Department of Statistics

COURSE STATS 330

Model answers for Assignment 1, 2008

The data sets for this assignment were in the files **romania.txt** (for Question 1) and **weights.txt** (for Question 2) which were available on the course web page.

Question 1.

1. *Load the data into R, and make a data frame **chirot.df** to contain the data. Check for any typographical errors. [5 marks]*
The code

```
chirot.df = read.table(file.choose(), header=T)
```

will read in the data. Alternatively, to grab the data directly from the web page, use

```
chirot.df =
read.table("http://www.stat.auckland.ac.nz/~lee/330/datasets.dir/romania.txt",   header=T)
```

There were no typos in the data set, in the sense that the data in the computer file matched the data given on the assignment sheet.

2. *Are either of these theories supported by the data? What is the relationship between intensity and the other variables? Is the relationship planar? Draw suitable plots to answer this question. Don't try and fit any models. [20 marks, 10 for the plots and 10 for the discussion]*

As a first step, a pairs plot will tell us which of the variables are related to the response:

```
pairs(chirot.df)
```

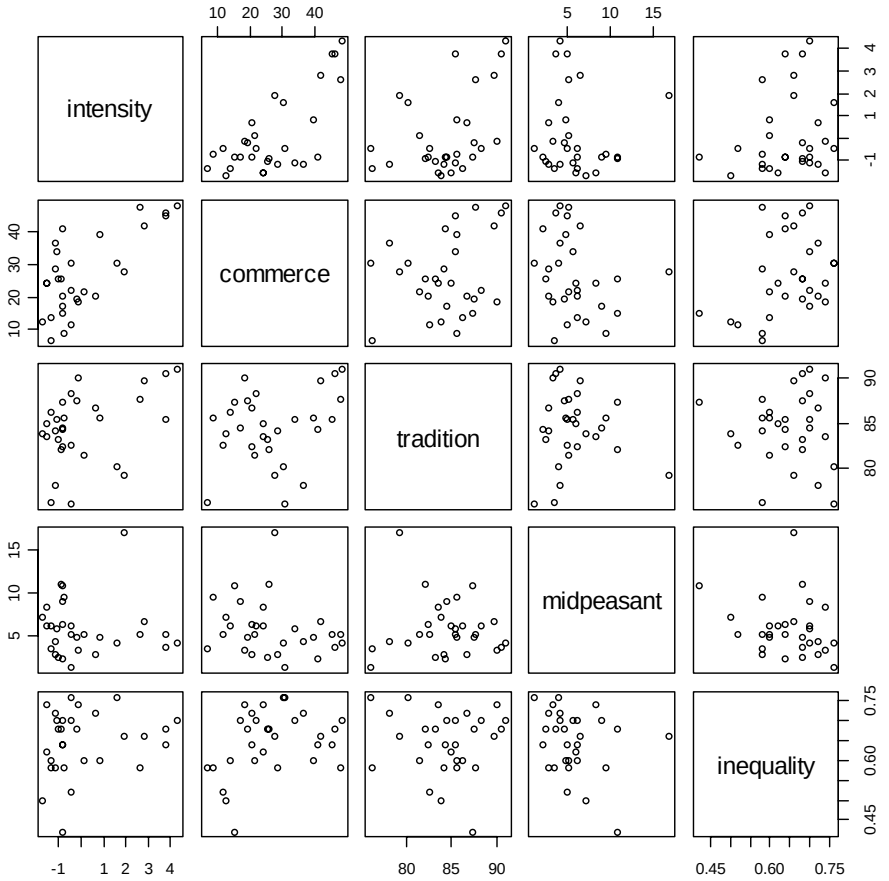
The plots involving the variable midpeasant are not very clear due to the two counties having large values for the variable midpeasant. We can redraw the plot without these two counties:

```
pairs(chirot.df[-c(31,32),]) # omit rows 31 and 32
```

From the plot (shown below) the plot and it seems clear there is not much relationship between intensity and midpeasant. There do seem to be relationships between intensity, commerce and tradition , with a very weak relationship between intensity and inequality. Computing a correlation matrix shows the same thing:

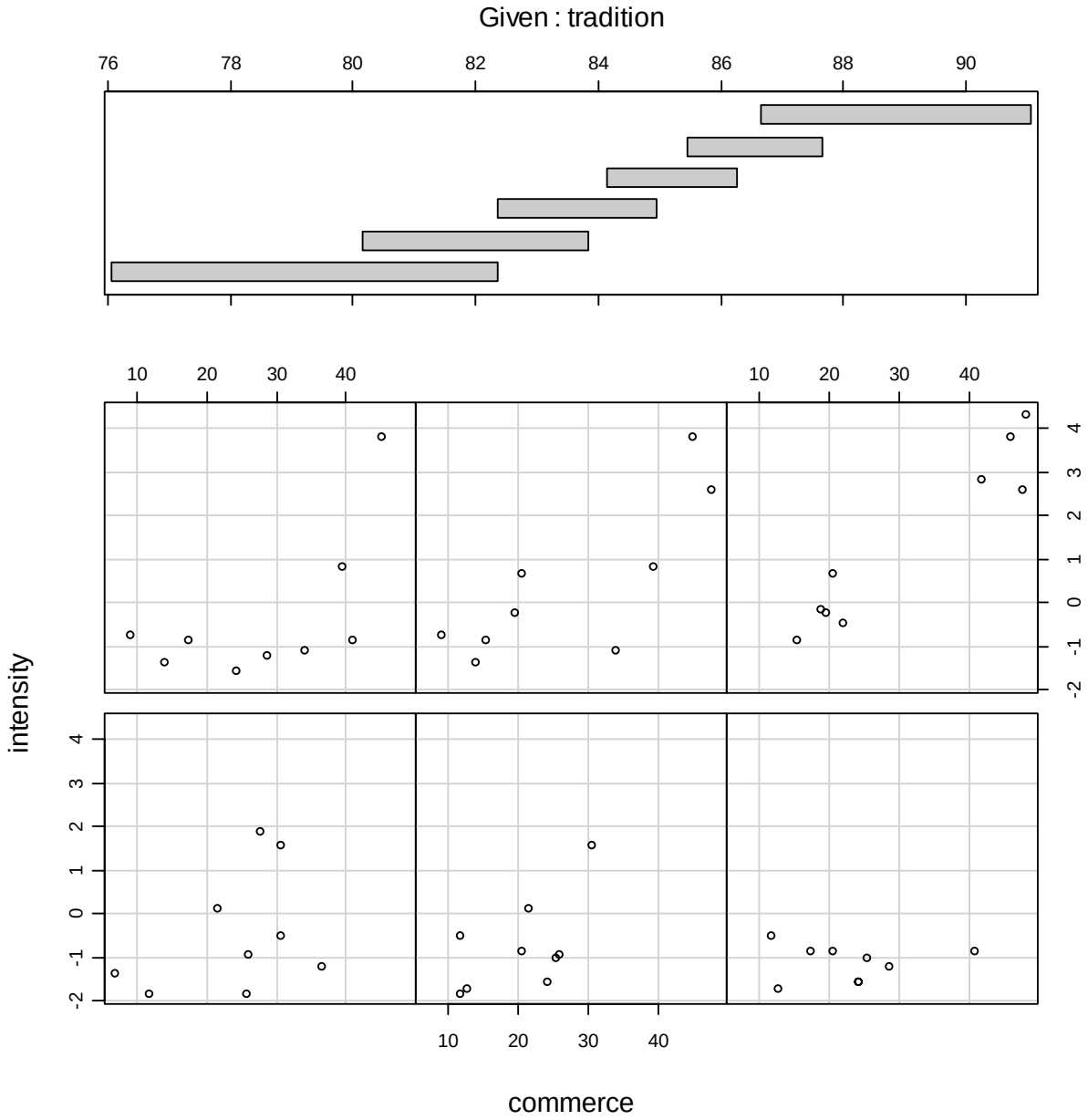
```
> round(cor(chirot.df[-c(31,32),]),2) # omit rows 31 and 32
```

	intensity	commerce	tradition	midpeasant	inequality
intensity	1.00	0.71	0.42	-0.04	0.18
commerce	0.71	1.00	0.28	-0.25	0.33
tradition	0.42	0.28	1.00	-0.05	-0.06
midpeasant	-0.04	-0.25	-0.05	1.00	-0.25
inequality	0.18	0.33	-0.06	-0.25	1.00



The first theory states that there is a relationship between intensity and commerce and tradition. To investigate this further, we can do some coplots. After some experimentation, we obtained the following plot, using the code

```
coplot(intensity~commerce|tradition,data=chirot.df)
```



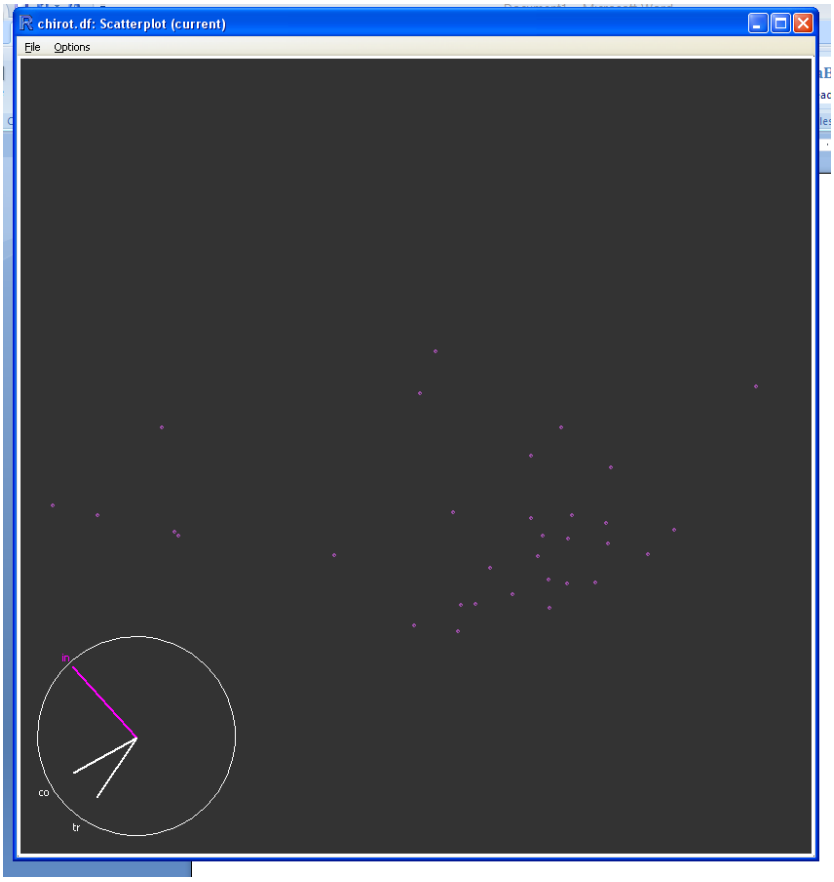
This shows that the relationship between intensity and commerce is quite strong when tradition is high, but is not as strong when tradition is low. The relationship is not planar, since the linear relationships in the coplot are not parallel.

We can draw a trellis plot as a substitute, but have to use the function equal count to turn the continuous variables into “shingles” (categorical variables). This is a bit more complicated, as we have to define new variables to store the shingles.

Also, we can use the Ggobi spinner. We can use Ggobi as a stand-alone program, or alternatively, if the R package rggobi is loaded, we can start Ggobi from within R using the code

```
library(rggobi)
ggobi(chirot.df)
```

A screen capture of the Ggobi plot using intensity, commerce and tradition is shown on the next page. There seems to be a definite curved relationship between these three variables.



We can sum up the discussion as follows.

There is no relationship between midpeasant and intensity.
There is a weak curved relationship between inequality and intensity. The few counties having low inequality all had low intensity. For the bulk of the counties having inequality over 0.6, there was no relationship. There is only very weak evidence for the “structural” theory.

There was a definite curved relationship between intensity and tradition and commerce, with intensity increasing as commerce and tradition increase. It would seem that the “transitional” theory has some merit.

3. *Are there any features of the data that might make fitting a regression model difficult?* [5 marks]

As we noted in the pairs plot, there are two large outliers in the midpeasant variable, which would most likely have a bad effect on the analysis. We will study how they affect the analysis in later lectures. To fit a regression to the variables intensity, commerce and tradition (i.e. to test the transitional theory) is made difficult by the fact the data are not planar. We would have to transform the data in some way.

[5 marks, two for the outlier comment and 3 for the planar comment]

Question 2

1. *Load the data into R, and make a data frame **weights.df** to contain the data. Check for any typographical errors. [3 marks]*

```
   weights.df =
read.table("http://www.stat.auckland.ac.nz/~lee/330/datasets.dir/weights.txt",
header=T)
```

There are no typos but there is one strange data value in the printed sheet, where the 12^{th} entry in the data file seems to have weight and height swapped.
We will swap them back:

```
weights.df[12,2]=57
weights.df[12,3]=166
```

2. *Is there any evidence in these data of systematic over- or under-reporting of height and weight? Does the sex of the subject make a difference? As in Question 1, don't fit any models, confine yourself to graphical methods. [7 marks, 3 for graph and 4 for discussion]*

A very effective graph to answer this question is to draw a trellis plot of weight versus reported weight separately for males and females. If there is no tendency for over and under reporting of weights, the plots should cluster around the 45^{o} line (representing the fact that the reported and actual weights are approximately equal.)

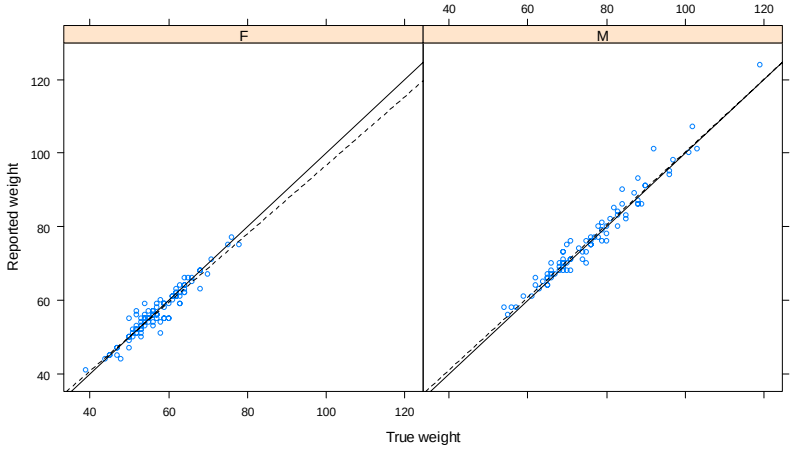
Drawing males and females separately allows the relationship between real and reported weights for males to be easily compared to that for females. It is also helpful to show a least squares line to show the actual relationship between heights/weights and reported heights/weights.

The code for weights is

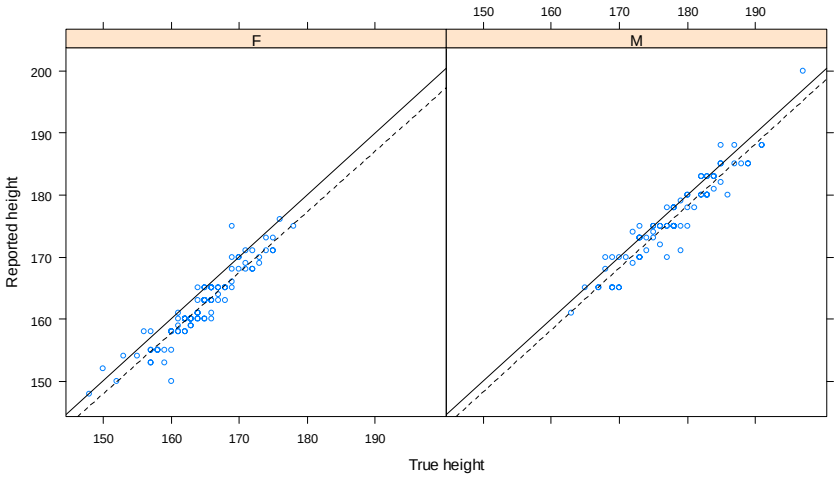
```
xyplot(repwt~weight|sex,
panel=function(x,y){
panel.xyplot(x,y)
panel.abline(0,1)
panel.lmline(x,y, lty=2)},xlab="True weight",ylab = "Reported
weight", data=weights.df)
```

Note the use of xlab and ylab to give nice axis labels, and the "panel function" to draw the 45^{o} line (solid) and ls line (dashed). See the Tour of Trellis graphics on p 12 for more on panel functions.

The trellis graph is shown overleaf and clearly shows that there is no substantial under-or over-reporting for females or males: the points cluster about the 45^{o} line. We have also added a least-squares line to the plots. These are very similar to the 45^{o} lines. There is a very slight tendency for light and heavy males to over-report their weight.



The equivalent picture for height is shown below. Both males and females tend to slightly under-report their heights, females more so than males.



Notes for markers

Question 1

Question 1.1 Read in data – give 3 marks for evidence of doing this, 1 marl for saying there are no typos, one mark for commenting on the two counties with large values of midpeasant.

Q1.2 Give 3 marks for discussion of the relationships, 4 if the theories seem true, and 3 for commenting on the planarity (it isn't).
Give 10 marks if these comments are supported by suitable plots (pairs, coplots with sensible choice of conditioning and relationship variables, ggobi.)

Q.13 two marks for outlier comment and 3 for planarity comment.

Question 2

Q2.1 Reading in data: 1 marks. No typos 1 mark. Comment on strange value 1 mark.

Q2.2 Plots: give 3 for proper trellis plot (including 45^o line, -1 if no line). Give 2 for plot of differences. Discussion: no under/over estimation of weight 2 marks, Slight underestimation of height 1 mark, different for males and females 1 mark.