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

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?

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

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.

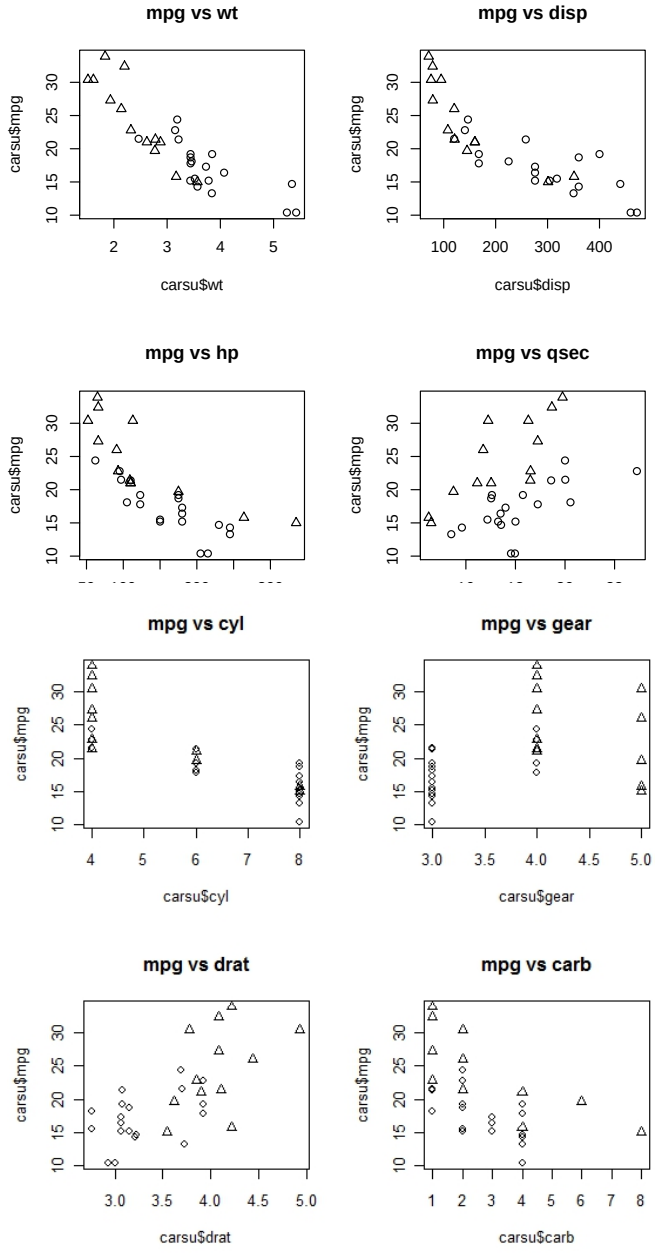

Figure 1: Scatter plots - question 3a

(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
```

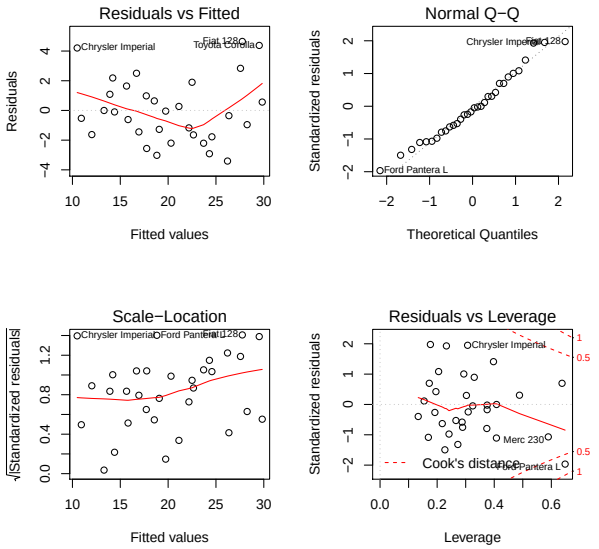

Figure 2: Diagnostic plots - question 3a

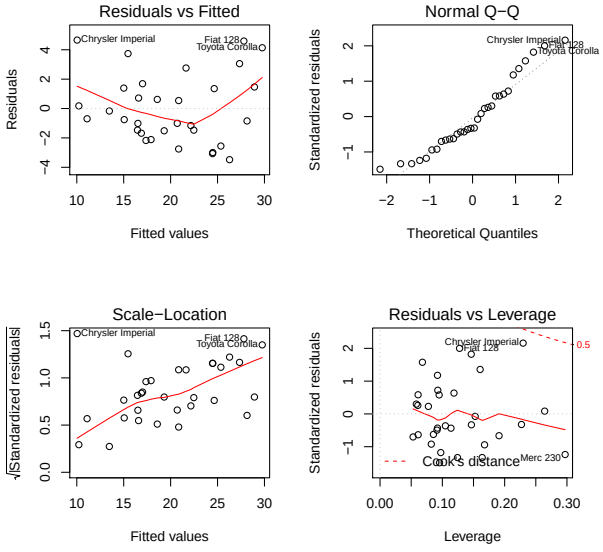

Figure 3: Diagnostic plots, reduced model - Question 3a

(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?

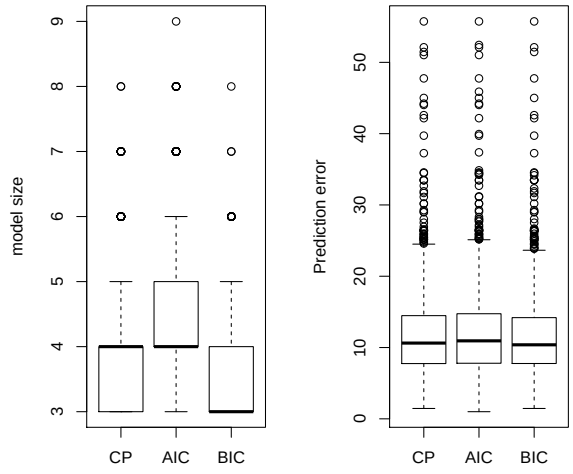

Figure 4: Randomsplits - model sizes and prediction errors - Question 3b. Mean PE is 12, 12.5 and 11.7 for Cp, AIC and BIC respectively.

(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"
```

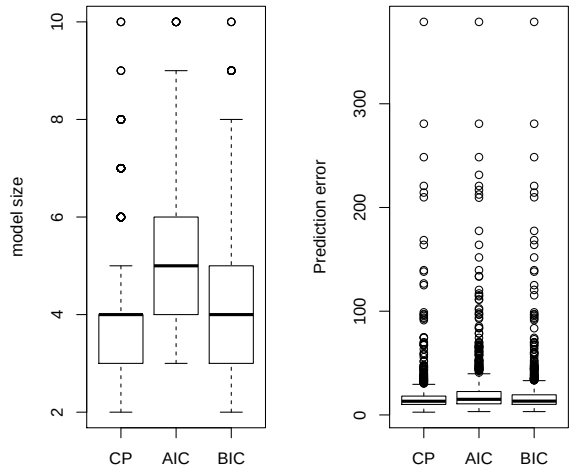

Figure 5: Randomsplits - model sizes and prediction errors - Question 3c. Mean PE is 18, 23 and 20 for Cp, AIC and BIC respectively.

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).

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"
```

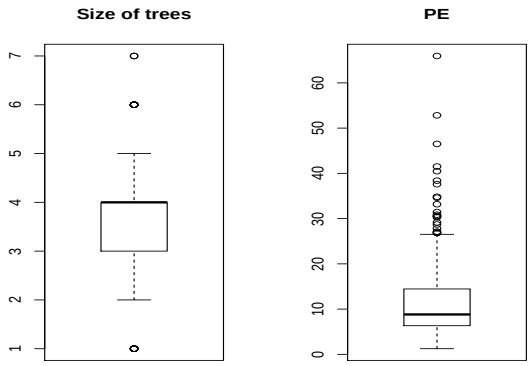

Figure 6: Question 4: Size of trees and Prediction error. Mean PE is 10.9

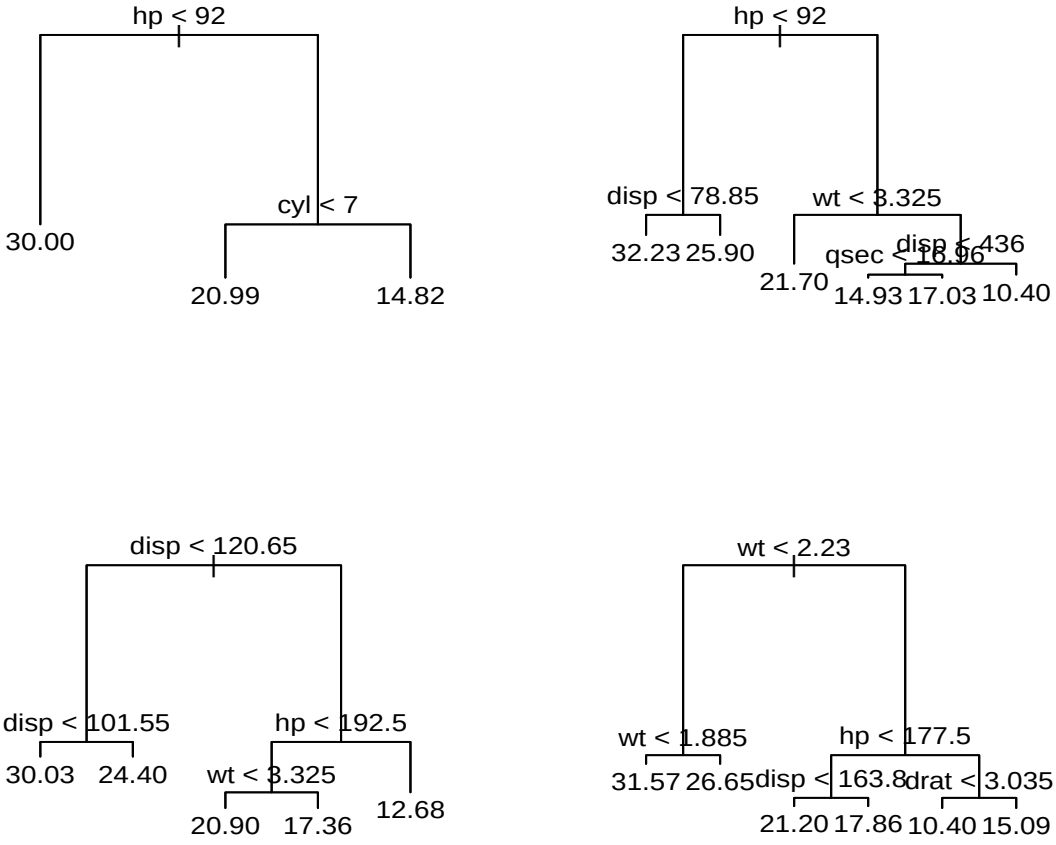

Figure 7: 4 regression trees.

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:

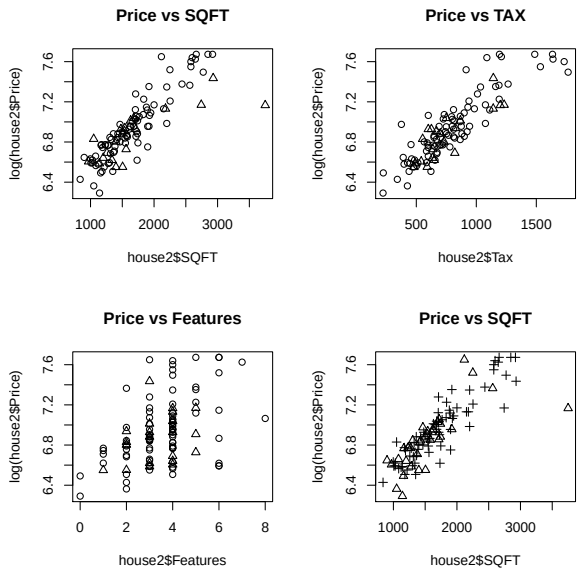

Figure 8: Question 5: Scatter plots. First 3 plots, triangles denote corner plots. Last panel, + denotes NE location.

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
```

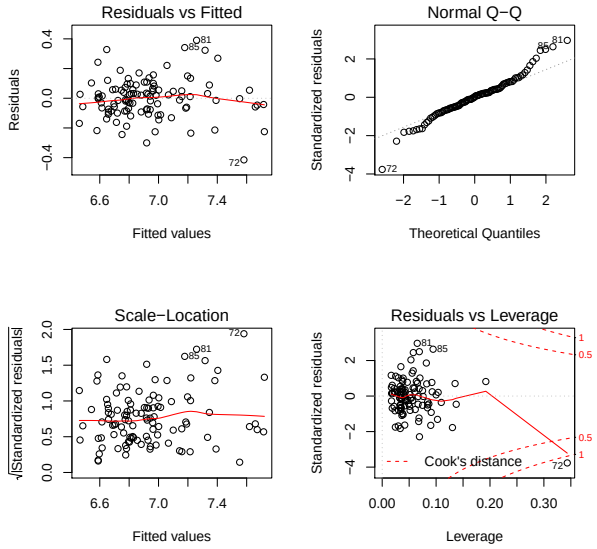

Figure 9: Diagnostic plots - Question 5

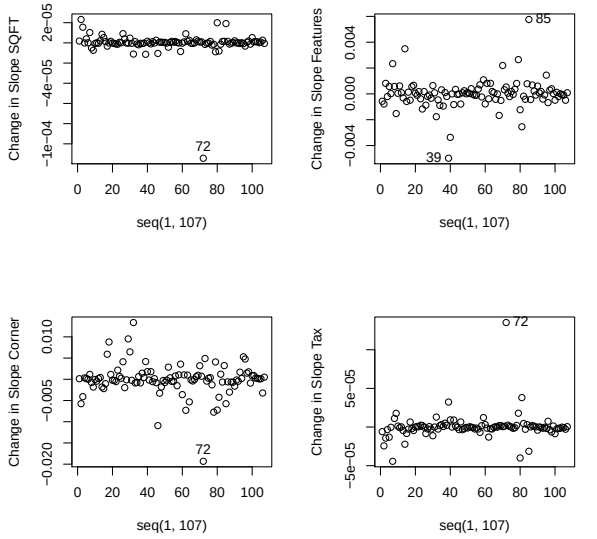

Figure 10: Diagnostic plots - Question 5 - Change in slopes

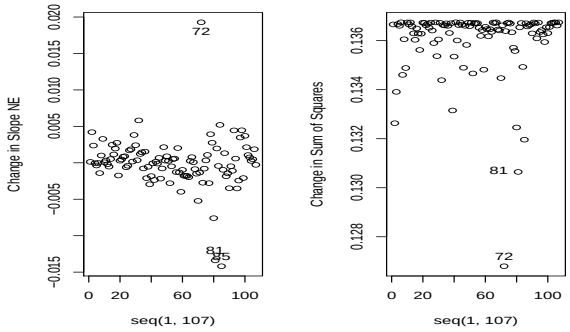

Figure 11: Diagnostic plots - Question 5 - Change in slope and Sigma

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

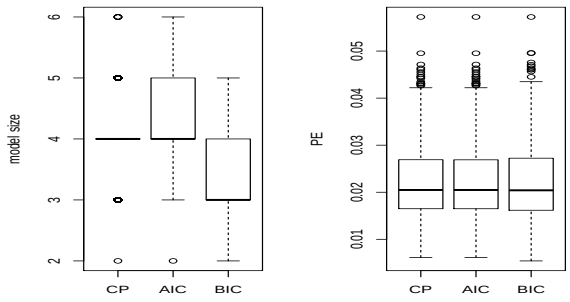

Figure 12: Question 5(c): Model sizes and PE - no interactions

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

```
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"
```

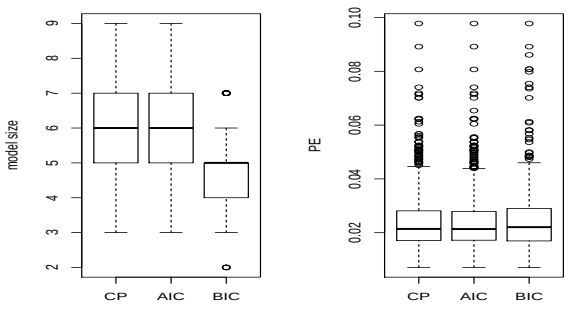

Figure 13: Model sizes and PE - interactions allowed

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