MVE190/MSG500 **Solutions to Final 2018-01-11**

Examiner: Rebecka Jörnsten, 0760-491949

Remember: To pass this course you also have to hand in a final project to the examiner.
You can use the text book and the MVE190/MSG500 lecture notes - but no calculators, computers or old exams allowed

Make sure to give detailed and specific answers. Avoid yes/no answers. You should also provide a motivation. Good Luck!

Question 1 (25p=5+5+5+5+5)

a) A data set (x, y) is collected in a lab. Two pieces of equipment is used to save time and money, the same number of observations collected from each. From a calibration study it is known that equipment A has twice the level of measurement error compared to equipment B but equipment B has an offset problem for measuring y (adds bias B to each y value) when x-values exceed a level L. Explain how you would utilize the data set from both sets of equipment to obtain unbiased and efficient estimates for the linear model coefficients for y given x.

Since you know which equipment each observation comes from you can simply align the B data by subtracting the offset for observations over level L in x. To get the most efficient estimates you can use weighted least squares where observations from data A are weighted half of the observations in B. You could also use a dummy variable $1\{x > L\}$ *for the B data. This would estimate the bias via regression.*

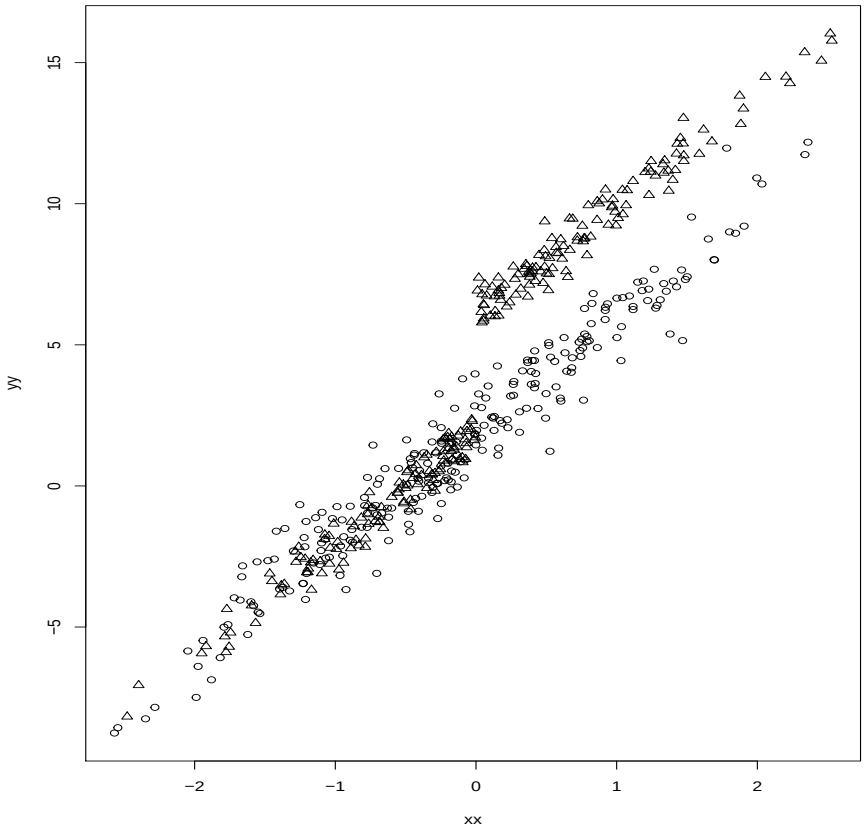


Figure 1: Question 1a solution. Triangles are equip B observations

b) A sloppy graduate student messes up the log-book and it is no longer known which observations were obtained from which equipment nor what their calibration settings were. Draw a sketch of what the data set looks like when the A/B information is unknown. What if instead it was B that also had the higher level of measurement noise?
Now bias B, level L and equipment label are unknown. Depending on the sample size, the size of the bias and the noiselevel, this problem may be easy or hard to detect. It can give the appearance of a skewed error distribution if the bias is not big but instead produces a long-tailed behaviour above level L. It can also give the appearance of nonlinear dependency. If the equipment B also has a larger noise level it makes for the appearance of skewed errors. Without knowing about the problem it would not be easy to spot it without a huge bias.

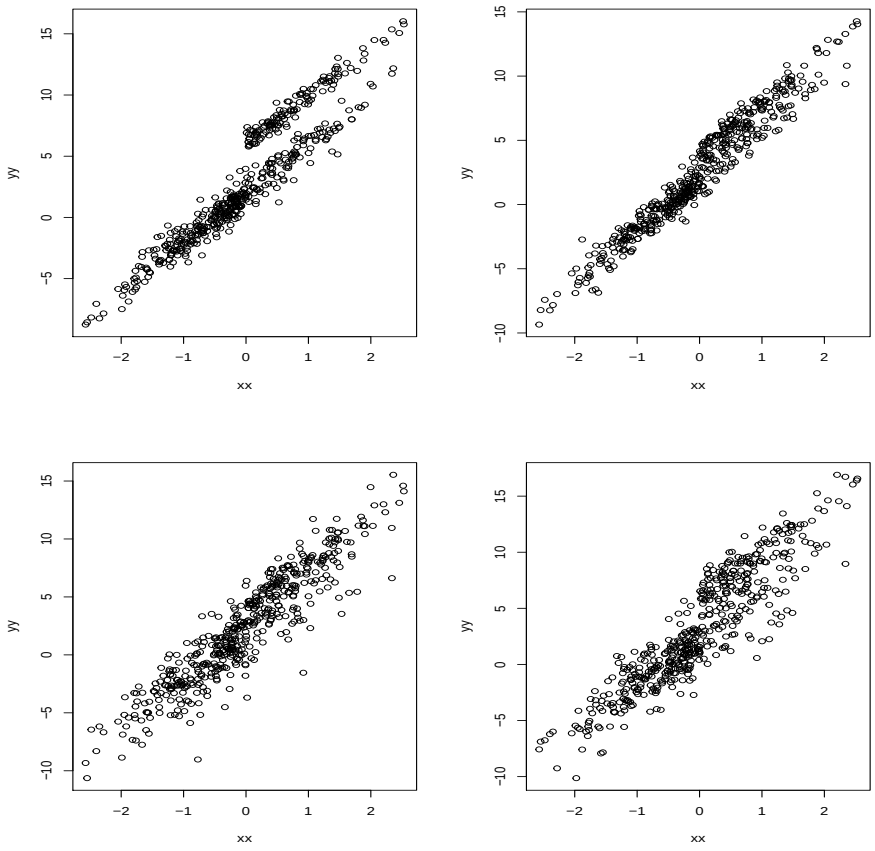


Figure 2: Question 1b solution: top left-right:less bias. bottom left: more noise and less bias, bottom right: more noise

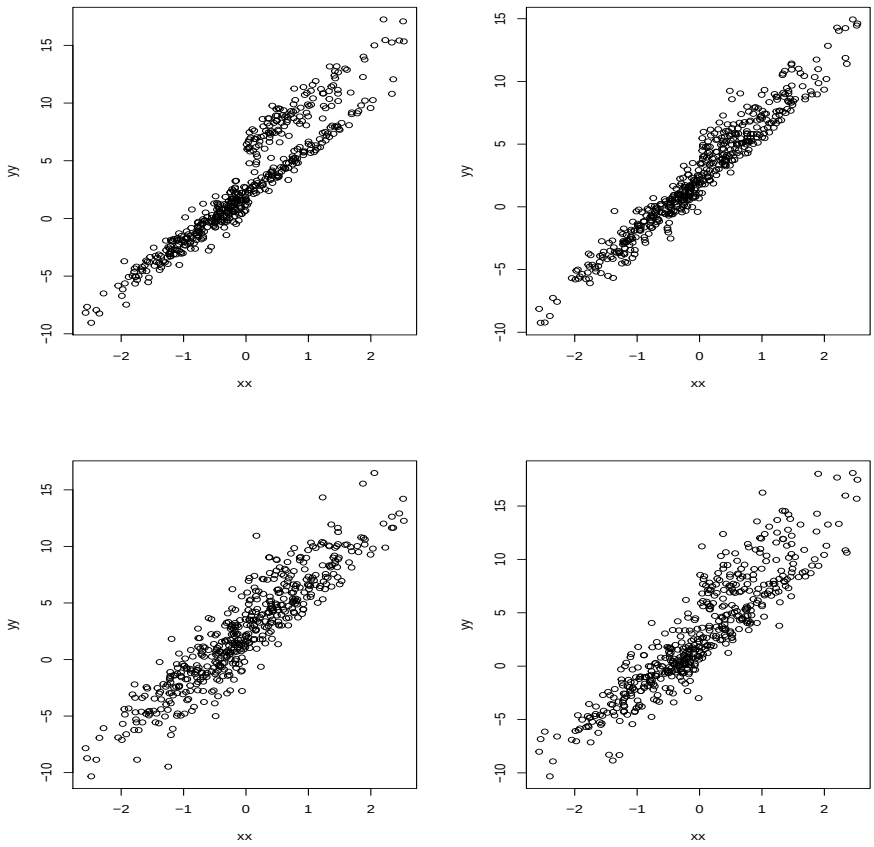


Figure 3: Question 1b solution, with equipment B more noise: top left-right:less bias. bottom left: more noise and less bias, bottom right: more noise

c) How would you go about analyzing the data under these circumstances? What is the impact of not knowing the A/B information on the model fit - can you recover the true model when A/B is unknown?
If you ignore the problem it will tend to lead to bias in the estimate of β (the coefficient for x) in the direction of the bias B. Also, ignoring the differing variance of the data sets leads to higher variance estimate of the regression coefficient. Without knowing what the problem is you cannot easily fix it. However, if you knew there was a level L and bias B and two sets of equipment, you're just missing the precise information, you could look at models for groups in data. You could try to identify the group above above the level L via residual diagnostics and then use dummy variable coding as in a).
d) What if the offset bias B occurs not for fixed levels of x but randomly with probability p. A clever graduate student in the lab has rigged an alarm that detects if a biased recording has occurred. Draw what the data might look like. Explain how you would go about modelling the data in this case.
This is just like a). You can subtract the bias from the B data and use weighted least squares otherwise.
e) Same as d) but now the sloppy graduate has lost the log-book again so you don't know the A/B information or when the biased offset occurred. What is the impact of the modelling when A/B information is unknown?
This will give the appearance of skewed errors (in the direction of the bias) and will lead to a biased estimate of the intercept mainly.

Question 2 (25p=5+10+10)

a) Data is generated from a linear model $y = \alpha + \beta x + \epsilon$. One statistician models the data with values (y, x) and the other decides to round off the values of the x prior to modeling (see Figure 1 for one example of such data). Discuss the impact of this strategy on average: effect on estimated coefficients, estimated noise level, R-squared, significance, etc.
From the figure and thinking about the effect of rounding, the impact is that the noise level around the regression model now appears larger than it is. That is because several points on the true regression line are now forced to be located at the same x-location so in fact comes from different mean-distributions which results in a bigger spread. The impact on the regression coefficient estimate is minor, but the R-squared decreases as the noiselevel increases, p-values are larger since the noiselevel increases. In general, the strategy can produce problems with leverage as the extremes of x might comprise very few observations.

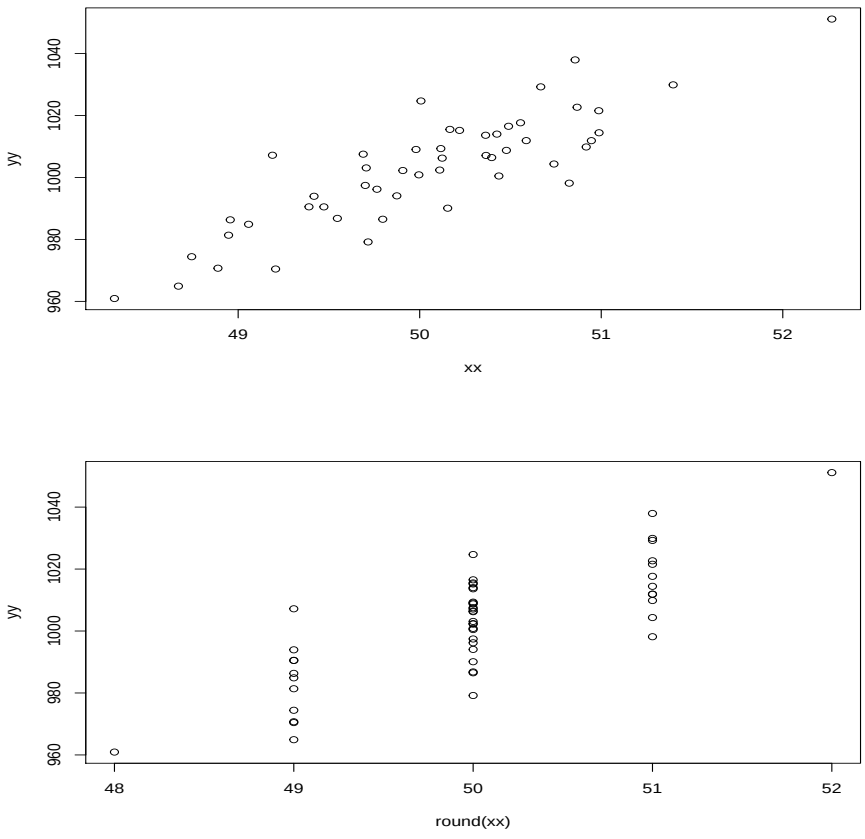


Figure 4: Question 2a: top=raw data, bottom=rounded x-values

b) A group of 10 international researchers collect data to study the relationship between variables x and y. Each of the 10 researchers conduct a subset of the study in their country, collecting a set of 10 data points each, and they plan to then aggregate the data for a full analysis. Unfortunately, they have very poor computer skills on their team so they decide to summarize each of the individual studies with the average x and y for a total of 10 average values to be jointly analyzed. Illustrate by drawing least 3 examples of the impact of such an aggregation (changing(how)/not changing the detected x-y relationship).

Groups in data is a big problem, and aggregating group-level data can obscure important information. Lots of things can go wrong. In the figure below I illustrate a couple of cases. Each ellipse represents the data cloud of 10 points from each study. A: there is a trend in each of the data sets, but the strength of the association is exaggerated by the aggregation. The noiselevel will appear to be tiny. B: Here, the trend is the same in each data set and there is an offset in x and y that is group specific. Aggregation will lead to about the same trend conclusion as each separate data set and the aggregate noise level will be about the same as the noiselevel of each data. C: Here, the association between x and y is group-specific (interaction) and aggregation obscures this. I this example, it will also produce a downward trend instead of the various upward trends we see for each data set. D: There is no association between x and y. By aggregating the data we create a trend due to the offset in x and y that is group specific. E: Here, all groups have an upward trend but the aggregation produces a downward trend instead. F: The assumption made by the people in the example: that each data is generated from the same distribution so there are no groups in the data. Even so, it may not be the best way to handle the data. There are alternatives in terms of repeated-measures, mixed effects models (see experimental design classes if you want to learn more).

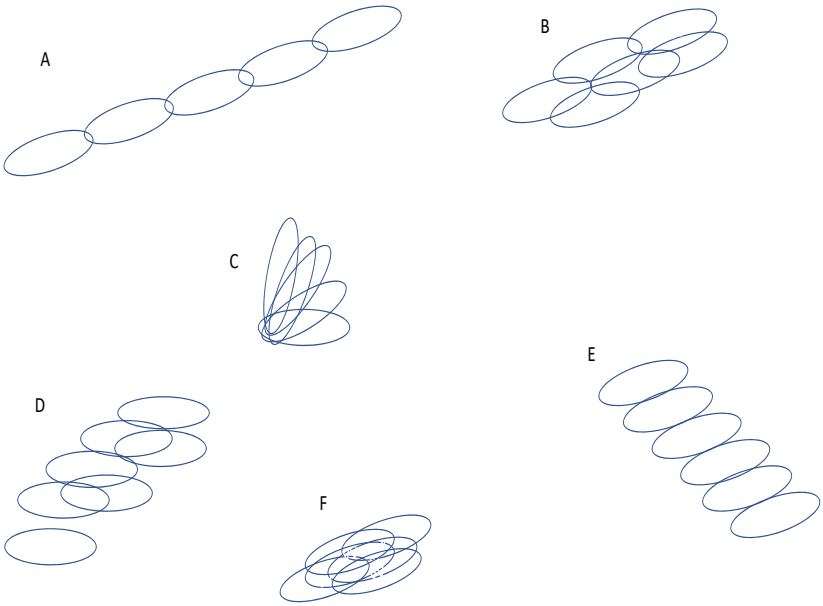


Figure 5: Question 2b

c) For the data sets in Figure 2 and regression trees in Figure 3 - couple each data set with the corresponding tree and draw what the tree model looks like in the scatter plot.
The way to go about this is try to work through an elimination process. The easiest is to look at the various scales first. Figure iv) shows a data with range 0-25 and the only tree with that range is IV. Double check by looking at the effect of x and z: For z=0, we have a constant y-level of 7 (matches the data iv) and oterhwise we have repeated splits on x, i.e. a slope, which also matches the data. iv=IV established. Data iii has a range -30 to 40 with z=1 having a weaker slope than z=0, z=1 values below z=0 values for the same value of x. Range-wise, tree I looks like a good candidate. Verify: whenever there is a split on z, the z=1 has smaller values in the leaf-node. iii=I. The two remaining dat sets, i and ii, are similar in range but for i we see a strong trend for z=1, weaker for z=0 and y-values overlapping for low values of x and then moving apart, for ii we see the same trend for both but z=1 have higher y-values that z=0 for the same x. Looking at the two trees II and III: in tree III z is present everywhere in the tree, for the full range of x, producing lower fitted values for z=0 than z=1 in each split. In tree II, z appears in the tree only for larger x-values and producing lower values for z=0 data in that range. Conclusion: ii=III, i=II

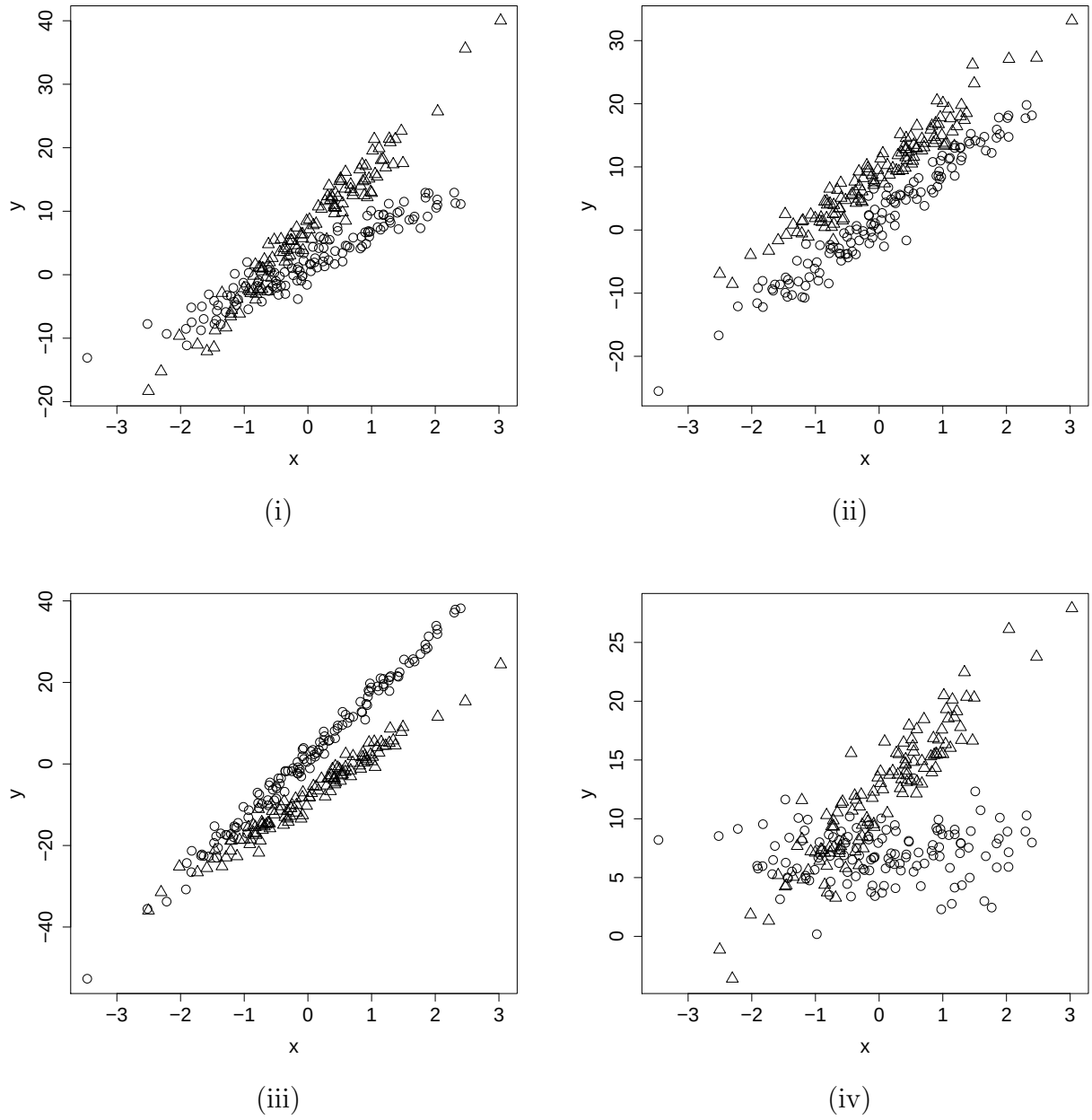
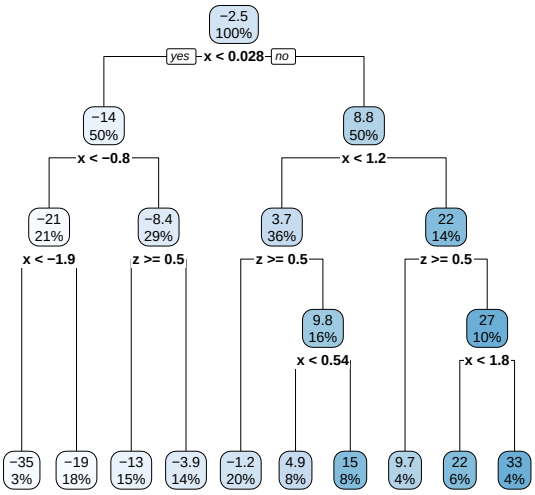
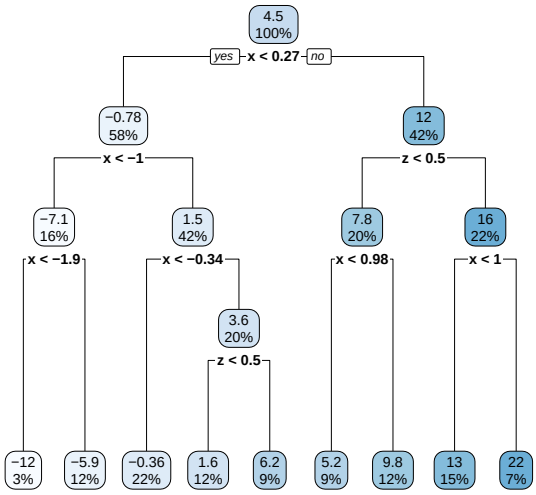


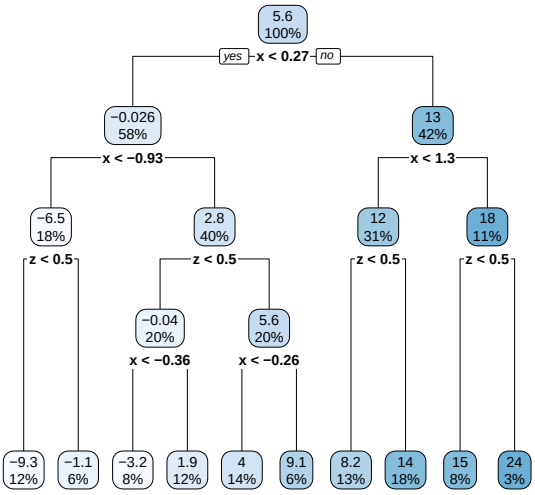
Figure 6: Question 2c: data sets. z=1 triangles, z=0 circles



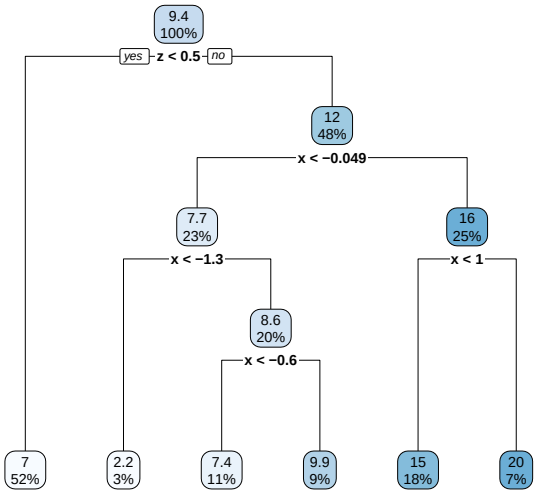
(I)



(II)



(III)



(IV)

Figure 7: Question 2c: Regression trees

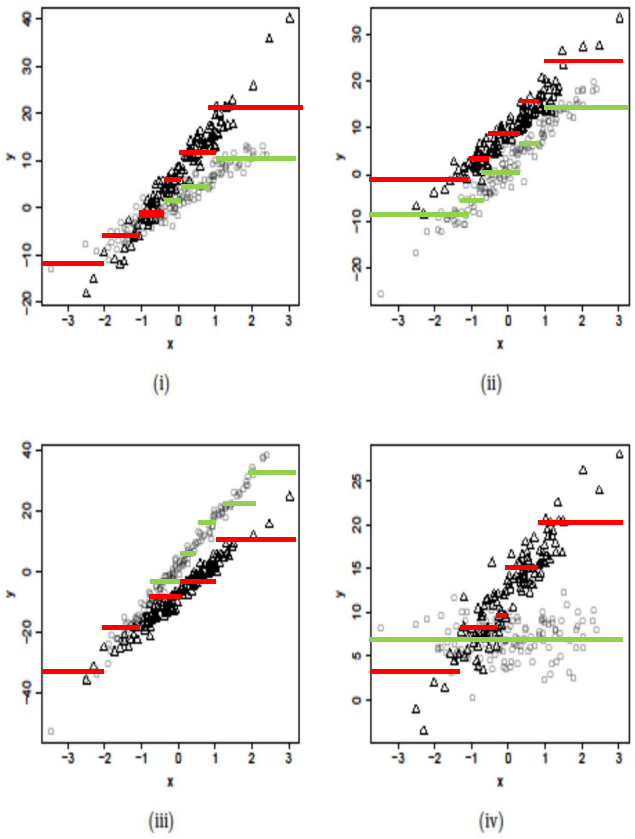


Figure 8: Question 2c: Trees in scatter plot

Question 3 25p=(5+5+5+5+5)

A data set on risk factors for low birthweight comprises 188 women. The outcome variable is birthweight (bwt) in grams. Other variables include;

age	mother's age in years.
lwt	mother's weight in pounds at last menstrual period.
race	mother's race (1 = white, 2 = black, 3 = other).
smoke	smoking status during pregnancy.
ptl	previous premature labours (0 = no, 1 = yes)
ht	history of hypertension (0 = no, 1 = yes)
ui	presence of uterine irritability.
ftv	number of physician visits during the first trimester (0,1 or 2+).

In figures 4 and 5, scatter plots of birthweight on the other variables are shown.
a) In figure 6 the residuals diagnostics from a linear model fit of bwt on the other variables are shown. Below is the summary of the fit. Comment and interpret. I treated race and ftv as a 3-level factor. Comment. ptl and ftv were actually numerical variables with observed range 0-4 (number of previous premature births) and 0-6 (number of visits) respectively, with about 10-20 observations in the range ptl 2+ and ftv 2+. Comment.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2981.972    311.186   9.583  < 2e-16 ***
age               -10.842      9.960  -1.089 0.277791
lwt                 4.752      1.701   2.794 0.005786 **
as.factor(race)2 -474.094    147.035  -3.224 0.001504 **
as.factor(race)3 -305.417    115.492  -2.644 0.008917 **
smoke            -290.415    107.833  -2.693 0.007757 **
ptl              -202.771    136.045  -1.490 0.137880
ht               -591.350    197.760  -2.990 0.003185 **
ui               -483.091    134.935  -3.580 0.000443 ***
as.factor(ftv)1    96.414    121.012   0.797 0.426673
as.factor(ftv)2   -33.473    121.368  -0.276 0.783027
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 635.1 on 177 degrees of freedom
Multiple R-squared:  0.2545,Adjusted R-squared:  0.2124
F-statistic: 6.043 on 10 and 177 DF,  p-value: 7.049e-08
```

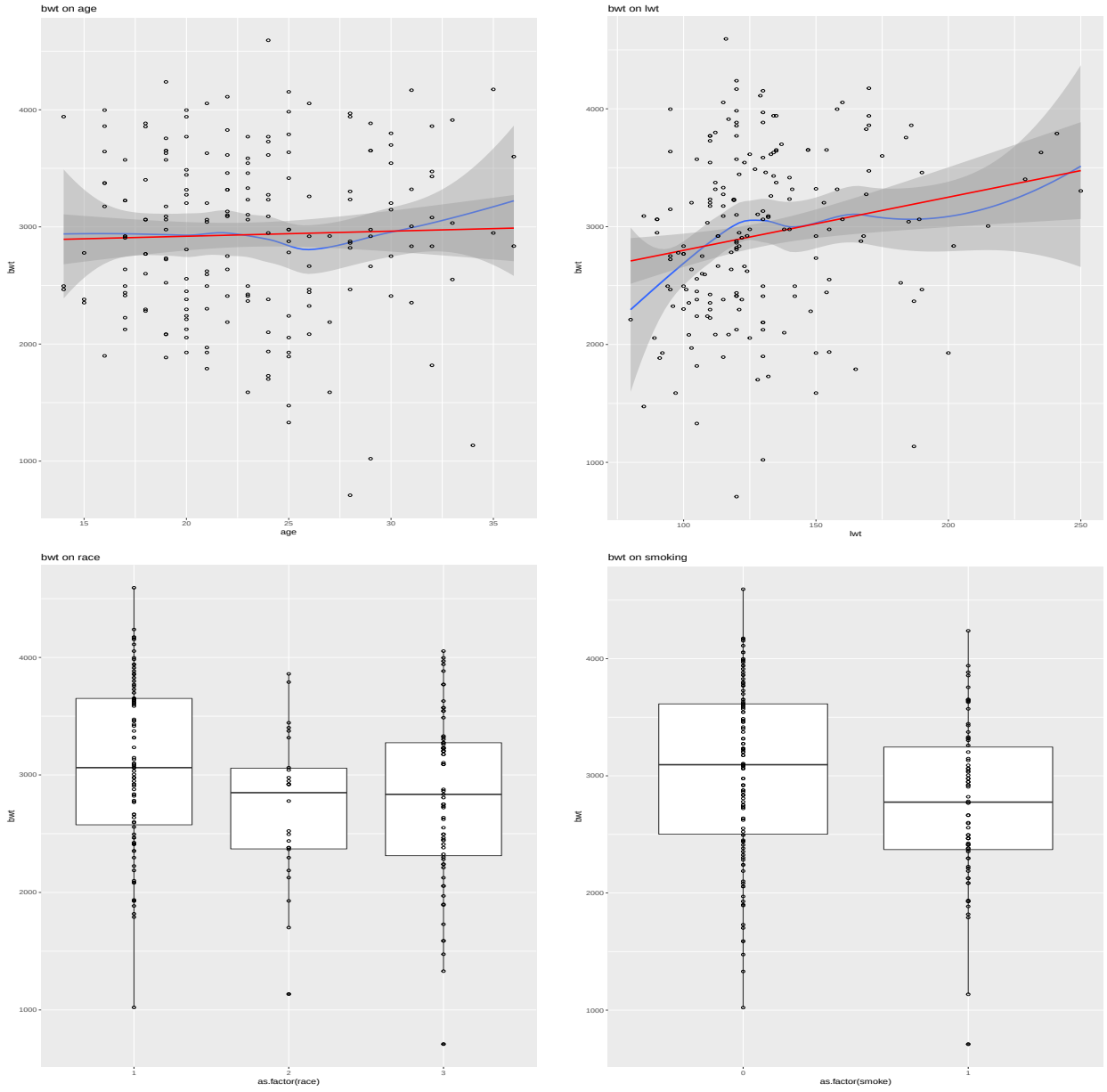


Figure 9: Question 3: Baby birthweight data (top: bwt on age and lwt. bottom: bwt on race and smoke

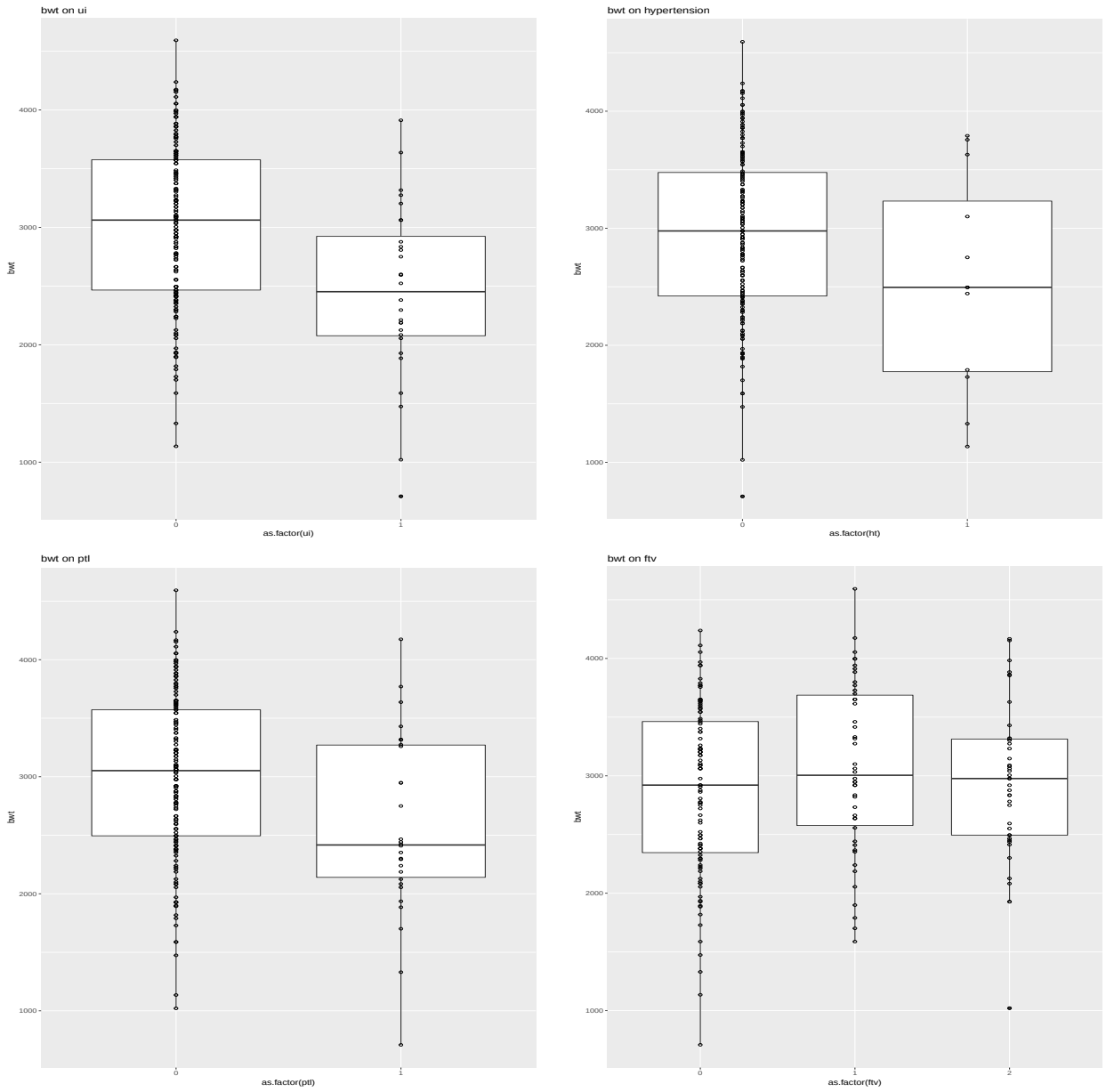


Figure 10: Question 3: Baby birthweight data (top: bwt on ui and ht. bottom: bwt on ptl and ftv

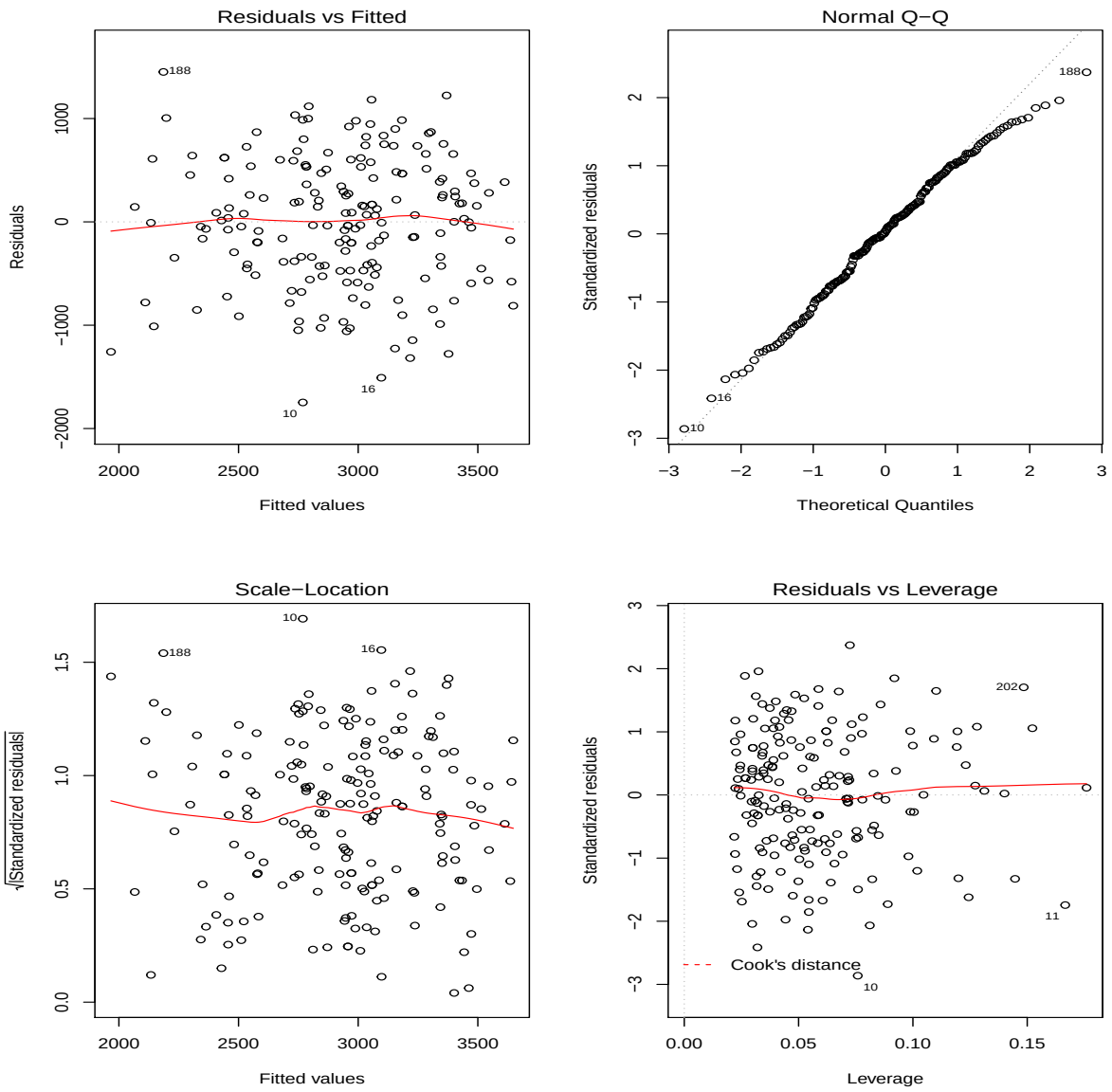


Figure 11: Question 3a: Residual diagnostics

First, some general comments about the fit. The R-squared is not very high 20-25%. It is pretty clear from the scatter plots that the noiselevel is high. In terms of the model: based on the summary table it suggests that race is an important factor with bwt being lower for both black and other. In addition, smoking is associated with a lower bwt as is hypertension and uterine irritability. Previous premature labor is not significant and neither is age. A higher weight for the mother (lwt) is associated with a higher bwt. We might wonder if this is correlated with another variable. ftv is not significant. The boxplot shows that no visits is associated with a slightly lower bwt as is more than 2 visits. Perhaps many visits is due to a high-risk pregnancy? The residual diagnostics do not look too bad. No obvious problems. Race as a 3-level factor makes sense - there is no ordering. One could consider having a non-white factor and then an additional difference between black and other. ftv as a 3-level factor might be based on having about the same number of observations in each group or also can be argued as above. There is a risk factor with no doctor's visits (a socio-economic factor) but also with many visits since this is probably due to complications. Turning ftv and ptl into factors was perhaps necessary since very few observations were in the upper range and that might lead to extreme leverage if treating these variables as numerical.

b) I removed one observation from the data set prior to modeling: a 45-year old white woman, non-smoker, with lwt 123, ftv 1 and whose baby weighed 4990grams. Comment.
This observation is an extreme x in age and an extreme y, high bwt. A clear outlier.
The data set above contained 188 observations. A larger data set is available (1000 women) but for this data we have many missing values for lwt and ftv. Discuss how you would proceed.
We might first want to assess how important lwt is in the model for the 188 women. How much lower is Rsquared without lwt. fvt was not significant, but had also been rounded off due to few observations with more than 2 visits. In general, we would expect to be able to fit a better model with 1000 observations, especially if lwt does not contribute much in the 188 data. However, you should try to figure out why this information is missing. Are high risk groups linked to few doctor's visits and less likely to know lwt? How might this then influence the results? You don't have any variables here that are proxies for fvt and lwt at first glance. You could code missingness and a dummy variable and see if that is predictive of bwt.
I run backward selection using AIC and end up with the following model. Comment and discuss.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2821.281    240.775  11.717  < 2e-16 ***
lwt                  4.188      1.649   2.540 0.011938 *
as.factor(race)2  -447.115    143.118  -3.124 0.002079 **
as.factor(race)3  -307.309    111.055  -2.767 0.006245 **
smoke             -302.706    103.731  -2.918 0.003970 **
ptl               -204.432    131.491  -1.555 0.121769
```

```
ht               -569.665    195.960  -2.907 0.004107 **
ui               -482.883    133.913  -3.606 0.000403 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 633.6 on 180 degrees of freedom
Multiple R-squared:  0.2455,Adjusted R-squared:  0.2162
F-statistic: 8.367 on 7 and 180 DF,  p-value: 7.699e-09
```

AIC selection removed fvt, age as one might expect from the result in a). The p-values here are not "honest" - selection bias - so too good to be true.

c) I create dummy variables for race and for the 3-level ftv variable. I split the data into 150 observations for training and 38 for testing and obtain the following model selection results (see also Figure 7):

```
       R.sq test      age lwt smoke ptl ht ui ftv ftv2 black other
cvmod  0.116           1   1     1   1  1  1   1    0     1     1
cpmod  0.134           0   1     1   1  1  1   1    0     1     1
aicmod 0.134           0   1     1   1  1  1   1    0     1     1
bicmod 0.045           0   0     1   0  1  1   0    0     1     0
```

Comment and discuss.
The selected models are quite big. The R-squared on the test data is really low! The best result (and that's a pretty unimpressive result) is AIC and CP models that use lwt, smoke, ptl, ht, ut ftv and race.
I also use modelaveraging, aggregating the top-10 models based on each selection criterion and obtain the following results;

```
           R.sq test      age   lwt   smoke  ptl   ht   ui   ftv   ftv2  black other
[1,] "CV"  "0.172"       "0.4"  "1"   "1"   "0.2"  "1"  "1" "0.6"  "0.5"  "1"   "1"
[2,] "CP"  "0.187"       "0.4"  "1"   "1"   "0.5"  "1"  "1" "0.4"  "0.2"  "1"   "1"
[3,] "AIC" "0.18"        "0.4"  "1"   "1"   "0.6"  "1"  "1" "0.5"  "0.2"  "1"   "1"
[4,] "BIC" "0.11"         "0"   "0.5" "0.4" "0.3"  "1"  "1" "0.3"  "0"    "0.3" "0.2"
```

Comment and discuss.
Modelaveraging improved the R-squared on the test data (i.e. the prediction performance) by 5%. It appears top models that are aggregated use either age or ptl or ftv.

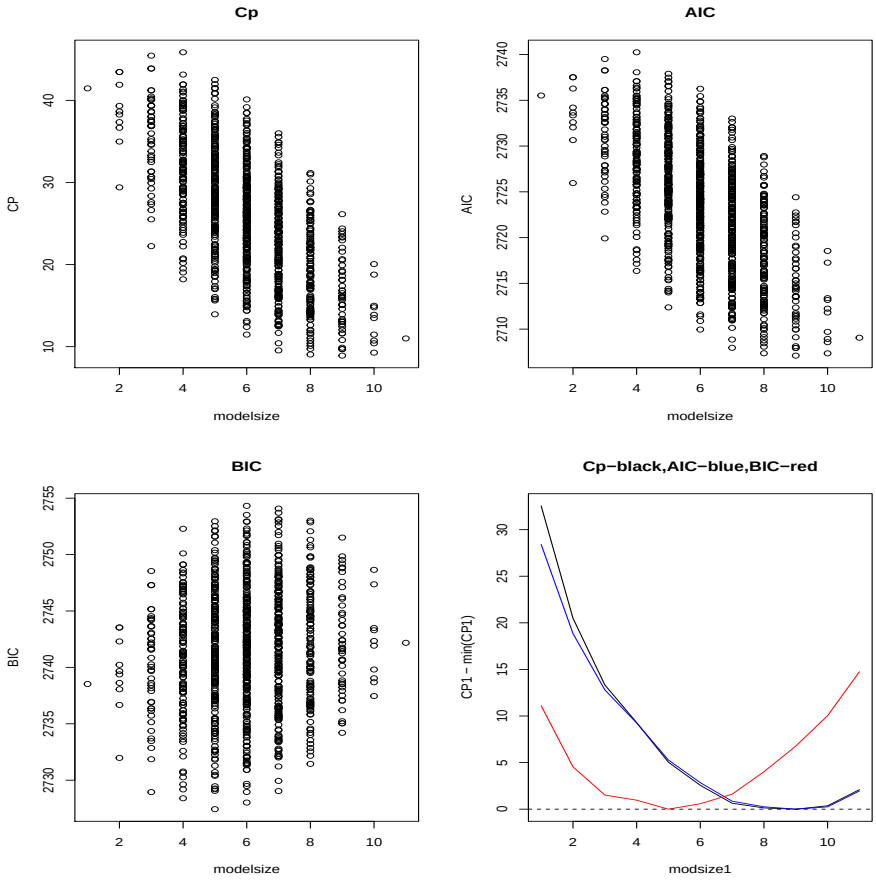


Figure 12: Question 3c: Modelselection results

d) I repeat the above 100 times, repeatedly splitting the data into 150 for training and 38 for testing. I obtain the following results;

```
Model selection - best model
       R.sq test     age  lwt smoke  ptl   ht ui  ftv ftv2 black other
cvmod  0.10976      0.41 0.93  0.94 0.38 0.84  1 0.11 0.05  0.94  1.00
cpmod  0.11881      0.36 0.97  0.97 0.48 0.97  1 0.09 0.05  0.97  0.99
aicmod 0.11958      0.41 0.97  0.99 0.51 0.97  1 0.12 0.05  0.99  0.99
bicmod 0.03583      0.00 0.52  0.59 0.25 0.60  1 0.02 0.00  0.53  0.56
Model selection - average of top-10 models
       R.sq test     age   lwt    smoke   ptl    ht    ui   ftv   ftv2  black other
CV     0.12161      0.384  0.909  0.916  0.525  0.844 0.989 0.311 0.243 0.891 0.982
CP     0.13161      0.395  0.929  0.949  0.571  0.938 0.998 0.315 0.228 0.939 0.982
AIC    0.13357      0.412  0.936  0.956  0.578  0.948 0.998 0.324 0.253 0.953 0.986
BIC    0.07396      0.046  0.541  0.589  0.346  0.611 0.962 0.058 0.016 0.500 0.594
```

In Figure 8 I depict the rank-statistics based on the Rsq on the test data for each of the selection criteria across the 100 runs. That is, if a criteria was always producing the best results it would have rank 8 (comparing 8 strategies: CV, CP, AIC, BIC and top-10 average of models chosen by CV, CP, AIC and BIC).
Comment and discuss. What about the stability of modelselection? Is it easy to identify the most important predictors. Write a few sentences that would be like a "press release" about the main risk factors for low birth weight in babies.
Modelaveraging does improve the prediction. CP and AIC are the best models in terms of prediction and for those criteria the averaging boosted their performance the most. BIC performs the worst. lwt, smoke, ht, ui and race are important predictors.
The model selection stability is not great. lwt, smoke, ht, ui and race are almost always in the AIC models, but there are also other variables that appear quite frequently (age, ptl).
"Smoking, history of hypertension and uterine irritability is associated with an increased risk of having a low birthweight baby, but race is a factor as well."

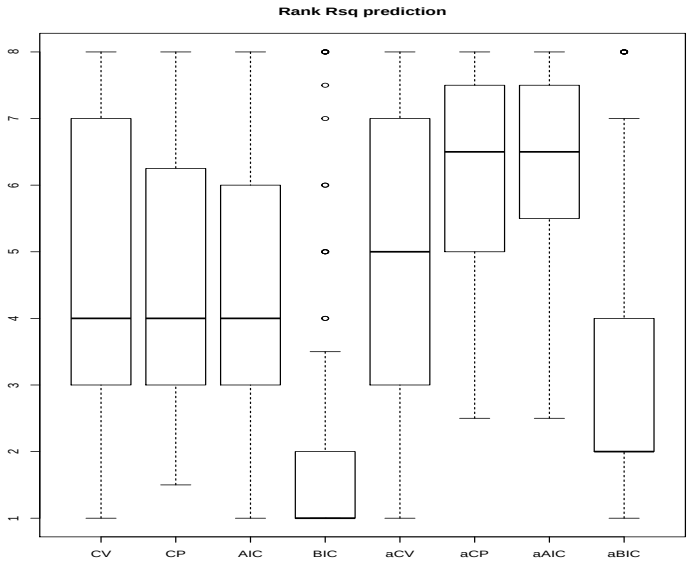


Figure 13: Question 3d: Ranks of selection methods

e) I model the birth weight data using lasso-regression. Figure 9 shows the CV-results. In the table below I summarize the least-squares fit, the fit from the CV-minimum lasso model and the solution with penalty factor λ within 1SE of the CV-minimum result (as indicated in Figure 9).

	Least-Squares	minCV	minCV+1SE
(Intercept)	2981.972423	2948.265478	2857.982191
age	-10.842488	-8.804754	.
lwt	4.752387	4.449404	1.961779
smoke	-290.415145	-273.643514	-117.575155
ptl	-202.771088	-197.943166	-127.251988
ht	-591.349894	-557.483330	-247.282353
ui	-483.091350	-469.321486	-339.387835
ftv	96.414074	89.387151	.
ftv2	-33.472787	-19.357551	.
black	-305.416814	-282.651668	-79.592859
other	-474.093767	-439.380726	-128.211904

I apply lasso 10 times on random 150 training and 38 testing observations. I obtain the follow results;

```
"R.sq"   "R.sq"     "age" "lwt" "smoke" "ptl" "ht"  "ui" "ftv" "ftv2" "black" "other"
 test   debiased
Lasso - min CV lambda
"0.1184" "0.1131"   "0.7"  "1"   "1"    "0.9" "1"   "1"  "0.7"  "0.4"  "1"     "1"
Lasso - min CV+1SE lambda
"0.0721" "0.0934"   "0.2" "0.8" "0.9"   "0.8" "0.9" "1"  "0.5"  "0"    "0.7"   "0.7"
```

Comment and discuss. Compare with the results using least squares and modelselection in question 3a-3d above.

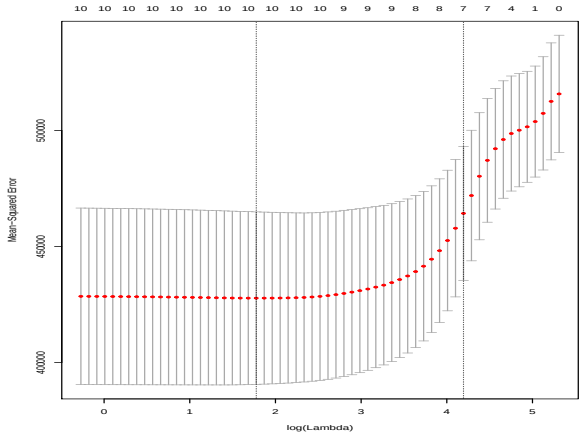


Figure 14: Question 3e: Modelselection results

The minCV performance is on par with AIC and the minCV+1SE model is on par with BIC. From the summary table for the coefficients we see the effect of the l1-penalty on the

coefficients, shrinking them toward 0. But as with the selection criteria, the CV-based lasso model is quite big, with age and ftv the only variables being removed by the selection. With repeated selection we see that as well and that the selection is not very stable.

Question 4 (25p=(10+10+5)

In this question you will discuss building a predictive model for ozone exposure. Ozone exposure in the population is a major health concern. High levels of ozone have been linked to e.g. respiratory illnesses such as asthma. To estimate the risk for the population researchers have come up with predictive models for personal ozone exposure. Predictors have to include "activity type" and "risk behavior" information of individuals, such as outdoor occupation, outdoors during the day, where you live, etc. Many studies have been conducted to assess population risk and come up with predictive models for personal exposure.

Here we will discuss a data set from a study conducted in the summer of 1991. Our personal exposure to ozone is difficult to predict because it is lifestyle specific. If you spend a lot of time outdoors, especially during midday, you can be highly exposed. But you can also be exposed if the ozone penetration of your house is high. Your exposure depends on where you live as well, altitude, rural/urban etc. For practical purposes, goal being risk assessment, it's not feasible to equip every single person with an ozone sampler. The study in 1991 used personal ozone samplers on a set of 19 children, and tried to relate these personal measures to more easily obtainable measures such as indoor and outdoor ozone measurements near the home of the children. The passive personal ozone samplers were little badge clips that were put on the children's backpacks. The children also turned in activity sheets stating how much time they had spent indoors at home, outdoors.

The data: Ozone concentrations were measured in State College, PA Aug 7 - Aug 27 1991. Ozone samples were collected on days with very varied levels of ozone concentration, to see the full range of possible exposure. Personal samples were collected for 19 children (ages 10-11), living in non-smoking homes is six different residential regions. Personal exposures were measured during the day (8am-8pm). Regions 1, 2, 4 and 5 are densely populated, whereas regions 3 and 6 are less developed. Outdoor concentrations were measured a one stationary site (State College National Dry Deposition Net- work site, 6 km west of downtown State College). At this stationary site 12 hr average samples were collected twice daily (day: 8am-8pm, night: 8pm-8am). In addition, indoor samples (home) were taken with passive samplers twice daily (12 hr averages, for day and night). Outdoor sam- ples were collected near the homes (24 hr averages, beginning 8am).

Let's introduce some notation. For a given day and child (sid = id number (1-19) for each child), y is the personal exposure mea- surement, x1:d and x1:n are the stationary site measurements, day and night time. x1:o is the outdoor measurement near the home. x1:di and x1:ni are the indoor measurements in the individual homes, day and night time averages. We denote by x2:o and x2:i the fraction of time a child spent anywhere outdoors and at home indoors on a given day, and this may sum to less than 1 if the child spent a lot of time indoors but not at home. The number of observations in total is 69 with 1-6 measurements obtained for each of the 19 children. Scatterplots of the measurements data are shown on the next page.

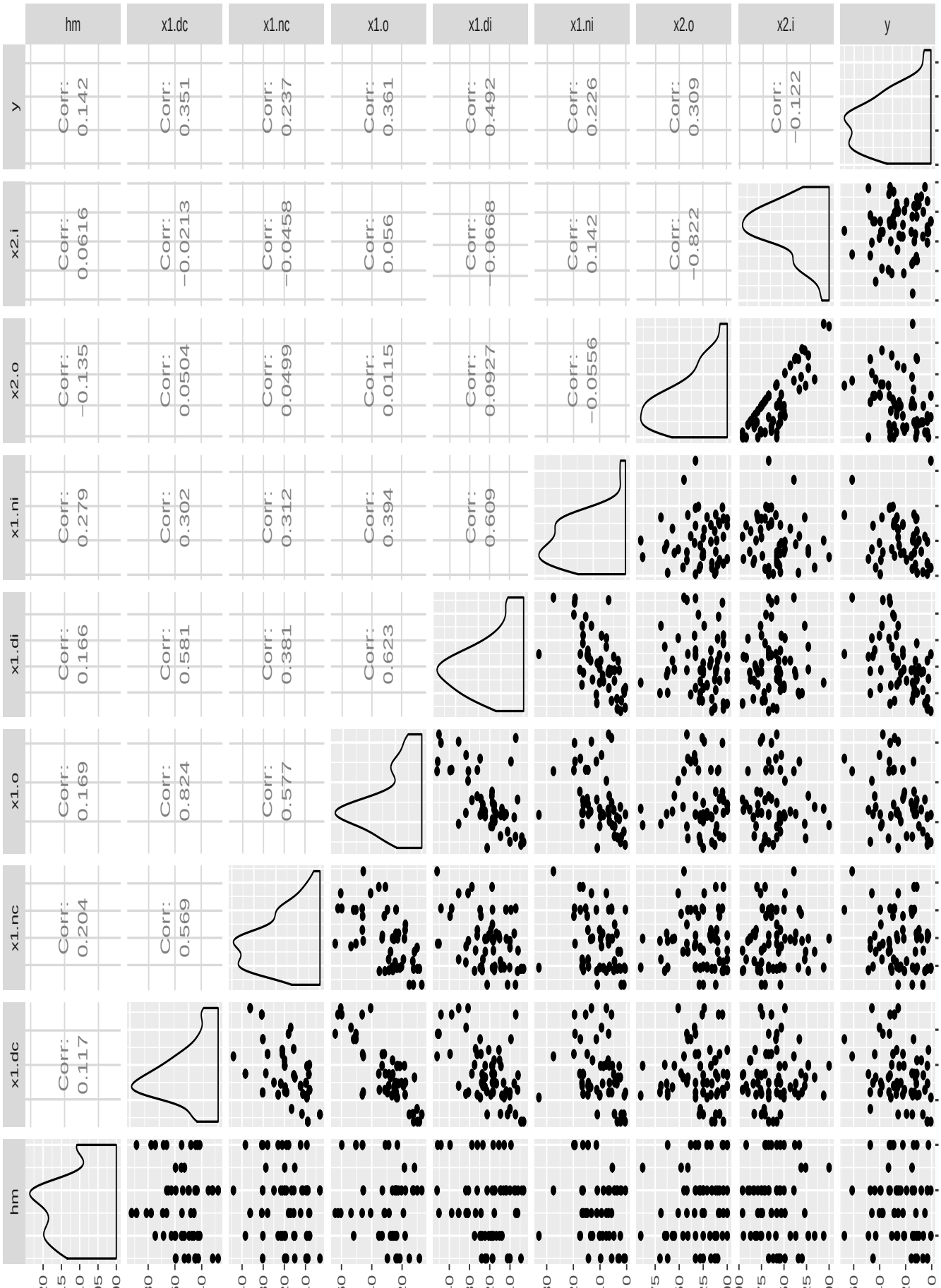


Figure 15: Question 4: Scatterplots

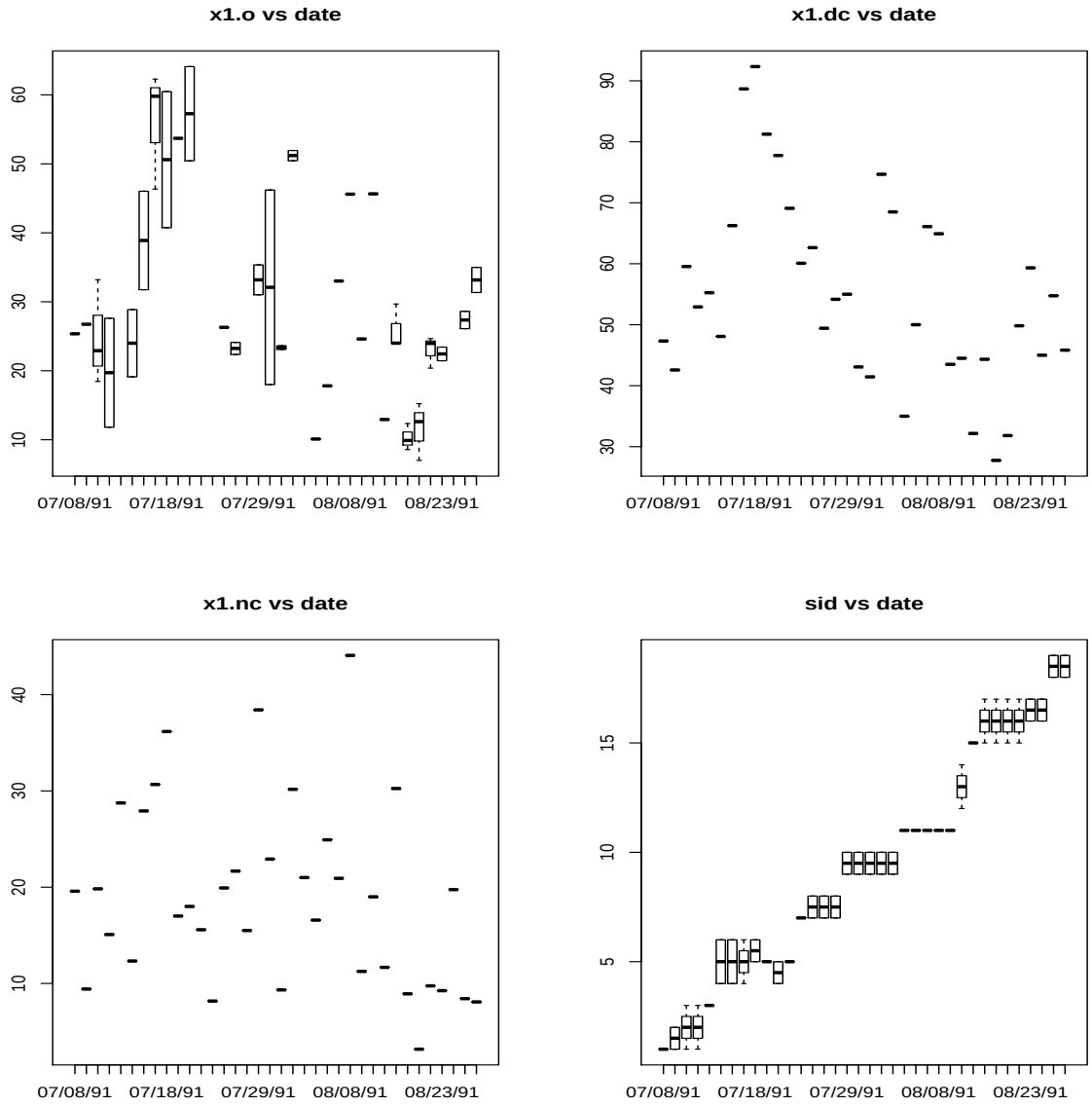


Figure 16: Question 4: Scatterplots

a) Discuss how you would go about modeling this data and how the outcome from such a model might be interpreted; concerns you might have regarding assumptions, how you would consider incorporating the information from the stationary site, the individual measurements and the activity information.
The scatter plots show that the daytime indoor is most closely linked to the personal exposure - this makes sense if the children spend most of their time at home and indoors. You also see a correlation between y and the daytime ozone levels at the stationary site and with the outdoor levels outside the home. The outdoor levels are 24hr averages though so not daytime only. What these suggest though is that the stationary site levels, while the same for all kids, are related to the personal exposure and overall day+night near the home is as well. That probably means that there is a strong day-to-day variation in overall ozone levels correlates with personal exposure. The day-to-day variation is confirmed by the boxplots. Now, to come up with a good predictor for personal exposure we might want to way the time spent outdoors with the outdoor measurements. Similarly, we may want to look at the time spent indoor at home and see if we can by weighing enhance the correlation between the indoor measurements and the personal exposure this way. The region variable seems to be related to y (not linearly of course). Where you live could impact your exposure if there is a pollution effect. However, I am a bit concerned with the apparent link between region and the stationary site measurements: this could only happen if kids from different regions participated in the study on different days, region 3 and 6 on particulary high ozone level days. That the kids were not randomly assessed over the days in the study period is clear from the boxplots on sid. This is a huge concern. Another concern is that we are following the same kids over a few days. This means that the errors are not uncorrelated and you should probably look into a mixed effects model. Another concern; The sampler is attached to their backpack - there is a risk that the kids don't keep the backpack always with them, e.g. when playing outdoors or even in a different location in the home.
b) There are a lot of missing values in the data set, especially for the indoor measurements at the children's houses (x1.di, x1.ni). 3 strategies were considered for imputing data for these missing values; (i) Replacing the missing values with the stationary site data, (ii) Replacing the missing values with the average measurements taken at the child's home, (iii) Replacing the missing values with the average indoor and outdoor values respectively measured at other children's homes in the same region on the same day. Comment on these strategies and what assumptions are implicitly made in each case. Propose one more strategy and motivate why you think that's a reasonable approach.
i) Assume that there is not much variation between regions and houses and day-to-day variation dominates instead. ii) Assumes that the house is the most importance factor and that this variation dominates over day-to-day variation. iii) assumes day-to-day variation dominates but that region can also have an effect and that this variation dominates the variation between individual houses in the region. iv) You could use a regression model for the indoor and outdoor temperatures from the other variables (excluding y) to fill in the missing values. A good candidate for a variable to use for imputation is the outdoor ozone at the home for example - based on the correlations in the scatter plot.
c) What if you could collect more data; of a similar kind or perhaps collecting additional information? Construct a short proposal - motivate why you think this additional information would be helpful. Also comment on how costly/complex your proposed study would be/not

be compared to the original study. Why is your proposal possibly improving/assisting model building for predicting individual ozone exposure.
It would be useful to have information about ozone levels indoors where the kids spend a lot of time that's not at home. Measuring ozone at the school should be relatively cheap. You could collect both outdoor and indoor measurement.
Weather data would also be helpful: sunny, rainy, windy: these data already exist so just need to be joined with the current data set.
A more expensive, but probably needed, improvement to the study is to increase the number of kids. An increased sample size would allow us to construct a predictive model with more confidence and we would want a better overlap of multiple kids on multiple days so that we see the personal exposure for both high and low ozone level days for each kid.
The activity report is rather limited and probably not that accurate - outdoor where? indoor where? This is an older data set. Now we could probably use GPS and mobile apps to get a better accuracy of reporting.