Solutions 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?

No, association is not the same as causation. There may be another underlying factor that causes depression and increased bmi. Or perhaps depression causes increased bmi?

(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?

Controlling for years of experience essentially means that you compare salaries for men and women with the same amount of experience. We run regression of income on years of experience and then run regression of those residuals on gender. This approach "corrects" the salary based on years of experience and then check if gender is a predictive variable of the remaining variation. After this correction the gender difference is not significant.
However, to better understand what's going on you have to check the data. Is years of experience and gender correlated? Does your data have men and women for all ranges of years of experience, but more men with more experience? Perhaps this company has senior management that is all male and junior management is mixed - are salaries matched among junior management? Think about what regression does if the groups are not represented across the full range of experience. We do know that it not the case that all senior management are men and senior administration staff are all women since then controlling for years of experience would not have changed the outcome.

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 $X' = 10X$.

With $X' = X - 10$, the regression coefficient is unchanged since

$$Y = \beta_0' + \beta_1' X' + \epsilon = \beta_0' - 10\beta_1' + \beta_1' X + \epsilon$$

so $\beta_1' = \beta_1$, the X coefficient, but the intercept has changed. The MSE and R-squared are also unchanged since the error is additive and not affected by translation of X.
If $X' = 10X$, then $\beta_1' = \beta_1/10$ but MSE and R-squared are still unchanged since this transformation does not affect the additive errors.

(b)Suppose you transform the response as $Y' = Y + 10$. Answer the same questions as in (a). What if $Y' = 5 * Y$?

If $Y' = Y + 10$ we again see only an effect on the intercept, not on MSE or R-squared.
If $Y' = 5 * Y$ you have changed the scale of Y. This means that coefficients in the model are scaled by 5 as well as is the error. So MSE is now 25 times as large! The R-squared is unchanged since both the scale of Y and the residuals changed by the same factor (cancels out in the calculation for R-squared).

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.

$$income = \beta_0 + \beta_g * I_{gender} + \beta_{edu} * education + \epsilon$$

where β_g is the income difference between women to men (if β_g positive, then women have a higher income and v.v.), β_{edu} is the income increase you would expect for every additional "unit" of education.

$$income = \beta_0 + \beta_g * I_{gender} + \beta_{edu} * education + \beta_{int} gender * education + \epsilon$$

The coefficient β_{int} is the additional income increase per unit of education increase for women.
The model for women states:

$$income = \beta_0 + \beta_g + (\beta_{edu} + \beta_{int}) * education$$

and for men

$$income = \beta_0 + \beta_{edu} * education$$

(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).

$$income = \beta_0 + \beta_g * I_{gender} + \beta_{edu} * education + \epsilon$$

Now, income for women can be written as $\beta_0 - \beta_g + \beta_{edu} * education$ whereas for men $\beta_0 + \beta_g + \beta_{edu} * education$. So the income difference between men and women is now $2\beta_g$ and $-\beta_g$ is the

difference in income for women from the population average (β_0).
Similarly, if we include an interaction the model for women states

$$income = \beta_0 - \beta_g + (\beta_{edu} - \beta_{int}) * education$$

and for men

$$income = \beta_0 + \beta_g + (\beta_{edu} + \beta_{int}) * education$$

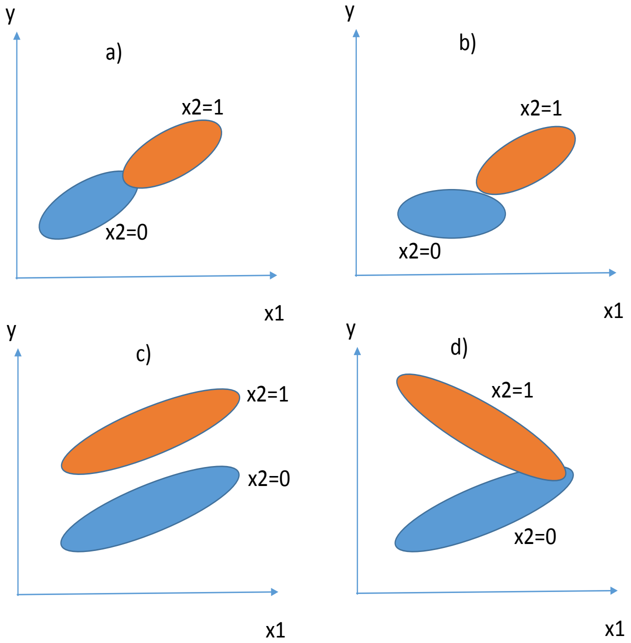
(c) Does this alternative coding adequately capture the effect of gender? Can we conclude that any model works as long as the regressor has two different values for men and women? Why would you prefer one coding to another?

Both models work and can be used for investigating the income as a function of gender and education and to assess the significance of the gender effect and the interaction (captured by the β_g and β_{int}. In a) the baseline population is the male population. In b) the baseline is the average of women and men. It depends on the case at hand how you want to formulate the model, if the baseline has meaning or if only the group differences are of interest.

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

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.


y
a)
x2=1
x2=0
x1
y
b)
x2=1
x2=0
x1
y
c)
x2=1
x2=0
x1
y
d)
x2=1
x2=0
x1

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?

This statement is meaningless since the degrees of freedom is not provided. Without this information you cannot tell if the residual deviance is large or not. You can tell that the second model is a worse fit, but you have no idea if it significantly worse without the degrees of freedom, nor do you know if an ANODEV is appropriate since it is not stated if the models are nested or not.

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.
a) The outlier will not affect the coefficient estimates but it will inflate the MSE. Therefore, CIs will be wider since the SEs will be larger.
b) Since case A (outlier removed) will have smaller SEs and more narrow CIs this corresponds to a larger t-value. So the power of the t-test is greater for case A.

(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).
a) The coefficient estimates may not be effected. Let's say that the outlier pulled the regression line toward the x-axis for a regression coefficient. Then the slope estimate in A will be larger than that in B (and v.v if the outlier pulled the regression line away from the x-axis).
Since the outlier pulls the regression line away from the mass of data, the MSE will be larger in B than in A. However, the SE for a regression coefficient also depends on the spread of x. In case A this spread is reduced. The SE has MSE in the numerator and spread of x in the denominator. In case A both quantities are smaller which could lead to the SE being bigger or smaller. Therefore it

is not clear if the CI will be wider in B or in A.
b) The outcome of the t-test depends on if the outlier pulls the line away or toward the x-axis. So the t-test can lead to an erroneous rejection of a true null or the failure to reject a false null in case A.

(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?
case B generally leads to wider CIs because you have added a correlated variable. However, if x2 is in fact related to y and can improve the regression fit, then MSE will decrease and the CIs may become more narrow. This depends therefore on the degree of correlation between x1 and x2 and the strength of the relationship between y and x2.
b) How do expect the outcome of a t-test for β_1 to compare for case A and B? Motivate your answer.
Including a correlated x-variable unrelated to y in the model will reduce the power of the t-test for β_1. *In general therefore case B will have less power. Again, if x2 is related to y it's more complicated.*
c) How would interpretation of the sign and magnitude of the estimated β_1 compare for case A and B?
You have to be careful about interpreting β_1 *in the presence of collinearities! The sign and magnitude of* β_1 *can be almost random if x1 and x2 are highly correlated. Of course, if x2 is strongly related to y it is equally misleading to omit it from the model since then* β_1 *will be capturing the effect of x2 as well as x1 on y.*
To summarize case III: it depends on the degree of correlation between x1 and x2 and their relationship to y!

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
```

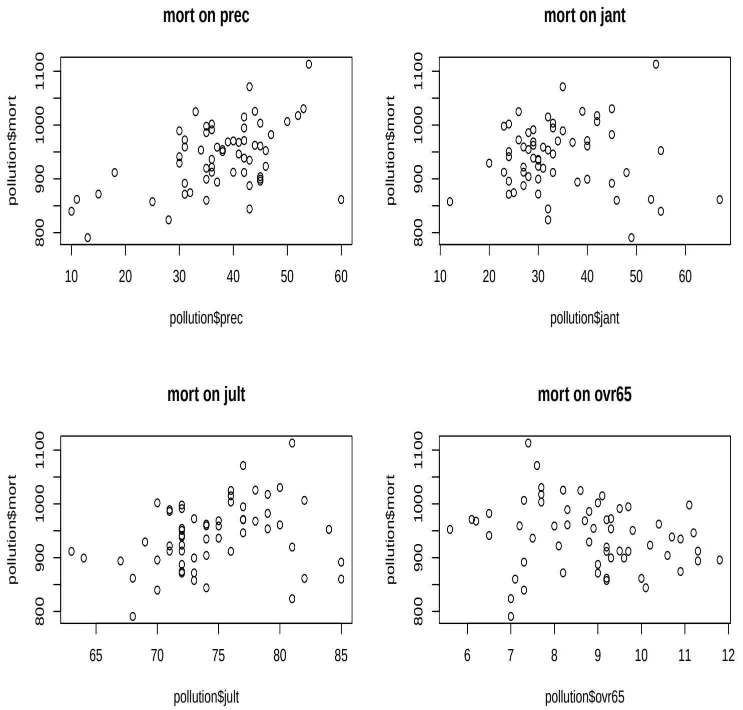


Figure 1: Scatter plot 1

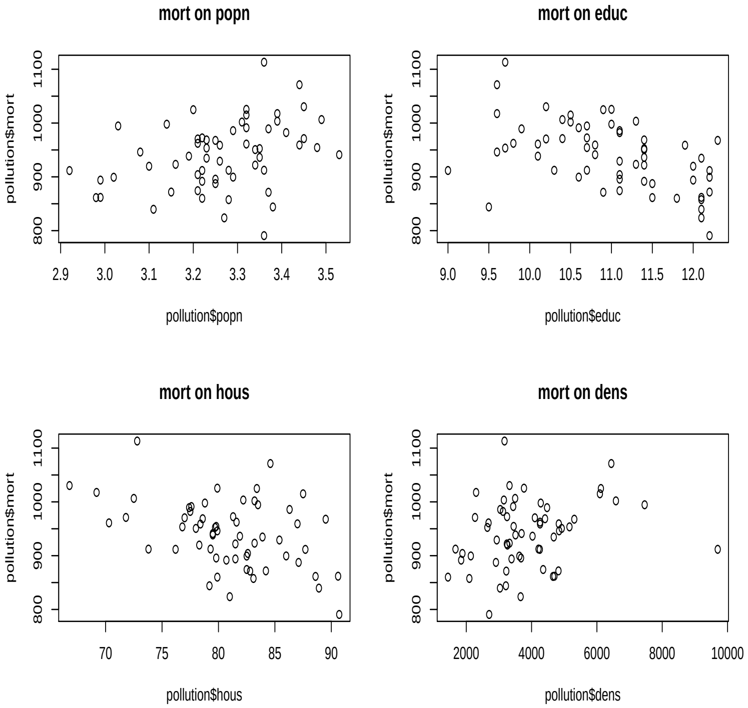


Figure 2: Scatter plot 2

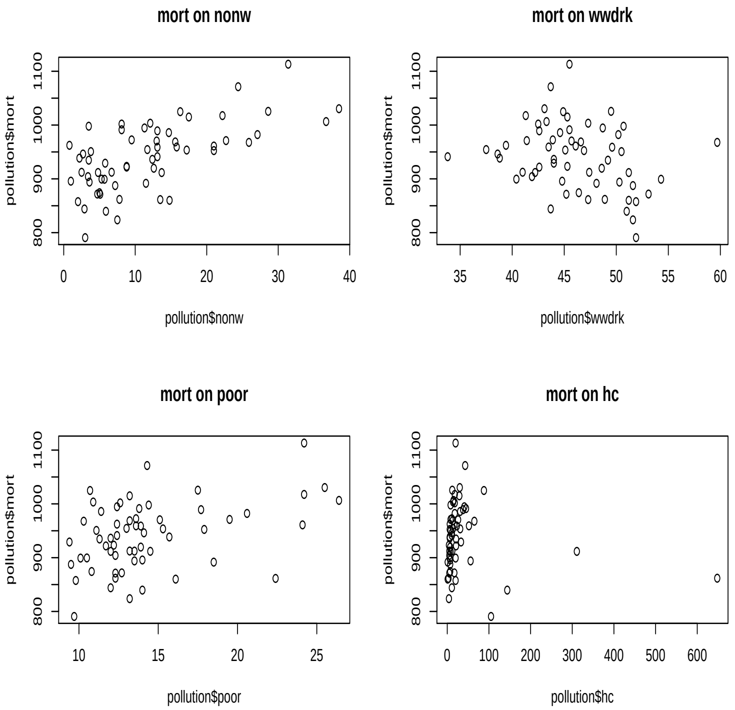


Figure 3: Scatter plot 3

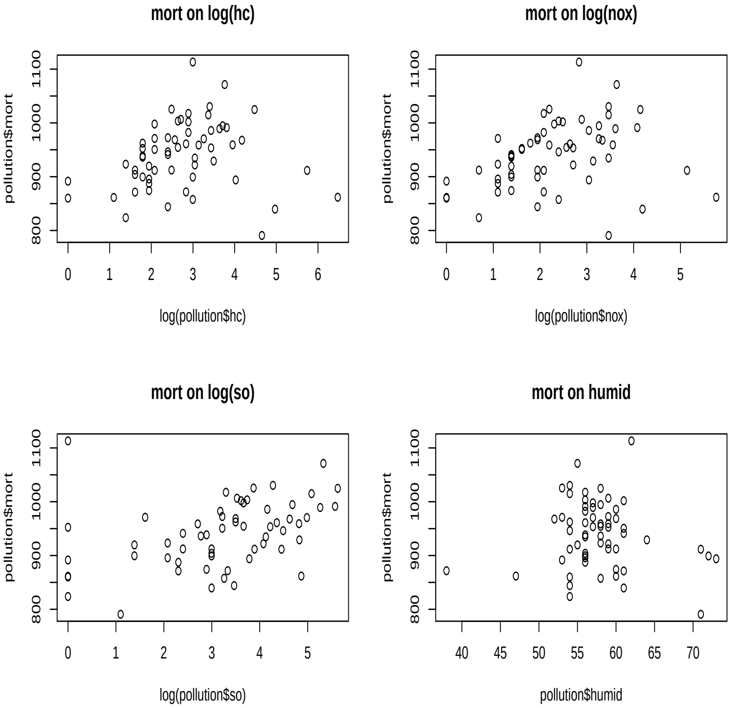


Figure 4: Scatter plot 4

```
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?

In plot 3 bottom right panel you see that it was necessary to transform hc since the spread was too uneven otherwise. In figure 4 top left we see after transformation an association between mortality and log(hc). There are no plots provided for nox and so, so we cannot tell if the log made the association more clear.

In the correlation matrix we can see that the log-transform increased the correlation between nox and mortality and for hc and mortality but actually decreased the correlation with so somewhat. In addition, prior to transformation hc and nox were highly correlated (0.98) and moderately so with so. After transformation hc and nox are still highly correlated (.95) and the correlation with so has increased.

Summary: the scatter plot for hc indicates that transformation of this variable at least was the way to go. BUT we can see from the correlation matrices that we have also created a bigger collinearity problem than before.

(b). I fit a linear model to the data, summarized in the table below. In Figure 5, 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?

There appears to be a few outliers present, with both high leverage and high residual value. But there is no trend in the residuals and no problem with nonconstant error variance.

The R-squared is quite large (adjusted 82%). The are a few coefficients that are significant. We see that higher temperatures are associated with lower mortality, whereas higher proportion of non-whites in a district is associated with higher mortality. There is a problem with the correlated hc and nox variables that appear in the model with opposite sign for their coefficients. This is a clear sign of a collinearity problem!

```
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
```

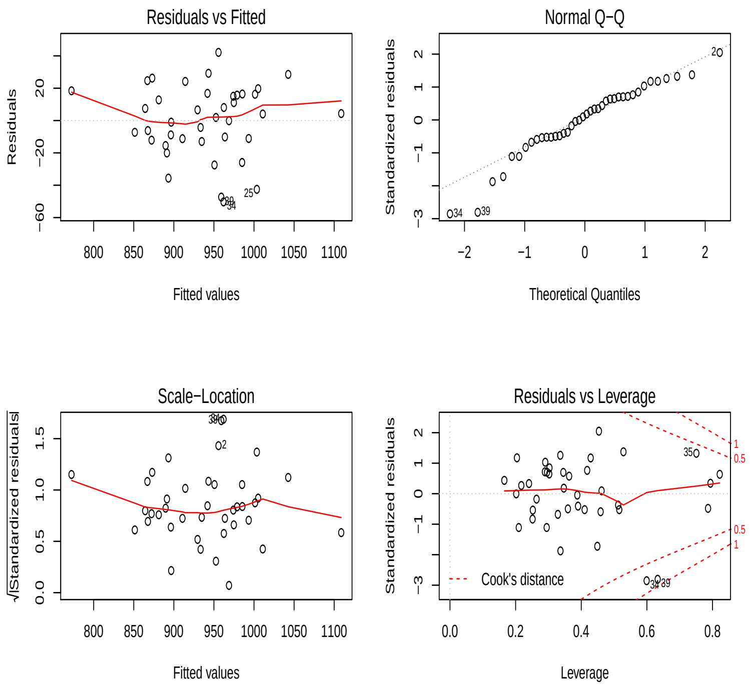


Figure 5: Diagnostic plots

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.

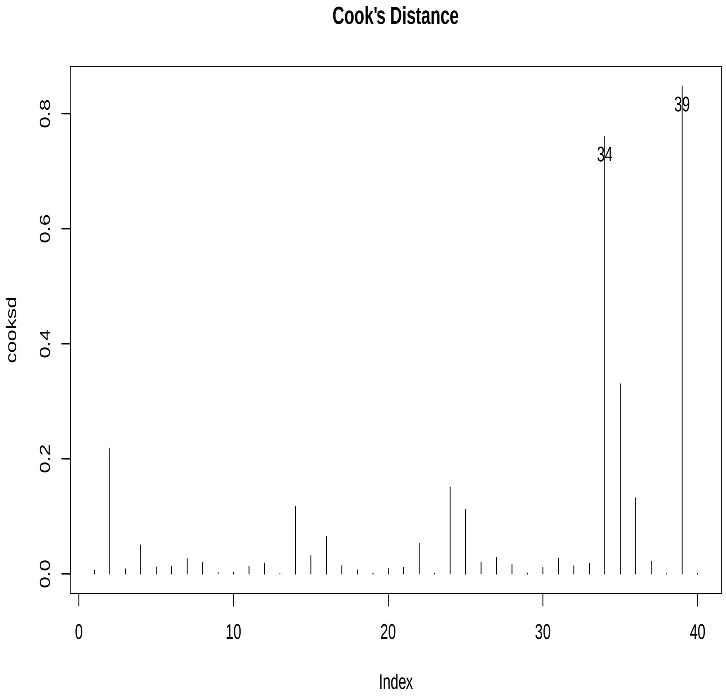


Figure 6: Cook's distance

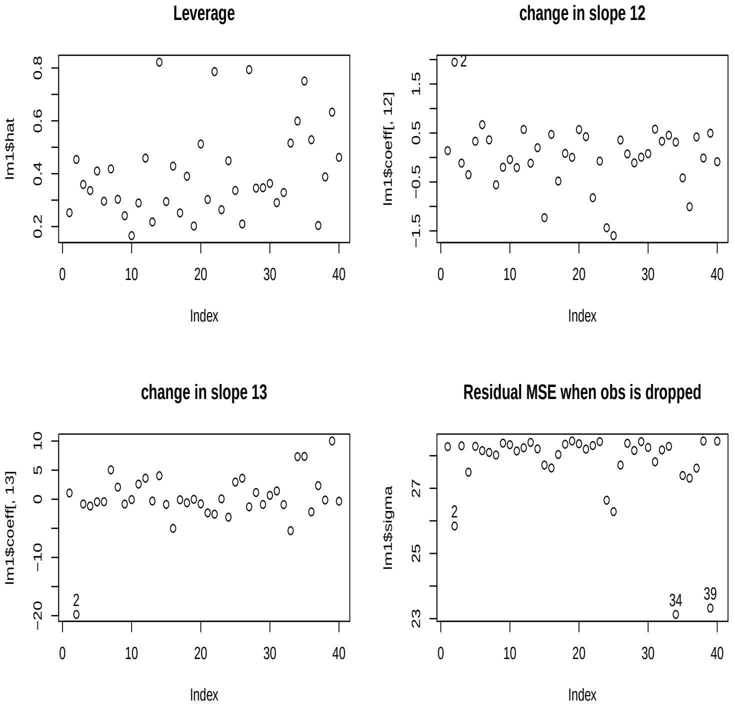


Figure 7: Diagnostic plots 2

The Cook's distance indicates that observations 34 and 39 have high leverage and/or large residuals. The leverage plot does not identify extremes, but the change in slope indicates that observation 2 has a large impact on at least these two slope parameters (an influential outlier). The change in MSE identifies 34 and 39. That is, these observations have such large residuals that they inflate the MSE estimate (but don't affect the slopes so much).

The figure illustrates the impact of two different kinds of outliers. An influential outlier pulls the regression line towards it. This changes the slope and what's illustrated in the diagnostic figure is the difference between the slope of the solid and dashed line as an observation is removed. If the outlier is influential and draws the regression line away from the mass of data, this will also increase the MSE. The change in MSE illustrates the MSE after and observation is dropped. Observations that are not influential or aren't outliers will lie near a line centered at the MSE for the model including all observations.

A pure outlier is like the one in the right panel. It has no influence on the slope and only a modest influence on the intercept. It does lead to a large MSE because it has a large residual.

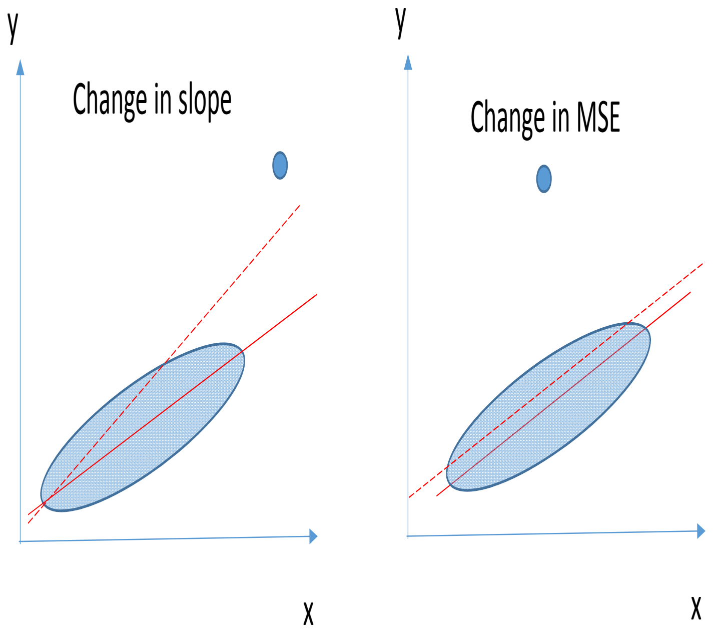


Figure 8: Diagnostic plots

(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.

The R-squared increased by about 10% after we drop 3 outliers. That is a good trade off. As a *result of the decreased MSE (27 to 18), many more coefficient estimates are now significant. We still have a problem with the collinearity between the pollutant variables and between some of the other variables also so many of the coefficient signs are in conflict with the marginal correlations. Interpretation of this model is difficult. There are many weak effects and a complex correlation structure between the predictors.*

The residual diagnostic plots look OK but we should investigate the potential outlier 14 to make sure it is not driving the fit.

```
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
```

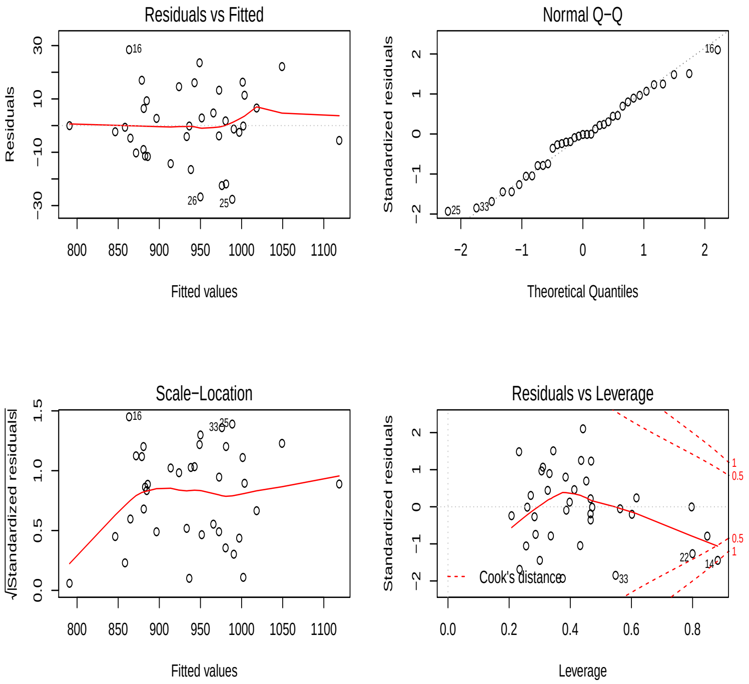


Figure 9: Diagnostic plots

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.
The small sample size (compared with number of parameters), many weak correlations between the outcome and predictors and many strong correlations between predictors make for a complicated selection problem. Looking at the marginal correlations between the outcome and each predictor, there is no single predictor with a strong correlation with mortality and the high R-squared does indicate that the many weak relationships lead to a strong predictive model. So, difficult model selection problem but probably a large set are needed to predict mortality with any great accuracy.

```
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?

This is a difficult selection problem! It is clear from the RSS picture that RSS levels of quite slowly with the number of parameters. The Cp and AIC suggest that very big models are best for prediction. BIC is a bit more conservative, but even here the increase in BIC beyond 9 variables is quite slow. So, ”flat” selection criteria curves as a function of number of variables makes a difficult selection problem. You also see that there are many models of similar or same size that are equally good and thus not possible to choose between based on prediction performance only.

BUT that being said, it is relatively easy to come up with a good model for prediction. Any of the models with 9-12 variables perform really well in terms of prediction (I base this on the R-squared above). Model selection for interpretation and model selection for prediction are not the same tasks.

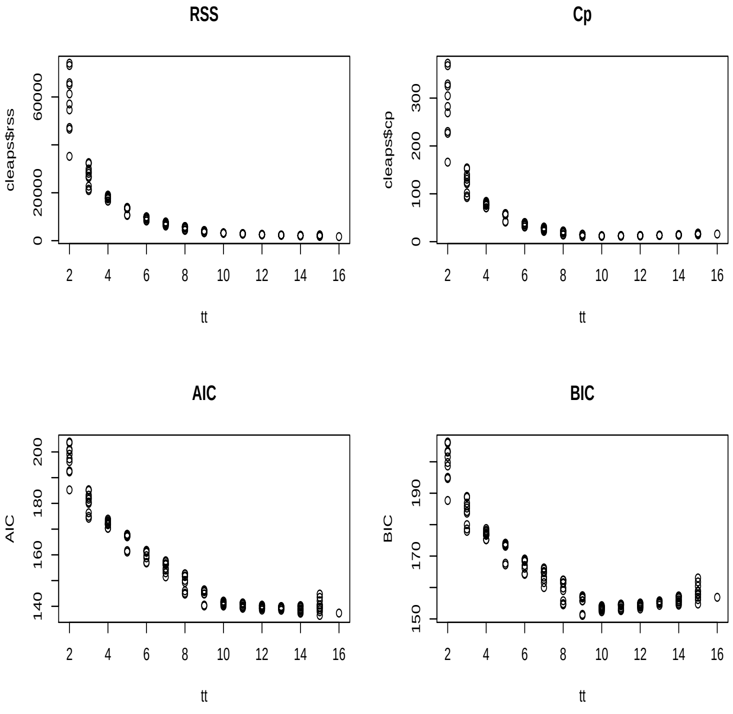


Figure 10: Selection criteria

```
"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 11 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.

BIC picks smaller model on average and also leads to smaller prediction errors. So BIC looks like the better choice here. Looking at the selection results, it is again indicating that model selection is difficult. We see that the temperature variables, population size, density, nonwhite proportion are almost always in the model, followed by so and nox. The other variables are present with in a non-negligible proportion indicating they are taking turns to be in the selected model. Therefore it is difficult to write down a ”final model” beyond the variables with around 90% presence. It is unclear which and how many of the other variables are needed.

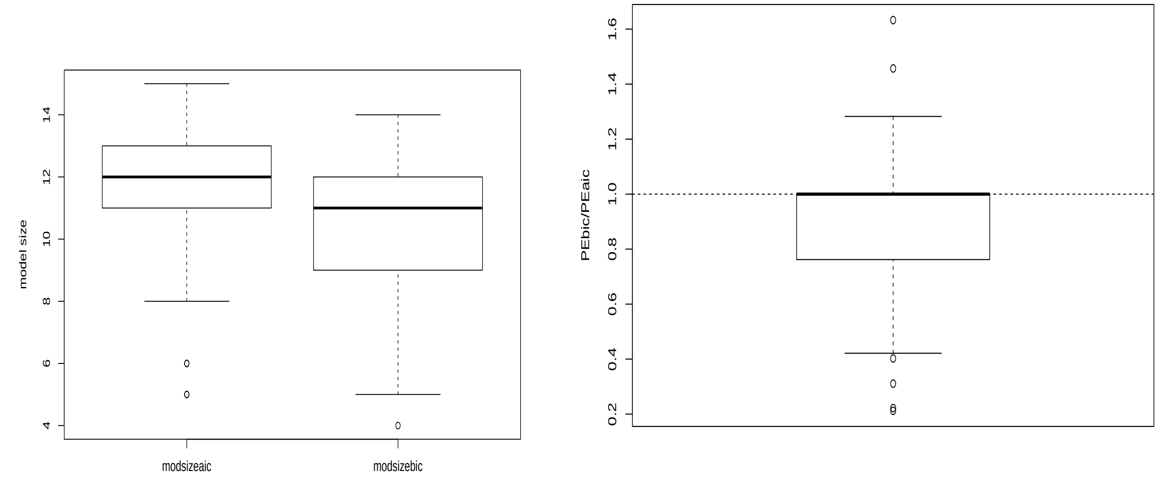


Figure 11: Left: Model size aic and bic. Right: Ratio of PE with AIC models compared with BIC models

```
             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.

The CART model first splits on nonw, which was also significant in Q7. The second split is on so, also significant in Q7. Futher splits on dens, educ and hous are not retained after cross-validation. Temperature and other polutants were not selected by CART.
The pruned tree model can be interpreted as follows: If the nonwhite proportion of the population is high, this is associated with the highest mortality. If the nonwhite proportion is low but the so level is high, mortality levels are higher than for low nonwhite and low so levels.

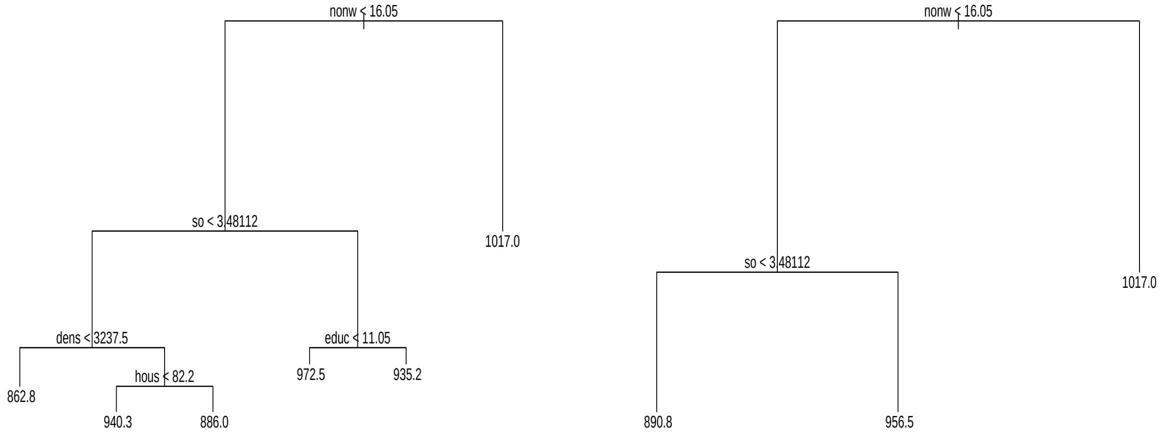


Figure 12: CART model and pruned tree

(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).

The selection outcome is much more stable with CART (this is not a general feature of CART though, so don't assume this will be case for all data sets). Here, nonwhite is clearly the most important feature but we sometimes split on education instead. In the tree, nonw is almost always used, whereas prec, educ, nox and so are sometimes in the model (one or two of them). This is a much smaller model than Q9 shows is selected for a linear regression model. The average model size is around 10 in Q9 and the average model size (in terms of selected variables) is around 3! However, there are no prediction error results provided so we cannot *say that CART is a better model that linear regression. It is more stable and utilizes fewer variables.*

```
              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"
```