Statistical modelling: practical 2 solutions

1   *Simple linear regression*

1. CONSIDER THE data in table 1 for ice cream sales at Luigi Minchella's ice cream parlour.

   (a) Perform a linear regression of y on x. Should temperature be included in the model?

   ```
   x = c(4,4,7,8,12,15,16,17,14,11,7,5)
   y = c(73, 57, 81, 94, 110, 124, 134, 139, 124, 103, 81, 80)
   m = lm(y~x)
   summary(m)

   ##
   ## Call:
   ## lm(formula = y ~ x)
   ##
   ## Residuals:
   ##     Min      1Q  Median      3Q     Max
   ## -10.312  -2.656  -0.016   2.880   7.240
   ##
   ## Coefficients:
   ##               Estimate Std. Error t value Pr(>|t|)
   ## (Intercept)   45.5200     3.5026    13.0 1.38e-07
   ## x              5.4480     0.3186    17.1 9.88e-09
   ##
   ## (Intercept) ***
   ## x           ***
   ## ---
   ## Signif. codes:
   ## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
   ##
   ## Residual standard error: 5.038 on 10 degrees of freedom
   ## Multiple R-squared:  0.9669,Adjusted R-squared:  0.9636
   ## F-statistic: 292.3 on 1 and 10 DF,  p-value: 9.879e-09

   ##The p-value for the gradient is 9.9e-09
   ##This suggests temperatue is useful
   ```

   (b) Calculate the sample correlation coefficient r. Perform a hypothesis test, where $H_0 : \rho = 0$ and $H_1 : \rho \neq 0$.[1] Compare this p-value to the p-value for testing if $\beta_1 = 0$.

   ```
   ##The p-value for the correlation is also 9.9e-09
   cor.test(x, y)

   ##
   ##  Pearson's product-moment correlation
   ```

```
##
## data:  x and y
## t = 17.098, df = 10, p-value = 9.879e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9397526 0.9954575
## sample estimates:
##      cor
## 0.983323
```

(c) Construct a graph of the data. Add a dashed red line indicating the line of best fit.

```
plot(x, y, xlab="Temp", ylab="Sales")
abline(m, col=2, lty=2)
```

(d) Using the text function, add the text *r = 0.983* to your plot.

```
text(5, 130, "r=0.983")
```


Figure 1: Scatterplot with the earnings data. Also shows the line of best fit (question 1c).

(e) Plot the standardised residuals against the fitted values. Does the graph look random?

```
##Model diagnosics look good
plot(fitted.values(m), rstandard(m))
```

(f) Construct a q-q plot of the standardised residuals.

```
qqnorm(rstandard(m))
##Model diagnosics look good
```

2. IN A STUDY of the effect of temperature x on yield y of a chemical process, the data in table 2 was obtained.

(a) Perform a linear regression of y on x.

```
m = lm(y~x)
```

(b) Calculate the sample correlation coefficient r.

```
cor(x, y)

## [1] 0.983323
```

(c) Plot the data and add the line of best fit to your plot.

```
plot(x, y)
abline(m)
```

(d) Plot the standardised residuals against the fitted values.

```
plot(fitted.values(m), rstandard(m), ylim=c(-2.5, 2.5))
abline(h=c(-2, 0, 2), lty=3, col=4)
```
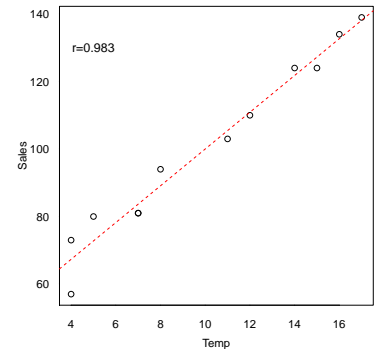
(e) Construct a q-q plot of the Standardised residuals.

```
qqnorm(rstandard(m))
qqline(rstandard(m), col=4)
```

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
x, °C	4	4	7	8	12	15	16	17	14	11	7	5
y, £000's	73	57	81	94	110	124	134	139	124	103	81	80

Table 1: Monthly sales data from Luigi Minchella's ice cream parlour.

x	25	26	27	28	29	30	31	32	33	34	35	36
y	10	16	13	17	20	18	19	23	25	22	29	26

Table 2: Twelve measurements from a study of temperature on yield.

2   Multiple linear regression

1. THE DATA

```
library("nclRmodelling")
data(graves)
```

are from 101 consecutive patients attending a combined thyroid-eye clinic. The patients have an endocrine disorder, Graves' Ophthalmopathy, which affects various aspects of their eyesight. The ophthalmologist measures various aspects of their eyesight and constructs an overall index of how the disease affects their eyesight. This is the Ophthalmic Index (OI) given in the dataset. The age of the patient and their sex are also recorded. In practice, and as this is a chronic condition which can be ameliorated but not cured, the OI would be monitored at successive clinic visits to check on the patient's progress. However, these data are obtained at presentation. We are interested in how OI changes with age and gender. The data can be obtained from

   (a) Fit the multivariate regression model predicting OI from age and gender.

   (b) Examine the Standardised residual plots. Is there anything that would suggest you have a problem with your model? What do you do?

2. DR PHIL comes to see you with his data. He believes that IQ can be predicted by the number of years education. Dr Phil does not differentiate between primary, secondary and tertiary education. He has four variables:

   • IQ - the estimated IQ of the person (the response variable);

   • AgeBegin - the age of the person when they commenced education;

   • AgeEnd - the age of the person when they finished education;

   • TotalYears - the total number of years a person spent in education.

   The data can be obtained from:

```
data(drphil)
```

   Read the data into R and fit the linear regression model:

   $$IQ = \beta_0 + \beta_1 AgeBegin + \beta_2 AgeEnd + \beta_3 TotalYears + \epsilon$$

   Explain what is wrong with this model? Suggest a possible remedy.

```
(m = lm(IQ ~ AgeBegin + AgeEnd + TotalYears, data=drphil))

##
## Call:
## lm(formula = IQ ~ AgeBegin + AgeEnd + TotalYears, data = drphil)
##
## Coefficients:
## (Intercept)      AgeBegin         AgeEnd
##   101.86617      -0.20122        0.07377
##  TotalYears
##          NA

#The problem is TotalYears = AgeEnd - AgeBegin
#Solution: remove TotalYears
```

3   One way ANOVA tables

1. A PILOT STUDY was developed to investigate whether music influenced exam scores. Three groups of students listened to 10 minutes of Mozart, silence or heavy metal before an IQ test. The results of the IQ test are as follows

Mozart	109	114	108	123	115	108	114
Silence	113	114	113	108	119	112	110
Heavy Metal	103	94	114	107	107	113	107

Table 3: Results from the study on how music affects examination performance.

(a) Construct a one-way ANOVA table. Are there differences between treatment groups?

```
x1 = c(109, 114, 108, 123, 115, 108, 114)
x2 = c(113, 114, 113, 108, 119, 112, 110)
x3 = c(103, 94, 114, 107, 107, 113, 107)
dd = data.frame(values = c(x1, x2, x3), type = rep(c("M", "S", "H"), each=7))
m = aov(values ~ type, dd)
summary(m)

##              Df Sum Sq Mean Sq F value Pr(>F)
## type          2  193.1   96.57   3.401 0.0559 .
## Residuals    18  511.1   28.40
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##The p value is around 0.056.
##This suggests a difference may exist.
```

(b) Check the standardised residuals of your model.

```
plot(fitted.values(m), rstandard(m))
## Residual plot looks OK
```

(c) Perform a multiple comparison test to determine where the difference lies.

```
TukeyHSD(m)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = values ~ type, data = dd)
##
## $type
##            diff        lwr       upr     p adj
## M-H   6.5714286 -0.6981512 13.841008 0.0804419
## S-H   6.2857143 -0.9838655 13.555294 0.0970627
## S-M -0.2857143 -7.5552941  6.983865 0.9944700
```
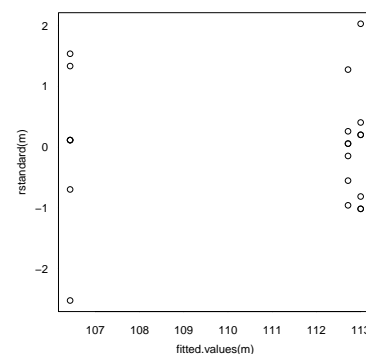

Figure 2: Model diagnosics for the music data.

THE FOLLOWING SECTIONS use the results of the Olympic heptathlon competition, Seoul, 1988. To enter the data into R, use the following commands

```
data(hep)
##Remove the athletes names and final scores.
hep_s = hep[,2:8]
```

4   Hierarchical clustering

USING THE HEPTATHLON data set, carry out a clustering analysis. Try different distance methods and clustering functions.

```
plot(hclust(dist(hep_s)), labels=hep[,1])
```

5   Principal components analysis

1. Calculate the correlation matrix of the hep data set.

```
##Round to 2dp
signif(cor(hep_s), 2)

##          hurdles highjump  shot run200m longjump
## hurdles   1.0000  -0.8100 -0.65    0.77   -0.910
## highjump -0.8100   1.0000  0.44   -0.49    0.780
## shot     -0.6500   0.4400  1.00   -0.68    0.740
## run200m   0.7700  -0.4900 -0.68    1.00   -0.820
## longjump -0.9100   0.7800  0.74   -0.82    1.000
```

```
## javelin  -0.0078   0.0022  0.27   -0.33     0.067
## run800m   0.7800  -0.5900 -0.42    0.62    -0.700
##           javelin run800m
## hurdles  -0.0078    0.78
## highjump  0.0022   -0.59
## shot      0.2700   -0.42
## run200m  -0.3300    0.62
## longjump  0.0670   -0.70
## javelin   1.0000    0.02
## run800m   0.0200    1.00
```

2. Carry out a PCA on this data set.[2]

   • Keep `score` in your PCA analysis. What happens and why? Do you think you should remove `score`?

   ```
   ##Remove:
   ##1st column: athletes name
   ##Last column: It's a combination of the other columns
   dd = hep[ ,2:8]

   ##Run principle components
   prcomp(dd)

   ## Standard deviations:
   ## [1] 8.3646430 3.5909752 1.3856976 0.5857131
   ## [5] 0.3238168 0.1471221 0.0332496
   ##
   ## Rotation:
   ##                    PC1            PC2         PC3
   ## hurdles    0.069508692 -0.0094891417  0.22180829
   ## highjump  -0.005569781  0.0005647147 -0.01451405
   ## shot      -0.077906090  0.1359282330 -0.88374045
   ## run200m    0.072967545 -0.1012004268  0.31005700
   ## longjump  -0.040369299  0.0148845034 -0.18494319
   ## javelin    0.006685584  0.9852954510  0.16021268
   ## run800m    0.990994208  0.0127652701 -0.11655815
   ##                    PC4            PC5         PC6
   ## hurdles   -0.32737674 -0.80702932  0.424850883
   ## highjump   0.02123856  0.14013823  0.098373568
   ## shot      -0.42500654 -0.10442207 -0.051744802
   ## run200m   -0.81585220  0.46178680  0.082486244
   ## longjump   0.20419828  0.31899315  0.894592570
   ## javelin   -0.03216907  0.04880388  0.006170438
   ## run800m    0.05827720  0.02784756 -0.002987043
   ##                    PC7
   ## hurdles   -0.083123145
   ## highjump  -0.984881131
   ## shot      -0.015649644
   ## run200m    0.051312974
   ```

```
## longjump  0.142110352
## javelin   0.005033005
## run800m   0.001041451
```

- Do you think you need to scale the data?

  ##Yes!. run800m dominates the loading since
  ##the scales differ

- Construct a biplot of the data.

  ```
  prcomp(dd, scale=TRUE)
  ```

  ```
  ## Standard deviations:
  ## [1] 2.1119364 1.0928497 0.7218131 0.6761411
  ## [5] 0.4952441 0.2701029 0.2213617
  ##
  ## Rotation:
  ##                    PC1         PC2         PC3
  ## hurdles    0.4528710 -0.15792058 -0.04514996
  ## highjump  -0.3771992  0.24807386  0.36777902
  ## shot      -0.3630725 -0.28940743 -0.67618919
  ## run200m    0.4078950  0.26038545  0.08359211
  ## longjump  -0.4562318  0.05587394 -0.13931653
  ## javelin   -0.0754090 -0.84169212  0.47156016
  ## run800m    0.3749594 -0.22448984 -0.39585671
  ##                    PC4         PC5         PC6
  ## hurdles    0.02653873 -0.09494792 -0.78334101
  ## highjump  -0.67999172 -0.01879888 -0.09939981
  ## shot      -0.12431725 -0.51165201  0.05085983
  ## run200m   -0.36106580 -0.64983404  0.02495639
  ## longjump  -0.11129249  0.18429810 -0.59020972
  ## javelin   -0.12079924 -0.13510669  0.02724076
  ## run800m   -0.60341130  0.50432116  0.15555520
  ##                    PC7
  ## hurdles   -0.38024707
  ## highjump  -0.43393114
  ## shot      -0.21762491
  ## run200m    0.45338483
  ## longjump   0.61206388
  ## javelin    0.17294667
  ## run800m    0.09830963
  ```

  ```
  biplot(prcomp(dd, scale=TRUE))
  ```

6   Survival analysis

This final section, is meant to give you a taste at other statistical techniques available. R has many packages[3] available. To install a package, we use the command

[3] A package is just an "add-on" that provides new functionality.

```
install.packages("survival")
```

Once the package is installed, we load it using `library` function

```
library(survival)
```

The data set we will use is the `lung` data set[4] - this comes with the `survival` package. To load this data set, we use the command

```
data(lung)
```

We the use the `dim` function to extract the number of rows and columns

```
dim(lung)
```

```
## [1] 228  10
```

This creates a data frame with the following columns

- `inst`: Institution code

- `time`: Survival time in days

- `status`: censoring status 1=censored, 2=dead

- `age`: Age in years

- `sex`: Male=1 Female=2

- `ph.ecog`: ECOG performance score (0=good 5=dead)

- `ph.karno`: Karnofsky performance score (bad=0-good=100) rated by physician

- `pat.karno`: Karnofsky performance score as rated by patient

- `meal.cal`: Calories consumed at meals

- `wt.loss`: Weight loss in last six months

To begin, we create a Surv object:[5]

```
Surv(lung$time, lung$status)
```

We can fit a Kaplein-Meier curve using the `surfit` function:

```
survfit(Surv(lung$time, lung$status)~1)
```

and also plot survival curve

```
plot(survfit(Surv(lung$time, lung$status)~1))
```

To fit the model using an additional covariate, we just alter the formula

```
plot(survfit(Surv(lung$time, lung$status)~lung$sex))
```

Task: Load the heart data set:[6]

data(heart)

and make a `Surv` object

```
Surv(heart$start, heart$stop, heart$event)
```

Why are there three arguments in the above function? Construct a Kaplein-Meier plot. Look at the `coxph` function. Try fitting a cox proportional hazard function.

Solutions

Solutions are contained within this package:

```
library("nclRmodelling")
vignette("solutions2", package="nclRmodelling")
```

[6] Look at the help page: `?heart`