Department of Statistics

COURSE STATS 330

Model answers for Assignment 1, 2005

Question 1.
The data set in the file `gradrate.csv` contains historical data on student graduation rates for 264 colleges in the United States. This file (in the form of a csv file) can be downloaded from the course web page. The data are also reproduced at the end of this assignment.

In the early 80's, there was concern over the graduation rates of student athletes. With some justification, it was thought that college admission boards were placing insufficient weight on the academic caliber of potential recruits, and admitting them purely on atheletic ability. Accordingly, the National Collegiate Athletic Association (NCAA) introduced rules (Proposition 48) that set minimum admission standards for student athletes. These rules came into force in 1986.

The data in the file gives graduation rates (as percentages) for all students (as) and student athletes (sa) for students entering college in the indicated year, with the exception of sa83 and as83, which refer to all students entering college between 1983 and 1985 inclusive. Thus the variables sa83 and as83 refer to students admitted **BEFORE** the new rules came into effect, and the other variables refer to students admitted **AFTER** the rules came into effect.

Load the data into R. Then answer the following by drawing suitable graphs. Do not fit any models for this question.

The data were in the file `gradrate.csv.` You can read it in and make a data frame called `gradrate.df` using the code

```
gradrate.df<-read.table(file.choose(), header=T, sep=",")
```

or, a bit more simply as

```
gradrate.df<-read.csv(file.choose(), header=T)
```

[5 marks]

1. Is there any evidence that the rules have had the effect of raising the graduation rates of student athletes? Draw a graph to support your conclusion.

We need to draw a graph that shows if the distribution of the graduation rates changes between 1983 and subsequent years. Side by side boxplots are a good choice. Use the code

```
boxplot(sa83, sa86, sa87, sa88, sa89, sa90, sa91,
names=c("1983","1986","1987","1988","1989","1990","1991"),
ylab="Graduation rate",
main="Graduation rate of student atheletes")
abline(h=median(sa83), lty=2)
```

This gives the plot below. Note the strange value for 1991 (Maryland-Eastern Shore). It is the same as in the printed sheet so we will retain it even though it is probably wrong.

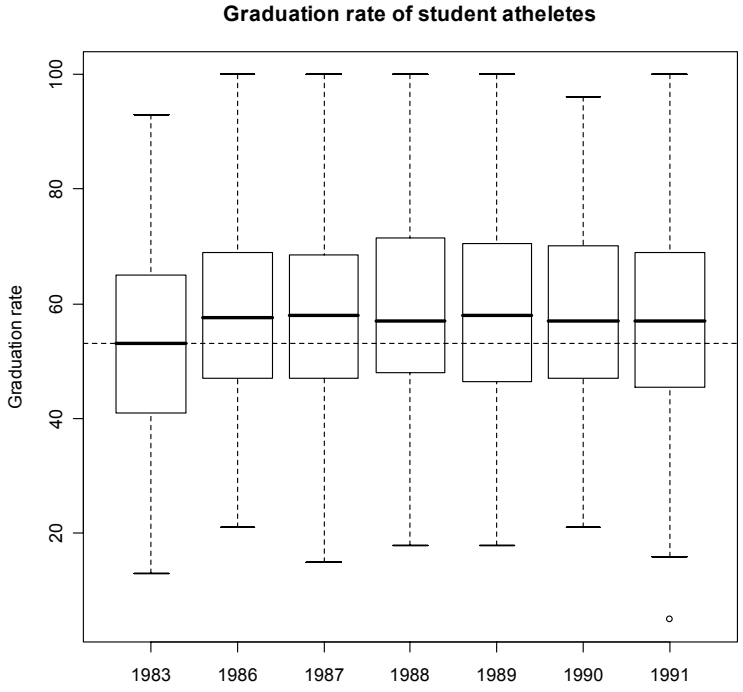


Figure 1. Plot for Part 1.

Note how we have labeled the axes, given the plot a title, and drawn a dotted reference line. This makes it easier to compare 1983 to the other years.
Conclusion: The graduation rates seem to have jumped after 1983 and remained at a higher level. It would appear that the rules might have had some effect, but it could be due to a rise in the rate for all students.

[5 marks. Marks were deducted for not labeling axes or drawing an inappropriate plot, and not commenting on the outlier]

2. *Is there any change in the relative graduation rates of student athletes (i.e. relative to all students)? Again, draw a graph to support your conclusion.*

We interpret "relative to all students" as the ratio of athlete graduation rate to "all student" graduation rate. The difference would also have been OK.

The following code calculates the ratios and draws boxplots as in part 1.

```
ratio83 <- sa83/as83
ratio86 <- sa86/as86
ratio87<- sa87/as87
ratio88 <- sa88/as88
ratio89 <- sa89/as89
ratio90 <- sa90/as90
ratio91 <- sa91/as91

boxplot(ratio83, ratio86, ratio87, ratio88, ratio89, ratio90,
ratio91,
names=c("1983","1986","1987","1988","1989","1990","1991"),
ylab="ratio of athelete to all student rates",
main="Ratio of athelete to all student graduation rates by
year")
abline(h=median(ratio83), lty=2)
```

The boxplots are shown overleaf. We see that the ratio is consistently higher by a small margin after the rule change. Thus, athletes seem to be doing better relative to all students after the rule change.

[5 marks. Marks were deducted for not interpreting "relative" correctly, and for not labeling the graph.]

3. *What is the relationship between student athlete graduation rates and all student graduation rates? Does this relationship change from year to year?*

Here we want to see if the relationship between athlete and all student graduation rates changes over the years. For each year, the relationship is best shown by a scatter plot of athlete rate versus all student rate (remember that we use a scatter plot to show the relationship between two continuous variables). We can use trellis graphics and condition on year. To do this, we need to rearrange the data as three variables, namely athlete rate, student rate and year. We use the R functions `c` and `rep` to do this, noting that there are 256 observations on each variable for each year. We use the code

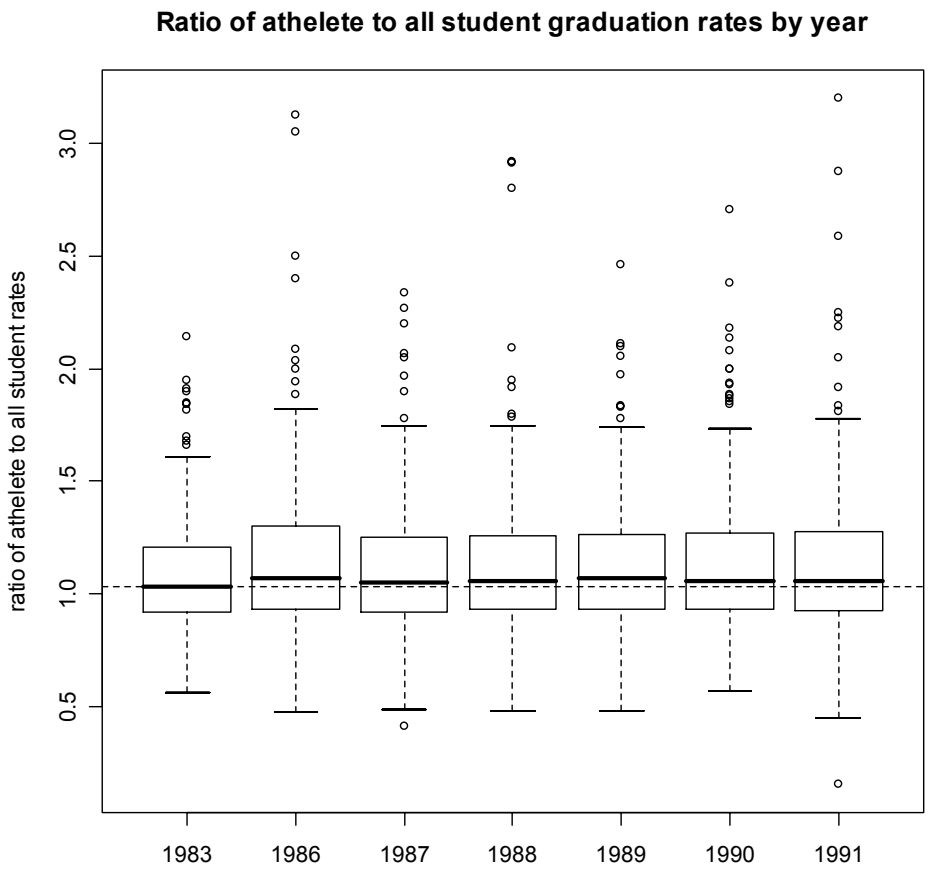


Figure 2. Graph for part 2.

```
athletes<-c(sa83, sa86, sa87, sa88, sa89, sa90, sa91)
students<-c(as83, as86, as87, as88, as89, as90, as91)
years<-rep(1:7, rep(256,7))
```

We can draw a basic trellis graph using the code

```
trellis.par.set(theme = col.whitebg())
xyplot(athletes~students|years)
```

This produces a picture like Figure 3, but without the lines. Note that the first line is used to change the background of the plot to white to make it easier to read after it is printed.

Its not easy to see if the relationships change after 1983. We could enhance the plot by drawing the fitted least squares line on each plot. This requires the use of "panel functions". You can find out more about these by going to the website for the course STATS 760 and downloading the "trellis tutorial". The URL is http://www.stat.auckland.ac.nz/~lee/760/links.php.

The code is

```
xyplot(athletes~students|years,
panel=function(x,y){
   panel.xyplot(x,y)
   panel.lmline(x,y)})
```

which produces the picture shown in Figure 3.

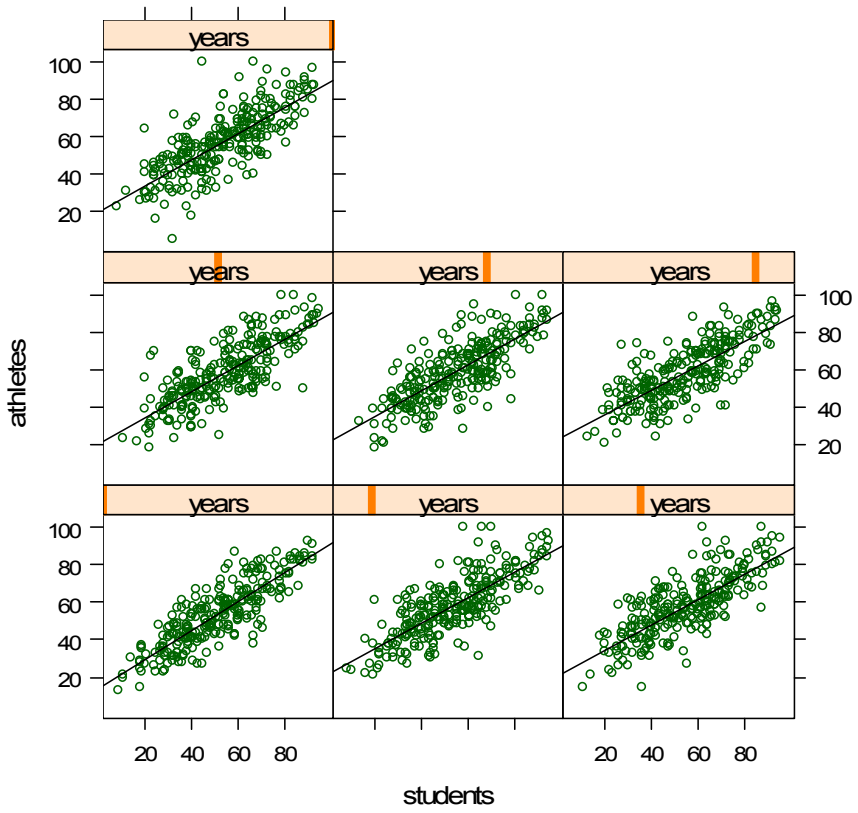


Figure 3. Trellis plot with least squares lines added.

It seems that the line for the post 1983 plots is a bit higher than for the first plot, indicating that the athletes grad rates are higher for a fixed all student rate after 1983. We can calculate the slope and intercept for each line using the code

```
for(j in 1:7){
   y<-athletes[years==j]
   x<-students[years==j]
   reg.stuff<-lm(y~x)
   print(coef(reg.stuff))
}
```

This produces the output

```
(Intercept)           x
  13.830859    0.768915
(Intercept)           x
 21.2056147   0.6862438
(Intercept)           x
 21.0116678   0.6790637
(Intercept)           x
 20.4113968   0.6974881
(Intercept)           x
 21.1074548   0.6836214
(Intercept)           x
 22.4391756   0.6575261
(Intercept)           x
 20.1496292   0.6925427
```

Thus, the first intercept (the one for 1983) is clearly less than the others. The slope is also greater. To compare the lines we could also plot them all on the same graph, as is shown in Figure 4 on the next page. From this, we see that the effect of the rule change is greater in colleges where the graduation rate for all students was lower.

Code for Figure 4 was

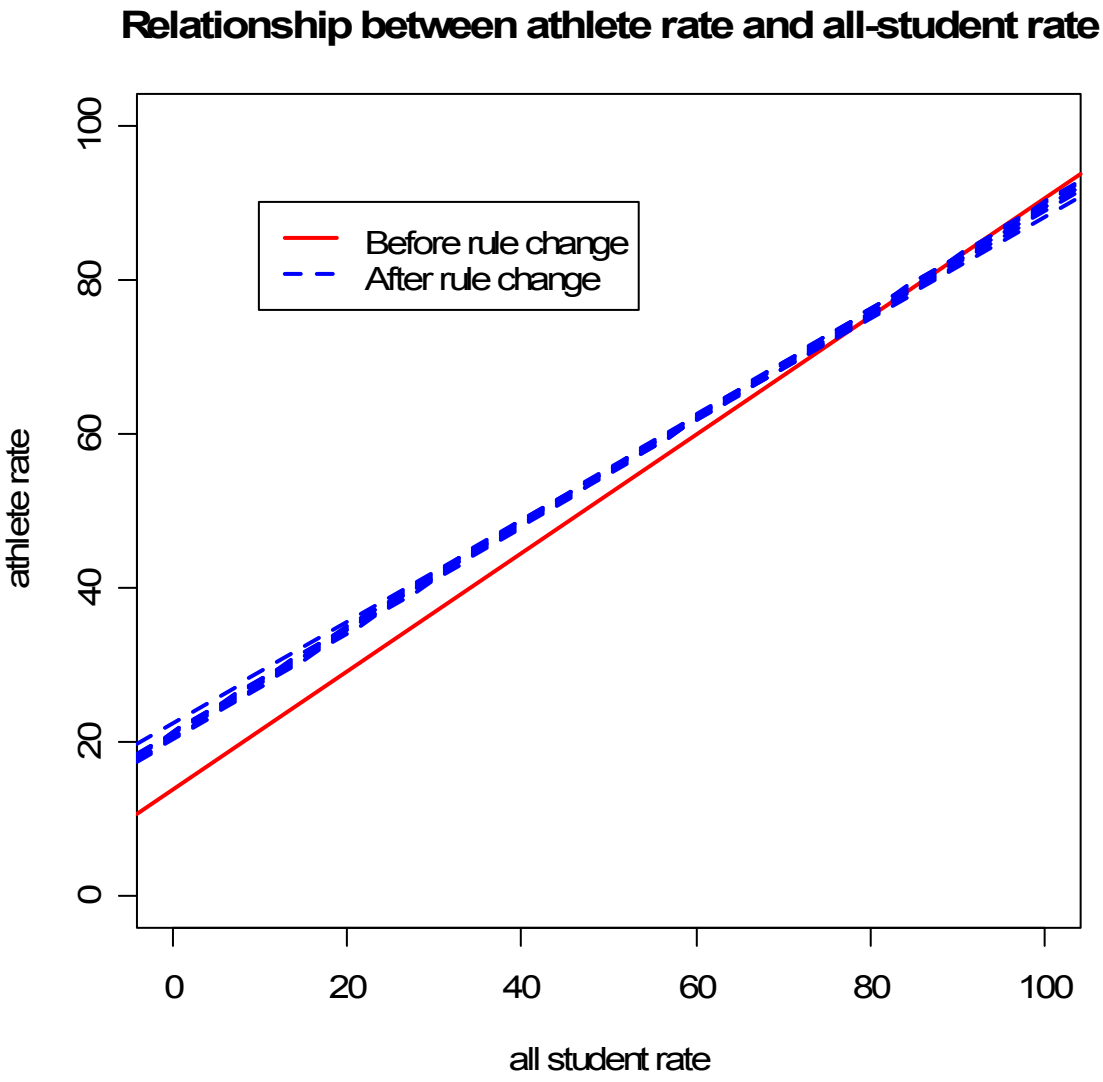
```
plot(c(0,100),c(0,100), type="n", xlab="all student rate",
ylab="athlete rate",
main="Relationship between athlete rate and all-student rate")

colour=c("red","blue","blue","blue","blue","blue","blue")
linetype=c(1,2,2,2,2,2,2)
for(j in 1:7){
y<-athletes[years==j]
x<-students[years==j]
reg.stuff<-lm(y~x)
abline(coef(reg.stuff), col=colour[j], lwd=2, lty=linetype[j])
}

legend(10,90,legend=c("Before rule change","After rule
change"), col=c("red", "blue"), lwd=2, lty=1:2)
```

Note the use of the function `abline` to draw the lines, and a loop to draw each line in turn. The function `coef` extracts the slope and intercept from the least squares fit.

Overall conclusion: the rule change seems to have improved the performance of athletes relative to all students.


Relationship between athlete rate and all-student rate
Before rule change
After rule change
athlete rate
all student rate
0
20
40
60
80
100

Question 2.

The data in the data set **clocks.txt** relate to auction sales of antique clocks. The variables in the data set are the sale price in $$ (variable **price**), age in years of the clock (variable **age**) and number of bidders (variable **bidders**). Download the file (which is a text file); read the data into R and then answer the following questions by fitting a suitable regression model, and drawing any suitable graphs.

The file can be read into a data frame clocks.df using the code

1. *Is there any evidence that the number of bidders has an effect on the price? If so, how?*

   We fit the regression and examine the coefficient of bidders:

```
clocks.df <- read.table(file.choose(), header=T)
clocks.lm<-lm(price~age+bidders, data=clocks.df)
summary(clocks.lm)

Call:
lm(formula = price ~ age + bidders, data = clocks.df)

Residuals:
    Min      1Q  Median      3Q     Max
-206.48 -117.34   16.66  102.55  213.50

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1338.9513   173.8095  -7.704 1.71e-08 ***
age            12.7406     0.9047  14.082 1.69e-14 ***
bidders        85.9530     8.7285   9.847 9.34e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 133.5 on 29 degrees of freedom
Multiple R-Squared: 0.8923,     Adjusted R-squared: 0.8849
F-statistic: 120.2 on 2 and 29 DF,  p-value: 9.216e-15
```

   The coefficient of bidders is significant (the p-value is very small). There is strong evidence that the number of bidders makes a difference to the price. For a fixed age of clock, each extra bidder increases the price by about $86.

   [5 marks. Marks deducted if the effect not properly explained]

2. *Is there any evidence that the age of the clock has an effect on the price? If so, how?*

The coefficient of age is also significant (the p-value is again very small). There is strong evidence that the age makes a difference to the price. For a fixed number of bidders, each extra year of age increases the price by about $12.74

[5 marks. Marks deducted if the effect not properly explained]

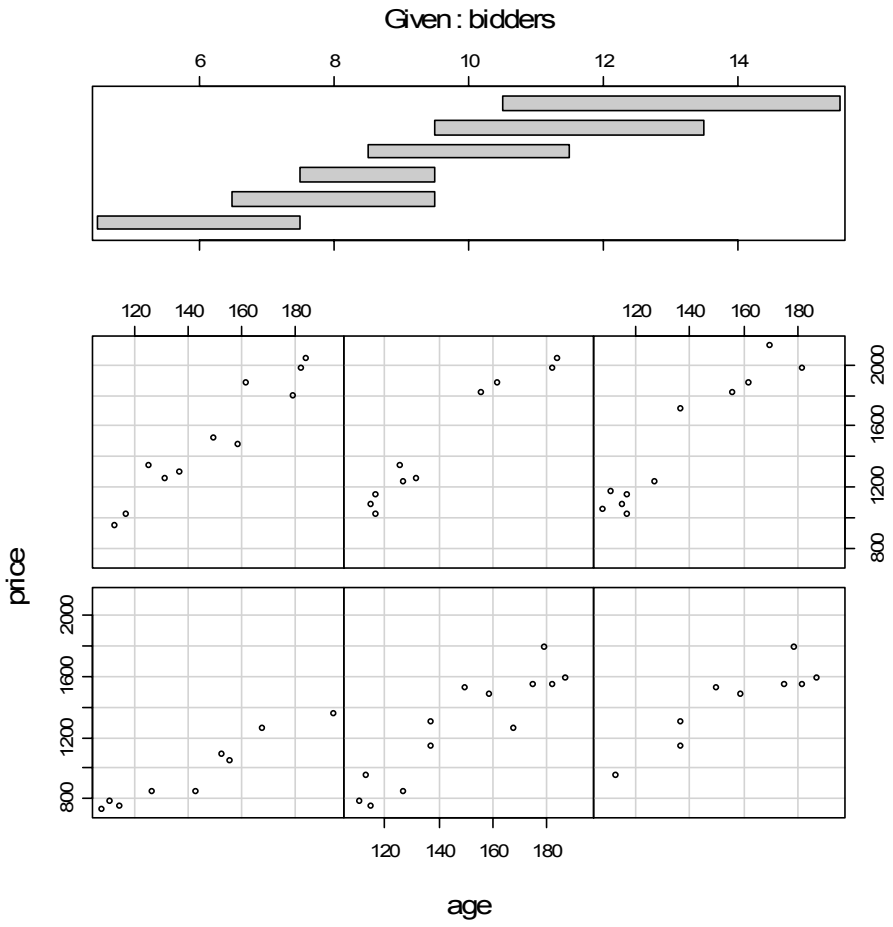
3. *Do you think these data are suitable for a regression analysis (i.e. randomly scattered about a plane)?*

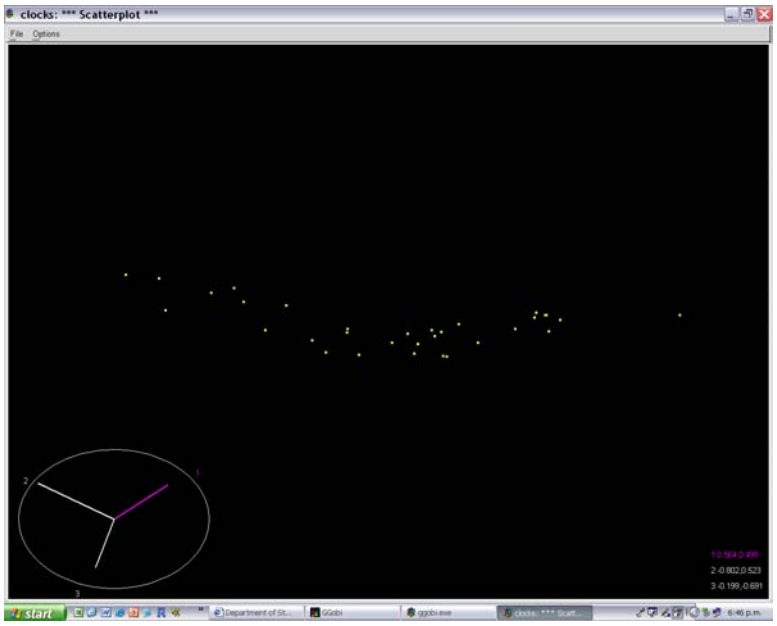
The fit is good (the R2 is 89%). I didn't expect students to do residual plots, as we haven't discussed them in class yet. I expected the use of coplots and spinners.

The coplot (on next page) shows approximately parallel straight bands of points, which implies a reasonable set of approximately planar data.

However, the spinner shows a definite curve, suggesting that the data are not quite planar. See next page.

[4 points for one or the other, 5 points for both if correctly interpreted.


Given : bidders
6
8
10
12
14
120
140
160
180
800
1200
1600
2000
price
age


clocks: *** Scatterplot ***
File
Options