Final MSG500/MVE190 January 14, 2016

Examiner: Rebecka Jörnsten, 0760-491949.
REMINDER: You need to complete the mandatory project to pass this course. Submit to the course administrator Rebecka Jörnsten (jornsten@chalmers.se).

Open book, open notes!
Motivate your answers! Avoid simple yes/no answers - provide explanations.

Question 1 a-d: 10p

Below I provide a scatter plot of a simulated data example. In subproblem a-d, please indicate what you think the fitted model would look like.
a) A least-squares fit of a first order polynomial model.
b) A least-squares fit of a second order polynomial model.
c) A regression-tree model with two terminal nodes (leaves).
d) A regression-tree model with three terminal nodes.

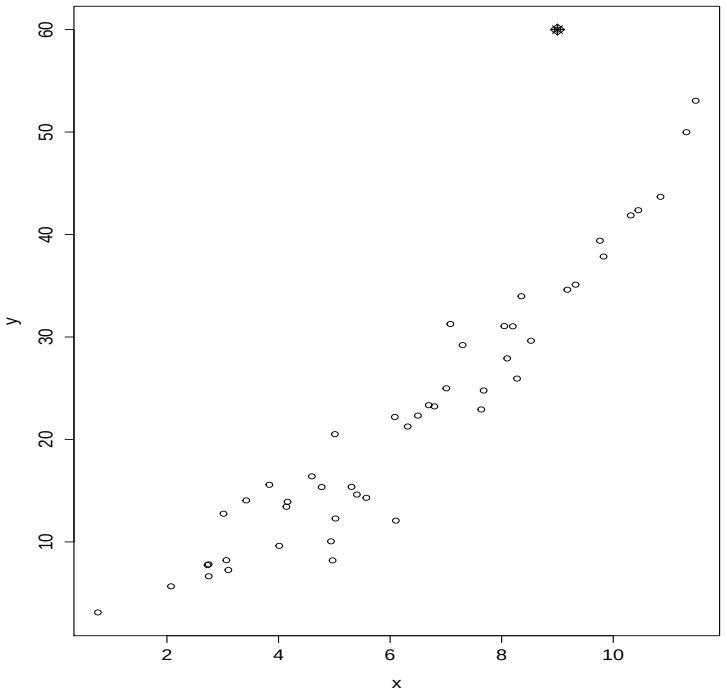


Figure 1: Question 1: scatter plot

Question 2 a-d: 10p

Repeat question 1, discussing how your answer changes as the y-position of the highlighted observation at $(x, y) = (9, 60)$ moves from 0 to 60. Does this x-position have high leverage or low leverage? Which model (a-d) is most affected by the changing y-position of this observation and why?

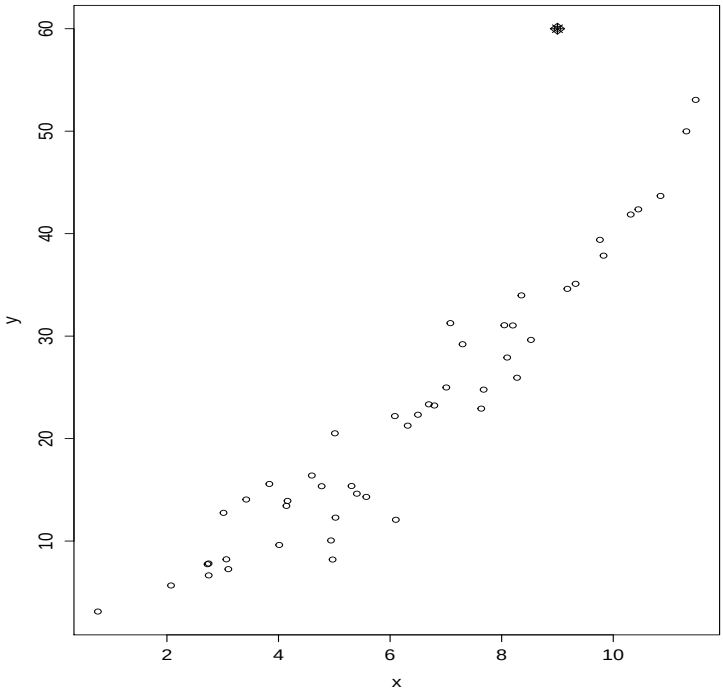


Figure 2: Question 2: scatter plot

Question 3: 10p

Regarding the data set in Question 1: Discuss the predictive performance of each of the models (consider both bias and variance) in each of the following cases:
a) Predicting at x=5
b) Prediction at x=9
c) Prediction at x=15
d) Prediction at x=0
e) If you employ cross-validation for model selection, which of the 4 models do you think would be selected and why?

Question 4: 20p

Consider a regression data set with 6 covariates $(x_j, j = 1, \cdots, 6)$ and continuous outcome y. Furthermore, we have $Cor(x_i, x_j) > 0.95, i, j = 1, 2, 3$ and $Cor(x_k, x_l) < 0.2, k = 1, \cdots, 6, l = 4, \cdots, 6$. Consider case I: $n = 25$ observations and case II: $n = 250$ observations.
In each of cases I and II, discuss how you would proceed to
a) construct a model with optimal predictive performance
b) best identify the underlying true model for y.
Discuss how you think sample size affects the problem.

Question 5: 20p

A company wants to optimize the packing of its products. The packing machine consists of a moving conveyor belt where boxes pass underneath a number of drop-points for finished products. One can alter the speed of the conveyor belt, and the number of products released at each drop-point. This process is not entirely precise so there is some randomness as to the exact number of products released each time, perhaps different precision for different speeds of release and/or quantities released. In addition, at higher speeds there is some risk that products will not drop into the box underneath. Therefore, for quality control, boxes that contain less than 90% of correct number of products or more than 110% are removed by weight control of the boxes at the end of the conveyor belt.

The company investigates how to optimize the settings in terms of three different speed settings for the conveyor belt, and two different drop quantities for each drop-point (set to not release any more product at later drop-points if the total number of released units exceeds 120% of the goal quantity).

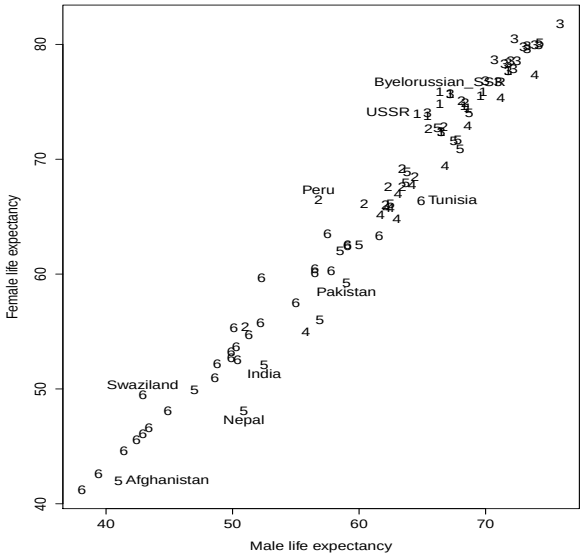
The end-point (result measured) is the time it takes to obtain 100 boxes that pass quality control.

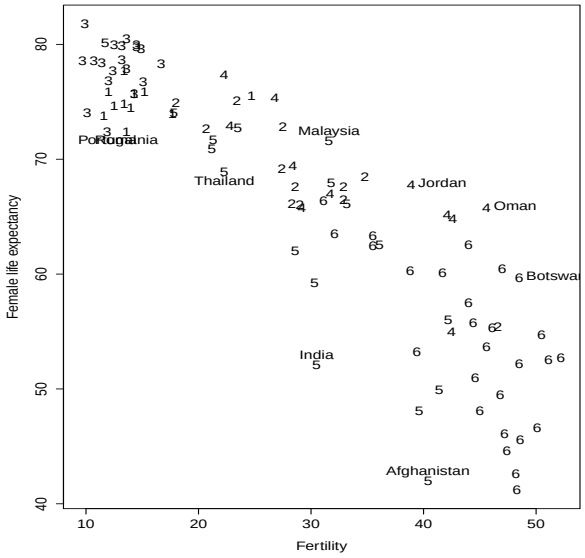
Discuss how you would set up the experiment, the data you would collect and state the model you would use to determine the optimal setting, how you would fit this model and use it to determine the optimal setting (I expect you to define variables, write a model equation and say what the parameters in your model mean).

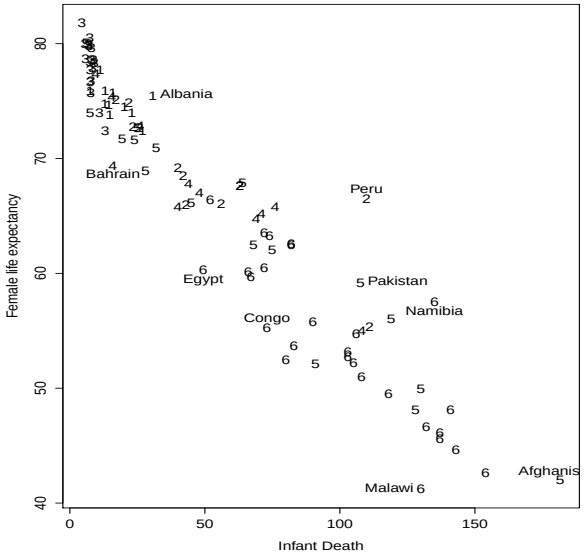
Question 6 a-f: 30p (5+5+5+5+5+5)

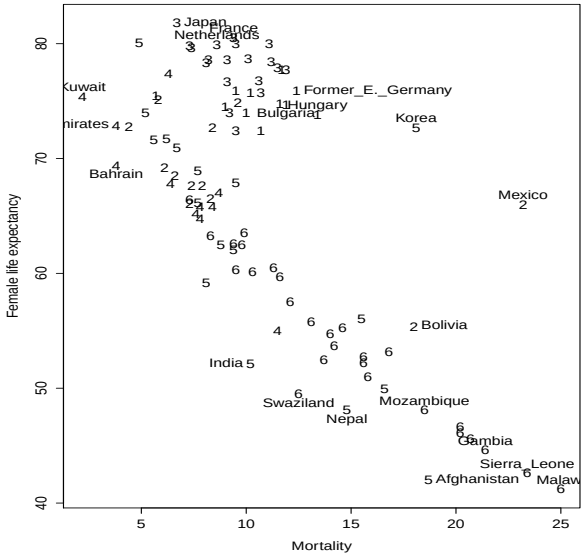
Statistics from 91 countries are collected. Variables include mortality rate per 1000 (Mort), birthrate per 1000 (Fertility), infant deaths per 1000 in population under 1 years old (InfDeath), life expectancy of males and females (LifeM, LifeF) and gross national product (GNP). In addition, countries are grouped into 6 groups; 1 = Eastern Europe, 2 = South America and Mexico, 3 = Western Europe, North America, Japan, Australia, New Zealand, 4 = Middle East, 5 = Asia, 6 = Africa. We will consider female life expectancy to be our outcome variable. We are interested in finding predictive factors that influence female life expectancy.

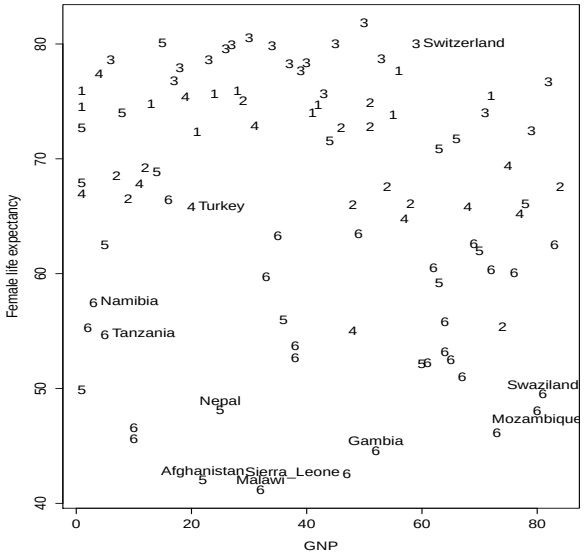
a) The data is summarized in 6 scatter plots on the next page. Discuss the plots: do the basic assumptions for regression modeling via least squares hold? Concerns? What would your preliminary steps be prior to modeling? Explain why.

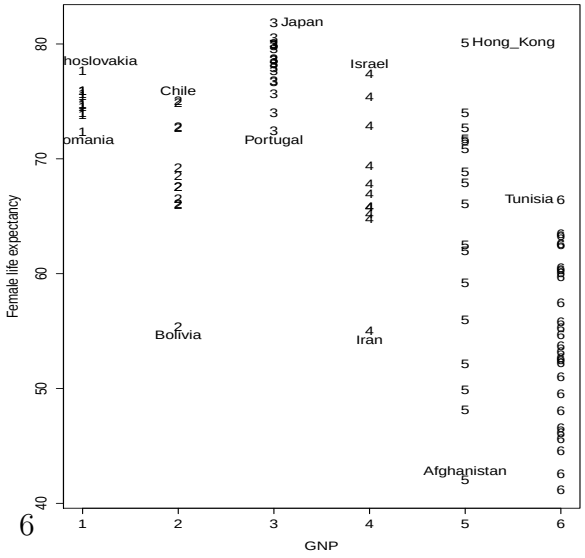

Female life expectancy
Male life expectancy
Byelorussian_SSR
USSR
Peru
Tunisia
Pakistan
India
Swaziland
Nepal
Afghanistan


Female life expectancy
Fertility
Malaysia
Thailand
Jordan
Oman
India
Afghanistan


Female life expectancy
Infant Death
Albania
Bahrain
Peru
Egypt
Pakistan
Namibia
Congo
Malawi


Female life expectancy
Mortality
Japan
Netherlands
Former_E._Germany
Hungary
Bulgaria
Korea
Bahrain
Mexico
Bolivia
India
Swaziland
Mozambique
Nepal
Gambia
Sierra_Leone
Afghanistan


Female life expectancy
GNP
Switzerland
Turkey
Namibia
Tanzania
Swaziland
Nepal
Mozambique
Gambia
Afghanistan
Sierra_Leone
Malawi


Female life expectancy
GNP
Japan
Hong_Kong
Israel
Chile
Portugal
Tunisia
Bolivia
Iran
Afghanistan

b) I proceed with modeling without taking any additional steps (but I'm not saying that is optimal - this is simply because of the exam setting leaving part a) for you to discuss....). I summarize the fit with the table below and the diagnostic plot in Figure 3. Discuss and interpret the results. What have we learnt about factors that influence female life expectancy? Any limitations/concerns? Any particular information missing you think is relevant for the discussion?

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       24.987488   4.575020   5.462 4.41e-07 ***
Fert              -0.066132   0.033515  -1.973  0.05164 .
Mort              -0.075795   0.054926  -1.380  0.17114
InfDeath          -0.044620   0.009787  -4.559 1.67e-05 ***
LifeM              0.776396   0.058938  13.173  < 2e-16 ***
as.factor(Group)2 -0.207352   0.707762  -0.293  0.77024
as.factor(Group)3 -0.405374   0.542161  -0.748  0.45666
as.factor(Group)4 -2.571451   0.847751  -3.033  0.00319 **
as.factor(Group)5 -2.918847   0.678447  -4.302 4.41e-05 ***
as.factor(Group)6 -1.656934   0.876791  -1.890  0.06212 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.391 on 87 degrees of freedom
Multiple R-squared:  0.9855,Adjusted R-squared:  0.984
F-statistic: 658.1 on 9 and 87 DF,  p-value: < 2.2e-16

Correlation matrix:
               Fert        Mort   InfDeath      LifeM      LifeF         GNP
Fert      1.0000000  0.48619655  0.8583534 -0.8665189 -0.8944140  0.16081216
Mort      0.4861966  1.00000000  0.6546232 -0.7334666 -0.6930331  0.01092786
InfDeath  0.8583534  0.65462322  1.0000000 -0.9368384 -0.9553516  0.10214488
LifeM    -0.8665189 -0.73346661 -0.9368384  1.0000000  0.9825578 -0.14217731
LifeF    -0.8944140 -0.69303311 -0.9553516  0.9825578  1.0000000 -0.15060940
GNP       0.1608122  0.01092786  0.1021449 -0.1421773 -0.1506094  1.00000000
```

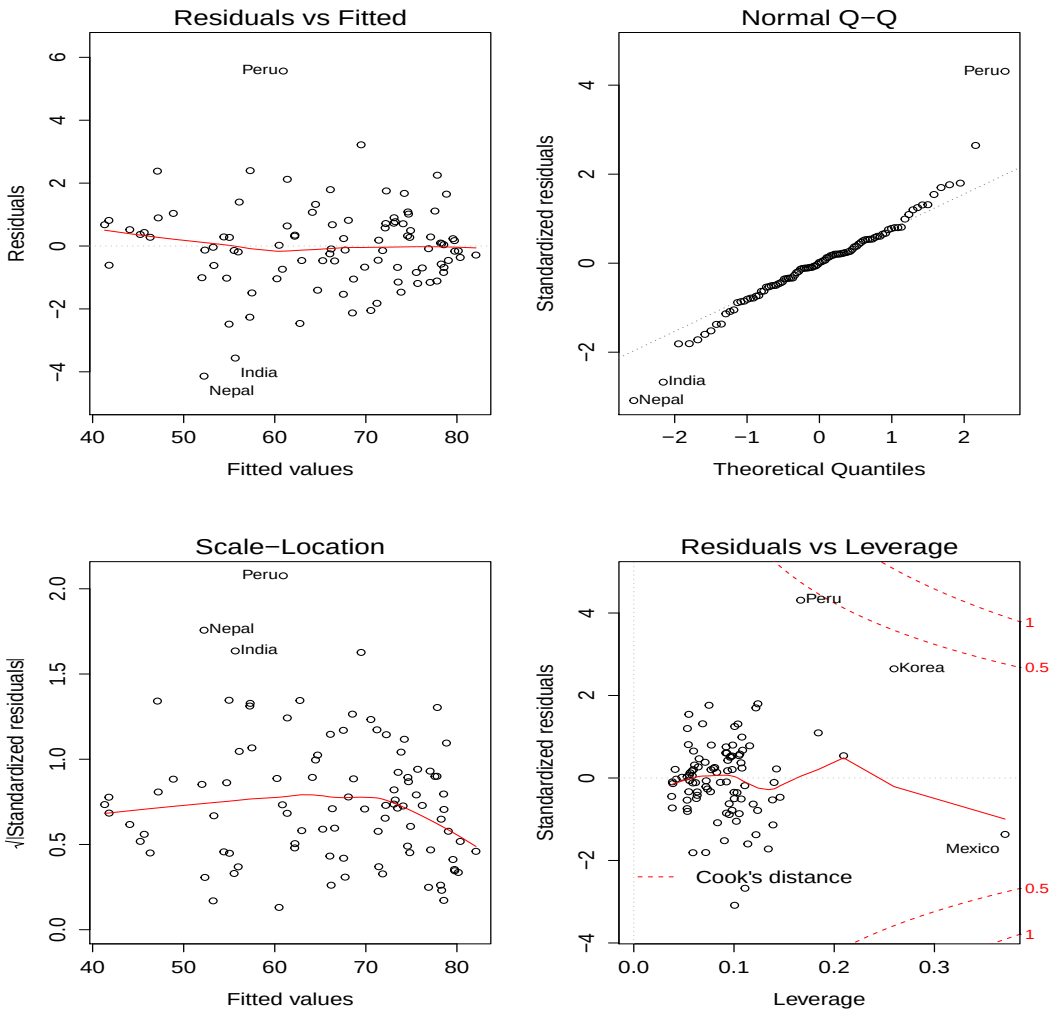


Figure 3: Question6: Diagnostic plot

c) I perform model selection used all-subset comparisons and Cp, AIC and BIC. The results are presented below. Does these results align with the model summary in b) above? Any surprises? Discuss and compare.

```
            CP  AIC BIC
Fert         T   T   F
Mort         T   T   F
InfDeath     T   T   T
LifeM        T   T   T
GNP          T   T   F
gr2          F   F   F
gr3          F   F   F
gr4          T   T   T
gr5          T   T   T
gr6          T   T   T
```

d) I perform random splits, holding out 25% of the countries for testing. The following results are obtained, with the numbers in the table stating percentages of bootstrap models that contains each variable in question. Comment and interpret the results.

```
                modselcp modselaic modselbic
 [1,] "Fert"     "71"     "74"      "51"
 [2,] "Mort"     "57"     "63"      "27"
 [3,] "InfDeath" "100"    "100"     "100"
 [4,] "LifeM"    "100"    "100"     "100"
 [5,] "GNP"      "79"     "82"      "40"
 [6,] "gr2"      "7"      "8"       "6"
 [7,] "gr3"      "1"      "2"       "0"
 [8,] "gr4"      "99"     "99"      "91"
 [9,] "gr5"      "100"    "100"     "100"
[10,] "gr6"      "83"     "83"      "68"
[1] "mean PE for cp, aic and bic"
[1] "PEcp="    "2.6219"   " PEaic="  "2.624"    " PEbicK=" "2.7182"
[1] "mean model size for cp, aic and bic"
[1] "sizecp="   "7.97"      " sizeaic=" "8.11"      " sizebic=" "6.83"
```

e) I repeat the analysis, holding out 50% of the data for testing. The results below are obtained. Comment, interpret and compare with part d).

```
                modselcp modselaic modselbic
 [1,] "Fert"     "55"     "59"      "49"
 [2,] "Mort"     "43"     "48"      "24"
 [3,] "InfDeath" "95"     "96"      "90"
 [4,] "LifeM"    "100"    "100"     "100"
 [5,] "GNP"      "45"     "48"      "24"
 [6,] "gr2"      "19"     "22"      "17"
 [7,] "gr3"      "8"      "11"      "4"
 [8,] "gr4"      "91"     "92"      "74"
 [9,] "gr5"      "100"    "100"     "96"
[10,] "gr6"      "62"     "63"      "46"
[1] "mean PE for cp, aic and bic"
[1] "PEcp="    "2.73"     " PEaic="  "2.715"    " PEbicK=" "2.8195"
[1] "mean model size for cp, aic and bic"
[1] "sizecp="   "7.18"      " sizeaic=" "7.39"      " sizebic=" "6.24"
```

f) How do you think CART would perform on this data set? Which do you think is the more appropriate model, regression or CART?
Propose at least three more steps that you would undertake at this point of the analysis.