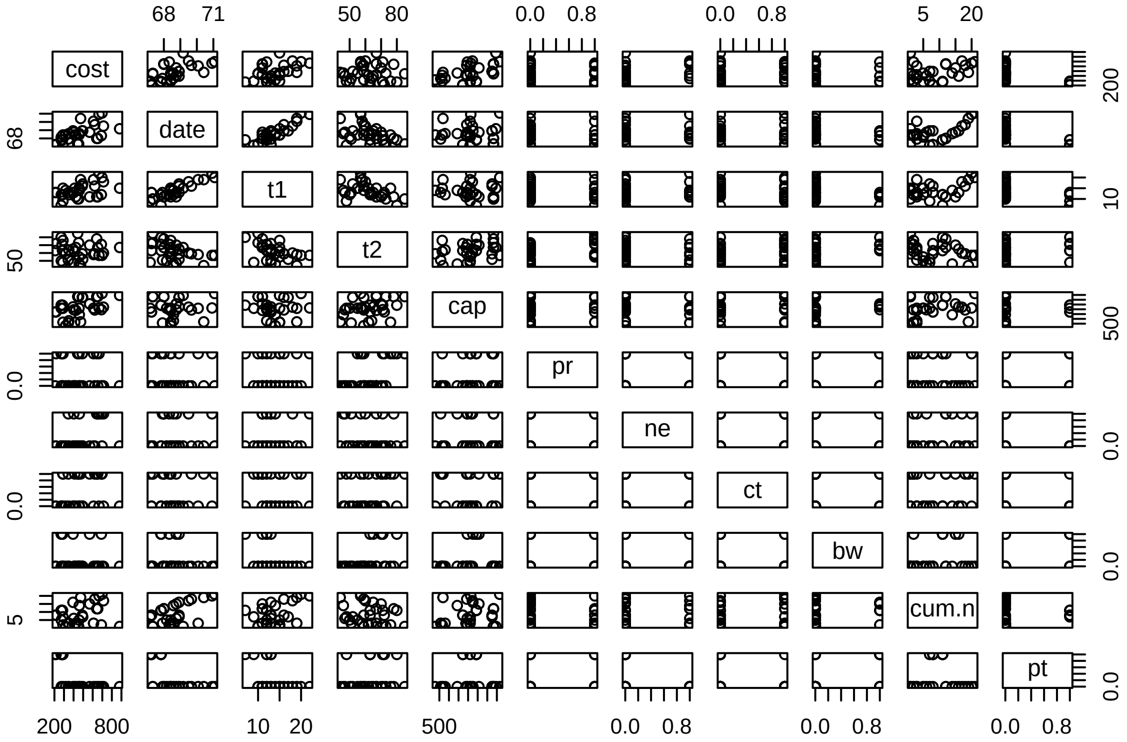
a)

Loading the dataset

```r
datadir = "http://www.uio.no/studier/emner/matnat/math/STK2100/data/"
nuclear = read.table(paste(datadir, "nuclear.dat", sep=""), header=T)
```

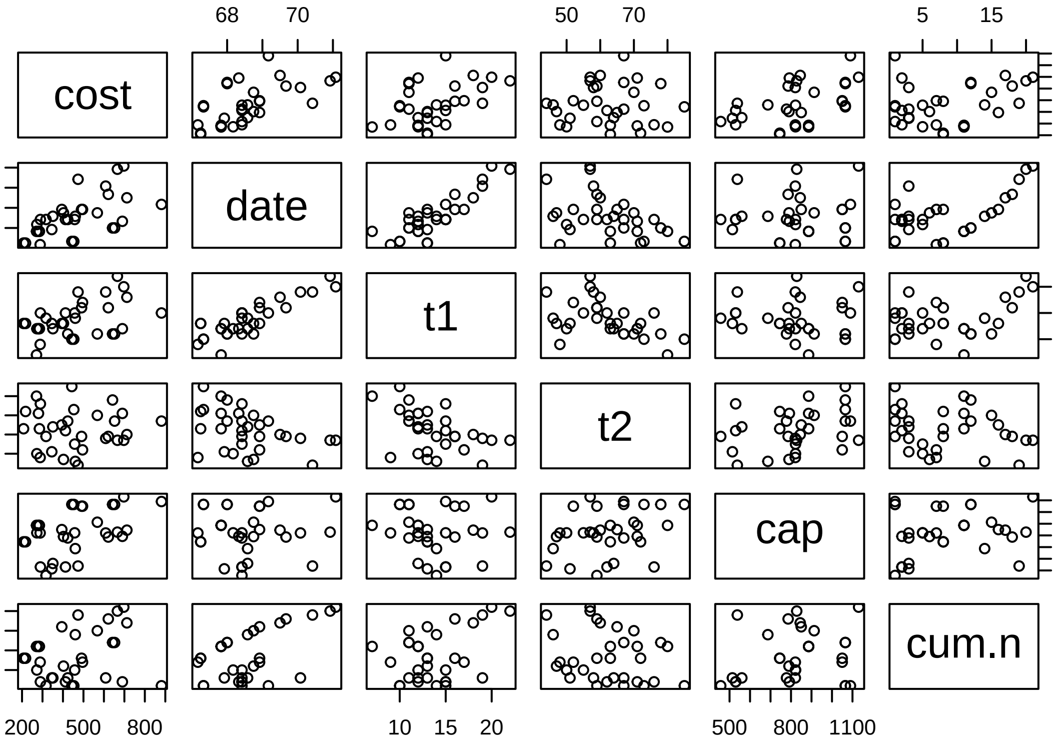
Generating some plots to gain an intuition for the dataset.

```
plot(nuclear)
```



This is very cluttered and hard to grasp, much due to categorical variables not giving any immeadiate insight. Therefore we will plot the dataset only regarding the continous variables.

```
cont_var <- c("cost", "date", "t1", "t2", "cap", "cum.n")
cat_var <- c("pr", "ne", "ct", "bw", "pt")
plot(nuclear[cont_var])
```



\

From this figure we can see some clear trends. We can see the cost of the LWRs are increasing over time, which can be attributed to inflation. A very positive correlation is also spotted in the t1 vs date indicating that the time between application and issue of the construction permits is steadily increasing and almost doubles at the end of the dataset. t2 vs date shows however a negative correlation, indicating that it takes shorter and shorter time from obtaining operating license to gaining construction permit.

The net capacity of the LWR, cap, does show a slight increase with cost. This is expected as the capital cost of the investment should match it's returns.

NOTE: still not sure how the cum.n "The cumulative number of power plants constructed by each architect-engineer." — what architect-engineer? – is construed.

b) Constructing a model

Look at the model

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{p,i} + \epsilon_i$$

where Y_i is cost at log scale for observation i.

- What are the standard assumptions about the noise term ϵ_i? Discuss also which of these assumptions that are the most important? The assumptions of the noise term ϵ_i is that it is normally distributed.
- Fit this model including all the observations with log(cost) as response and all the other variables as covariates. Discuss the results

```r
#fit with all covariates
model = lm(log(cost)~., data=nuclear)
summary(model)
```

```
##
## Call:
## lm(formula = log(cost) ~ ., data = nuclear)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.284032 -0.081677  0.009502  0.090890  0.266548
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.063e+01  5.710e+00  -1.862  0.07662 .
## date         2.276e-01  8.656e-02   2.629  0.01567 *
## t1           5.252e-03  2.230e-02   0.236  0.81610
## t2           5.606e-03  4.595e-03   1.220  0.23599
## cap          8.837e-04  1.811e-04   4.878 7.99e-05 ***
## pr          -1.081e-01  8.351e-02  -1.295  0.20943
## ne           2.595e-01  7.925e-02   3.274  0.00362 **
## ct           1.155e-01  7.027e-02   1.644  0.11503
## bw           3.680e-02  1.063e-01   0.346  0.73261
## cum.n       -1.203e-02  7.828e-03  -1.536  0.13944
## pt          -2.220e-01  1.304e-01  -1.702  0.10352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1697 on 21 degrees of freedom
## Multiple R-squared:  0.8635, Adjusted R-squared:  0.7985
## F-statistic: 13.28 on 10 and 21 DF,  p-value: 5.717e-07
```