Department of Statistics

COURSE STATS 330/762

Model Answers for Assignment 1, 2011

Question 1.

Instructions

1. *Load the data into R, and make a data frame **oats.df** to contain the data. Check for any typographical errors (the data below may be taken to be the correct data, but the data on the web may have been corrupted). Print out the last 10 lines of the data file.* [5 marks]

The following code will read in the data, using the web address syntax:

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

You can check that the data are correct by comparing this file with the one printed on the question sheet. You could also draw boxplots of the variable Y to check for any gross errors. There are in fact none. To print out the first 10 lines, use the R subsetting facility []

```
oats.df[1:10,]
```

```
   B           V      N   Y
1  I     Victory 0.0cwt 111
2  I     Victory 0.2cwt 130
3  I     Victory 0.4cwt 157
4  I     Victory 0.6cwt 174
5  I Golden.rain 0.0cwt 117
6  I Golden.rain 0.2cwt 114
7  I Golden.rain 0.4cwt 161
8  I Golden.rain 0.6cwt 141
9  I  Marvellous 0.0cwt 105
10 I  Marvellous 0.2cwt 140
```

[5 marks]

2. *What is the relationship between the amount of fertilizer applied and the yield? Does the relationship depend on the block or the variety? If so, how? Draw suitable plots to answer this question* [5 marks]

We can draw a trellis plot using Y and N as relationship variables and Vand B as conditioning variables:

```
dotplot(Y~N|V*B, data=oats, xlab="Amount", Ylab = "Yield")
```

The plot (not shown) is enhanced somewhat by joining up the plotted points. This is done by typing

```
dotplot(Y~N|V*B, data=oats, xlab="Amount", Ylab = "Yield",
type="l")
```

The resulting plot is shown overleaf. From the plot it is clear that the yield goes up with the amount of fertilizer, as we would expect. The slope and height of the plots don't seem to depend much on the variety, but the blocks seem to differ: for example block I has higher yields than block IV.

The effect of the blocks can be made more obvious by plotting the relationships on the same graph, one for each variety, making sure the Y-scales are the same for each graph. We can do this using code similar to that used in the rat example.

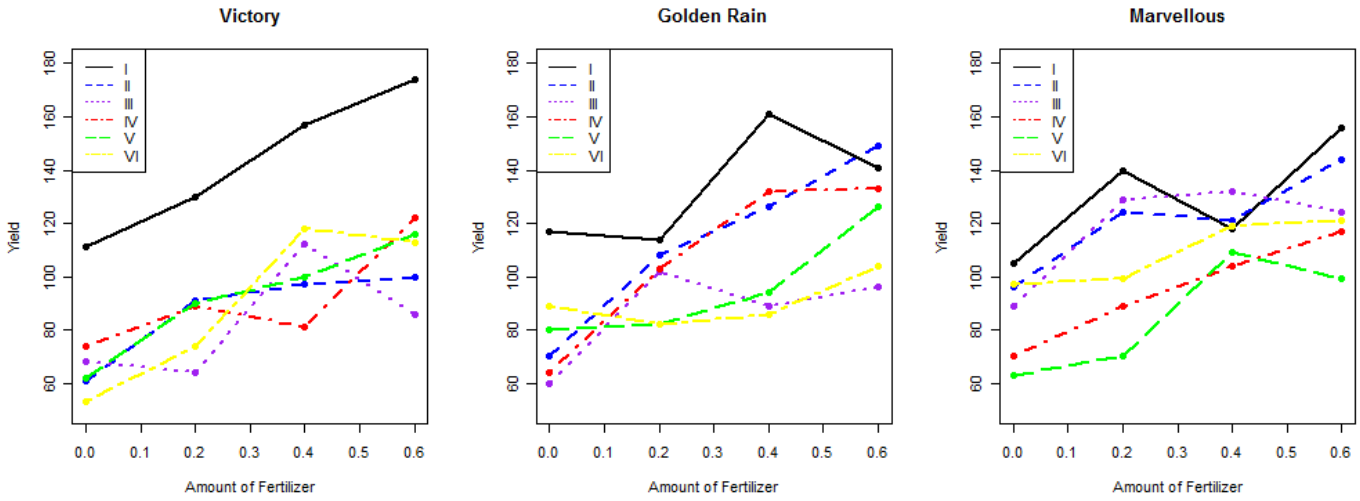
The code is

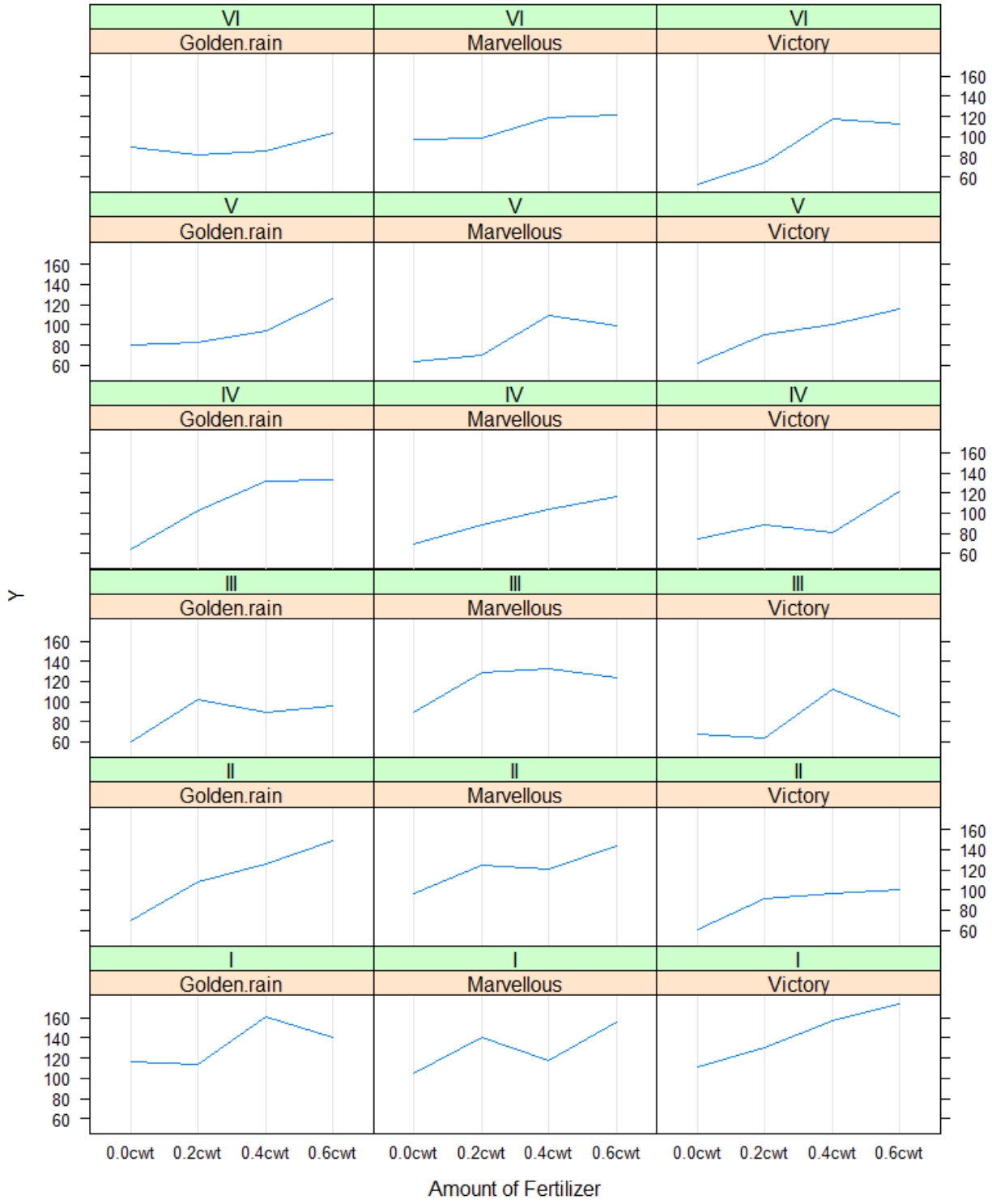
```
n.amount = c(0,0.2,0.4,0.6)
yield = oats.df$Y
mycol = c("black", "blue", "purple", "red", "green",
"yellow")

# set up 3 plots in the one graph

par(mfrow=c(1,3))

# first draw plot for the victory variety

temp = oats.df[oats.df$V == "Victory",]
plot(c(0,0.2,0.4,0.6), yield[1:4], ylim = c(50,180),
xlab = "Amount of Fertilizer",
ylab = "Yield",
main = "Victory",
type="n")   # note this does not plot anything, just sets up
the axes
# now do the plots
use = 1:4
for(j in 1:6){   # loop over the 6 blocks to draw 6 lines
    points(n.amount, yield[use], pch = 19, col = mycol[j])
    lines(n.amount, yield[use], lty=j, lwd=2, col = mycol[j])
    use = use + 4   # pick out the next group of 4
```

```
}

# make a legend
legend.text = c("I","II","III","IV","V","VI")
legend("topleft", legend = legend.text, lty = 1:6,
col = mycol)
```

Repeat for Golden Rain and marvelous.

The resulting picture makes it much clearer which are the best blocks: I (black) tends to have a higher intercept than the others. The slopes (i.e. the response to the fertilizer) seem to be similar for the three varieties.





3. *Which variety had the highest yield overall?[* 5 marks]

From the graphs it is hard to see if any variety is best: there is too much variation. We can get a better picture by averaging over the blocks:

```
> tapply(oats.df$Y, list(oats.df$V, oats.df$N),mean)
              0.0cwt    0.2cwt   0.4cwt   0.6cwt
Golden.rain 80.00000  98.50000 114.6667 124.8333
Marvellous  86.66667 108.50000 117.1667 126.8333
Victory     71.50000  89.66667 110.8333 118.5000
```

Now it is clear that Marvelous is the best at every level of fertilizer. To determine if this is a real finding, or just a chance result for this experiment, we have to fit some models. More later!

4. *Which block had the highest yield? Did some varieties do better in some blocks than others?*[5 marks]

   We have seen from the graphs that Block 1 had the highest yield for all varieties, but that for the other blocks, the question of what block was best depended on the variety. If we average over the fertilizer amounts, we get the following

```
> tapply(oats.df$Y, list(oats.df$V, oats.df$B),mean)
                 I     II    III    IV     V     VI
Golden.rain 133.25 113.25  86.75 108.0 95.50  90.25
Marvellous  129.75 121.25 118.50  95.0 85.25 109.00
Victory     143.00  87.25  82.50  91.5 92.00  89.50
```

Thus, Block II is second best for Marvellous and Golden Rain, but not for Victory. So, some varieties do better in some blocks than others.

Question 2

In his classic book on statistical graphics (*The Elements of Graphing Data*, Wadsworth, Monterey, California, 1985) , William Cleveland argues that several principles should govern the drawing of graphs. One of these principles relates to the placing of labels on scatter diagrams. The two figures overleaf illustrate this.

The data file **Animals.txt** contains a set of data very similar to that graphed in Cleveland's Figure 2.23 reproduced overleaf. There are two variables**, body** and **brain** containing the body and brain weights of 28 different animal species.

Instructions

1. *Load the data into R. You may assume that there are no errors in this data set.* [5 marks]

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

Note that there are row labels included in this data set. To read the data in, let R figure this out, by omitting header=TRUE.

2. *Draw a scatter plot of the data that resembles as much as possible Cleveland's Figure 2.22 shown overleaf. The key idea here is to avoid the bad features in Fig 2.23 and implement good labeling of the species. Pay careful attention to the labeling of the points and the axes.* [15 marks]

Required features of the plot are

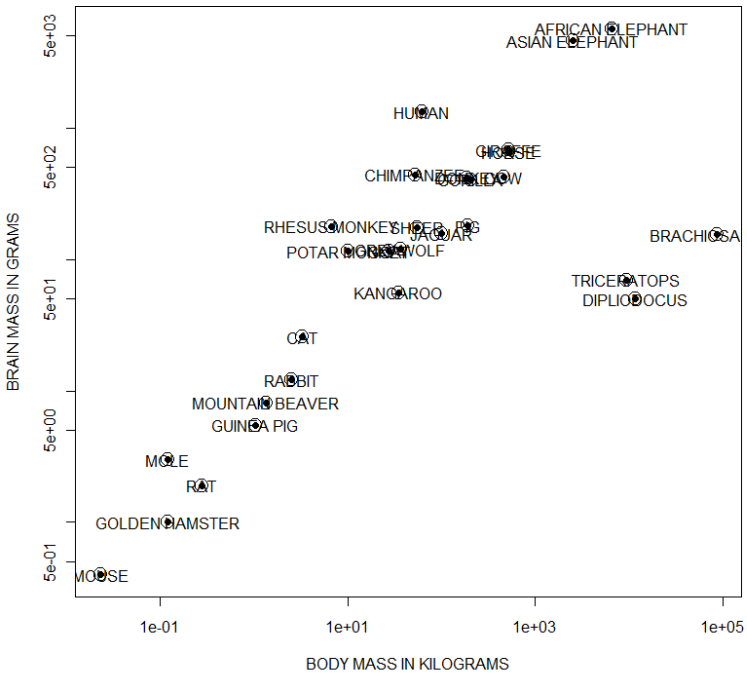
I. Labeled axes
II. Bold plotting points
III. Subdued labeling
IV. Plot on log scale.

The following code implements these.

The first attempt uses the log argument to plot, and duplicates the plotting symbol (the dot withing the circle) by plotting twice (once with pch=19 for the dot), and then with an enlarged circle for the outer circle (pch=21 for the circle, and cex=2 to make it bigger).

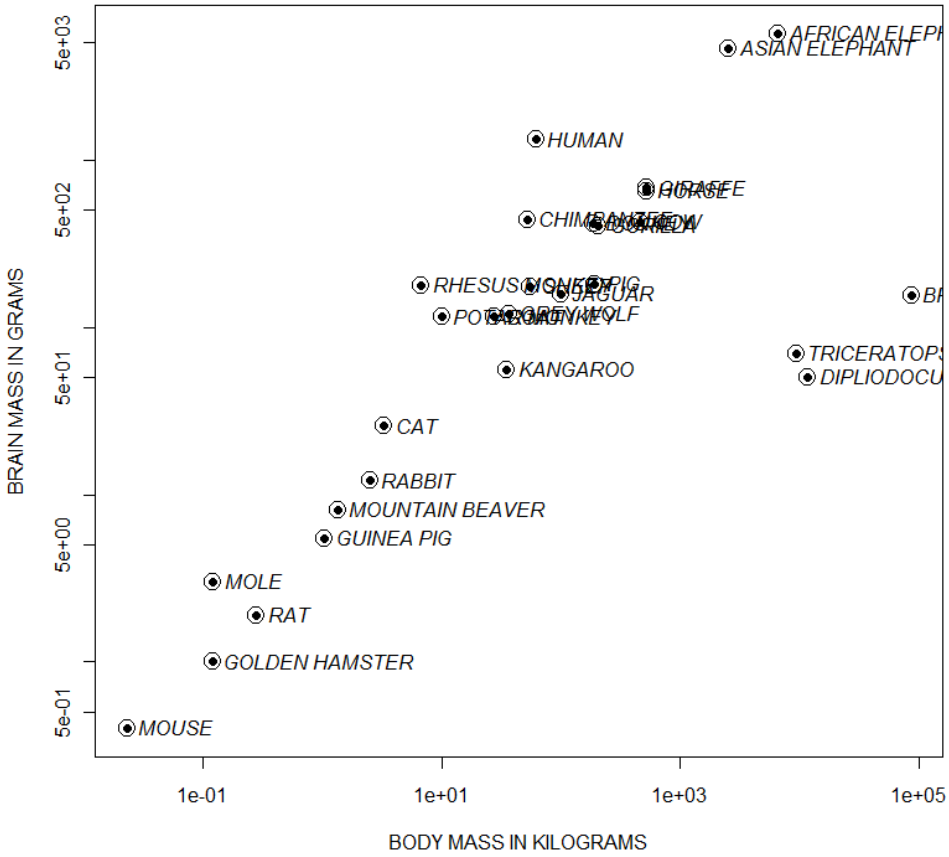
The code is

```
plot(body, brain, log="xy",
xlab = "BODY MASS IN KILOGRAMS",
ylab = "BRAIN MASS IN GRAMS",
pch=19)
points(body, brain, pch=21, cex=2)
text(body, brain, toupper(rownames(animal.df)))
# toupper converts to upper case
```



The result is not very pretty – the names overprint the points, and the axis labeling is not very nice. The secret is to position the labels alongside the points using the pos argument to plot. Setting pos=4 will put the labels to the right. Setting them in italic also helps (font=3)

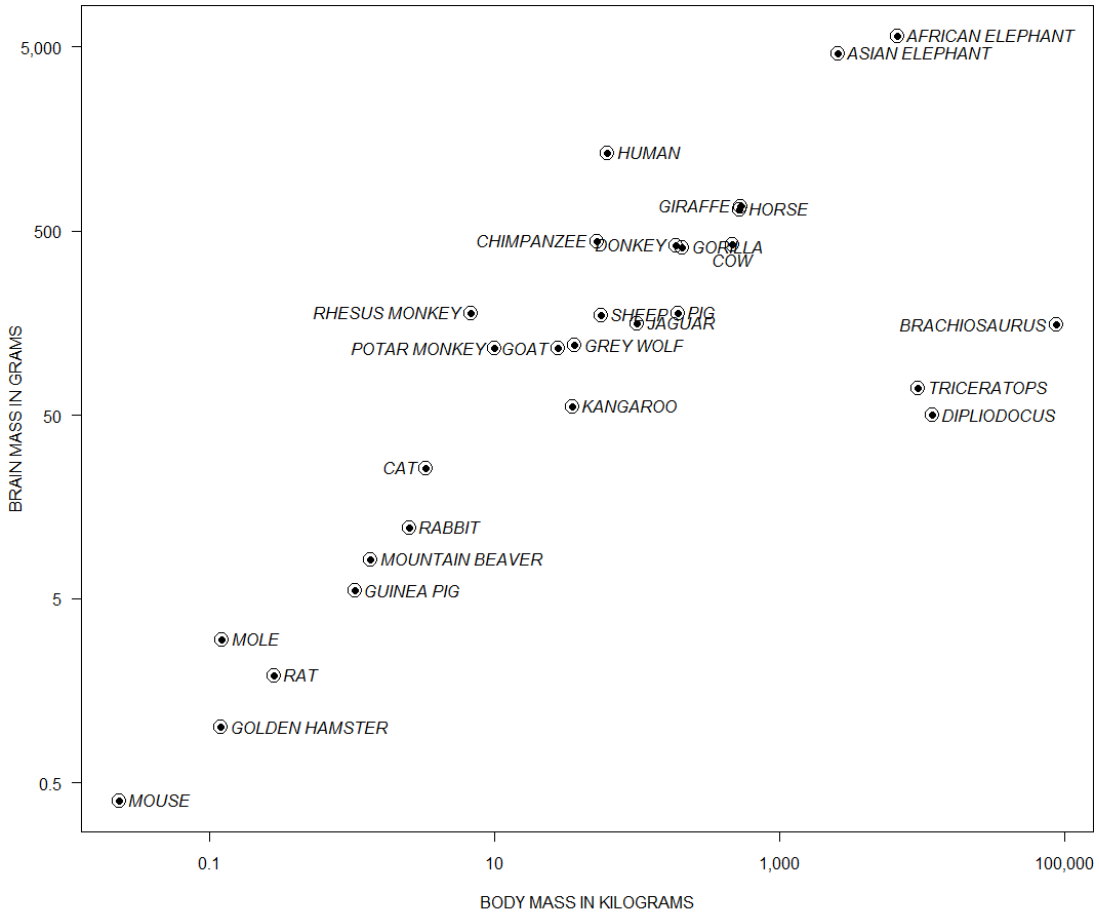
```
plot(body, brain, log="xy",
xlab = "BODY MASS IN KILOGRAMS",
ylab = "BRAIN MASS IN GRAMS",
pch=19)
points(body, brain, pch=21, cex=2)
text(body, brain, toupper(rownames(animal.df)),pos=4, font=3 )
```



This is better, but some names should be to the left, not the right. The fix is to set the pos argument to be a vector, with 28 elements indicating whether the label should be to the right (4), the left (2) or below (1). We can also pretty up the axis labels using the R facility for hand-crafting labels. This results in the code

```
plot(body, brain, log="xy",xlab = "BODY MASS IN KILOGRAMS",
ylab = "BRAIN MASS IN GRAMS", pch=19, axes=FALSE)

axis(2, at = c(0.5,5,50,500,5000),
 labels = c("0.5", "5", "50","500","5,000"), las=1)
axis(1, at = c(0.1,10,1000,100000),
 labels = c("0.1", "10", "1,000","100,000"), las=1)
# las=1  makes the labels horizontal
box()   # drws the box around the graph
points(body, brain, pch=21, cex=2)
pos = rep(4,28)
pos[c(4,8,10,11,12,17,24,26)]=2
pos[2]=1
text(body, brain, toupper(rownames(animal.df)),pos=pos,
font=3 )
```

The graph is



[15 marks]

If you got as far as the first graph, you got 8 marks. Fixing the position of the labels got you an extra 4. Prettying up the axis labels was worth 3.