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 $\pm$ 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

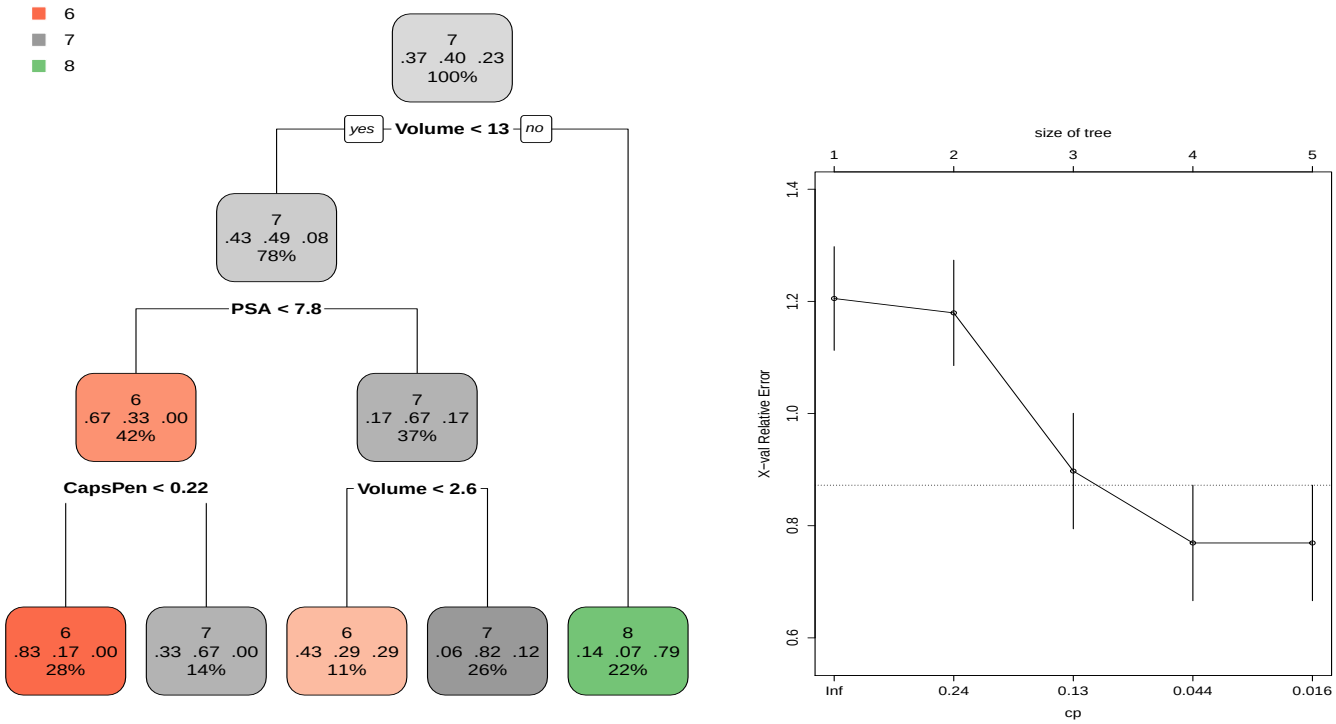

Figure 1: Left: Full tree (in each leaf; the majority label, the proportions of each label and the percentage of the total number of observations in each leaf. Right: Cross-validation of the classification tree.

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"
```

Solution

a) Cancer volume is the first variable we split on and for cancer volume exceeding 13 the majority of tumors (79%) are scored with the highest Gleason score (in this data set), namely 8. If we further refine the smaller tumors (¡ 13) by splitting on PSA, tumors with smaller PSA are majority Gleason score 6 (the smallest in this data set). Tumors with volume less than 13 and PSA more than 7.8 can further be split by looking at tumors with volume more than 2.6 which are 82% Gleason score 7 whereas the smaller tumors are a mix of all Gleason scores but majority is Gleason score 6. For the tumors with PSA less than 7.8 we split on the invasive variable Capsular Penetration. If Caps Pen is less than .22, 83% of the tumors have Gleason score 6. If the Caps pen is more than .22, 67% have Gleason score 7.
Summary: Cancer Volume, PSA and Capsular penetration separate tumors' Gleason score, with large Volume separating out the worst tumors (grade 8) and PSA and Volume and Capsular Penetration separating tumors of grade 6 and 7. We also note that none of the leaves are completely "pure" due to the tree building control settings preventing further splits (default is to stop splitting once a node contains less than 20 observations or when the resulting node contains less than 20/3 observations - here our data set is rather small so we don't get nodes smaller than 11% (7 observations)).

b) As stated above, the tree building stopped because the nodes contained too few observations (as default setting) which is also indicated by the cross-validation error - for the larger trees investigated the cross-validation has not started to increase yet. (If you run the tree building relaxing the defaults to allow for smaller leaf-nodes, we can get pure nodes and see the cross-validation error start increasing for larger trees.)
What does the plot tell us? After the 1st split, the cross-validation error is barely improving on an empty tree. It's only after the second split (size 3 tree) that things start to improve. From the results we have available to us here, the largest tree investigated is also the tree with the smallest cross-validation error. However, the tree with size 4 has a cross-validation error that is smaller than the minimum error + 1SD for the size 5 tree. The size 3 tree is pretty close too - but its cross-validation error is just above the min error + 1SD line. We therefore select the tree of size 4. (However - if this was your data set, you would probably want to rerun this with different control parameters!).

c) To prune the tree to size 4, consider which split removed would increase the error the least. Here, the second split on volume leads to a fairly small leaf (11% of observations) that are a mix of the majority label at the node above and other labels. If you remove this split, it won't change the predictions by much. The size 4 tree is thus the tree without the second split on Volume less than 2.6.
To see this, consider the change in error rate due to removing the split. For the second split on volume you go from error $.11 * (.29 + .29) + .26 * .18$ to $.37 * (.17 + .17)$ - an increase from about 11% to 13%. For the split on Capsular penetration removed you go from error .28*.17+.14*.33 to .42*.33, i.e. from about 9% to 13% which is a larger increase.

d) You can estimate the training error from the proportions in the leaves that the bottom of the tree. The percentage of observations in the leaf times the proportion of observations that don't belong to the majority labels are the contributions to the training error from each leaf.

$$.28 * .17 + .14 * .33 + .11 * (.29 + .29) + .26 * (.06 + .12) + .22 * (.14 + .07) \simeq .25$$

So the error rate is about 25%.

e) The results agree with the findings from a-d. PSA and Volume are the most important factors for explaining Gleason score. However, it is clear that each individual tree can look very different! About 2/3 of the trees first split on Volume whereas 1/3 split on PSA! Roughly 1/3 of the trees use Capsular penetration as well. The model selection clearly identifies PSA and Volume and indicate that Capsular penetration may be important, but there is no stable tree-model (trees that split first on PSA or first on Volume seem to work equally well). Occasionally, another variable is used in the tree model (like prostate weight).

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.

Solution

a) log(PSA) is explained to a relatively high degree (69%) by the predictors. Based on the marginal t-tests, we see that log(PSA) is significantly positively related to Volume and Seminal Invasions and Gleason score. That is, the larger the tumor is, the more invasive it is and the higher the Gleason score, the higher we expect log(PSA) to be. Caveat: we have to check for collinearity problems before we read too much into each individual p-value.
b) From the basic plots we observe a trend in the residuals (downward). This could indicate a lack of fit for the model. However, we also see that these trends can be attributed to mainly 1 observation which has high leverage and a large negative residual. Without it the trend problems will probably go away. The main violation is thus the presence of an outlier! We would want to see the residual vs predictor plots to understand what the source of the high leverage is (extreme Volume perhaps)?
The absolute value of the residuals also exhibits a trend, in part from the outlier but also from a group of other observations with large fitted values. It looks like we have increasing residual variance with fitted values - perhaps we need to investigate another transformation of PSA and/or the other variables (e.g. Volume, Prostate weight as treating Gleason as a categorial variable)? We need to look at the residual plots vs other predictor variables to resolve this.
From the data description we know that we are analyzing 65 different men - it is not exactly clear how they were sampled (e.g. cluster sampled) - we should check this to be sure we can assume uncorrelated errors.
The errors are fairly symmetric around 0 though we have some larger negative residuals than large positive ones - perhaps also due to the outlier presence or a suboptimal data transformation as mentioned above.
Outliers - one is clear - BUT we probably want to re-evaluate this after trying some other transformations.
c) I would first try removing the obvious outlier and see if this solved all the problems. However, if any of the residual plots vs other variables exhibited trends etc I would explore

different transformations of the predictor variables to see if I could surpress the non-constant error variance and fix the lack-of-fit. These actions are motivated by the presence of the high-leverage outlier in the top-left plot, the long-left-tail error distribution in the top-right plot and the non-constant error variance in the bottom-left plot.

```
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
```

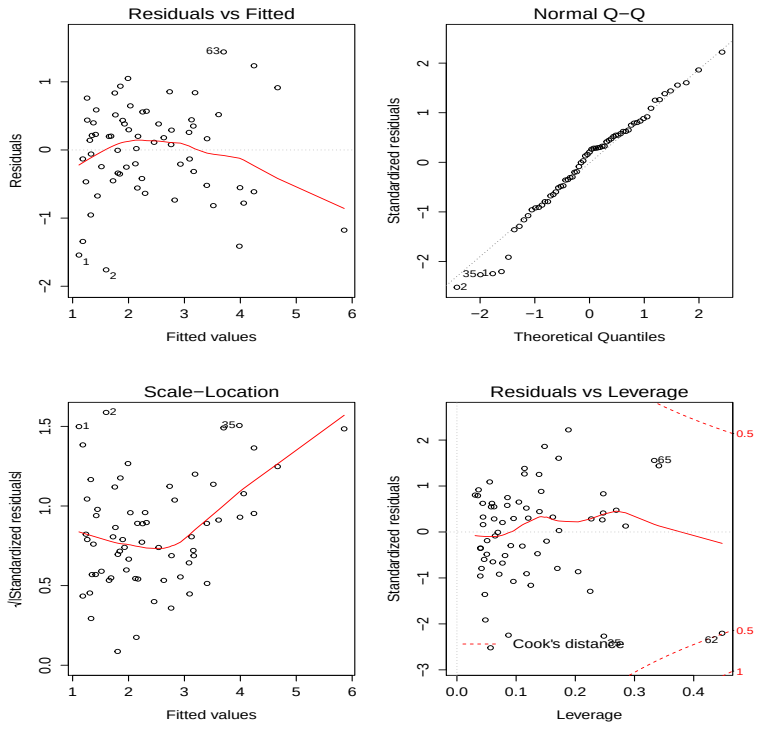

Figure 2: Diagnostic plots

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?

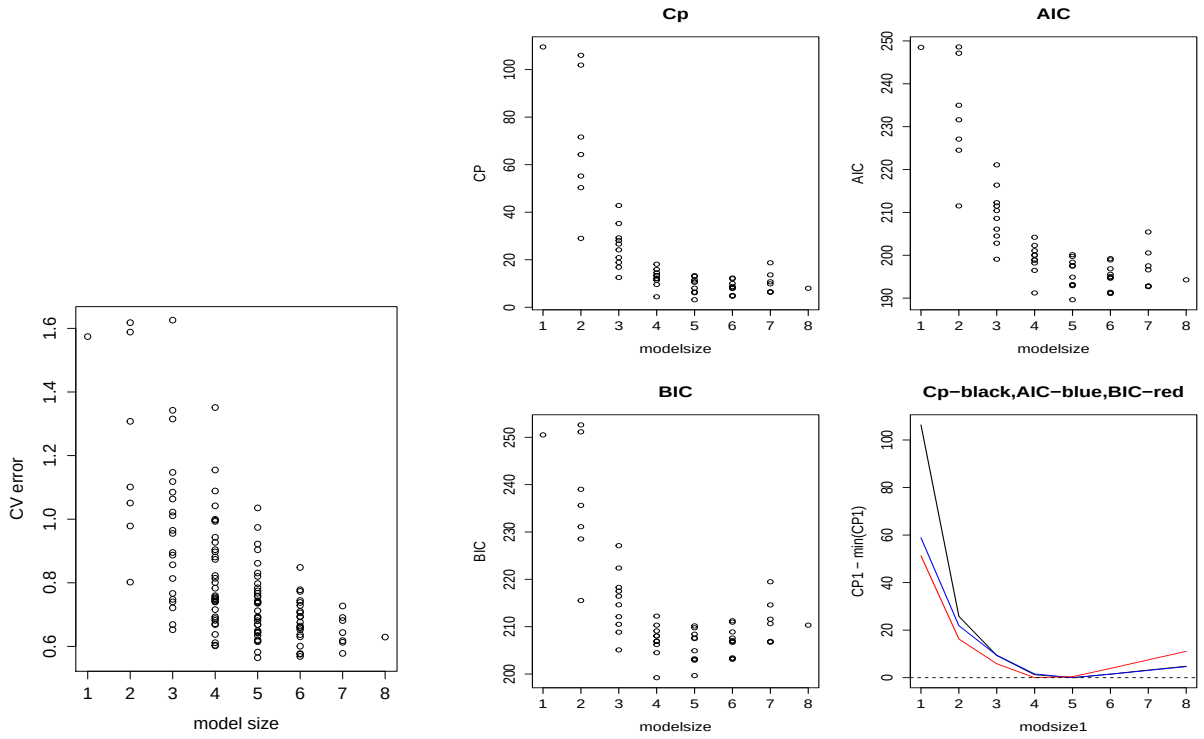

Figure 3: Model selection results

		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?

Solution

e)The model size 4-6 seems to be the best across criteria. CV, Cp and AIC all select a model of size 5 whereas BIC selects a model of size 4. For the Cp, AIC and BIC selection it looks like there's one model of size 4 and one of size 5 that are pretty close in terms of performance. When we use CV there are 3-4 models of size 5-6 that have near-identical cross-validation errors. CV, Cp and AIC picks the same model: including Volume, prostate weight, Seminal invasion and Gleason score. BIC reduces this model by not including Gleason.
Summary: there is no obvious winning model - but several candidates. Model size 4-6 are almost equally good so it not clear that there is best size model either - many competing models produce similar results!

e) All selection criteria identify Volume and Gleason score as important predictors. CV, Cp and AIC also include Prostate weight and Seminal invasion almost always. The average model size for CV, Cp and AIC is a bit over 5 whereas the average model size for BIC is a bit over 4. Age and Capsular penetration is almost never included in a model using these criteria. BPH is used in roughly 1/3 of the models which is not very stable. CV, Cp and AIC models are relatively stable otherwise but BIC is more unstable, sometimes including Prostate weight and/or Seminal Invasion.
From the average test errors, CV produces the best prediction model. However, from the one-time run in d) we see that prediction errors can vary a lot! We should probably compare the test errors using boxplots. From the results we have available it looks like BIC performs worse than the other criteria. 55 observations to train a model with 7 predictors - BIC might be too conservative here and picking models that are too small!

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.

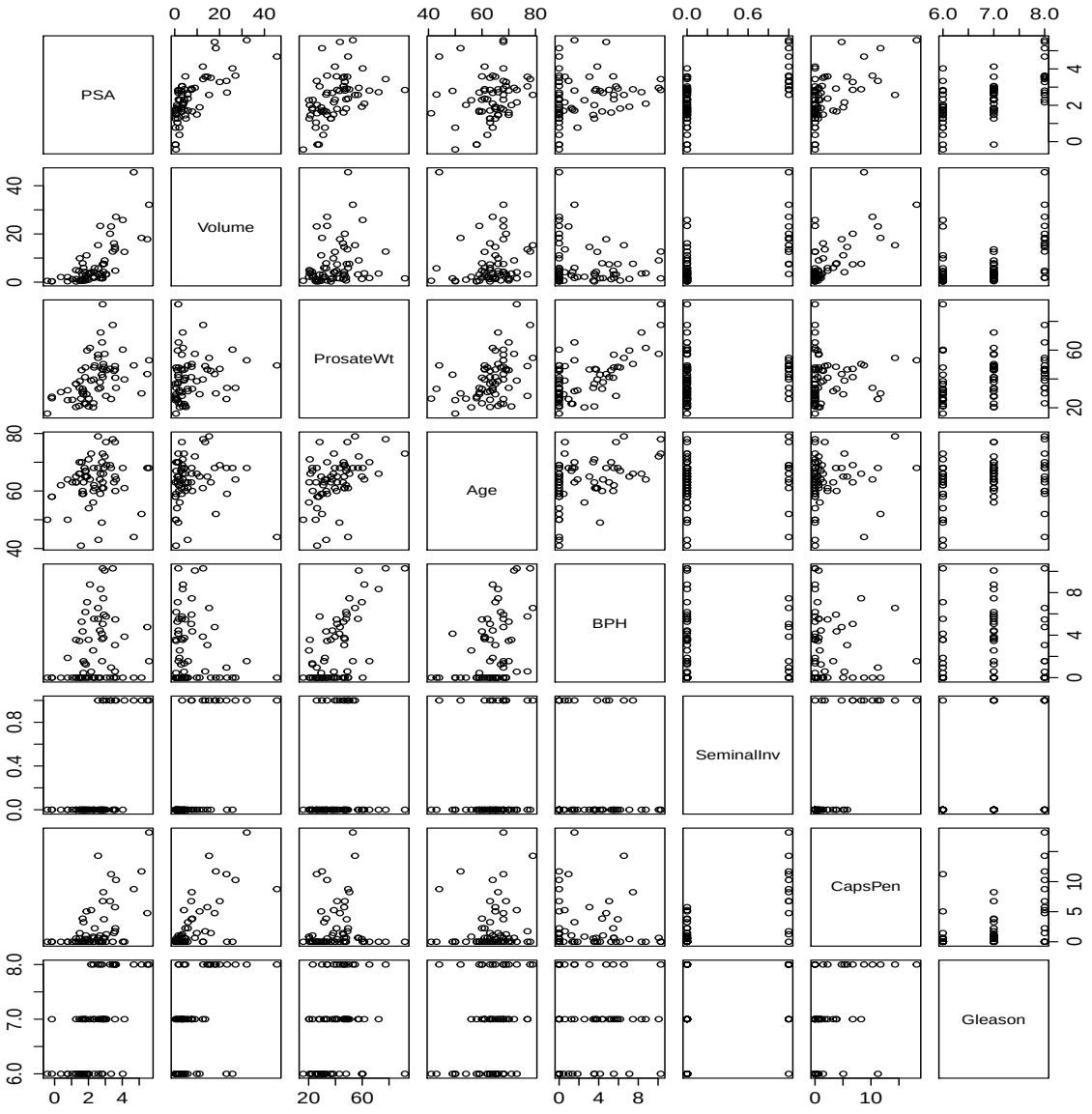

Figure 4: Scatter plots

Solution

a) From the top-row in the panel: log(PSA) is obviously not linearly related to Volume - we need to consider a transformation (e.g. log) of Volume. Likewise, we might enhance the relationship with prostate weight if we transform this variable (also log to bring in the largest prostates). Gleason score is an ordinal variable. While it looks like a linear treatment of the scores is not bad, we might want to explore using it as a categorical variable.
Both BPH and Capsular penetration are "weird" - there's clearly a lot of 0s but also some non-0 values. This is not strange - either there is a penetration you can measure which produces a number or there isn't, which is coded as 0. Either you have BPH which can be measured or you don't, which is coded 0. That means 0 has a special meaning in this variable. We should explore including these variables as dummies in addition to their numerical values.
Additional plots: I would like to look at some coplots to see if there are synergies between variables that better explain PSA. If the coplots between numerical and/or categorical variables indicate that PSA depends on variables differently depending on another feature, we should include interactions in our model. It would be interesting to explore such interactions between e.g. the invasion variables and volume and weight.

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
```

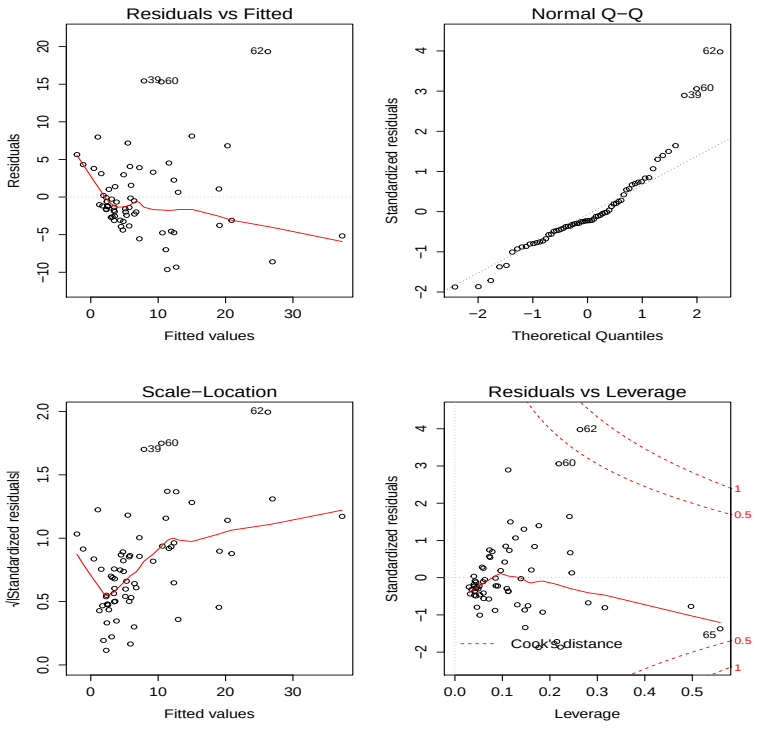

Figure 5: Diagnostic plot, Question 3

Solution

Cancer volume is relatively well summarized by the other features (R-squared about 60%). The significant predictors are prostate weight, BPH (with a negative coefficient) and Capsular penetration. What does this suggest? Well, a larger prostate is a predictor of a large tumor but a larger BPH (a benign growth) is inversely related to cancer volume (makes sense). In addition, if there is invasion of surrounding tissue this is associated with a larger tumor also. PSA is borderline significant with large PSA associated with larger tumors.
However, looking at the diagnostic plots we should probably be very careful about interpreting this model. There are lots of issues with this fit! There is a trend in the residuals and there are a couple of large positive residuals and/or high leverage observations. I would go back and try to transform the data set to better spread out the Volume so that the overall fit is improved. There is clearly a "clump" of observations closely together in terms of fitted values and a much smaller percentage of observations spread out around this mass - they are now dominating the diagnostics and fit.

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
```

Solution

At a first look, the removal of outliers have done a great job. The R-squared is now over 80%. However, it's dangerous to sequentially remove outliers without considering the source of their "outlyingness". From the rank-statistic we see that it is the top-4 largest PSA values that have been removed and one of the top-4 Volume values. The fact that it's the largest values in PSA and Volume that are removed is a very unorthodox way of fixing a lack-of-fit - it's like trying to cut off the data where a linear trend doesn't fit instead of including a nonlinear trend or transforming the data to allow for a linear fit. If we transform Volume, PSA and prostate weight things might look very different. By "cleaning" the data, PSA is now highly significant and Prostate weight is no longer significant but a better way of doing

this is to transform the data to fix the lack-of-fit directly instead of "cleaning" - after all, you could create a model that really doesn't work for future data by doing this. The model you train by selectively removing large PSA values won't work for large PSA in the test data either!

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?

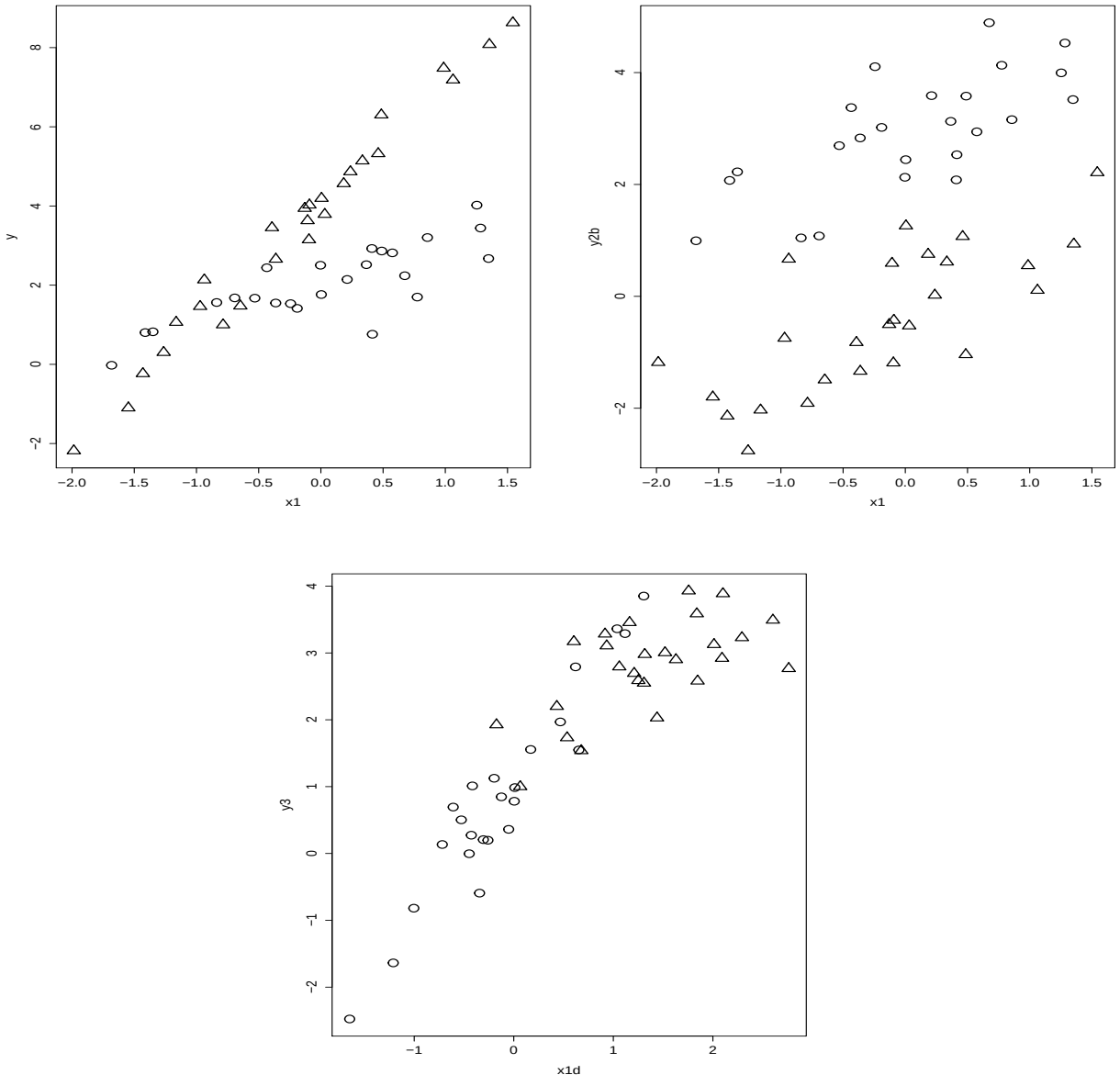

Figure 6: Scatter plots y vs x1, circles correspond to x2=0 and triangles to x2=1.

Solution

d) You can do rough estimation of slopes and intercepts from the plots as well as estimate the noise level, σ.
Top left: circle intercept (check y-value when x1=0) 2, slope (check y-values for x1 ± 1 and divide the difference by 2) $= (3-1)/2 = 1$. For the triangle data, the intercept is 4 and the slope $(7-1)/2 = 3$. The model equation is thus $y = 2+x_1+2x_2+2x_1*x_2+\epsilon$. The noise-level $V(\epsilon)$ can be estimated by looking at the spread around a regression line - a vertical box of height 2σ should cover almost all of the data, 4σ definitely all. Here, vertical slices of height 2 covers everything and slices of height 1 almost everything - so $\sigma \simeq 0.5$.
Top right: circles intercept 3, slope (4-2)/2=1. Triangles intercept -1, slope (0-(-2))/2=1. Model equation $y = 3 + x_1 - 4x_2 + \epsilon$ (i.e. an additive model looks good here). Estimate of σ around 1.
Bottom panel: circles intercept 1, slope (3-(-1))/2=2. Triangles intercept 2, slope (3-2)/2=.5. Model equation $y = 1 + 2x_1 + x_2 - 1.5x_1 * x_2 + \epsilon$ with σ around 0.5.
e) If $x2$ was unknown the pattern of y vs $x1$ looks nonlinear. I would be tempted to try to transform the $x1$ or include a polynomial term. This highlights the risk of missing variables....

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.

Solution

a) False. We can't say this without knowing the number of observations and the number of predictors. The R-squared is not a test-statistic though it is related to the F-test. We can get an R-squared quite big just by chance if the sample size is small and/or the number of predictors is large. The correct statement is that the R-squared is 60% and to this should be added an F-test result to be able to say anything about significance.
b) True. Such a large R-squared for such a large data set and such a small model is contradicted by the lack of significance of both coefficients.
c) False. We can't conclude this - because of the small data set, it could be that none of the coefficients are related to the outcome and the large R-squared is just due to chance OR because the predictors are correlated. To fix this statement we should check the correlation between the two variables - or redo the fit without one of the variables to see if this leads to a significant result. This is quite a small data set to try to draw conclusions from...
d) Partially true. However, the R-squared of 60% is for the training data so we wouldn't expect to explain the test data quite this well, even for the 85% where the model "applies". In addition, how would you know which 85% of test data to predict for? If outlier screening can be turned into a filter based on x-values we could apply this to the test data first but otherwise this is quite tricky. What you could write - first that the R-squared for test data will be somewhat smaller (can estimate this via cross-validation - prediction error reduction due to using a regression model compared with the intercept model). In addition, since the

statement is partially true you should just be very clear about the ease/difficulty in identifying the 85% of data where prediction works!
e) False. The sampling distribution for the coefficient estimates is not the same as the error distribution. A long-tailed distribution can be compensated for by a large sample size and the coefficient sampling distribution is essentially normal for large samples. For small samples and extreme long-tailed noise it would be more a question of "outliers" - where will the large errors be located? Depending on their placement you can have a very non-robust model fit that affects the testing.