Department of Statistics

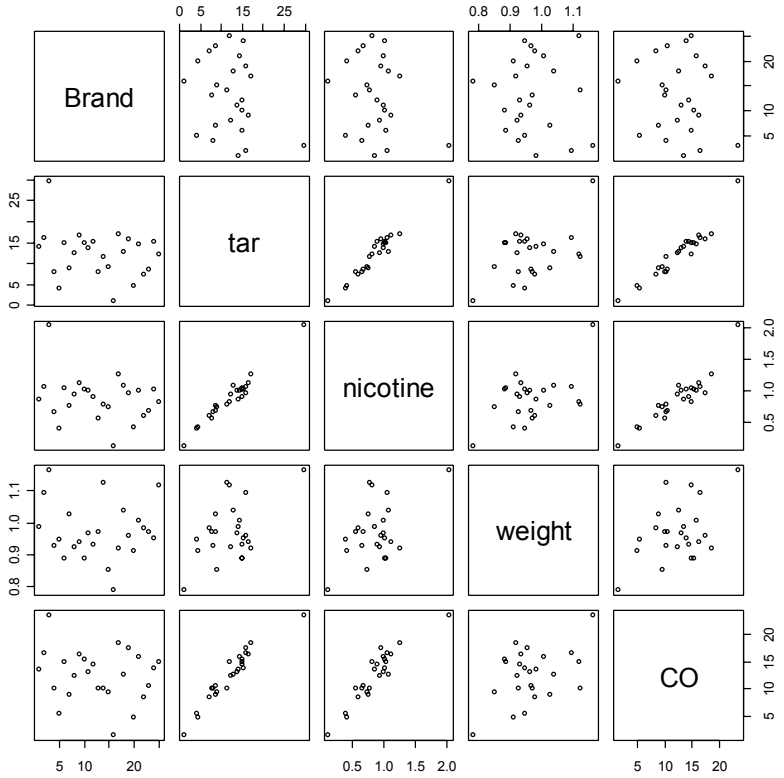
COURSE STATS 330

Assignment 1, 2003: Model Answers

1. *Read the data into R and make a data frame* ***cigs.df*** *. Inspect the data for any unusual values and if you find any, make a new data frame with the outliers deleted. Delete no more than 2 brands. Print out the new data frame in R.*

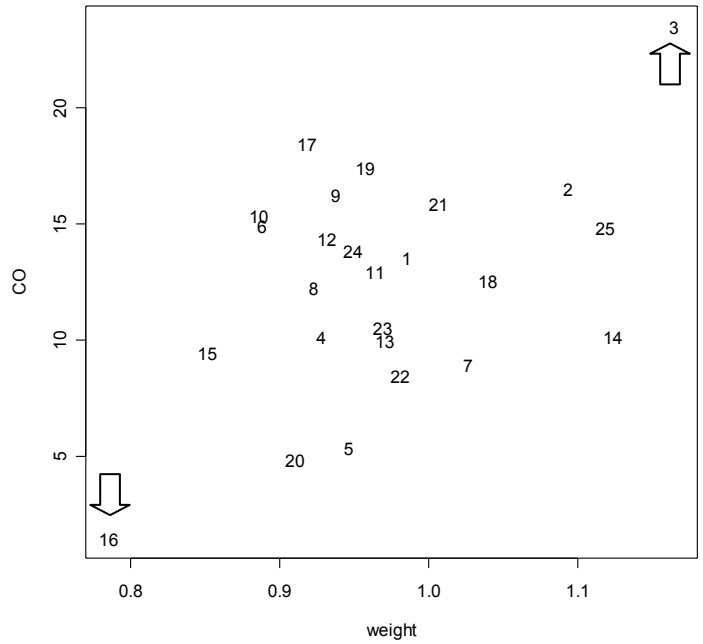
   This can be done using the code

```
cig.df<-read.table(file.choose(),header=T)
pairs(cig.df)
```



   There seem to be 2 outliers: to identify them we could plot CO vs weight, with the cigarette label (1-25) shown on the plot:

```
attach(cig.df)
plot(weight, CO, type="n")
text(weight, CO, 1:25)
```



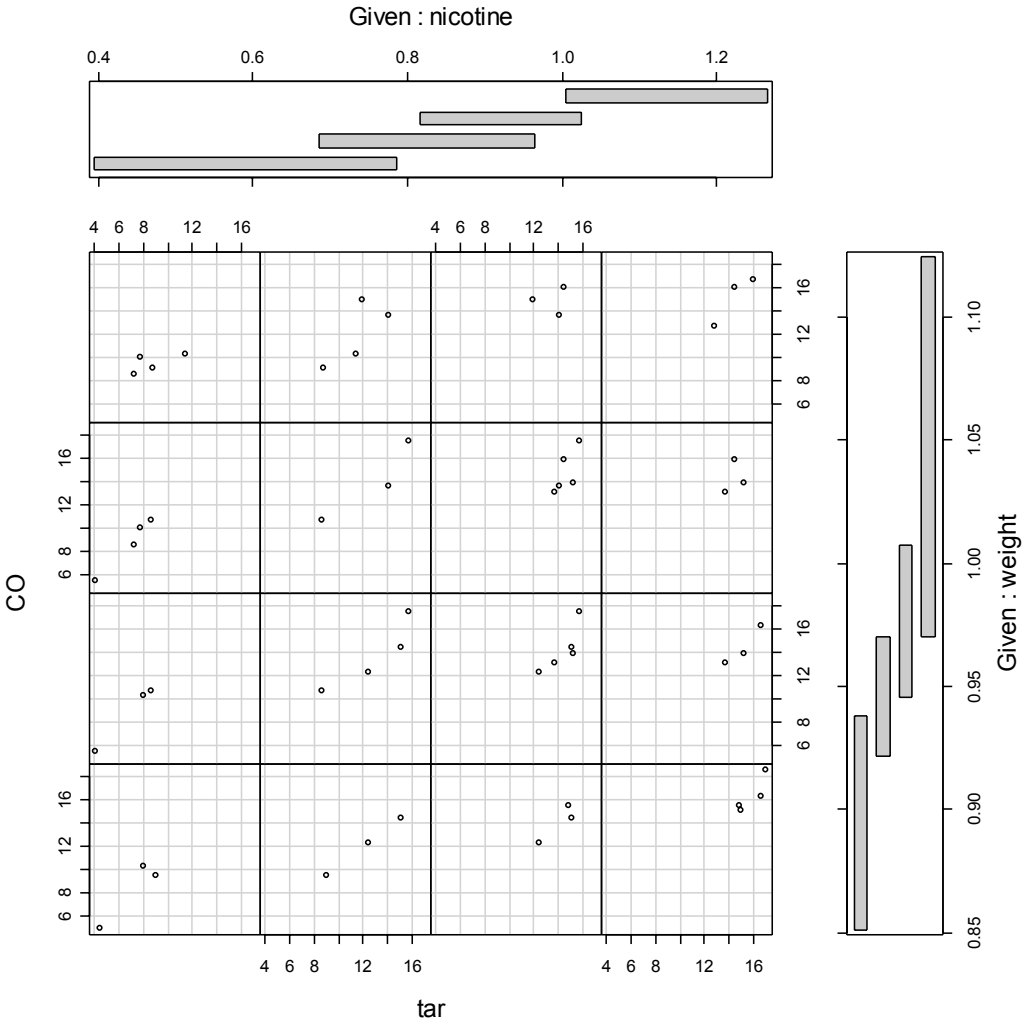
Thus, the outliers are points 3 (Bull Durham) and 16 (Now). We can delete them from the data frame and make and print a new data frame using the code

```
> newcig.df <- cig.df[c(-3,-16),]
> newcig.df
                Brand  tar nicotine weight   CO
1             Alpine 14.1     0.86 0.9853 13.6
2     Benson&Hedges 16.0     1.06 1.0938 16.6
4       CamelLights  8.0     0.67 0.9280 10.2
5           Carlton  4.1     0.40 0.9462  5.4
6      Chesterfield 15.0     1.04 0.8885 15.0
7      GoldenLights  8.8     0.76 1.0267  9.0
8              Kent 12.4     0.95 0.9225 12.3
9              Kool 16.6     1.12 0.9372 16.3
10              L&M 14.9     1.02 0.8858 15.4
11       LarkLights 13.7     1.01 0.9643 13.0
12         Marlboro 15.1     0.90 0.9316 14.4
13            Merit  7.8     0.57 0.9705 10.0
14      MultiFilter 11.4     0.78 1.1240 10.2
15    NewportLights  9.0     0.74 0.8517  9.5
17          OldGold 17.0     1.26 0.9186 18.5
18    PallMallLight 12.8     1.08 1.0395 12.6
19           Raleigh 15.8     0.96 0.9573 17.5
20       SalemUltra  4.5     0.42 0.9106  4.9
21         Tareyton 14.5     1.01 1.0070 15.9
22             True  7.3     0.61 0.9806  8.5
23 ViceroyRichLight  8.6     0.69 0.9693 10.6
24    VirginiaSlims 15.2     1.02 0.9496 13.9
25    WinstonLights 12.0     0.82 1.1184 14.9
```

2. *Suppose we want to construct a linear regression model to explain the carbon monoxide content in terms of the other variables. Using the new data, draw some plots to explore the suitability of the model. Do the plots suggest that fitting a linear regression model might be appropriate? Give reasons.*

   The pairs plot suggests linear relationships among the variables. In particular, tar and nicotine are strongly related to CO (and to each other). Lets try a coplot:

```
coplot(CO~tar|nicotine*weight, data=newcig.df,
number=4)
```



   The lines are roughly parallel, which suggests the data are roughly planar, and the regression model is probably suitable.

3. *Fit a regression model to the data, using all the variables. Comment on the goodness of fit. Comment on the significance of the variables. Do you think the*

*variable **weight** is required in the model? i.e. does it help explain the amount of carbon monoxide emitted?*

Use the code

```
cig.lm<-lm(CO~tar+nicotine+weight, data=newcig.df)
summary(cig.lm)

Call:
lm(formula = CO ~ tar + nicotine + weight, data =
newcig.df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0376 -0.8192 -0.1837  1.0497  2.1307

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6675     3.6039   0.185 0.855024
tar           0.9080     0.2013   4.511 0.000239 ***
nicotine     -0.1587     3.4801  -0.046 0.964095
weight        1.1947     3.5307   0.338 0.738783
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.'
0.1 ` ' 1

Residual standard error: 1.178 on 19 degrees of freedom
Multiple R-Squared: 0.9112,     Adjusted R-squared:
0.8972
F-statistic: 65.03 on 3 and 19 DF,  p-value: 3.526e-10
```

The R^2 is large, indicating a good fit. The variable weight is not significant (p = 0.738783) so is not required in the regression. From the pairs plot, weight seems unrelated to CO.

4. *Fit a regression using carbon monoxide as the response and nicotine as the only explanatory variable. Contrast the regression coefficient of nicotine and its significance in this model with that of nicotine in the model fitted in Q3. Can you explain the difference?*

Fitting the regression of CO on nicotine is done using the code

```
> nico.lm<-lm(CO~nicotine, data=newcig.df)
> summary(nico.lm)

Call:
lm(formula = CO ~ nicotine, data = newcig.df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1910 -1.1690 -0.0803  1.0770  3.4770
```

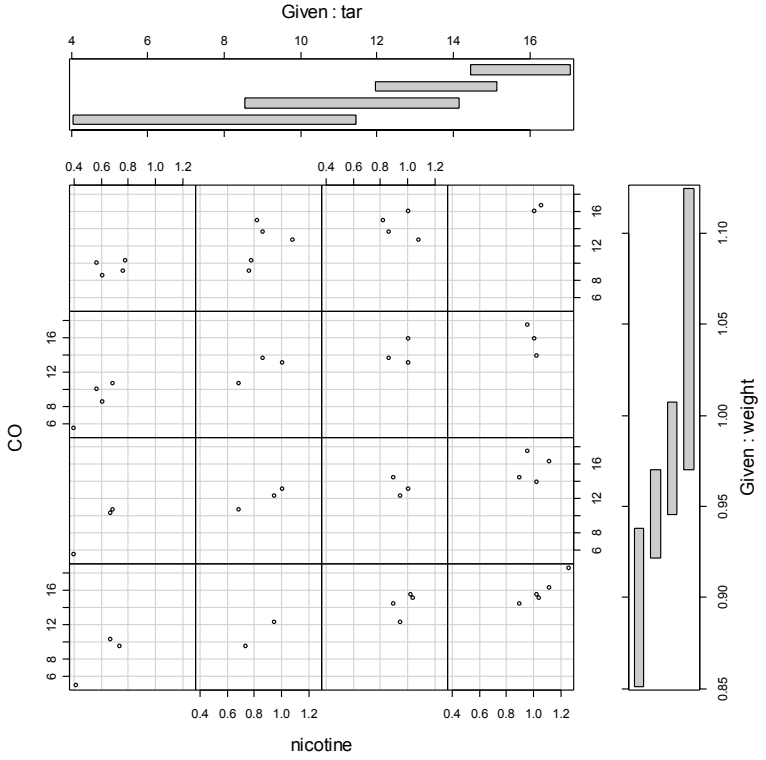
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.121      1.365  -0.089     0.93
nicotine      14.733      1.540   9.567 4.18e-09 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` '
1

Residual standard error: 1.625 on 21 degrees of freedom
Multiple R-Squared: 0.8134,     Adjusted R-squared: 0.8045
F-statistic: 91.53 on 1 and 21 DF,  p-value: 4.179e-09
```

The cofficient of nicotine in the full model is **not** significant (t = -0.046, p = 0.964095). In the model with nicotine alone, nicotine **is** significant (t = 9.567 4.18e-09). Reason: in the full model, we can dispense with nicotine, since tar (which is strongly correlated with nicotine)
is retained in the model. The significance of nicotine in the "nicotine only" model is due to its strong relationship with CO, which is evident in the pairs plot.

Put another way, the coefficient of nicotine in the full model is the slope when CO is plotted against nicotine in a coplot, with weight and tar held fixed. This coefficient is roughtly zero, as there is no trend in the coplot (see below).

On the other hand, the coefficient of nicotine in the "nicotine only" model is the slope when CO is plotted against nicotine in a scatterplot (see the pairs plot above).



Mark Scheme

Q1:

For creating the data frame and showing the pairs plot: 4 marks (2 for the creation, 2 for the plot)
For deleting the correct 2 data points: 4 marks (2 marks if delete only one, unless justified)
For printing out the new data frame: 2 marks

Q2:
3 marks for a sensible coplot, 3 marks for comments (parallel lines) (6 marks in all)
2 marks for comment on the pairs plot
2 marks for other sensible plots, with comments
(Some people might have fitted a model and done some residual analysis. You will have to use your judgement in this case.)

Q3:
Goodness of fit: comment on the R^2. (2 marks)
Significance of the variables: comment on the p-values and t-values (6 marks)
Weight required: no, justify by low t (high p-value) 2 marks.

Q4:
The students should explain the difference between the coefficient in the simple regression (which measures the relationship between CO and nicotine) and the coefficient in the full regression (which measures the extra predictive power of nicotine, given that weight and tar are already in the model). Give 6 marks if they make this clear. Give two marks if they mention the slope in the scatterplot (the coefficient in the simple regression) and the slope in the coplot (the slope in the full regression). Give a further two marks if they emphasize that a variable can have a non-significant t in the full regression, but still have a significant relationship with the response, and say that this happens when the variable is correlated with other explanatory variables.

Thus, 10 marks per question, total 40.