Final MSG500 Dec 15 2011

Question 1: 10p

(a) Based on survey data, a psychologist runs a regression in which the dependent variable is a measure of depression and the independent variables include marital status, employment status, income, gender, and body mass index (weight/height2). He finds that people with a higher body mass index are significantly more depressed, controlling for the other variables. Has he shown that being overweight causes depression? Why or why not?
(b) In a large organization, the average salary for men is \$47,000 while the average salary for women is \$30,000. A t-test shows that this difference is significant at the 0.001 level. When we control for years of work experience, the p-value for the effect of gender changes to 0.17. What would you conclude?

Question 2: 10p

(a) Suppose that the independent variable are transformed according to the equation $X' = X - 10$, and that the response Y is regressed on X'. Expression the regression coefficient estimate you would obtain after transformation in terms of the ones before transformation. Also report the MSE (RSS/n-p) and R-squared after the transformation as a function of it before. What happens if you transform $X' = 10X$.
(b)Suppose you transform the response as $Y' = Y + 10$. Answer the same questions as in (a). What if $Y' = 5 * Y$?

Question 3: 10p

(a) Consider the model for response variable "income" with independent variables "gender" and "education". Suppose you use a dummy variable (1=women, 0=men). Explain the meaning of (interpret) the regression coefficients in an additive model, and in a model with an interaction between gender and education included.
(b) What if you used a variable with (-1=women, 1=men)? Write down the equation with this alternative variable coding and interpret the regression coefficients in this model. Compare with (a).
(c) Does this alternative coding adequately capture the effect of gender? Can we conclude that any model works as long as the regressor has two different values for men and women? Why would you prefer one coding to another?

Question 4: 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 independent variable. Let

Y be the dependent variable. Draw plots (you choose how to make your point) of the following situations:
(a) X1 and X2 are correlated, and there is no interaction between X1 and X2
(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 5: 10p

In a similar class to this one, students were given a lab on GLMs: analyzing binomial data with several predictors. The students fit two candidate models provided by the instructor. The following sentence is a quote from a students lab report: The residual deviances for the two models investigated were quite large, 98 and 117 respectively, indicating that the models did not fit well. Please comment on this statement. Do you agree/disagree? Do you think the statement is informative/not informative? Why/why not? How do the two models compare?

Question 6: 10p

This is a multi-part question.
In each part I present different data scenarios. Regression models are fit to each data set using ordinary least squares. Think carefully about each part. Do questions a) and b) make sense in each scenario? There could be a trick to each question, so make sure to discuss the effect on CIs, coefficient estimates and SE estimates in each case, and how that influences a) and b).

I) Consider a data set to which you fit a regression model. Using residual diagnostic plots you detect a pure outlier (i.e. a large residual, with limited leverage). Consider case A: outlier removed and case B: outlier not removed.
a) Compare the CIs for a coefficient for case A and B.
b) How do expect the outcome of a t-test for a regression coefficient to compare for case A and B? Motivate your answer.
(II) Consider a data set to which you fit a regression model. Using residual diagnostic plots you detect an influential outlier (i.e. an observation with high leverage, and potentially small residual value). Consider case A: outlier removed and case B: outlier not removed. Answer a),b) from part (I).
(III) Consider a data set to which you fit a regression model. One of the explanatory variables is x1 (with corresponding coefficient β_1). Now, add a variable x2 to the regression fit, where x2 is correlated with x1. Case A: only x1 is included in the model. Case B: both x1 and x2 are included in the model.
a) Which case results in the wider CI for β_1? and why?
b) How do expect the outcome of a t-test for β_1 to compare for case A and B? Motivate your answer.
c) How would interpretation of the sign and magnitude of the estimated β_1 compare for case A and B?

Question 7:10p

The pollution data set consists of 60 observations and 16 variables. The outcome is the age-adjusted mortality rate (mort) in a neighborhood. The total set of variables are summarized in the table below.

```
PREC   Average annual precipitation in inches \\
JANT   Average January temperature in degrees F \\
JULT   Same for July\\
OVR65  % of 1960 SMSA population aged 65 or older \\
POPN   Average household size\\
EDUC   Median school years completed by those over 22 \\
HOUS   % of housing units which are sound & with all facilities\\
DENS   Population per sq. mile in urbanized areas, 1960\\
NONW   % non-white population in urbanized areas, 1960\\
WWDRK  % employed in white collar occupations\\
POOR   % of families with income < $3000\\
HC     Relative hydrocarbon pollution potential \\
NOX    Same for nitric oxides\\
SO     Same for sulphur dioxide\\
HUMID  Annual average % relative humidity at 1pm\\
MORT   Total age-adjusted mortality rate per 100,000\\
```

As you can see, there are several climate related variables (prec, humid, jant, jult) as well as several socio-economic factors (educ, hous, dens, nonw). The pollution variables hc, nox and so are of particular interest in this analysis. Does pollution affect mortality rates?

Below are some scatter plots of the data. I also provide the data correlation matrix.

```
Data correlation:
       prec  jant  jult ovr65  popn  educ  hous  dens  nonw wwdrk  poor    hc   nox    so humid  mort
prec   1.00  0.09  0.50  0.10  0.26 -0.49 -0.49  0.00  0.41 -0.30  0.51 -0.53 -0.49 -0.11 -0.08  0.51
jant   0.09  1.00  0.35 -0.40 -0.21  0.12  0.01 -0.10  0.45  0.24  0.57  0.35  0.32 -0.11  0.07 -0.03
jult   0.50  0.35  1.00 -0.43  0.26 -0.24 -0.42 -0.06  0.58 -0.02  0.62 -0.36 -0.34 -0.10 -0.45  0.28
ovr65  0.10 -0.40 -0.43  1.00 -0.51 -0.14  0.07  0.16 -0.64 -0.12 -0.31 -0.02  0.00  0.02  0.11 -0.17
popn   0.26 -0.21  0.26 -0.51  1.00 -0.40 -0.41 -0.18  0.42 -0.43  0.26 -0.39 -0.36  0.00 -0.14  0.36
educ  -0.49  0.12 -0.24 -0.14 -0.40  1.00  0.55 -0.24 -0.21  0.70 -0.40  0.29  0.22 -0.23  0.18 -0.51
hous  -0.49  0.01 -0.42  0.07 -0.41  0.55  1.00  0.18 -0.41  0.34 -0.68  0.39  0.35  0.12  0.12 -0.43
dens   0.00 -0.10 -0.06  0.16 -0.18 -0.24  0.18  1.00 -0.01 -0.03 -0.16  0.12  0.17  0.43 -0.12  0.27
nonw   0.41  0.45  0.58 -0.64  0.42 -0.21 -0.41 -0.01  1.00  0.00  0.70 -0.03  0.02  0.16 -0.12  0.64
wwdrk -0.30  0.24 -0.02 -0.12 -0.43  0.70  0.34 -0.03  0.00  1.00 -0.19  0.20  0.16 -0.07  0.06 -0.28
poor   0.51  0.57  0.62 -0.31  0.26 -0.40 -0.68 -0.16  0.70 -0.19  1.00 -0.13 -0.10 -0.10 -0.15  0.41
hc    -0.53  0.35 -0.36 -0.02 -0.39  0.29  0.39  0.12 -0.03  0.20 -0.13  1.00  0.98  0.28 -0.02 -0.18
nox   -0.49  0.32 -0.34  0.00 -0.36  0.22  0.35  0.17  0.02  0.16 -0.10  0.98  1.00  0.41 -0.05 -0.08
so    -0.11 -0.11 -0.10  0.02  0.00 -0.23  0.12  0.43  0.16 -0.07 -0.10  0.28  0.41  1.00 -0.10  0.43
humid -0.08  0.07 -0.45  0.11 -0.14  0.18  0.12 -0.12 -0.12  0.06 -0.15 -0.02 -0.05 -0.10  1.00 -0.09
mort   0.51 -0.03  0.28 -0.17  0.36 -0.51 -0.43  0.27  0.64 -0.28  0.41 -0.18 -0.08  0.43 -0.09  1.00
```

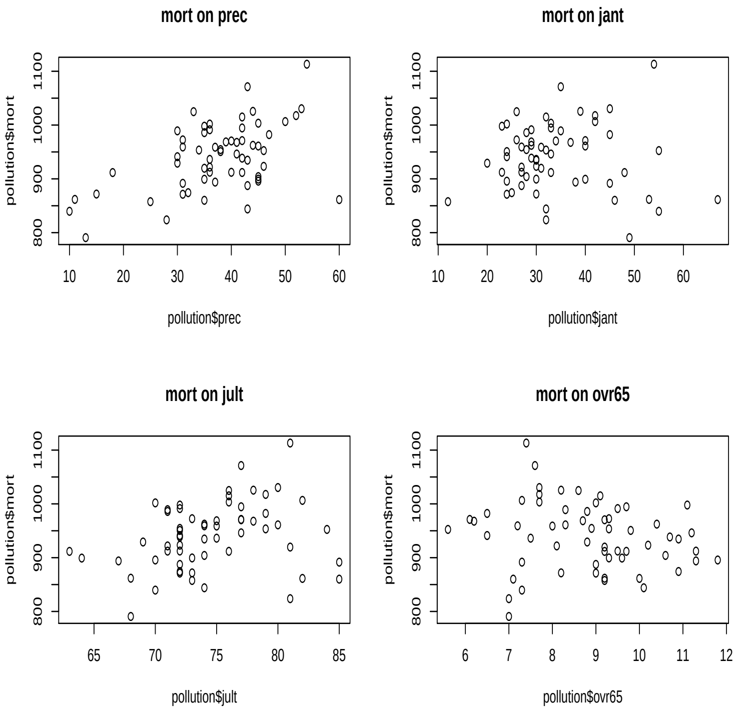


Figure 1: Scatter plot 1

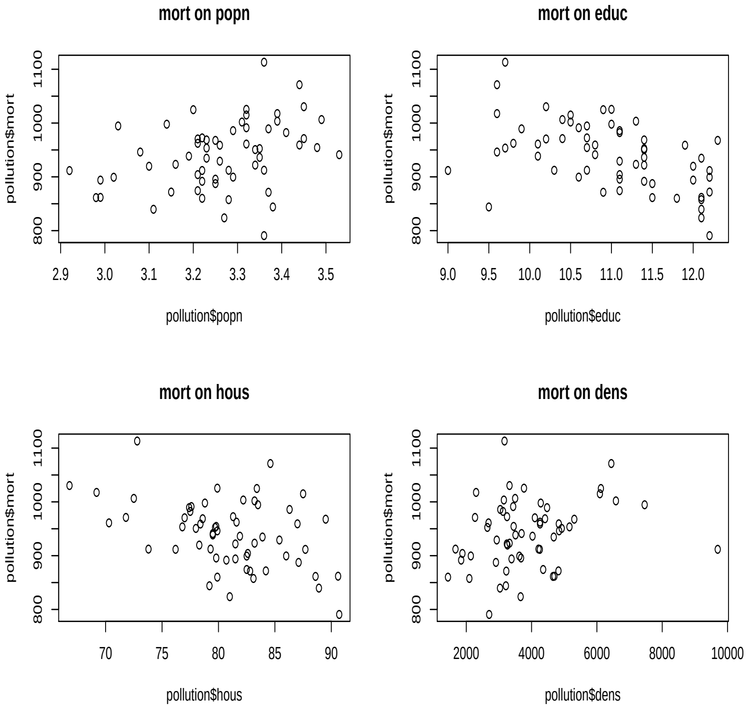


Figure 2: Scatter plot 2

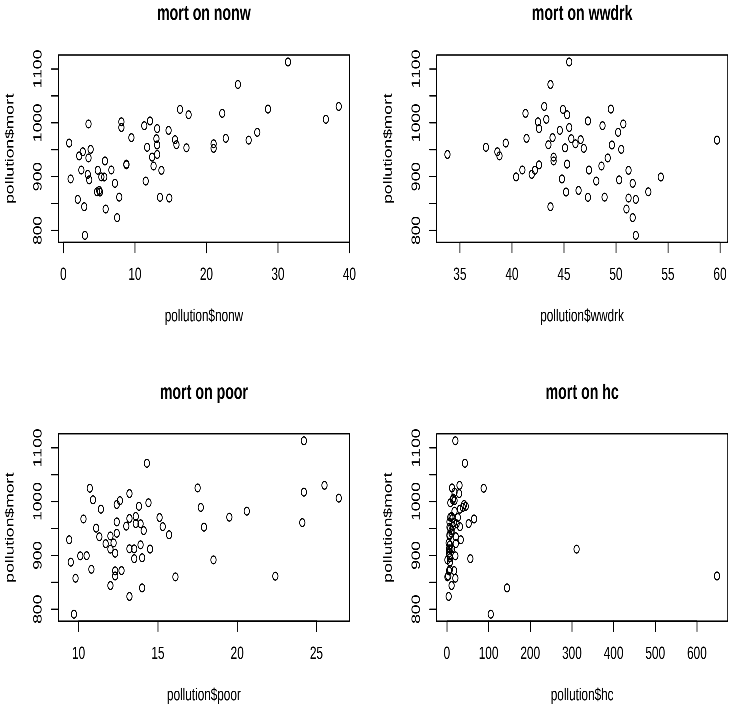


Figure 3: Scatter plot 3

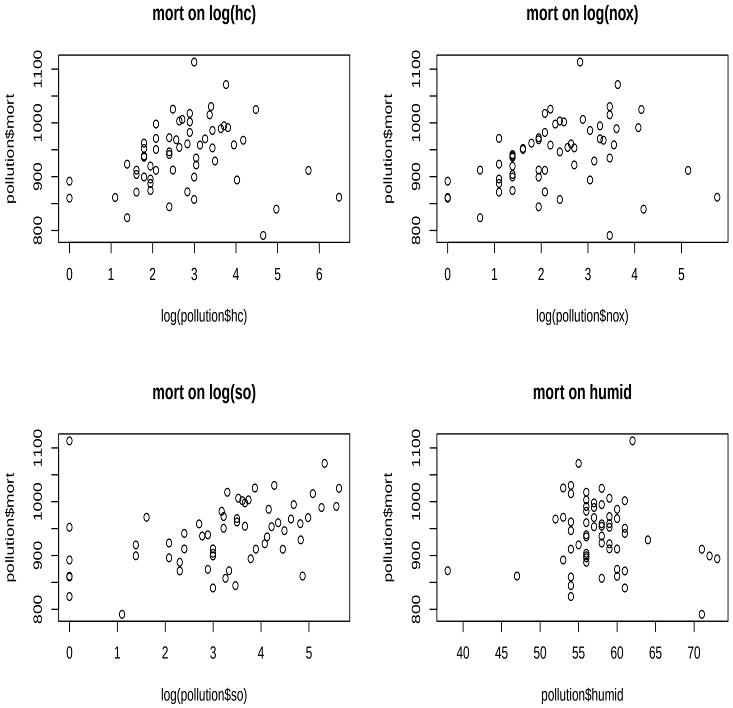


Figure 4: Scatter plot 4

```
Data correlation after variable transformation.
       prec  jant  jult ovr65  popn  educ  hous  dens  nonw wwdrk  poor    hc   nox    so humid  mort
prec   1.00  0.09  0.50  0.10  0.26 -0.49 -0.49  0.00  0.41 -0.30  0.51 -0.46 -0.37 -0.12 -0.08  0.51
jant   0.09  1.00  0.35 -0.40 -0.21  0.12  0.01 -0.10  0.45  0.24  0.57  0.17  0.13 -0.34  0.07 -0.03
jult   0.50  0.35  1.00 -0.43  0.26 -0.24 -0.42 -0.06  0.58 -0.02  0.62 -0.46 -0.36 -0.36 -0.45  0.28
ovr65  0.10 -0.40 -0.43  1.00 -0.51 -0.14  0.07  0.16 -0.64 -0.12 -0.31 -0.11 -0.07  0.22  0.11 -0.17
popn   0.26 -0.21  0.26 -0.51  1.00 -0.40 -0.41 -0.18  0.42 -0.43  0.26 -0.16 -0.10  0.00 -0.14  0.36
educ  -0.49  0.12 -0.24 -0.14 -0.40  1.00  0.55 -0.24 -0.21  0.70 -0.40  0.15  0.02 -0.26  0.18 -0.51
hous  -0.49  0.01 -0.42  0.07 -0.41  0.55  1.00  0.18 -0.41  0.34 -0.68  0.31  0.23  0.06  0.12 -0.43
dens   0.00 -0.10 -0.06  0.16 -0.18 -0.24  0.18  1.00 -0.01 -0.03 -0.16  0.30  0.35  0.48 -0.12  0.27
nonw   0.41  0.45  0.58 -0.64  0.42 -0.21 -0.41 -0.01  1.00  0.00  0.70  0.14  0.19  0.05 -0.12  0.64
wwdrk -0.30  0.24 -0.02 -0.12 -0.43  0.70  0.34 -0.03  0.00  1.00 -0.19  0.20  0.10 -0.12  0.06 -0.28
poor   0.51  0.57  0.62 -0.31  0.26 -0.40 -0.68 -0.16  0.70 -0.19  1.00 -0.15 -0.09 -0.19 -0.15  0.41
hc    -0.46  0.17 -0.46 -0.11 -0.16  0.15  0.31  0.30  0.14  0.20 -0.15  1.00  0.95  0.64  0.21  0.15
nox   -0.37  0.13 -0.36 -0.07 -0.10  0.02  0.23  0.35  0.19  0.10 -0.09  0.95  1.00  0.73  0.12  0.29
so    -0.12 -0.34 -0.36  0.22  0.00 -0.26  0.06  0.48  0.05 -0.12 -0.19  0.64  0.73  1.00 -0.07  0.40
humid -0.08  0.07 -0.45  0.11 -0.14  0.18  0.12 -0.12 -0.12  0.06 -0.15  0.21  0.12 -0.07  1.00 -0.09
mort   0.51 -0.03  0.28 -0.17  0.36 -0.51 -0.43  0.27  0.64 -0.28  0.41  0.15  0.29  0.40 -0.09  1.00
```

(a). After looking at the scatterplots (Figures 1-3) I decide to transform some of the variables (Figure 4). Comment on this choice of transform and the result. Compare the resulting data correlation matrix I obtain after transformation. Are there any 'big' changes you think will have an impact on modeling?

(b). I fit a linear model to the data, summarized in the table below. In Figure 8, I provide the basic diagnostic plots. Explain the results in the table and figures. Give me an interpretation of the mortality data based on this model - which factors influence mortality and how? Do the coefficient estimates (sign and magnitude) make sense to you? Why/why not? Are there any 'surprises' in the model and can you, if so, identify the source of these?

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.869e+03  4.850e+02   3.854 0.000762 ***
prec         1.915e+00  9.737e-01   1.967 0.060904 .
jant        -4.688e+00  1.116e+00  -4.202 0.000316 ***
jult        -4.720e+00  2.089e+00  -2.260 0.033174 *
ovr65       -8.218e+00  9.185e+00  -0.895 0.379794
popn        -1.548e+02  7.545e+01  -2.052 0.051225 .
educ        -7.515e+00  1.311e+01  -0.573 0.571902
hous         4.801e-01  1.765e+00   0.272 0.787915
dens         8.002e-03  4.036e-03   1.983 0.058935 .
nonw         7.031e+00  1.602e+00   4.388 0.000197 ***
wwdrk       -7.151e-01  1.922e+00  -0.372 0.713110
poor         5.665e+00  3.045e+00   1.861 0.075104 .
hc          -3.812e+01  1.668e+01  -2.285 0.031442 *
nox          5.851e+01  1.861e+01   3.145 0.004387 **
so          -1.628e+01  6.879e+00  -2.366 0.026380 *
humid       -1.853e-01  1.197e+00  -0.155 0.878224
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 27.85 on 24 degrees of freedom
Multiple R-squared: 0.888,      Adjusted R-squared: 0.818
F-statistic: 12.69 on 15 and 24 DF,  p-value: 5.938e-08
```

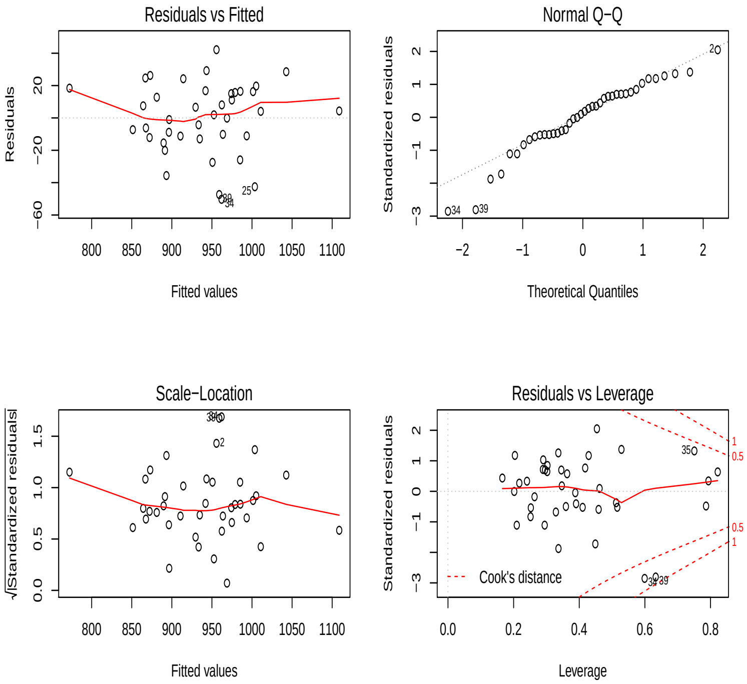


Figure 5: Diagnostic plots

Question 8: 10p

Continuing with question 7.

(a). In Figures 6 and 7, I depict the Cook's distance, change in slope after dropping an observation (for two select slope parameters) and the change in σ^2. Explain in what way these outlier detection measures identifies outliers. Draw a cartoon picture of how these measures are computed.

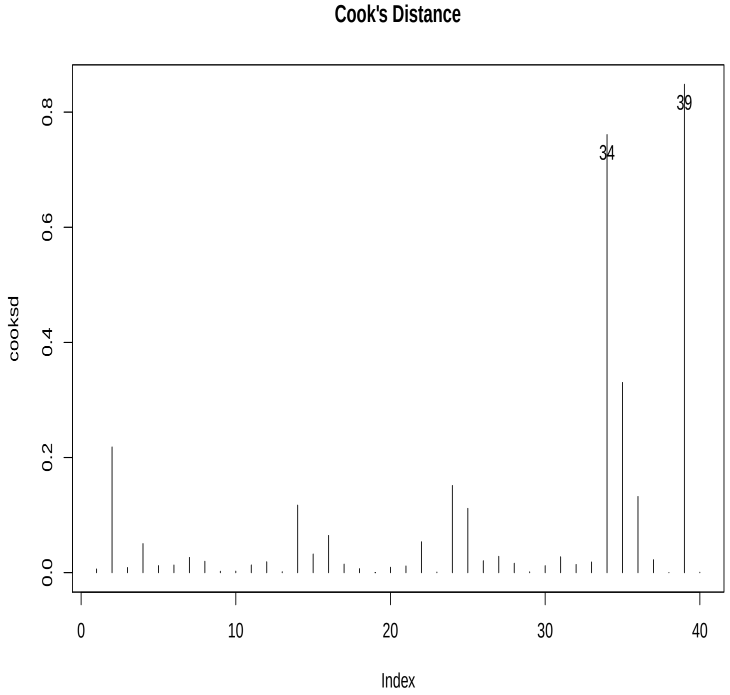


Figure 6: Cook's distance

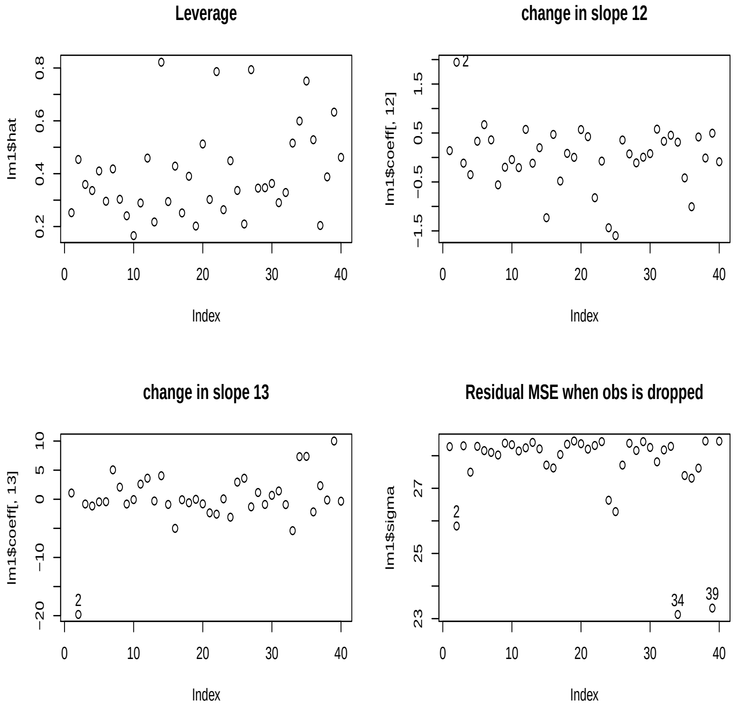
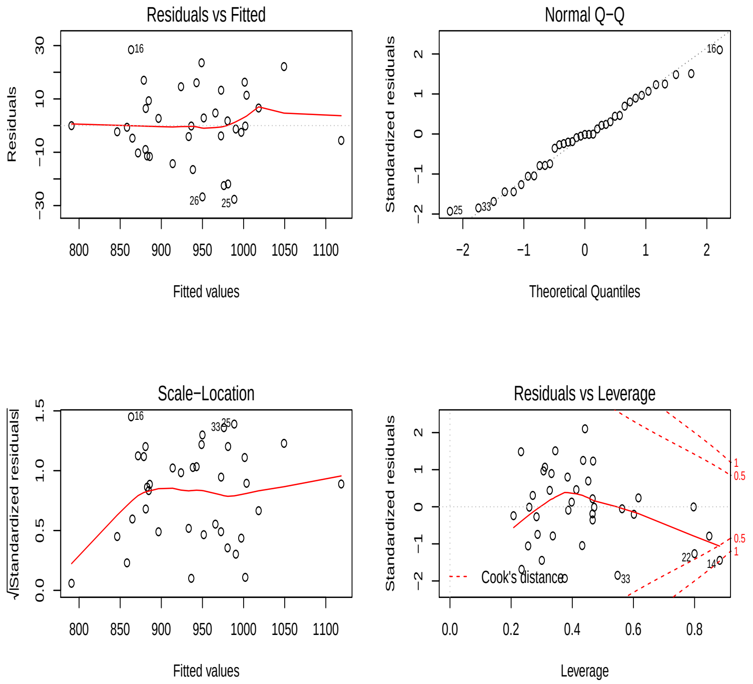


Figure 7: Diagnostic plots 2

(b). I removed three observations from the modeling and updated the fit. Compare the model summary and residual diagnostics below to the ones above. Do you spot any more problems that need to be addressed? Did the model interpretation change? If so, how? If not, explain how you arrive at that conclusion.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.699e+03  3.466e+02   7.786 1.27e-07 ***
prec         1.373e+00  6.788e-01   2.023 0.056034 .
jant        -3.802e+00  7.448e-01  -5.104 4.69e-05 ***
jult        -6.758e+00  1.480e+00  -4.566 0.000168 ***
ovr65       -1.566e+01  6.398e+00  -2.448 0.023233 *
popn        -2.502e+02  5.300e+01  -4.721 0.000116 ***
educ        -1.461e+01  8.660e+00  -1.687 0.106317
hous        -7.910e-01  1.214e+00  -0.651 0.521867
dens         1.233e-02  3.547e-03   3.476 0.002255 **
nonw         7.354e+00  1.249e+00   5.888 7.63e-06 ***
wwdrk       -2.620e+00  1.420e+00  -1.844 0.079287 .
poor         3.607e+00  2.074e+00   1.739 0.096635 .
hc          -3.604e+01  1.309e+01  -2.753 0.011919 *
nox          5.634e+01  1.437e+01   3.922 0.000783 ***
so          -1.920e+01  4.561e+00  -4.210 0.000394 ***
humid       -5.737e-01  8.023e-01  -0.715 0.482407
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 18.11 on 21 degrees of freedom
Multiple R-squared: 0.9573,     Adjusted R-squared: 0.9268
F-statistic: 31.39 on 15 and 21 DF,  p-value: 5.125e-11
```



Residuals vs Fitted

Normal Q−Q

Scale−Location

Residuals vs Leverage

Figure 8: Diagnostic plots

Question 9: 10p

Continuing questions 7 and 8.
(a). Stepwise model selection results in the model below. Most of the variables are kept in the model. Discuss why that might be.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.595e+03  3.157e+02   8.220 2.69e-08 ***
prec         1.308e+00  6.570e-01   1.991 0.058528 .
jant        -4.119e+00  6.308e-01  -6.530 1.16e-06 ***
jult        -6.509e+00  1.338e+00  -4.866 6.51e-05 ***
ovr65       -1.613e+01  6.202e+00  -2.600 0.015998 *
popn        -2.498e+02  5.022e+01  -4.974 4.98e-05 ***
educ        -1.682e+01  7.912e+00  -2.126 0.044438 *
dens         1.223e-02  3.091e-03   3.958 0.000625 ***
nonw         7.274e+00  1.175e+00   6.193 2.56e-06 ***
wwdrk       -2.296e+00  1.339e+00  -1.714 0.099983 .
poor         4.674e+00  1.559e+00   2.998 0.006422 **
hc          -3.680e+01  1.265e+01  -2.909 0.007912 **
nox          5.539e+01  1.395e+01   3.970 0.000607 ***
so          -1.841e+01  4.343e+00  -4.239 0.000310 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 17.64 on 23 degrees of freedom
Multiple R-squared: 0.9556,     Adjusted R-squared: 0.9305
F-statistic:  38.1 on 13 and 23 DF,  p-value: 1.99e-12
```

(b). I use model selection criteria Cp, AIC and BIC to select a model for the data. The results are provided below. Comment and compare to the findings in question 7. Would you say that model selection is "easy" or "difficult" for this data set? Why/why not? Would you say that coming up with a good prediction model is "easy" or "difficult" for this data set? Why/why not? Motivate your answers - on what basis do you arrive at your conclusion?

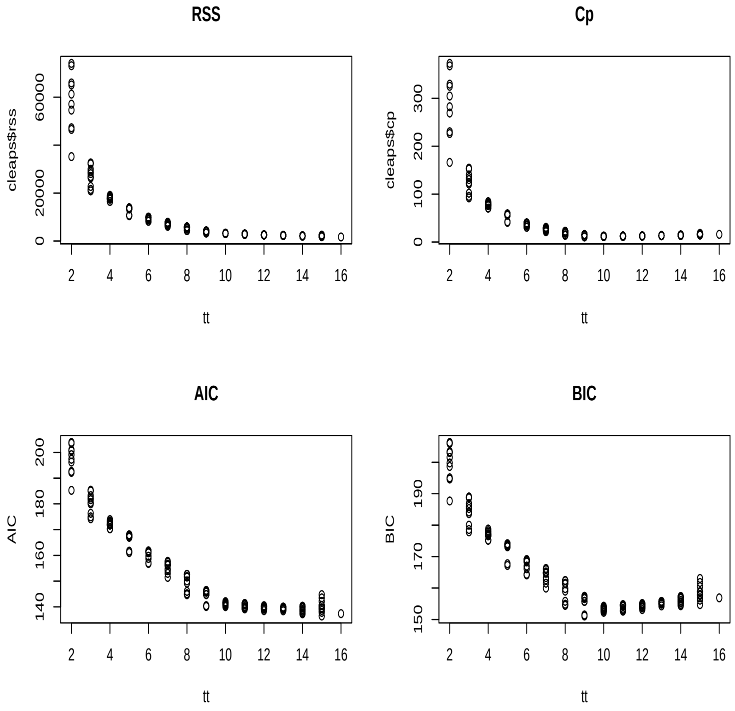


Figure 9: Selection criteria

```
"PE and size of selected models"
[1] "PECP=" "1699.9334" "size=" "9"
[1] "PEAIC=""1801.2986" "size=" "15"
[1] "PEBIC=""1699.9334" "size=" "9"
[1] "CP model" "jant" "jult" "popn" "educ" "hous" "dens" "nonw" "wwdrk"
[1] "AIC model""prec" "jant" "jult" "ovr65""popn" "educ" "hous" "dens" "nonw" "wwdrk" "poor" "
[1] "BIC model""jant" "jult" "popn" "educ" "hous" "dens" "nonw" "wwdrk"
```

(c). I use 2/3 random splits of the data and repeat the model selection 100 times. In Figures 10 I provide the relative model size obtained with AIC and BIC, and also the ratio of the prediction error I obtain with the AIC model over the one I obtain with the BIC model (for each random split). Below is also a table with selected variables. Interpret these results.

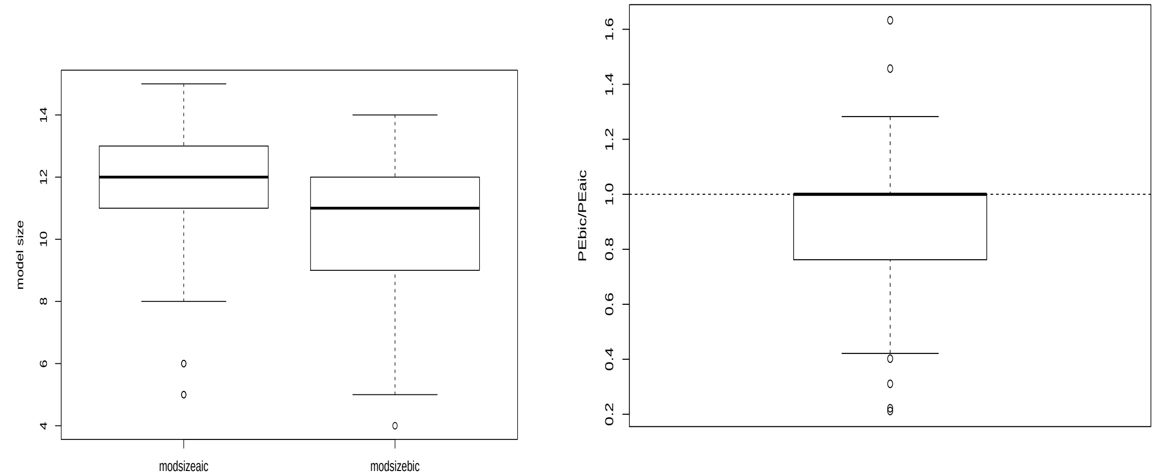


Figure 10: Left: Model size aic and bic. Right: Ratio of PE with AIC models compared with BIC models

```
        bic  aic
 [1,] "prec"  "38" "58"
 [2,] "jant"  "93" "98"
 [3,] "jult"  "90" "95"
 [4,] "ovr65" "64" "82"
 [5,] "popn"  "95" "99"
 [6,] "educ"  "58" "71"
 [7,] "hous"  "26" "38"
 [8,] "dens"  "90" "93"
 [9,] "nonw"  "98" "99"
[10,] "wwdrk" "69" "81"
[11,] "poor"  "55" "71"
[12,] "hc"    "52" "64"
[13,] "nox"   "77" "92"
[14,] "so"    "82" "93"
[15,] "humid" "30" "39"
```

Question 10: 10p

(a.) I use CART to model the pollution data. Below I depict a tree model and a pruned version, chosen to minimize the CV error. Explain the meaning of the tree models and interpret. Compare with the findings in Question 7.

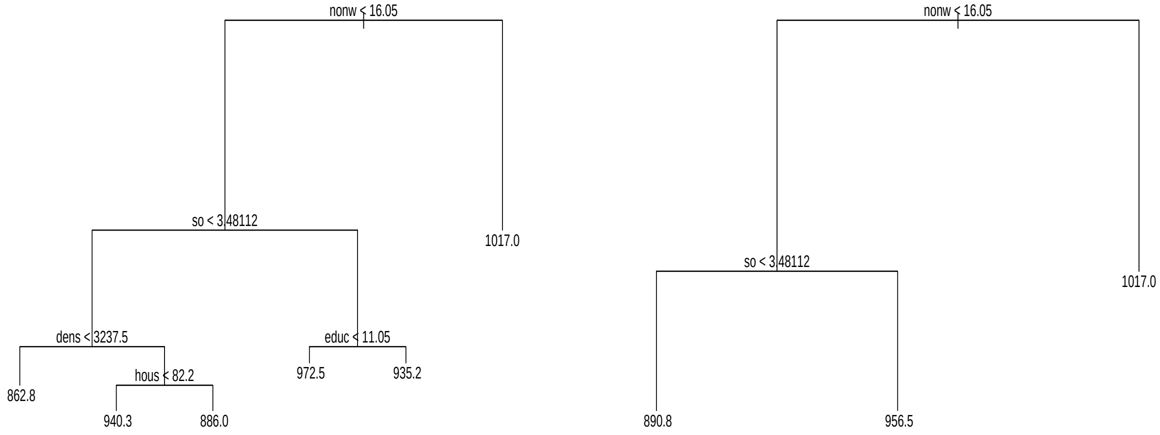


Figure 11: CART model and pruned tree

(b.) I repeat the exercise on random splits of data and repeat the CART modeling and pruning each time. I summarize the findings below. Explain the table and interpret the results. Comment on the similarities and differences you see between the results here and the model selection results using a linear model (Question 9).

```
              modfirst  modtab
 [1,] "prec"  "0.01"    "0.36"
 [2,] "jant"  "0"       "0.13"
 [3,] "jult"  "0"       "0"
 [4,] "ovr65" "0"       "0.13"
 [5,] "popn"  "0"       "0.04"
 [6,] "educ"  "0.13"    "0.39"
 [7,] "hous"  "0"       "0.12"
 [8,] "dens"  "0"       "0.16"
 [9,] "nonw"  "0.75"    "0.85"
[10,] "wwdrk" "0.02"    "0.18"
[11,] "poor"  "0"       "0.02"
[12,] "hc"    "0"       "0.07"
[13,] "nox"   "0"       "0.23"
[14,] "so"    "0.09"    "0.36"
[15,] "humid" "0"       "0.06"
```