TMA4255 Applied Statistics
Exercise 7

Observe: MINITAB commands in the end of the exercise, and R script from the course-www-page.

A manufacturer produces a specific pigment for use in the textile industry. The company delivers various amounts of the pigment to other manufacturers which make further use of it in their production.

Recently there have been several complaints from costumers who claim that the deliveries of the pigment contain an unacceptable level of impurities. The company management wants to control the production process such that the level of impurities is as low as possible. We will use regression analysis to answer these questions.

The following variables and data in Table 2 have been recorded for the process (and is available as a datafile on the webpage)

Variable name	Explanation
x_1	Weight of a particular ingredient
x_2	Age of catalyzer
x_3	Supply velocity of main ingredient
x_4	Start temperature of main ingredient
x_5	Cylinder velocity
y	Impurity

Table 1: Summary of the variables

a)As a first step of the analysis it is important to get a good overview of the data. Find the descriptive statistics (mean, standard deviation, minimum, maximum, quantiles etc) for the variables. Find the correlation between the variables. Plot various variables against each other. Give a first description of the dataset.

b) Try to fit a model to y with one of the five predictors. Which predictor should be chosen? Explain. Fit the model and check if the assumptions for the model are satisfied. Comment on the residual plots. Does the predictor variable influence the response y?

c) Look at the model with y and x_1. Find a 95% confidence interval for the expectation of y and then a prediction interval for a new observation for $x_1 = 1, 3$ and 5. Explain what these intervals describe.

Row	x_1	x_2	x_3	x_4	x_5	y
1	3	1	3	1	3	4
2	4	2	5	2	3	3
3	6	3	7	3	3	4
4	3	4	5	3	2	6
5	1	5	2	2	2	7
6	5	6	6	1	2	2
7	1	7	2	2	2	6
8	0	8	1	3	4	10
9	2	9	3	1	4	5
10	5	10	6	2	4	3
11	4	1	3	2	4	3
12	0	2	7	1	4	4
13	4	3	1	3	4	4
14	2	4	5	3	2	6
15	2	5	7	1	2	4
16	0	6	5	2	3	7
17	6	7	1	3	3	2
18	6	8	3	1	3	2

Table 2: Data for the process

d) Now try to fit a model between y and all the five predictors. Look at the variance-covariance matrix of the estimated regression coefficients and comment. Which of the five independent variables have a significant effect at 5% level?

e) Use best subset regression to find the best model. Comment.

f) (Optional) It can often be useful to include cross terms or second order terms in a regression analysis. In this situation we suspect that cross terms that contain x_1 could be relevant for the model. Define four new variables by setting $x_{12} = x_1 \cdot x_2$,$x_{13} = x_1 \cdot x_3$, $x_{14} = x_1 \cdot x_4$ and $x_{15} = x_1 \cdot x_5$. Try best subset regression again when you include these four predictors. What is the best model?

MINITAB

a)

Stat → Basic Statistics → Display Descriptive Statistics
Stat → Basic Statistics → Correlation
Graph → Matrix plot

b)

Stat → Regression→ Regression: Graphs x Residuals vs. fits
(x Normal plot of residuals)
Graph → Prob.plot: Options: Include confidence intervals in plot (Use this for the residuals)

c)

Stat → Regression→ Fitted Line Plot: Options: x Display confidence bands
x Display prediction bands
Stat → Regression → Options: Prediction intervals for new observations
(Do this for the column of new observations)

d)

Stat → Regression→ Regression: Storage: x X'X inverse
Display Data: (Choose the matrix of interest)

e)

Stat → Regression → Best Subsets

f)

Calc → Calculator: C1*C2 etc