Department of Statistics

COURSE STATS 330

Model answers for Assignment 2, 2006

Question 1 [30 marks]

1. Load the data into R and create a data frame. Make a pairs plot. Correct any obvious typos in the data. *[3 marks]*

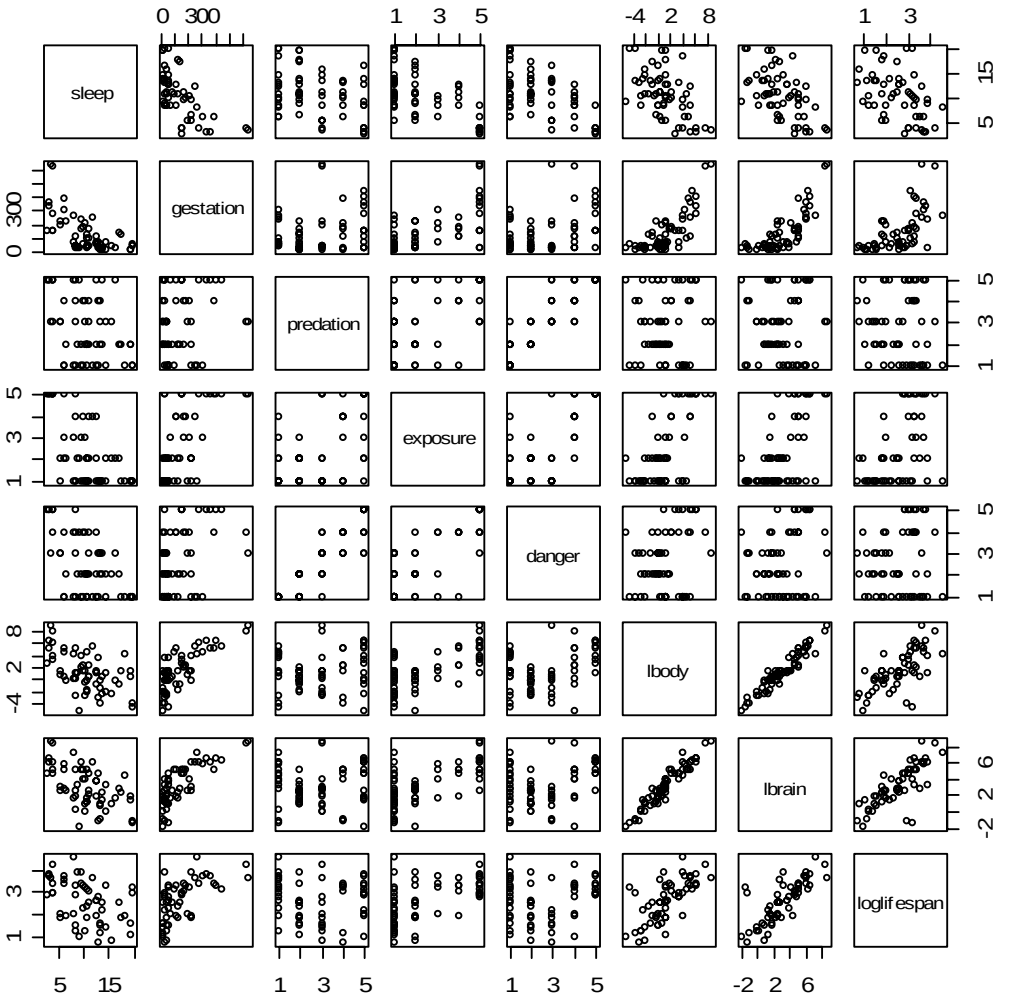
```
mammalsleep.df = read.table(file.choose(), header=T)
                                    # file is mammalsleep.txt
pairs(mammalsleep.df)
```

2. The three variables **body, brain** and **lifespan** have values that differ by orders of magnitude. Do you think the relationships between these variables and the variable **sleep** are linear? If we consider logged versions of these variables, would this difficulty go away? Would another transformation be suitable?

Its difficult to check the relationships because there are a few very large animals (elephants etc) that distort the plots. Let's try logging these variables, and make a new data frame. I also am deleting the nondream and dream variables, as we don't require thee for the rest of the assignment.

```
temp.df = data.frame(mammalsleep.df[,c(5, 7:10)],
lbody=log(mammalsleep.df$body),
lbrain=log(mammalsleep.df$brain),
loglifespan = log(mammalsleep.df$lifespan))

pairs(temp.df)
```



There seems to be some outliers but these seem not to be typos. The data are in fact correct. The log transformation has done a good job of making the relationships linear, although they aren't very strong.

3. *There are also many missing values in this data. When fitting regressions, the* **lm** *function drops out any cases that have missing values in any of the variables involved in the regression. However, this causes difficulties when we come to compare models, as we do in part 4 of this question. What should we do? See tutorial sheet 2 for some hints. Take the appropriate corrective action. [3 marks]*

To compare the results of different regressions, we need to have the same observations in each regression. We can do this by making a new data frame consisting of observations with complete data (i.e. no missing values). We use the technique described in Tutorial 2:

```
temp.mat = !is.na(temp.df)
use = apply(temp.mat, 1 ,all)
my.mammal.df = temp.df[use,]
```

4. *Fit a model to the data, using sleep as the response and taking into account suitable transformations of the variables body, brain and lifespan. Comment on the quality of the fit. [4 marks]*

   The residual plots (not shown) look OK, The R^2 is reasonably good, so the model seems OK. However there are some variables that can probably be deleted.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.408986   1.911654   8.061 3.89e-10 ***
gestation   -0.007282   0.004489  -1.622 0.112116
predation    2.374258   0.971349   2.444 0.018689 *
exposure     0.593245   0.568238   1.044 0.302315
danger      -4.459929   1.110479  -4.016 0.000233 ***
lbody        0.269102   0.453263   0.594 0.555823
lbrain      -1.201376   0.643287  -1.868 0.068649 .
loglifespan  1.142097   0.758355   1.506 0.139374

Residual standard error: 2.702 on 43 degrees of freedom
Multiple R-Squared: 0.7133,     Adjusted R-squared: 0.6666
F-statistic: 15.28 on 7 and 43 DF,  p-value: 7.613e-10
```

5. In their article, Allison and Cicchetti fitted a model with log **body** and **danger** as explanatory variables, using **nondream** as response . Do these variables adequately explain the variation in the variable **sleep**? (in other words, is the model with log **body** and **danger** as explanatory variables an adequate model?)

[3 marks]

```
model2 = lm(sleep ~  lbody + danger, data=my.mammal.df)
summary(model2)

Residuals:
     Min       1Q   Median       3Q      Max
-7.41865 -1.66773  0.03517  1.74371  6.26885

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.6902     0.8771  17.889  < 2e-16 ***
lbody        -0.6993     0.1368  -5.112 5.50e-06 ***
danger       -1.6830     0.3041  -5.534 1.28e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.966 on 48 degrees of freedom
Multiple R-Squared: 0.6142,     Adjusted R-squared: 0.5981
F-statistic: 38.21 on 2 and 48 DF,  p-value: 1.182e-10

> anova(model2,model1)
Analysis of Variance Table

Model 1: sleep ~ lbody + danger
Model 2: sleep ~ gestation + predation + exposure + danger + lbody +
lbrain +
    loglifespan
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     48 422.40
2     43 313.91  5    108.49 2.9724 0.02163 *
```

Both variables are significant, but the R2 has dropped considerably, and the significance test indicates that the smaller model is not adequate (p= **0.02163) . Seems like some more variables are required.**

6. *It would seem that both the size of the animal, and the danger/risk aspect effect sleep. Does adding other variables improve the model? Try the effect of variables log* ***brain*** *and* ***predation*** *only – I don't expect a full-scale model selection exercise. [5 marks]*

Lets add lbrain and predation to the model:

```
model3 = lm(sleep ~  lbrain + lbody + predation + danger,
data=my.mammal.df)
summary(model3)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  17.1019     1.5827  10.806 3.27e-14 ***
lbrain       -0.8303     0.5642  -1.472 0.147922
lbody         0.1018     0.4443   0.229 0.819715
predation     1.9160     0.9435   2.031 0.048097 *
danger       -3.6902     1.0096  -3.655 0.000658 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.803 on 46 degrees of freedom
Multiple R-Squared: 0.6699,     Adjusted R-squared: 0.6411
F-statistic: 23.33 on 4 and 46 DF,  p-value: 1.397e-10
```

The R2 has improved, but now both **lbrain** and **lbody** are insignificant. This is because they are very highly correlated (r = 0.96) , so we don't need both in the model. If we drop **lbody**, we get

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.8233     1.0034  16.766  < 2e-16 ***
lbrain       -0.7087     0.1903  -3.724 0.000525 ***
predation     1.9181     0.9339   2.054 0.045582 *
danger       -3.6826     0.9988  -3.687 0.000587 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.775 on 47 degrees of freedom
Multiple R-Squared: 0.6695,     Adjusted R-squared: 0.6484
F-statistic: 31.73 on 3 and 47 DF,  p-value: 2.307e-11
```

Comparison with the full model gives

```
Analysis of Variance Table

Model 1: sleep ~ lbrain + predation + danger
Model 2: sleep ~ lbody + lbrain + loglifespan + gestation + predation +
    exposure + danger
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     47 361.88
2     43 313.91  4     47.97 1.6428 0.1810
```

which seems OK.

7. *Subject you chosen model to the usual diagnostic checks. Are there any high-leverage observations or outliers? You should take your answer into Question 2 into account when answering this. [5 marks]*

The residual plots obtained from the plot command are OK. The largest Cooks D is 0.156 and the largest studentised residual is 2.62;. neither are cause for concern.

```
> max(abs(rstudent(mammal4.lm)))
[1] 2.620381
> max(cooks.distance(mammal4.lm))
[1] 0.1562440
> qf(0.1, 3, 47)
```

```
[1] 0.1939516
```

However, there are some quite influential points, as we can see from the output from the function **`influence.measures`**. The points 13 and 22 seem to be the worst. Removing them changes the significance of the variable **`predation`**. Refitting without these points and deleting the variable **`predation`** gives

```
model7 = lm(sleep ~ lbrain + danger, subset = (1:51)[-c(13,22)],
data=my.mammal.df)

Call:
lm(formula = sleep ~ lbrain + danger, data = my.mammal.df,
    subset = (1:51)[-c(13, 22)])

Residuals:
    Min      1Q  Median      3Q     Max
-5.6564 -1.2407 -0.2429  1.6216  4.9678

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.3512     0.8850  20.735  < 2e-16 ***
lbrain       -0.9692     0.1561  -6.207 1.42e-07 ***
danger       -1.8313     0.2748  -6.665 2.91e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.675 on 46 degrees of freedom
Multiple R-Squared: 0.6949,     Adjusted R-squared: 0.6817
F-statistic:  52.4 on 2 and 46 DF,  p-value: 1.382e-12
```

Moreover, this model is acceptable when tested against the full model. We will use it for prediction.

8. *The "Okapi" species did not have the variable **`sleep`** recorded in this data set. Use your model to make a prediction of what the value of **`sleep`** is for this animal. [4 marks]*

```
mammalsleep.df["Okapi",]
      body brain nondream dream sleep lifespan gestation predation exposure
danger
Okapi  250   490       NA     1    NA     23.6       440         5        5
5

new.data.df = data.frame(lbrain = log(490),  danger=5)
predict(model7, new.data.df, se.fit =T, interval="prediction")
$fit
          fit       lwr      upr
[1,] 3.191147 -2.431675 8.813969

$se.fit
[1] 0.8043077
```

```
$df
[1] 46

$residual.scale
[1] 2.675102
$df
[1] 46

$residual.scale
[1] 2.696313
```

The sleep is predicted as between **0** and **8.81** hours per day (can't be negative)

Question 2 [10 marks]

For our final model in Q 1, there were $n = 49$ animals in the final fit. The following code generates 10,000 "maximum residuals" from samples of size 49:

```
N=10000
results = numeric(N)
for(j in 1:N){
     results[j] = max(abs(rnorm(49)))
     }
```

We can draw a (relative frequency) histogram with added density curve:

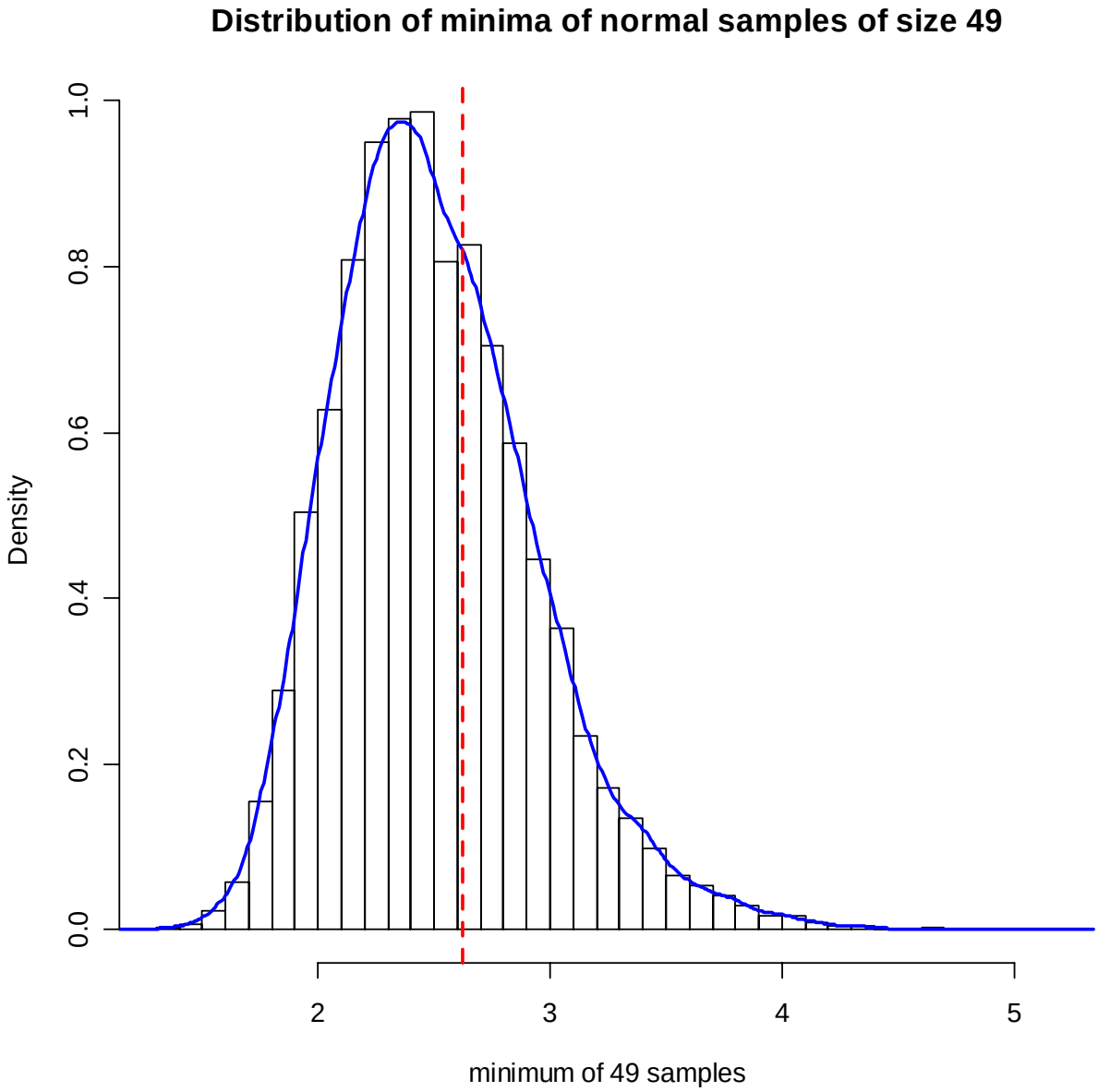
```
hist(results, nclass=30, freq=F, main = "Distribution of
maximum of normal samples of size 49", xlab = "minimum of 49
samples")

curve = density(results)
lines(curve$x, curve$y, lwd=2, col="blue")
```

Add a vertical red line at 2.62, the maximum residual from our regression:

```
library(MASS)
max(abs(studres(model7)))
[1] 2.232982
abline(v=2.232982, col = "red", lty=2, lwd=2)
```

Graph is on next page. Seems pretty typical, what we would expect.


Distribution of minima of normal samples of size 49
Density
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
minimum of 49 samples