Final MSG500/MVE190 January 14, 2016 - Solution

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.

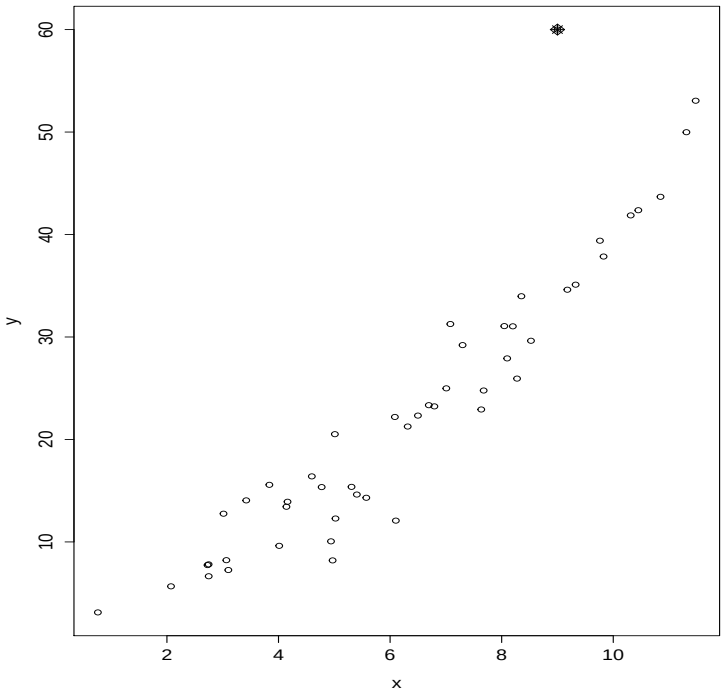


Figure 1: Question 1: scatter plot

Solution:
First thing to realize is that the highlighted observation is of moderate leverage since there

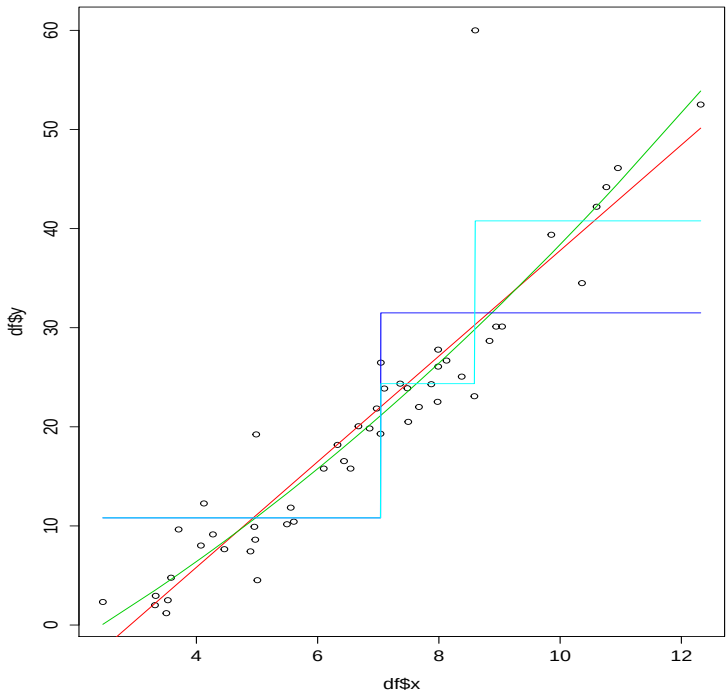


Figure 2: Question 1: scatter plot

are roughly 10/50 observations more extreme in x.
There is a slight hint at a curve-linear relationship but it's quite weak and the noise level is high.
a) The marked observation has limited leverage so the linear (red) model is not much affected.
b) The curve-linear model (green) looks almost like the linear one because the curvature in the data is modest.
c) There are 50 observations spread over x and since we are trying to minimize overall MSE the 2-leaf tree will split near the mean of x and similarly with roughly the mean value near the mid-point in each interval. The "outlier" is not that extreme in y in the upper range of x.
d) The 3-leaf model splits the data to minimize overall MSE. Since the outlier is near the high values of y for the largest x, the split is to the left of the outlier.
I'm looking for a general idea here. I know it's hard to eye-ball this. The main point is that the 2-leaf tree has to minimize overall MSE so therefore it won't split near the outlier but near the middle of the x range.

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?

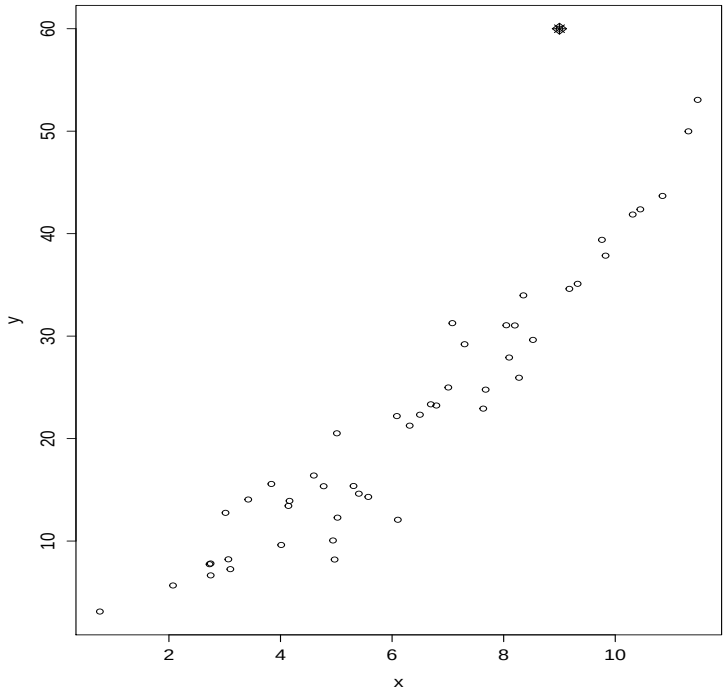


Figure 3: Question 2: scatter plot

Solution As you move the outliers from large y to small y there is a limited impact on a)

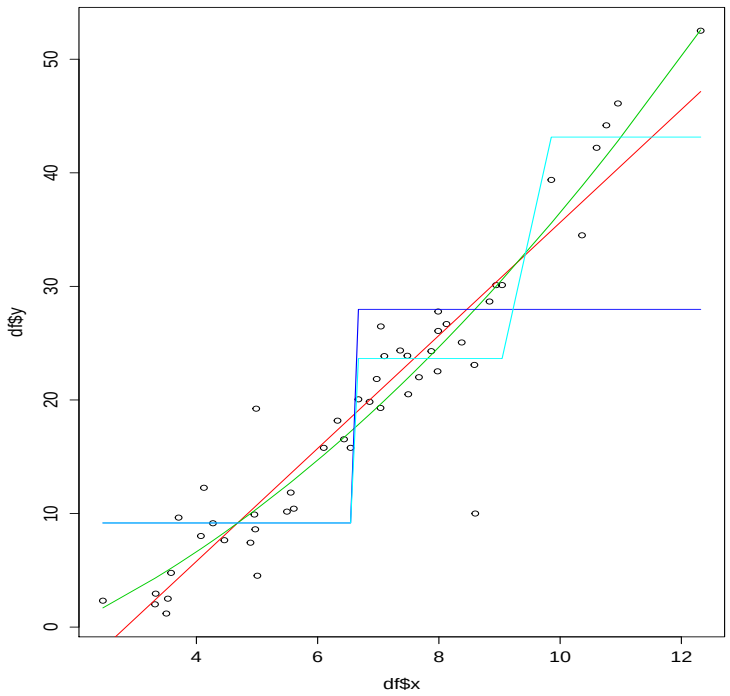


Figure 4: Question 2: scatter plot

and b). That's because the polynomial model are relatively rigid.
The 2-leaf tree is not much affected at all. The split is roughly at the same x and it's only the mean-level of the right-hand side that is affected.
The 3-leaf tree is more sensitive. Still, when the outlier is a large value for y and in the same range as ys above the 2nd split, the 3rd split keeps the outlier with it. Only when the outlier moves to the range that's more like the lower values of y does the 3 leaf model split on the other side of the outlier.
NB: I would accept many variants of answers here because I know this is difficult to do by eye. The rough idea is what I'm after. That is, the linear and polynomial models are not much affected, nor is the 2-leaf model because all three are quite rigid. The 3-leaf model can adapt more to the individual data and change the location of the split.

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?

Solution: To answer this you have to assume a correct model - let's say the quadratic (from eye-balling it this looks like the best bet).
a) Near $x = 5$ there is not much variability of the estimates really. Clearly, a linear or quadratic model fits the data better so would likely be less biased. The tree models round the data off at a mean value of y for observations below roughly the mean of x. That mean value is actually not too far from where the linear model and quadratic model would be at $x = 5$ so all models perform reasonably around $x = 5$.
b) Near $x = 9$ again the linear and quadratic models fit the data well (little or no bias) and since they are more rigid they will not exhibit much variance either. The 2-leaf model approximates y with the mean value of all observations in the upper range of x which probably comes reasonably close to there the linear and quadratic models lie for $x = 9$. The 2-leaf model is quite rigid to the estimate is not high variance. The 3-leaf model is a different story. Around $x = 9$ is where the location of the outlier, small changes to the data can lead to big changes in terms of the estimate - where will the 2nd split be? That means the estimate can vary a lot.
c) and d) You should not use your model for extrapolation! Especially bad for the tree models but even the linear and quadratic models should not be used outside the observed range. What if the y level off for big x? or what if y increases even more rapidly for big x - you don't know which model fits outside an observed range of data.
e) Since the noise-level is so high and the curvature quite modest, a linear model would probably be selected rather than the quadratic model but a close call. The tree models perform much worse in terms of prediction performance so would not be selected.

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.

Solution:
a) This is the easier problem (cmp to b). Collinearity does lead to higher estimation variance for the coefficients, but at the level of prediction the impact of high estimation variance at the coefficient level essentially cancels out. For the small sample size I would use LOOCV to select a predictive model. For the large sample size I would use random splits and split into training, validation and testing to obtain a good estimate of the final predictive performance of the selected model. In both cases one could consider using PCR as well. For the small sample you could consider using only one of the 3 correlated xs (assess via LOOCV which to include).
b) is a more difficult problem. You know from Labs and Minis that collinearity can make model selection very tricky. If only one of the highly correlated xs are in the true model, all selection procedures will have a problem unless the signal strength is really high. If several of the xs are in the model we may be able to determine the true model if, again, the signal strength is high.
So what to do? For large n I would use randomsplits paired with e.g. BIC. For small n we're probably not going to be able to determine the true model unless the signal strength is very high. The selection will be unstable and you will probably not be able to say which or how many of the correlated xs that are true predictors of y. I would compare Cp, LOOCV and AIC selected models and significance of coefficients. If these tend to agree I would feel a bit more confident, but I would also look at the LOOCV standard deviations not only the mean PE across folds and AIC and Cp values for all subset models to see how clear-cut the selection is.

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).

Solution:
This is a bit of a trick question... If it's the case that the packing machine has only these settings (6 total settings of 3 combined speeds and 2 dropping quantities) then all you have to do is run the machine a couple of times at each setting and see which one minimizes the total time. Formal testing can be done via ANOVA.
However, what if the machine has more potential settings and you want to determine how to optimize these from running current settings to get some ideas.
To set up the experiment, you want to run the machine at all 6 settings a couple of times, say $n = 10$ each. For each setting you record the time to get 100 boxes that pass quality control, call this y. There are two predictor variables; speed (a three level factor OR a numeric variable) and quantity (a two level factor OR numeric variable).
Once you have collected the data you plot to determine how to model; depending on the scatter plots you either use speed and quantity as factors or numerical variables.
The model would look something like this if you can use the variables as numerical.

$$y = \beta_0 + \beta_1 speed + \beta_2 drop + \beta_3 speed * drop + \epsilon$$

where I have included an interaction term as well. The coefficient β_1 tells you how the time is affected by the speed (probably negative) and β_2 how the time is affected by the drop-quantity (probably negative). However, the interaction could be such that for the lower drop quantity, perhaps at high speeds there is a lot of "misses" and so the slope with respect to speed may even be positive. This is all conjecture of course.
If the scatter plots indicate that you can't use the variables as numerical - perhaps because

there is a complex synergy between drop and speed - for a certain drop quantity there may be an optimal speed (say the 2nd setting) where everything that is dropped lands in the boxes whereas for a smaller drop quantity this speed is larger (or smaller). Then, you might not see a linear relationship of time wrt to speed and therefore have to use a factor coding.

$$y = \beta_0 + \beta_1 s2 + \beta_2 s3 + \beta_3 s2 * d2 + \beta_4 s3 * d2 + epsilon$$

where $s2$ and $s3$ are dummy variables for the 2nd and 3rd speed and $d2$ is a dummy for the second drop quantity. The baseline model is for drop quantity 1 and speed 1.
Provided that the scatter plots and residual diagnostics support this, you would obtain estimates using least squares. You now want to check signs and magnitudes and significance of each estimate to determine the optimal setting from the fitted values.

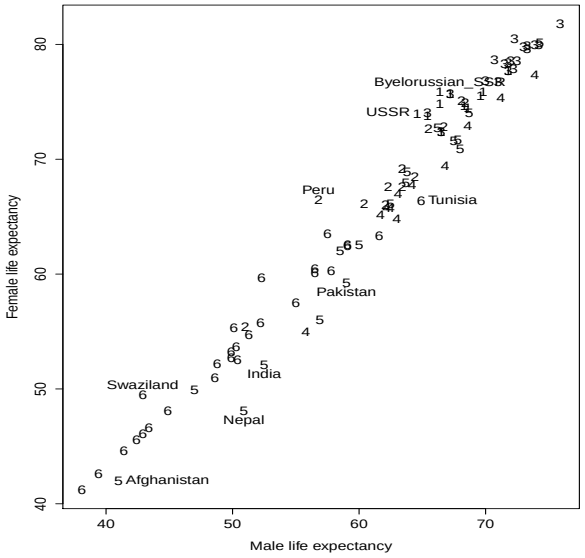
Question 6 a-f: 30p (5+5+5+5+5+5)

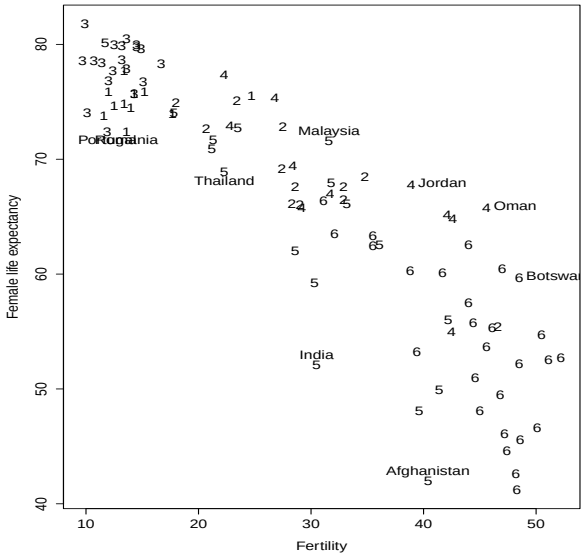
Statistics from 91 countries are collected. Variables include mortality rate per 1000 (Mort), birthrate per 1000 (Fertility), infant deaths per 1000 in population under 1 years old (InfDeath), life expectancy of males and females (LifeM, LifeF) and gross national product (GNP). In addition, countries 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.

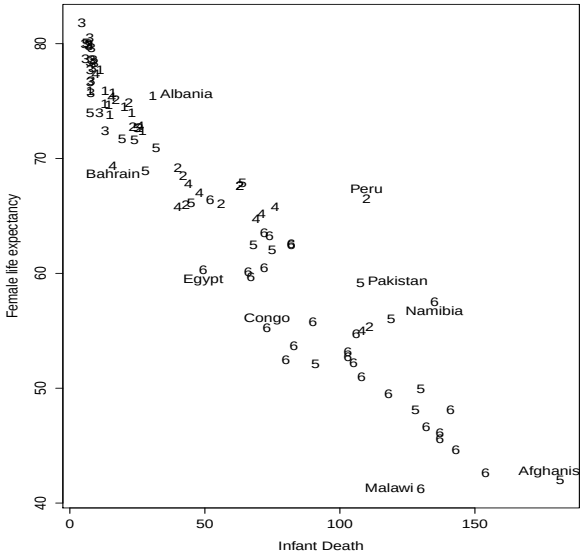
a) The data is summarized in 6 scatter plots on the next page. Discuss the plots: do the basic assumptions for regression modeling via least squares hold? Concerns? What would your preliminary steps be prior to modeling? Explain why.

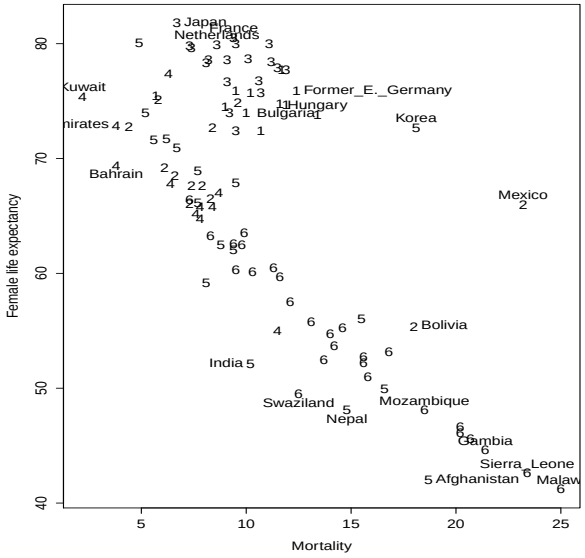
Solution: Top left looks good. Top right appears to be non-linear and perhaps non-constant error variance. Middle left looks nonlinear. Middle right looks like there's groups in data (but if the group variable is also included this might not be a problem - need to check the residuals plot vs mortality later). Bottom left looks very random. Bottom right non-constant error variance and some outliers.

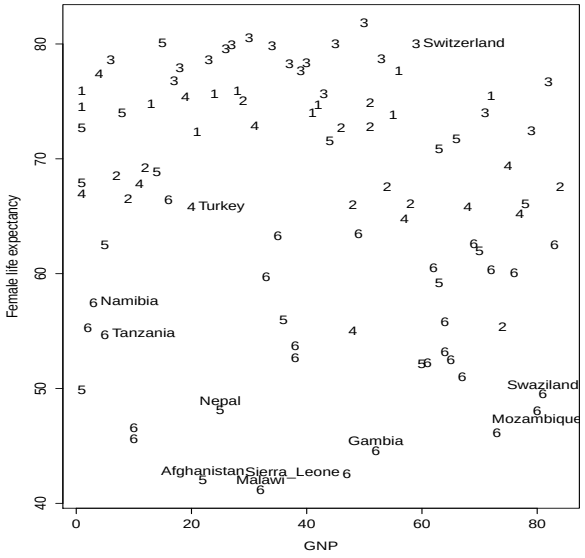
Concerns: Variable transformations are clearly needed - perhaps that will take care of both nonlinearities and non-constant error variance problems. Since the region effect appears to be considerable we should also explore interactions with coplots.

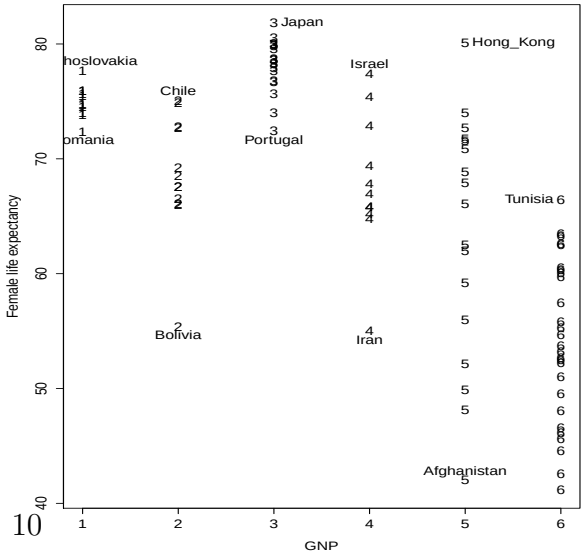

Female life expectancy
Male life expectancy
Byelorussian_SSR
USSR
Peru
Tunisia
Pakistan
India
Swaziland
Nepal
Afghanistan


Female life expectancy
Fertility
Malaysia
Thailand
Jordan
Oman
Botswa
India
Afghanistan


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


Female life expectancy
Mortality
Japan
France
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
```

Solution: The residual diagnostic plots look OK. No trend in the residuals or perhaps a small trend (nonlinear relationships?), a few outliers though. BUT we also have to check the residuals vs the other variables.
If I take the modeling results at face value, the model states that female life expectancy is significantly positively related to male life expectancy (expected) and negatively related to infant death (so if infant death is common in a country it tends to be associated with a lower female life expectancy). We also see that female life expectancy is significantly lower in regions 4 and 5 (middle east, asia). The R-squared is very high. This model can explain the variability in female life expectancy to a very high degree!

Limitations: You certainly have collinearity problems - e.g. Fertility and infant death and male life expectancy. From a) I also know that there are some relationships that appeared to be nonlinear which this model does not capture.

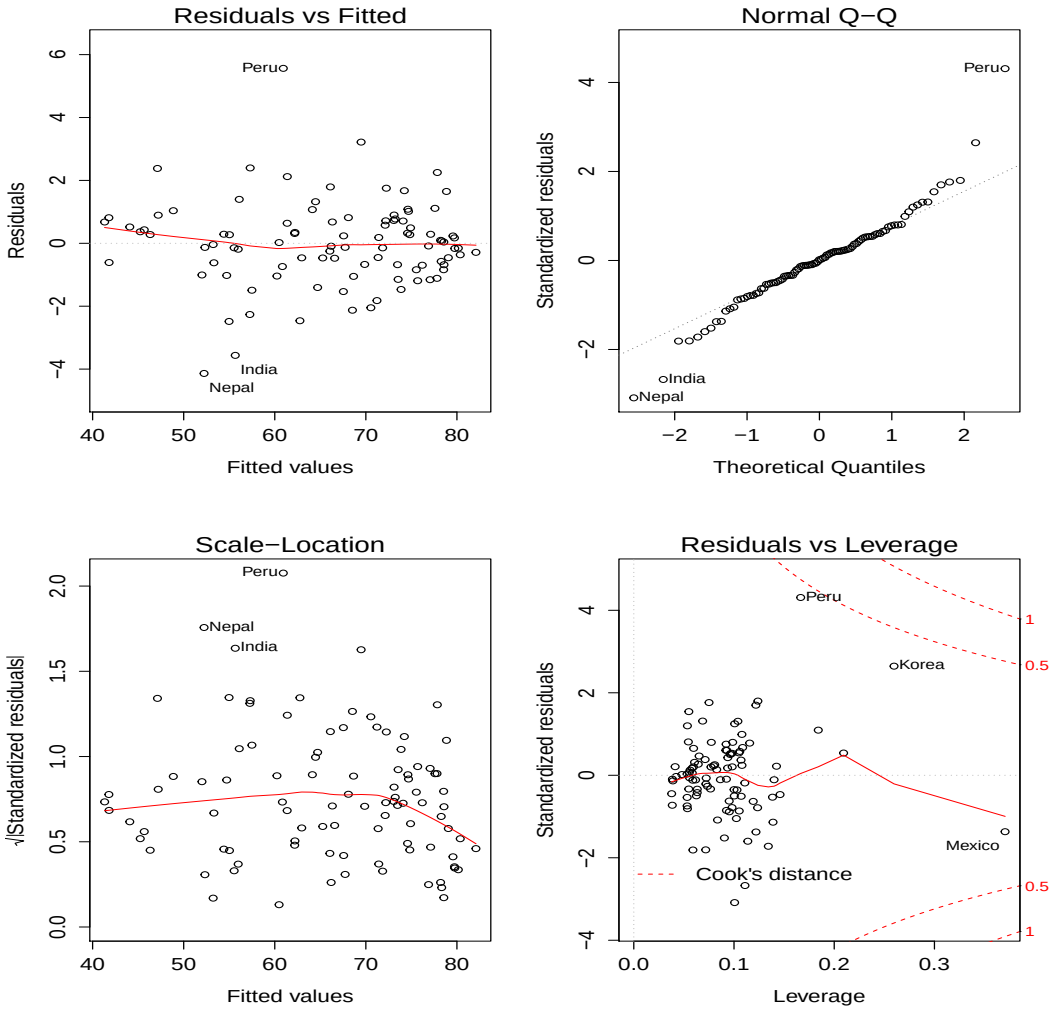


Figure 5: Question6: Diagnostic plot

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
```

Solution: BIC agrees well with the model summary, but has also added group 6 to the model. AIC and Cp result in bigger models, including also fertility (And GNP which was accidentally left out of question b) - typo in exam!).

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"
```

Solution: AIC selects the best models in terms of prediction error but it's also the biggest. BIC selects the smallest model. The results align quite well with b) but for some random splits fertility or mortality, gnp and region 6 may also enter into the model.

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"
```

Solution: These results are less clear. AIC is still best for prediction but the selection results are much more unstable. The model are somewhat smaller on average (natural for smaller training size) and the prediction errors somewhat larger (also due to worse training - remember PE is affected by the estimation variance).

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.

Solution: From the scatter plots, a linear model approach looks more appropriate. We see a nice linear or curve-linear relationship between female life expectancy and most of the other variables. PRO with CART would be that we don't have to worry so much about the variable transformations AND it can include complex interactions more easily.
What to do next? I would want to try different variable transformation and remove outliers (if they still are there after transformations). I would run CV to compare CART and the linear model. To deal with the collinearity, I would try PCR or regularized regression.