Name:
GU/Chalmers/PhD-student (circle one)
Personal ID number:
If PhD student, write your home department/institute:

Solutions: Final MSG500 Dec 15 2012

Motivate your answers!

Question 1:10p

A political scientist was taking a survey of a sample of registered voters in Iowa City in October 2012. Three of the variables she collected were:

- presidential preference: whether the voter prefers Obama or Romney.
- gender: whether the voter is female or male
- income: the income of the voter's household in 2011

The political scientist wished to use statistical methods to determine whether gender and income are significant predictors of presidential preference.
(a) What is the response variable and what is its data type?
(b) State an appropriate statistical method to use to answer the above question.
(c) If you also wanted to test whether the presidential preference dependency on income level is affected by gender, what would you do? Give a mathematical equation for your model and say in words what each parameters means.
(d) Iowa City is fairly racially homogeneous (90% white non-hispanic), but, like in most cities, there are diversities in terms of education, social background, etc. Is this a concern for you? How might this affect the statistical modeling?
(e) Many registered voters were found to be undecided (did not yet have a preference for either candidate). Is this a concern for you? Any thoughts on how you would address this in your analysis?

a) Presidential preference is the response variable and it is a 2-level categorical variable.
b) There are several ways to do this. One we've talked about in class is logistic regression. Here, the model is for the log-odds (logit) of the probability of preferring Obama to Romney with x-variables gender and income. The model can be summarized at the coefficient level with estimates, standard errors and p-values. Those p-values tell you if gender and/or income are significantly related to preference BUT remember that standard errors come from the last linear approximation step of Iterative Weighted Least Squares and that the z-test is an asymptotic (large sample size) approximation.
You could also use Analysis of Deviance (ANODEV) to check if a variable should be included or not in the model. This is also an asymptotic approximation (X2-test).
Of course, you could also use ANOVA (testing income differences as a function of preference) and

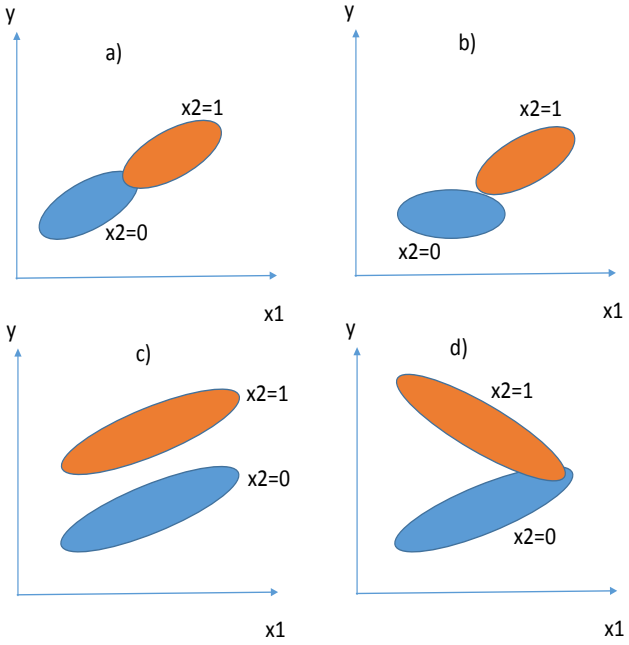
X2-test (testing gender-preference independence).
c) This is an interaction model. You would have a model

$$logit(\pi(Obama)) = \beta_0 + \beta_g gender + \beta_I income + \beta_{gI} gender \times income$$

where β_g is the increase in log-odds due to gender (say gender=1=female), β_I is the increase in log-odds for every unit increase of income and β_{gI} is the additional log-odds increase for women for every unit increase of income. Note, the impact of gender and income is additive on the log-odd and therefore multiplicative on the odds (or probability of preferring Obama to Romney).
d) The concern here would be "missing variables". If there are other factors that are important in determining voter preference (like education level, social background), then running a model excluding or ignoring those factors means that the model may misrepresent the actual associations. Perhaps it's not income but education that is important and income is made to look important because we don't take education into account? Or perhaps there is an interaction between income and education? Missing variables is a big problem in modeling. That being said, if all you want to do is build a prediction model, than income might be good enough for this - just be careful about interpreting the meaning of the coefficients if you have missing variables.
e) You can run the analysis ignoring all the undecideds, but of course that limits what the analysis tells you about the voting population. An option is to include the undecided as another group in your data and investigate how to predict memberships into each of these groups. Are the undecided evenly spread out over income levels and across gender?

Question 2: 10p

This question emphasizes the difference between interaction and correlation. Let Y be the dependent variable and $X1$ and $X2$ two independent predictors.
Let X1 be a quantitative independent variable, and X2 a dichotomous (2-level factor) independent variable. Let Y be numerical and continuous. Draw plots (you choose how to make your point, but you need to plot at least two figures for each case to answer the question) of the following situations:
(a) X1 and X2 are correlated, and there is no interaction between X1 and X2 in the model for Y
(b) X1 and X2 are correlated, and there is interaction between X1 and X2
(c) X1 and X2 are uncorrelated, and there is no interaction between X1 and X2
(d) X1 and X2 are uncorrelated, and there is interaction between X1 and X2



There are many ways you can do this but the figure above is an example. Uncorrelated x1 and x2 means there shouldn't be an association such that big x1 are linked to x2=1 etc. Interactions means that the relationship between y and x1 is affected by the level of x2.

Question 3: 30p

The *mtcars* data set comprises 32 observations and 10variables. The outcome variable is **mpg** (miles per gallon) and the input variables include **cyl** (Number of cylinders), **disp** (Displacement - relates to total volume of the cylinders) , **hp** (horsepower), **drat** (Rear axle ratio - relates to the efficiency of the gears at different speeds), **wt** (Weight (lb/1000)), **qsec** (1/4 mile time - relates to max acceleration), **am** (transmission 0 = automatic, 1 = manual), **gear** (Number of forward gears), **carb** (Number of carburetors).

I give you 8 scatter plots in the figures below. Circles means automatic transmission, triangles are manual. Note: mpg (miles per gallon) essentially measures how far you get on one tank of gasoline.

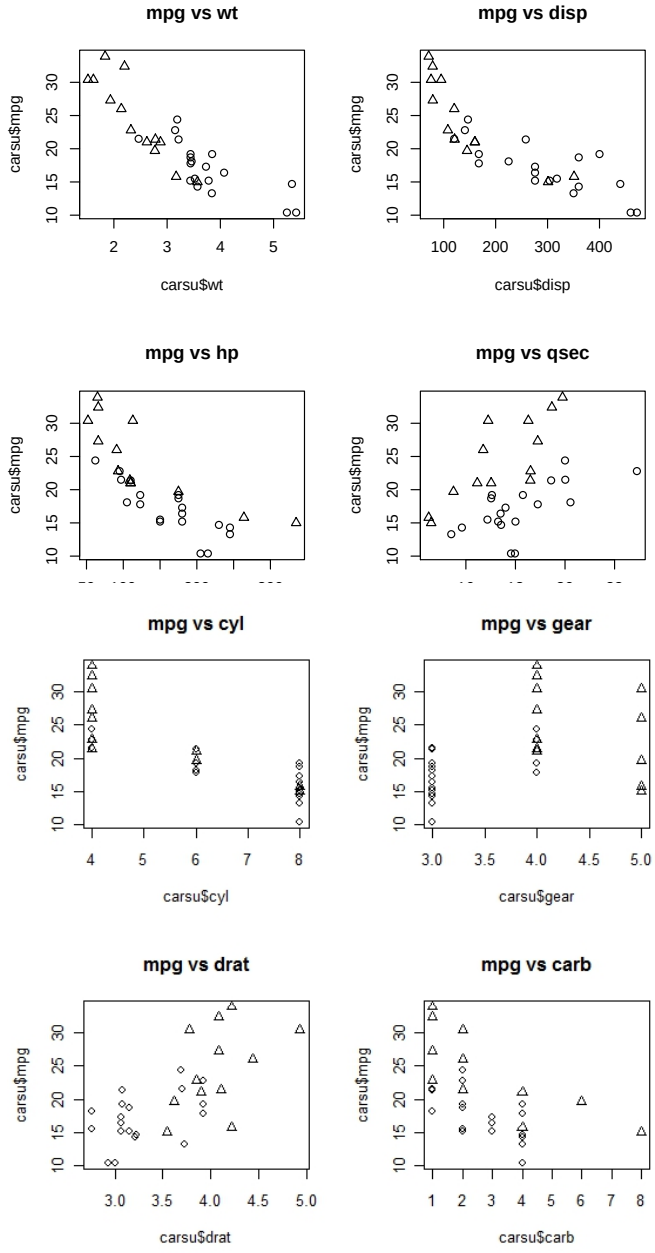


Figure 1: Scatter plots - question 3a

(a) I try a linear regression model to describe mpg as a function of the other variables, all treated as numerical. Below I provide a model summary and correlation matrix for all the 10 variables. I also give you the basic diagnostic plots. I perform stepwise backward selection and arrive at a reduced model. I provide its summary and diagnostic plots.
Please comment on the model overall. What can you say about mpg as a function of the other variables? Interpret this model.
Do you spot any problems with the data (you can also refer to the scatter plots)? If so, what additional plots (be specific) and approaches (be specific) would you use to resolve these issues? Any concerns regarding the fit (say based on what)?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.04177   18.21890   0.661    0.516
cyl         -0.16205    0.96757  -0.167    0.869
disp         0.01307    0.01737   0.752    0.460
hp          -0.02059    0.02048  -1.005    0.326
drat         0.79446    1.59793   0.497    0.624
wt          -3.73956    1.84519  -2.027    0.055 .
qsec         0.86134    0.66508   1.295    0.209
am           2.45510    1.96574   1.249    0.225
gear         0.66524    1.45833   0.456    0.653
carb        -0.21102    0.80665  -0.262    0.796

Residual standard error: 2.591 on 22 degrees of freedom
Multiple R-squared: 0.8689,     Adjusted R-squared: 0.8152
F-statistic:  16.2 on 9 and 22 DF,  p-value: 9.083e-08


Correlation matrix:
       mpg   cyl  disp    hp  drat    wt  qsec    am  gear  carb
mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.60  0.48 -0.55
cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.52 -0.49  0.53
disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.59 -0.56  0.39
hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.24 -0.13  0.75
drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.71  0.70 -0.09
wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.69 -0.58  0.43
qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00 -0.23 -0.21 -0.66
am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  1.00  0.79  0.06
gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.79  1.00  0.27
carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66  0.06  0.27  1.00



Reduced model:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6178     6.9596   1.382 0.177915
wt           -3.9165     0.7112  -5.507 6.95e-06 ***
qsec          1.2259     0.2887   4.247 0.000216 ***
am            2.9358     1.4109   2.081 0.046716 *

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared: 0.8497,     Adjusted R-squared: 0.8336
F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

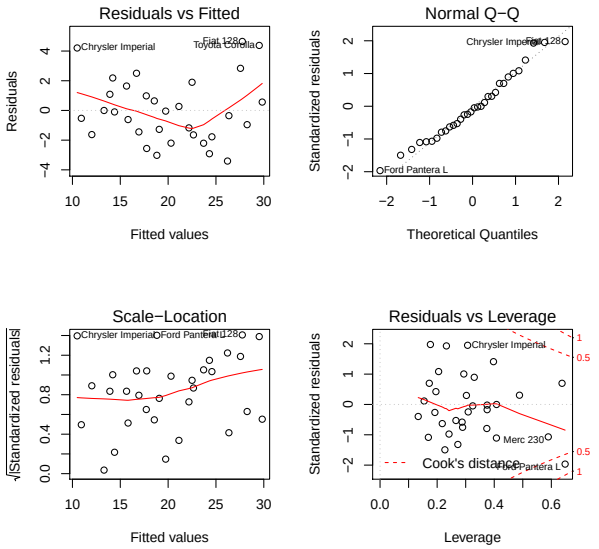


Figure 2: Diagnostic plots - question 3a

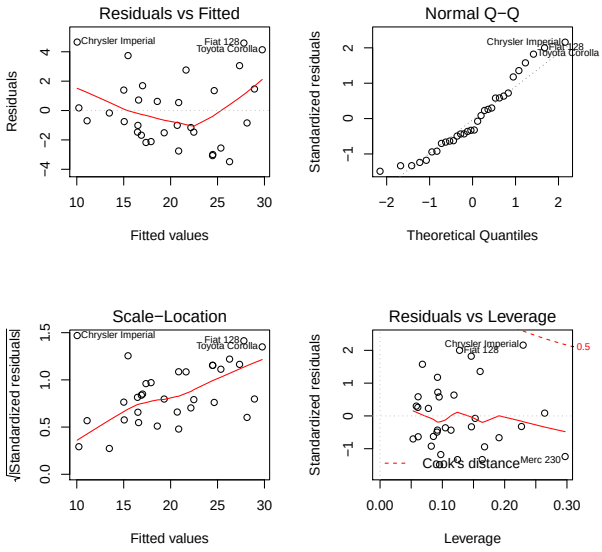


Figure 3: Diagnostic plots, reduced model - Question 3a

We start with the second part of the question: problems. The scatter plots clearly indicate that there is a nonlinear relationship between mpg and disp and hp, and perhaps some indication that this is true for wt as well. Gear is perhaps not best modelled with a linear trend either. There is a complex relationship between automatic transmission and gear that perhaps requires the construction of another variable (gears mean different things given the transmission type). The diagnostic plots show there is a trend in the residuals (due to the model insufficiency as described above) and for the reduced model there is also a hint at a non-constant variance. The strong correlations between variables is a concern for us as well - and we see that the collinearity has had a detrimental effect on the fit since none of the coefficients estimates are significant.
What to do? You need to try variable transformations. My guess would be that the inverse transform applied to mpg would solve most, if not all, of these problems. Of course, you would need to see the scatter and diagnostic plots after transformation! So, I would try inverse of mpg, redo the fit and plot diagnostic plots to see if the trend has been eliminated from the residual plot. If this doesn't work I would try a transformation on the x-variables where I see a nonlinear trend (disp and hp) - perhaps a quadratic term (remember to center the data so that you don't create yet another correlated variable).
What about the model fit? The R-squared is quite high, yet none of the x-variables have significant coefficient estimates!!! This seeming contradiction in terms is clearly a result of collinearity. In the correlation matrix we see just how bad things are (some x's correlated .9!). The extreme collinearity and the small sample size leads to enormous standard errors in the full model and so none of the coefficients are significant.
Stepwise model selection reduces the model to include three variables: wt, qsec and am. The R-squared is still high so this model is probably quite good for prediction. The interpretation is that mpg is reduced by the weight of the car (big cars less fuel efficient), and mpg is increased by having manual transmission or by having a very acceleration efficient engine. However, we should be careful about reading too much into this sine the collinearities may have made model selection unstable. Weight is highly correlated with some other variables, as are the other included variables.
The extreme collinearity problem means that you probably might want to try reducing the model first, either by pre-selecting some variables or by using principal components or some other dimension reduction approach.

(b) I perform 1000 randomsplits with training fraction .75. I obtain the following model selection results (using Cp, AIC and BIC):

```
           modselcp modselaic modselbic
 [1,] "cyl"  "266"    "254"     "270"
 [2,] "disp" "79"     "155"     "51"
 [3,] "hp"   "233"    "283"     "233"
 [4,] "drat" "159"    "224"     "133"
 [5,] "wt"   "885"    "901"     "882"
 [6,] "qsec" "432"    "529"     "406"
 [7,] "am"   "365"    "482"     "306"
 [8,] "gear" "135"    "208"     "113"
 [9,] "carb" "237"    "322"     "202"
```

I also provide boxplots with the model sizes and the prediction errors.
Question: Interpret and the discuss the results. Do these results agree with those of

question 3a? Why/why not? Any surprises?
Which of the model selection criteria would you recommend here? Why? Which final model, if any, would you recommend? Why?

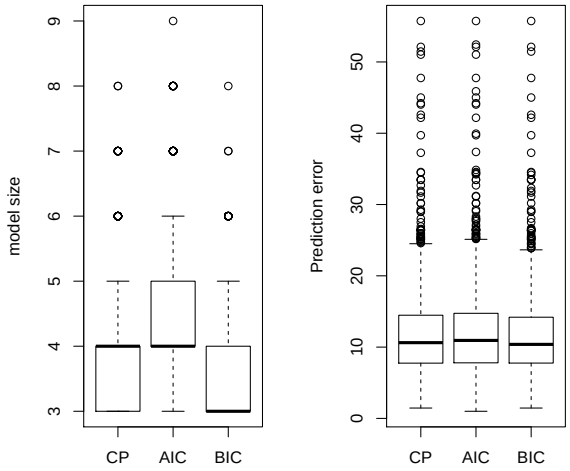


Figure 4: Randomsplits - model sizes and prediction errors - Question 3b. Mean PE is 12, 12.5 and 11.7 for Cp, AIC and BIC respectively.

Unfortunately, the random splits did not arrive at a conclusive result. As perhaps expected given the collinearities and the small sample size, there is much variability between selected models across random splits. Weight seems clear as a predictor, and disp and gear are not very good predictors. As for the rest, it's pretty much any subset thereof. BIC selects mostly 3-parameter models (2 variables), AIC and Cp 4-parameter models (3 variables). So BIC selects weight + one (sometimes two) more. AIC and Cp selects weight + two (sometimes 3-4) more. Since the prediction errors are on average about the same, I would recommend the BIC model it performs as good as the others with a more simple model. However, it would be important to report that there is much uncertainty in terms of model selection and you can't determine which, apart from weight, are the root cause of mpg performance.
The results agree with Q3a since that indicated that there was much uncertainty (full model fit) and the stepwise selection which identified weight as the most crucial variable.

(c) I repeat the exercise, but using a .5 training fraction instead. I obtain the following results:

```
            modselcp modselaic modselbic
[1,] "cyl"  "261"    "376"     "285"
[2,] "disp" "161"    "335"     "221"
[3,] "hp"   "321"    "414"     "348"
[4,] "drat" "242"    "384"     "294"
[5,] "wt"   "676"    "737"     "697"
[6,] "qsec" "397"    "522"     "437"
[7,] "am"   "294"    "444"     "355"
[8,] "gear" "273"    "426"     "321"
[9,] "carb" "353"    "479"     "401"
```

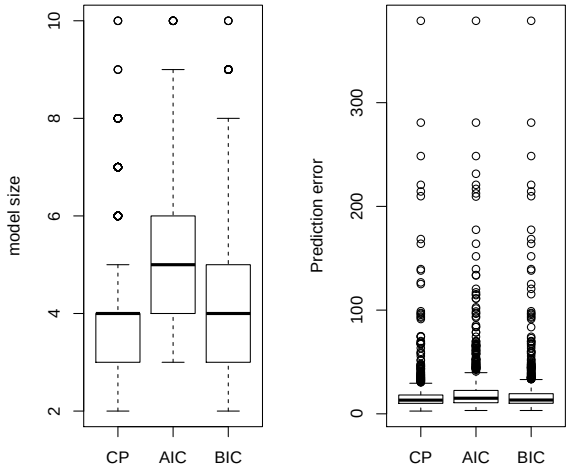


Figure 5: Randomsplits - model sizes and prediction errors - Question 3c. Mean PE is 18, 23 and 20 for Cp, AIC and BIC respectively.

Question: How do the results in 3b and 3c compare? Can you explain the differences? For example, why are the PE estimates higher in 3c do you think?
A surprising result is perhaps that there are several very large models selected here even though the training size is small, but also some smaller models than were observed in 3b. Any thoughts on why this might happen? (Look back to the description of the data, sample size, number of variables and correlation structure).

Smaller training sample made things even worse! Now model selection is almost random. You would expect smaller samples to lead to smaller models in general, but not if the training is so difficult that the estimation almost breaks down. With 16 observations to train on, 10 coefficients to estimate and .7-.9 correlations between variables, you are asking the impossible of modelselection. Just by chance, bigger models may be selected. Prediction suffers greatly from the estimation variance being so large here.

Question 4: 20p

I also try out a regression tree modeling of the cars data. I obtain the following results across 1000 randomsplits with training fraction .75 (first column tells you the number of times a variable is selected to be in a tree, the second column gives you the number of times a variable is selected to be at the top of the tree). I also give you an example of 4 trees (for 4 random splits) in a figure below. A second figure shows you the spread of selected tree sizes and corresponding prediction errors.
(a) Interpret each of the 4 trees in the figure (reminder: the equation at each node tells you the properties for the left branch of the tree).
(b) Compare the randomsplits results for CART to those of the regression model above. Discuss differences and similarities. There are some very obvious and perhaps surprising differences - any thoughts on their source?
(c) Based on the information here (model sizes, prediction performance, etc) and in question 3, if you were to recommend a model strategy for this data - would you recommend regression of CART? Why?
(d) What additional information, plots etc would you like to have access to, or analysis steps would you want to perform to make a final determination?

```
Selection results:
              %selected   selected first
 [1,] "cyl"    "520"          "165"
 [2,] "disp"   "918"          "266"
 [3,] "hp"     "837"          "252"
 [4,] "drat"   "231"           "0"
 [5,] "wt"     "684"          "317"
 [6,] "qsec"   "360"           "0"
 [7,] "am"     "19"            "0"
 [8,] "gear"   "15"            "0"
 [9,] "carb"   "12"            "0"
```

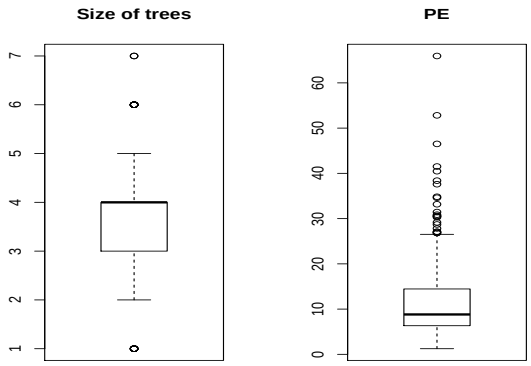


Figure 6: Question 4: Size of trees and Prediction error. Mean PE is 10.9

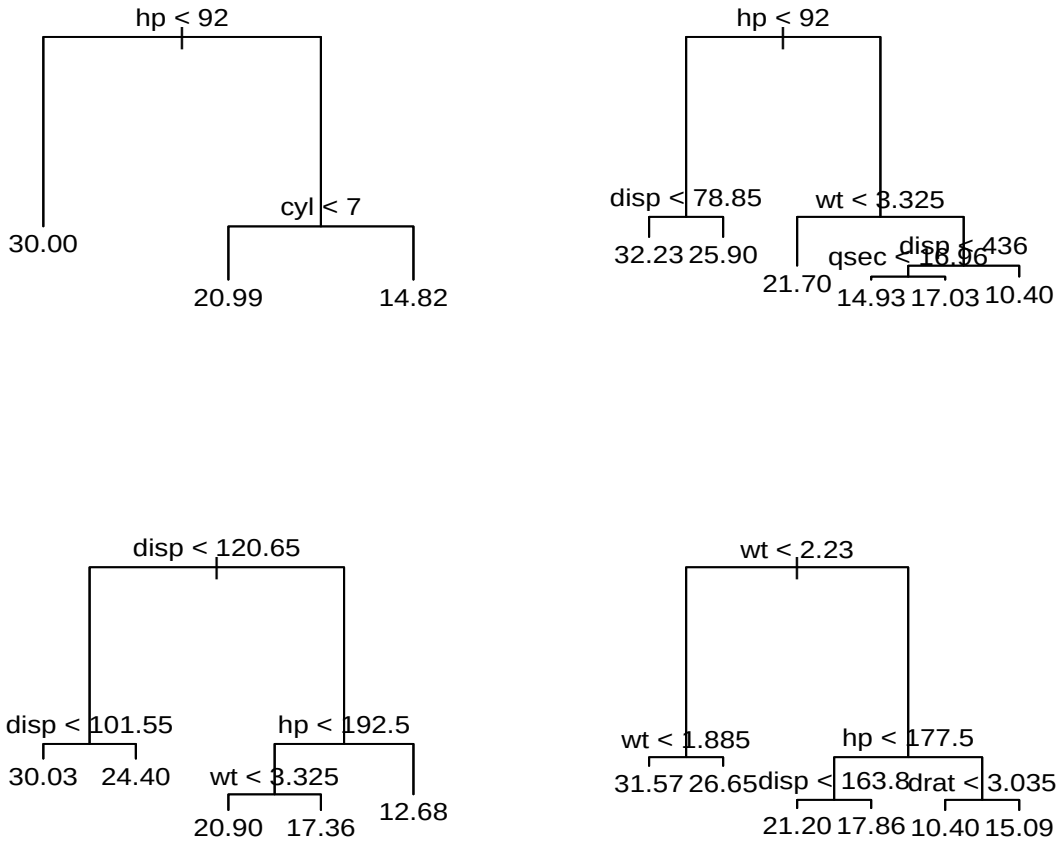


Figure 7: 4 regression trees.

a) Top left tree: If $hp < 92$ the mpg is estimated to be 30, if $hp > 92$ and $cyl < 7$, mpg is estimated 20.99 whereas if $cyl > 7$ it is 14.82. That is high horsepower and many cylinders makes for a fuel inefficient car!
Top right: If $hp < 92$ you then split on $disp < 78.85$ for two high (fuel efficient) levels of mpg. For $hp > 92$ you split the data on weight, displacement and qsec for a refined partition among less fuel-efficient cars.
Bottom left: You split on $disp < 120.65$ and again on $disp < 101.55$ to identify two sets of fuel efficient cars. If $disp > 120.65$ and $hp > 192.5$ you have the least fuel efficient cars with mpg 12.68. If $disp > 120.65$, $hp < 192.55$ you split cars into those weighing less than 3.325 (more fuel efficient) and those weighing more than this (less fuel efficient).
Bottom right: First split on weight. For cars weighing less than 2.23 you partition into even smaller cars (weight below 1.885) for the most fuel efficient (mpg 31.57). For heavier cars, you split on horsepower, displacement and drat. Etc etc.
Notice the branch lengths. The first split completely dominates such that most of the variation in mpg is explained by splitting once onf hp, disp or weight.

b) Different variables are picked using CART and it is a little more clear who's the important predictors. disp and hp are almost always picked, followed by weight and cylinder number (sometimes qsec and drat).
There is no clear "first split" decision though as it is almost randomly split between weight, hp, disp and cyl.
Similarities: difficult model selection problem since first splits and tree content vary between random splits.
Differences: hp and disp are now important predictors. This is because CART can handle the nonlinear structure in the data better than the linear model in Q3. However, the scatter plots indicate that a piecewise stepfunction is not a good approximation of the dependencies between x-variables and mpg and so CART is not the most efficient model (as demonstrated by repeated splitting on the same variable in trees in figure a)).

c) The prediction error for CART is slightly smaller than the prediction errors in Q3. And the model selection is a bit more stable. So at first glance, this suggests that CART is the better choice. However, we know from Q3 that there are many problems in the linear model that we could probably fix quite easily (variable transformations) and the scatter plots exhibit a nice (curve)linear relationship between mpg and the other variables. Therefore, I think a fixed linear model would be the best choice.

d) See above in c). I would want to redo regression after variable transformations to assess if this helps fix the trends in the residuals and reduces the prediction error.

Question 5:30p

The cars data set was rather small ($n = 32$) and contained a lot of variables (9) which made the analysis quite difficult. Here I will instead perform an analysis of a larger data set comprising house prices in Albuquerque. The variables are Price, SQFT (size of the house), Tax (the tax rate per year for the house), NE (an indicator variable for the location in Albuquerque - NE=NorthEast), Corner (an indicator that the house is located on a corner plot, and Features (0-11) which denotes the number of desirable features of the house (e.g. dishwasher, refrigerator, microwave, skylight(s), washer and dryer, handicap fit, cable TV, etc). The sample size is $n = 107$ and the number of variables $p = 4$. To enhance the linear dependency of Price on the other variables I have taken a log-transform. Here are some scatter plots:

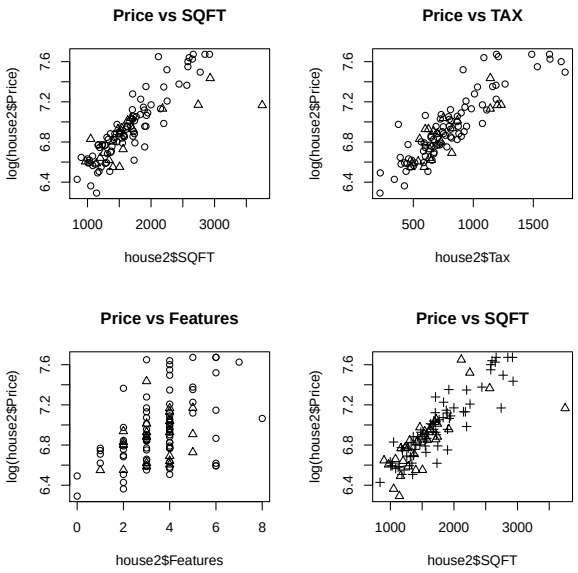


Figure 8: Question 5: Scatter plots. First 3 plots, triangles denote corner plots. Last panel, + denotes NE location.

I fit a linear model to the data. I present the model summary and the diagnostic plots below:

```
Correlation matrix:
          Price SQFT Features    NE Corner   Tax
Price      1.00 0.85     0.45  0.18  -0.10  0.88
SQFT       0.85 1.00     0.39  0.16   0.02  0.86
Features   0.45 0.39     1.00  0.24  -0.07  0.44
NE         0.18 0.16     0.24  1.00  -0.05  0.20
Corner    -0.10 0.02    -0.07 -0.05   1.00 -0.06
Tax        0.88 0.86     0.44  0.20  -0.06  1.00
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.082e+00  4.918e-02 123.671  < 2e-16 ***
SQFT         2.266e-04  4.908e-05   4.618 1.14e-05 ***
Features     1.569e-02  1.074e-02   1.460    0.147
NE          -2.460e-03  2.888e-02  -0.085    0.932
Corner      -5.901e-02  3.358e-02  -1.757    0.082 .
Tax          5.380e-04  8.683e-05   6.196 1.27e-08 ***

Residual standard error: 0.1361 on 101 degrees of freedom
Multiple R-squared: 0.8245,     Adjusted R-squared: 0.8158
F-statistic: 94.91 on 5 and 101 DF,  p-value: < 2.2e-16
```

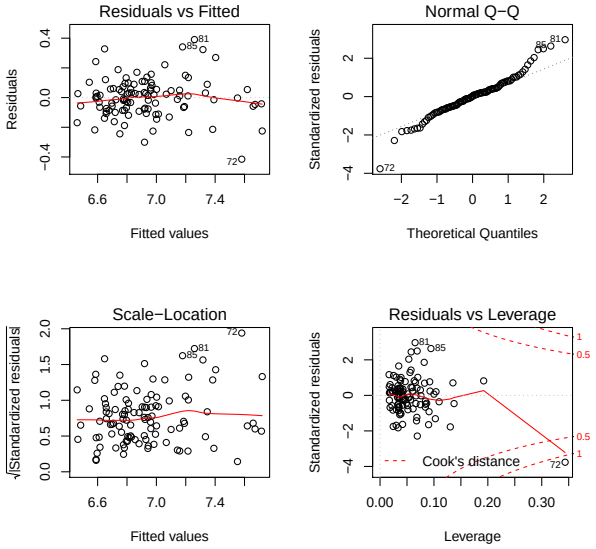


Figure 9: Diagnostic plots - Question 5

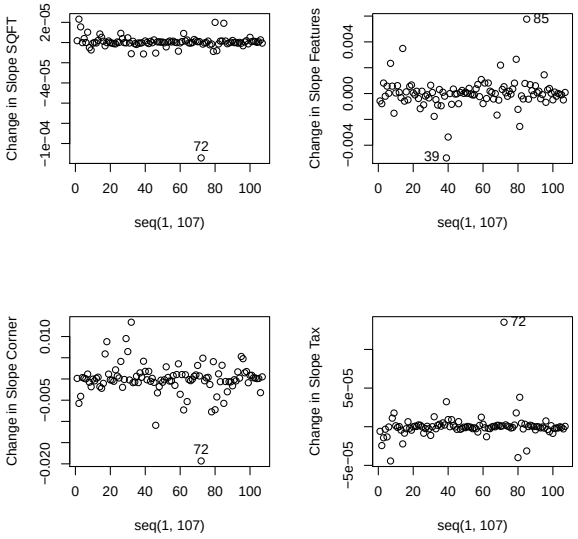


Figure 10: Diagnostic plots - Question 5 - Change in slopes

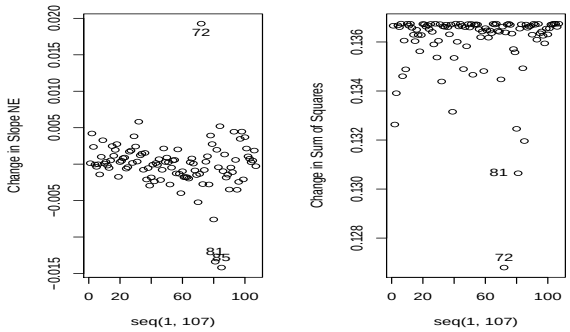


Figure 11: Diagnostic plots - Question 5 - Change in slope and Sigma

(a) Discuss the model fit. Do the basic assumptions hold (which can you verify from the given information)? Do you detect any problems, if so what are they and how would you address them?

The scatter plots look OK, but perhaps some outliers are present. The residual diagnostics indicate that there is no strong trend in the residuals so model sufficiency seems basically fulfilled. As for the uncorrelated errors, that is a design problem and we have to assume that the data set was sampled correctly. There is no clear trend in terms of residual variance, nor do the residuals appear to be asymmetric. The last basic assumption is clearly violated. There is at least one outlier present (72) which has an impact on slope estimates and the residual MSE estimate. From the Cook's distance plot we see it has both high leverage and a large residual value. Observations 81 and 85 also appear to be a concern since they affect the NE slope estimate.
It seems clear that observation 72 has to be dropped and the fit then further evaluated. Perhaps 81 and 85 will then be outliers or they are outliers now because of the effect of 72. New diagnostic plots will reveal this.

(b) Interpret the model. Say something about its expected usefulness to predict house prices from information such as features and size.

The model states that SQFT and Tax are significant predictors of Price. Tax and SQFT are correlated so one should interpret the coefficient values somewhat cautiously. Features and Corner have p-values near the .05 threshold we often use to declare significance. It is possible that after the outlier is removed that one or both of these variables would be significant. However, as it now stands Features does not significantly contribute to explaining Price.

(c) I perform 1000 randomsplits with training fraction .75. I summarize the results below. I also provide a summary of the model sizes and prediction errors in a figure. Interpret the results and discuss.

```
No interactions included:
              modselcp modselaic modselbic
[1,] "SQFT"     "999"    "999"     "995"
[2,] "Features" "390"    "412"     "104"
[3,] "NE"       "29"     "34"      "1"
[4,] "Corner"   "595"    "612"     "174"
[5,] "Tax"      "1000"   "1000"    "1000"
```

BIC picks the smallest model and in terms of average prediction error perform just as well as the larger models picked by AIC and Cp.
The model selection results are quite clear-cut. SQFT and Tax are almost always selected as predictors. BIC occasionally includes Corner Features, whereas AIC and Cp picks one or the other to always be in the model.
The random splits have confirmed that SQFT and Tax are predictors of Price.

(d) From the scatter plots, do you see any indication that interaction terms are needed in the model? Explain.

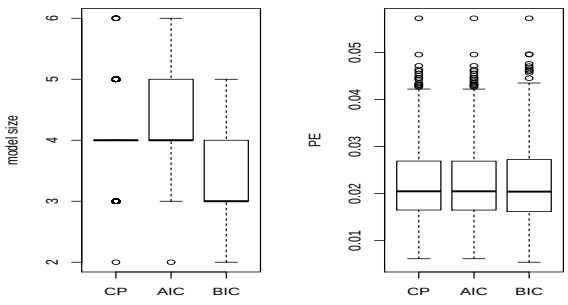


Figure 12: Question 5(c): Model sizes and PE - no interactions

If you really want to find an interaction, I would say SQFT with Corner is the only one (top left scatter plot) plot that suggests this. BUT it is completely driven by the few Corner plots with large SQFT!! In the other scatter plots I see no indication of differing slopes given the level of the Corner or NE variable.

(e) I create interaction variables between SQFT, Tax and Corner, as well SQFT, Tax and NE. I perform 1000 randomsplits with .75 training fraction and obtain the following results:

```
Interactions allowed:
                     modselcp modselaic modselbic
 [1,] "SQFT"          "942"    "948"     "881"
 [2,] "Features"      "270"    "297"     "71"
 [3,] "NE"            "117"    "134"     "30"
 [4,] "Corner"        "765"    "793"     "543"
 [5,] "Tax"           "993"    "994"     "994"
 [6,] "intSQFTNE"     "334"    "373"     "148"
 [7,] "intSQFTCorner" "855"    "867"     "700"
 [8,] "intTaxNE"      "332"    "379"     "138"
 [9,] "intTaxCorner"  "181"    "195"     "81"
```

Question (e): Compare the results when interactions are allowed in the model or not. Do the results indicate the interaction terms are needed? Compare model sizes, prediction errors etc.

Including interactions have only occluded the results. It is not much less clear who the essential predictors are. Moreover, the interactions SQFTCorner is present more often that the main effect Corner, which is a big no-no. Also, the presence of SQFT has been reduced.
The predictions errors are actually worse now (longer tail distribution in any case). Bigger models are selected. This indicates that we have created a more difficult model fitting and selection problem and prediction performance has been detrimentally affected by the increased estimation variance and uncertainty.

(f) Do you think the results from the randomsplits with interaction terms can be trusted? Why/why

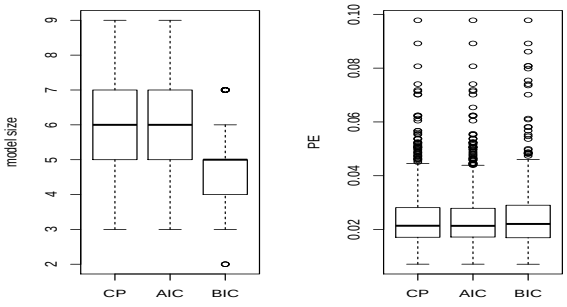


Figure 13: Model sizes and PE - interactions allowed

not? To assist you in answering this question I also provide the correlation matrix between the interaction variables and the other variables here.

```
Correlations between interaction terms and all variables:
              Price SQFT Features    NE Corner   Tax intSQFTNE intSQFTCorner intTaxNE intTaxCorner
intSQFTNE      0.52 0.52     0.38  0.88  -0.05  0.54      1.00          0.01     0.97          0.02
intSQFTCorner  0.01 0.24    -0.03 -0.04   0.92  0.06      0.01          1.00    -0.02          0.98
intTaxNE       0.57 0.54     0.38  0.83  -0.07  0.66      0.97         -0.02     1.00          0.00
intTaxCorner  -0.01 0.17    -0.03 -0.02   0.95  0.05      0.02          0.98     0.00          1.00
```

Including interactions created a collinearity problem. Check the correlations between the interaction intSQFTCorner, intTaxCorner with Corner!!! So, no you can't really trust the results. Model selection has become very difficult and essentially the interactions compete with the main effects SQFT and Corner making the model very difficult to interpret. The take-home message: be very careful about adding and automate selection of interactions!