MSG500 Final 2017-01-12

Examiner: Rebecka Jörnsten, 0760-491949

Remember: To pass this course you also have to hand in a final project to the examiner.
Open book, open notes but no calculators or computers allowed. Make sure to give detailed and specific answers. Avoid yes/no answers. You should also provide a motivation. Good Luck!

Question 1(25=5+5+5+5+5)

A university medical center urology group was interested in the association between a prostate- specific antigen (PSA) and a number of prognostic clinical measurements in men with advanced prostate cancer. Data were collected on 65 men who were about to undergo radical prostectomies (removal of the prostate). PSA is a marker for cancer but also other prostate problems. BPH (benign prostatic hyperplasia) is a non-cancerous enlargement of the prostate. Seminal vesicle invasion and capsular penetration give information about how invasive the growth is and is related to the rate of progression of the cancer. The Gleason score is a microscopic evaluation of a biopsy of the cancer cells and is used to score the severity (grade) of the disease.

Variable	Information
Identification number	1-97
PSA level	Serum prostate-specific antigen level (mg/ml)
Cancer volume	Estimate of prostate cancer volume (cc)
Weight	Prostate weight (gm)
Age	Age of patient (years)
Benign prostatic	Amount of benign prostatic hyperplasia (cm2) hyperplasia
Seminal vesicle invasion	Presence or absence of seminal vesicle invasion: 1 if yes; 0 o.w.
Capsular penetration	Degree of capsular penetration (cm)
Gleason score	Pathologically determined grade of disease (6,7,8)

In this question we will model the Gleason score using CART. In Figure 1 you see the CART (classification tree fit) and the cross-validation results. This is the `rpart` cross-validation result. You select the smallest model that has a cross-validation error within the minimum error + 1 standard deviation (errors ± 1SD are illustrated with vertical bars in the plot).

a) Interpret the tree - which clinical factors are associated with low or high Gleason scores?
b) Explain the cross-validation plot. What size tree is selected based on CV performance?
c) What does the pruned tree look like? (You can determine this from the information in the left panel of Fig 1).
d) What is the training error rate (You can determine this from the information in the left panel of Fig 1).
e) I randomly split the data into 55 observations for training and 10 for testing and repeat

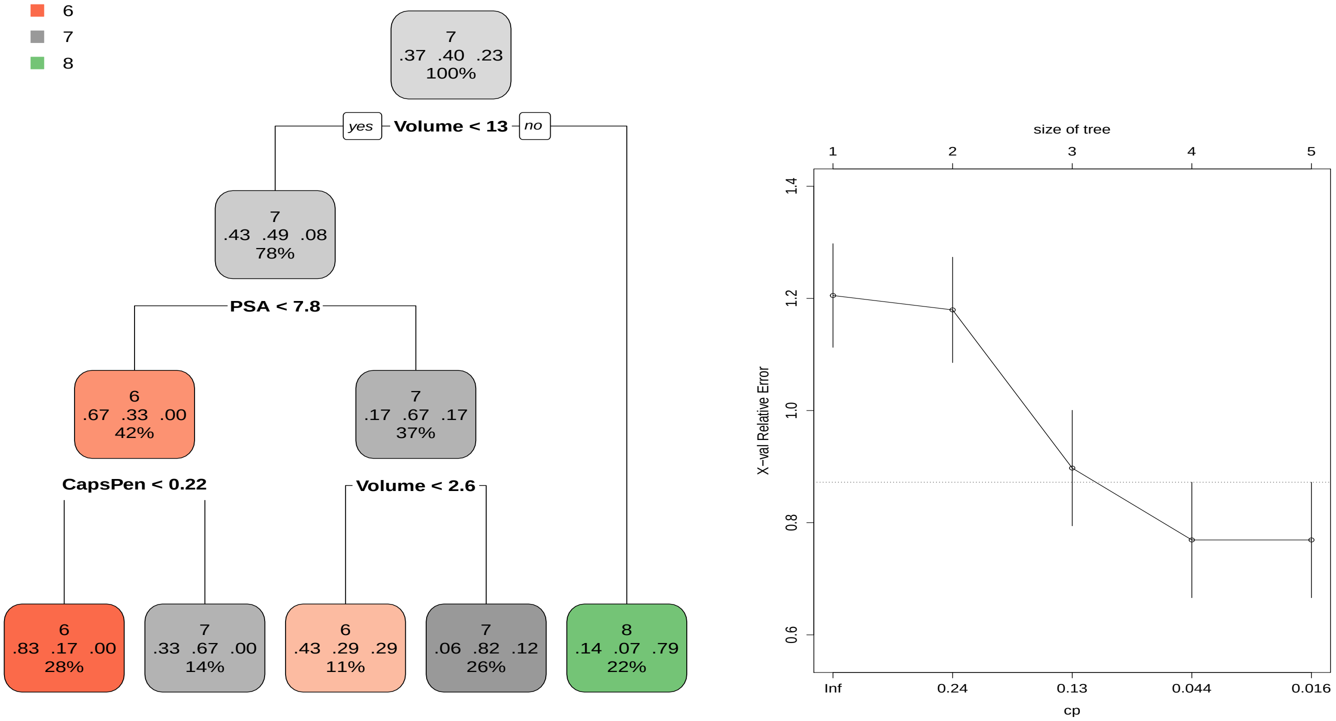

Figure 1: Left: Full tree (in each leaf; the majority label, the proportions of each label and the percentage of the total number of observations in each leaf. Right: Cross-validation of the classification tree.

this 100 times, using the rpart cross-validation error to select the model for each random split. Comment on these findings. Is the model selection problem ”easy” or ”hard” - motivate your answer.

```
                  modtab
[1,] "PSA"         "0.87"
[2,] "Volume"      "0.96"
[3,] "ProsateWt"   "0.12"
[4,] "Age"         "0.09"
[5,] "BPH"         "0.03"
[6,] "SeminalInv" "0"
[7,] "CapsPen"     "0.28"
```

```
                  modfirst
[1,] "PSA"         "0.37"
[2,] "Volume"      "0.62"
[3,] "ProsateWt"   "0"
[4,] "Age"         "0"
[5,] "BPH"         "0"
[6,] "SeminalInv" "0"
[7,] "CapsPen"     "0.01"
```

Question 2(25=5+5+5+5+5)

We continue to work with the PSA data. This time we will use a regression model to predict the (log)PSA level (think of this as an easily obtained measure and we want to see if it relates to other important disease markers). You can find the model summary and basic diagnostic plots (Figure 2) on the next page.

a) Interpret the model.
b) Comment on the diagnostic plots. Do the 5 basic assumptions hold - specify (which you can verify and which you need more information for).
c) Propose an action that you think might improve the fit. Be specific and back up your claim based on the results provided here.

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.7596 -0.4529  0.1421  0.4380  1.4388

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.090430   1.105634  -0.986 0.328179
Volume       0.062705   0.016262   3.856 0.000296 ***
ProsateWt    0.014745   0.009053   1.629 0.108912
Age         -0.008886   0.013304  -0.668 0.506873
BPH          0.065156   0.043089   1.512 0.136029
SeminalInv   0.846941   0.343141   2.468 0.016603 *
CapsPen     -0.030464   0.038868  -0.784 0.436417
Gleason      0.396301   0.143329   2.765 0.007657 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.7189 on 57 degrees of freedom
Multiple R-squared:  0.691,Adjusted R-squared:  0.6531
F-statistic: 18.21 on 7 and 57 DF,  p-value: 1.812e-12
```

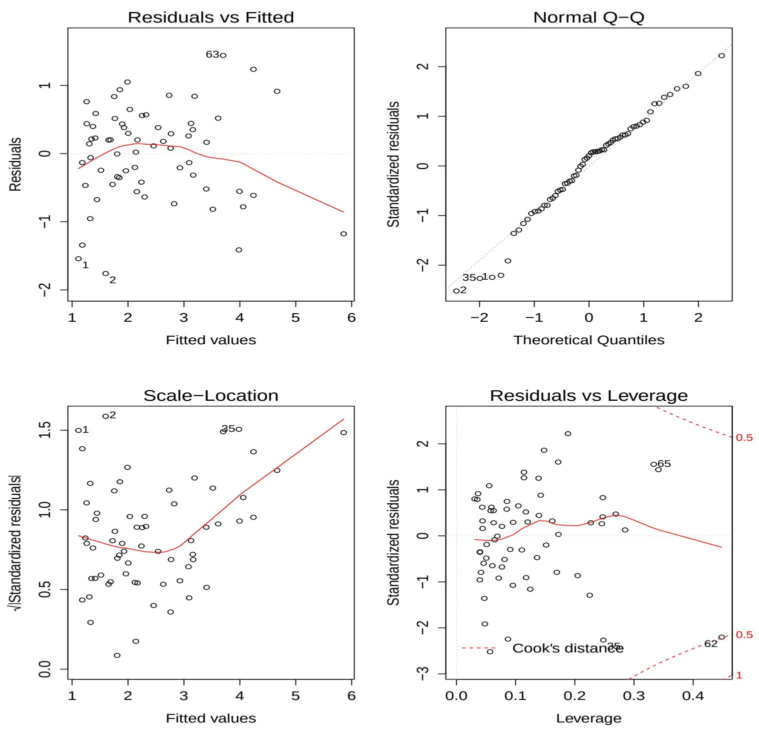

Figure 2: Diagnostic plots

d) I select a random sample of 55 observations for training and 10 for testing. The results using 10-fold cross-validation, Cp, AIC and BIC model selection are shown in Figure 3 and the table below. Comment on the results. Is there a clear "best model" - why/why not? Is there a clear "best model size" - why/why not?

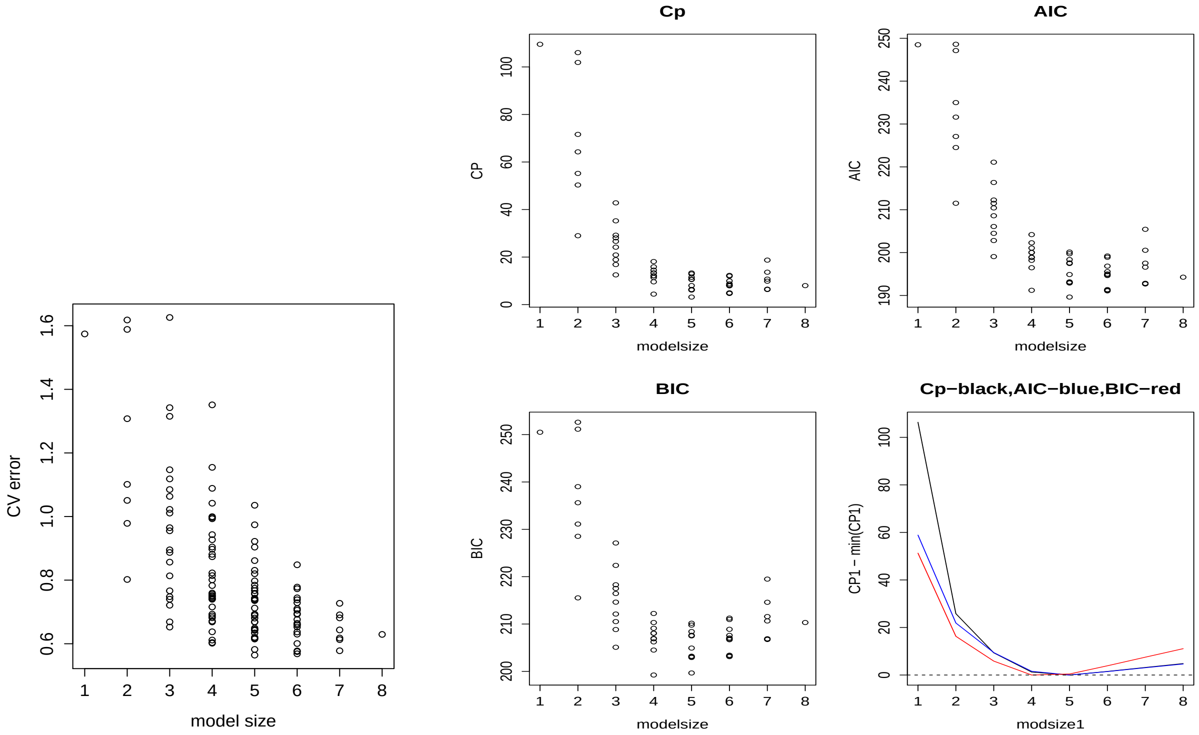

Figure 3: Model selection results

```
               Volume ProsateWt Age BPH SeminalInv CapsPen Gleason
cvmod   0.530       1         1   0   0          1       0       1
cpmod   0.530       1         1   0   0          1       0       1
aicmod  0.530       1         1   0   0          1       0       1
bicmod  0.721       1         1   0   0          1       0       0
```

e) I repeat the above 100 times and obtain the following results.

```
                 Volume ProstateWt  Age  BPH SeminalInv CapsPen Gleason
cvmod   0.64264       1       0.88 0.04 0.30       0.85    0.05    0.99
cpmod   0.66897       1       0.74 0.08 0.33       0.91    0.11    1.00
aicmod  0.66661       1       0.76 0.09 0.33       0.93    0.11    1.00
bicmod  0.69113       1       0.67 0.00 0.33       0.66    0.02    0.90
```

Comment on the model selection; which are the most important features? is it a stable selection problem? which method selects the best model?

Question 3(25p=5+5+5+5+5)

a) In Figure 4 I provide the scatter plots of log(PSA) and the other features. With this additional information, suggests some ways to improve and expand on the modelling of log(PSA). Give an example of an expanded model and explain how the result from fitting such a model could be interpreted. Pay specific attention to the characteristics of the different features (0/1 features, ordinal features, features that are 0 and non-0, nonlinear trends, outliers,....). Specify some additional plots you might want to look at and why.

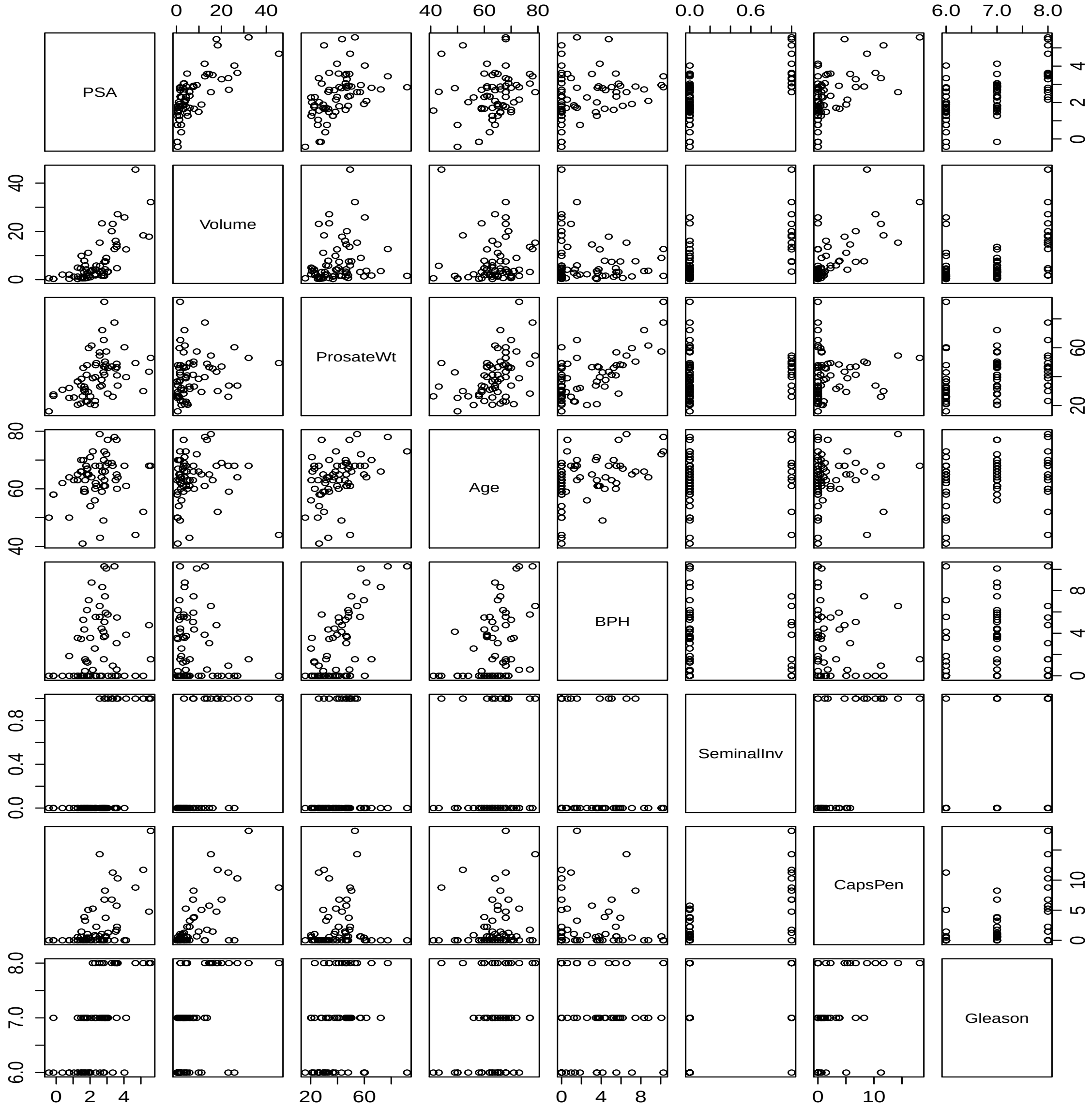

Figure 4: Scatter plots

b) I also run a regression model to predict cancer volume (a very important prognostic marker) from the other features. Below I provide the results and the basic diagnostic plots (Figure 5). Interpret the model. Which problems can you identify with the fit and propose actions you would undertake to remedy this (be specific and motivate based on results you see here).

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.46281    8.67172  -0.745  0.45917
PSA          0.03942    0.01984   1.987  0.05174 .
ProsateWt    0.18030    0.06728   2.680  0.00961 **
Age         -0.08080    0.10552  -0.766  0.44701
BPH         -0.91261    0.31766  -2.873  0.00570 **
SeminalInv   3.67226    2.70113   1.360  0.17933
CapsPen      0.80072    0.29134   2.748  0.00801 **
Gleason      1.55649    1.12407   1.385  0.17154
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.663 on 57 degrees of freedom
Multiple R-squared:  0.645,Adjusted R-squared:  0.6015
F-statistic:  14.8 on 7 and 57 DF,  p-value: 8.022e-11
```

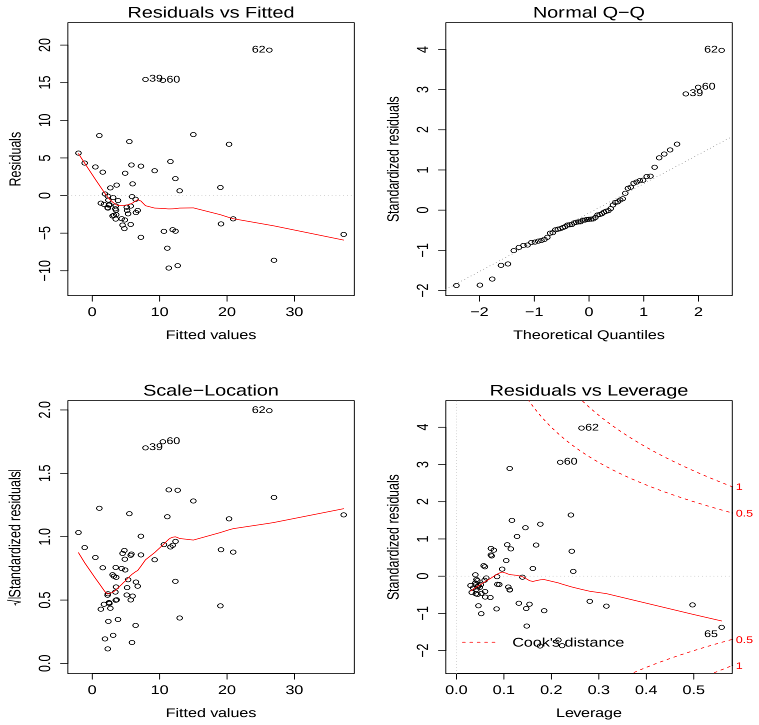

Figure 5: Diagnostic plot, Question 3

c) I remove outliers (defined as those with extreme leverage and/or Cook's distance) one-by-one until no such outliers remain. This resulted in the removal of 5/65 observations (observations 10,12,13,36,55). The new model summary is provided below and also a table of the top rank order (largest observation) for all features. Comment on this procedure and the results before and after outliers are removed.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.24490    4.52238   0.275  0.78420
PSA          0.34968    0.03666   9.539 5.16e-13 ***
ProsateWt    0.03484    0.03633   0.959  0.34207
Age          0.13604    0.05902   2.305  0.02520 *
BPH         -0.60214    0.16160  -3.726  0.00048 ***
SeminalInv  -3.78679    1.57308  -2.407  0.01966 *
CapsPen      1.32377    0.17270   7.665 4.32e-10 ***
Gleason     -1.54690    0.61656  -2.509  0.01527 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.787 on 52 degrees of freedom
Multiple R-squared:  0.8347,Adjusted R-squared:  0.8124
F-statistic: 37.51 on 7 and 52 DF,  p-value: < 2.2e-16


------------------------
rank
       PSA Volume ProstateWt Age BPH SeminalInv CapsPen Gleason
[61,]  42     55         35   49  16         36      37      52
[62,]  36     11          4   28  35         37      51      54
[63,]  13     37         16   30  43         42      13      55
[64,]  10     12         52   52  49         51      27      58
[65,]  12     36         49   27  52         59      12      62
```

d) In Figure 6 (next page) I provide scatter plots for 3 different data sets with y as the response and two independent variables ($x1$ and $x2$). For each of the data sets, state the model you think the data has been generated from (provide both the model equation AND explicit numbers for the coefficients and the noise level! you can get rough estimates from the figures).

e) In the bottom panel, what if you didn't have access to $x2$. Which model would you propose to fit to the data then?

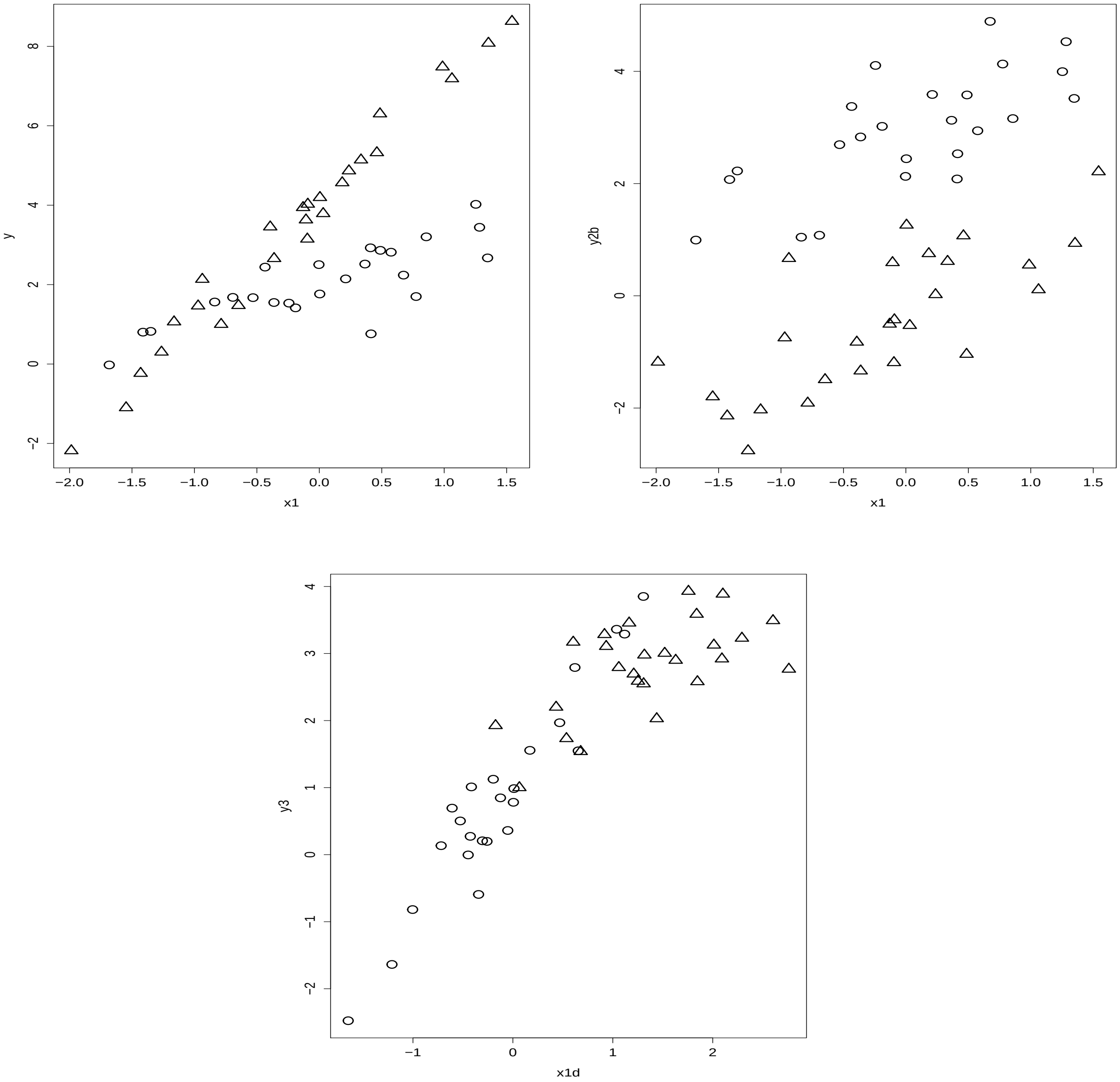

Figure 6: Scatter plots y vs x1, circles correspond to x2=0 and triangles to x2=1.

Question 4(25p=5+5+5+5+5)

Below I provide 5 different statements. I want you to state whether these are False, Partially False/True or True. For False and Partially False/True statements I want you to amend the statements so that they are True. Motivate your answers - explain.

a) A linear model was fit to the data using least squares. The R-squared was 60% so at least one of the predictors is significantly related to the outcome variable.
b) A linear model with 2 predictors was fit to the data set comprising 100 observations using least squares. The R-squared was 60% but none of the coefficients were significant at the 5% level. From this we can conclude that the two predictors must be highly correlated.
c) A linear model with 2 predictors was fit to the data set comprising 10 observations using least squares. The R-squared was 60% but none of the coefficients were significant at the 5% level. From this we can conclude that the two predictors must be highly correlated.
d) A linear model was fit to the data using least squares. In order to satisfy the 5 basic assumptions, 15% of observations were removed from the data. The resulting R-squared was 60%. We therefore expect that we can explain 60% of the variability on 85% of future observations.
e) A linear model was fit to the data using least squares. The residual diagnostics indicate that the error distribution is long-tailed. We conclude that the t-tests for the coefficient estimates are overly liberal, false rejecting the null hypothesis too easily.