Consumer evaluations of carrots

Data description

In a consumer study 103 consumers scored their preference, degree of bitterness, degree of sweetness and degree of crispiness for 12 danish carrot types using a scoring scale from 1 to 7. The carrots were harvested in autumn 1996 and tested in march 1997. The data set "carrots.txt" includes 1236 observations (rows) and 14 variables (columns):

Variable	Description
Consumer	Numbering identifying the consumers
Frequency	Valued 1-5 (see below)
Gender	Valued 1-2 (see below)
Age	Valued 1-4(see below)
Homesize	Valued 1 and 3 (see below)
Work	Valued 1-7 (see below)
Income	Valued 1-4 (see below)
Preference	preference score
Sweetness	Sweetness score
Bitter	Bitterness score
Crisp	Crispiness score
product	Product identification
cultivar	The cultivar of the carrot product
region	The growing region of the carrot product

In `R` it looks like:

```
carrots<- read.table("C:/Users/pbb/Documents/per/02418/carrots.txt",header=TRUE,sep=",")
head(carrots)
```

```
  Consumer Frequency Gender Age Homesize Work Income Preference Sweetness BITTER Crisp  product cultivar      region
1      168         1      2   4        3    7      2          4         2      4     2 Bolero_E   Bolero Ejstrupholm
2      168         1      2   4        3    7      2          5         5      2     6 Bolero_L   Bolero  Lammefjord
3      168         1      2   4        3    7      2          4         2      2     3  Major_E    Major Ejstrupholm
4      168         1      2   4        3    7      2          7         6      2     6  Major_L    Major  Lammefjord
5      168         1      2   4        3    7      2          5         4      2     4  Navar_E    Navar Ejstrupholm
6      168         1      2   4        3    7      2          6         6      2     5  Navar_L    Navar  Lammefjord
```

Note that 103 consumers times 12 products constitute the 1236 observations, so these two variables (factors) define the overall setup. The remaning 10 variables/columns can be categorized in three different types:

- 4 Response variables(Ys, scorings): Preference, Sweetness, BITTER, Crisp
- 6 Consumer background variables(factors): Frequency, Gender, Age, Homesize, Work, Income

- 2 Product characteristic variables: cultivar, region

The explanation of the consumer background information variables are:

Frequency: "How often do you eat carrots?"

1. Once a week or more
2. Once every 2 weeks
3. once every 3 weeks
4. At least once a month
5. Less than once a month

Gender:

1. male
2. female

Age:

1. -25 y
2. 26-40 y
3. 41-60
4. 61-

Homesize: (number of persons in the household)

1. 1 or 2 persons

3 3 or more persons

Work: (7 different types of employment)

1. Unskilled worker (no education)
2. Skilled worker (with education)
3. Office worker
4. Housewife (or man)
5. independent businessman/self-employed
6. Student
7. Retired

Income: (of the household)

1. <150.000 DKK
2. 150.000-300.000 DKK
3. 300.000-500.000 DKK
4. >500.000 DKK

Tasks/questions

Let Y be (any) one of the three response variables (do NOT use BITTER, which is used as example below).

1. Focus only the two main factors, Consumer and product
   (a) Are there any consumer and/or product differences wrt. average Y?
   (b) And if so, what differences are there?? (Do a "Post hoc" analysis for the product differences)
   (c) Challenges: How to do post hoc analysis in R. Think about correcting for "multiplicity"/"multiple testing". (E.g. by "Bonferroni" approach, see Miller&Freund's Probability and Statistics for Engineers (Textbook from Course 02402 on introduction to statistics), page 378 or by the "Tukey" approach:

      `http://en.wikipedia.org/wiki/Tukey's_range_test`

2. Focus on product characteristic variables
   (a) Decompose the product effect into the main effects and interaction effects between "cultivar" and "region". What can be concluded?
   (b) How would you summarize the results?
3. Focus on Consumer background variables
   (a) For background consumer demographic factors A and B: (chosen among the 6 possible ones) Decompose/explain the consumer effect by the main effects and interaction effects between A and B and the remaining Consumer effect. What can be concluded?
   (b) How would you summarize the results?
   (c) Do this analyses in two ways:
      i. On the 103 average values (averaging over products for each consumer)

ii. On the full data set - extending the model from step 1

iii. Compare the results (Try e.g. to multiply the SS's from the ANOVA table from the average data analysis by 12!) Which P-values do you believe more?

(d) Looking at either Income or Age together with another consumer background factor, e.g. Gender, is there a (linear) trend in the consumers responses related to any of these, and do this trend depend on the age group?

General advice on practical data analysis approach

Steps in a data analysis process:

1. Explorative data analysis, plotting of raw data
2. Modelling/hypothesis testing/identifying significant and important effects
3. Model diagnostics (Residual investigation, outliers, influential observations?)
4. Re-do Step 2, if step 3 calls for transformation of y-values, removal of observations etc
5. Post-hoc analysis, summary of results, estimates, confidence bands, plotting of important effects/messages in data.

Hints and R-starters

First make sure that (the relevant) factors are really factors, and let's have a summary look at the data:

```
carrots$Consumer=factor(carrots$Consumer)
carrots$product=factor(carrots$product)
summary(carrots)
```

A short analysis of BITTER corresponding to question 1, and following the 5 steps

```
# 1. Initial plotting:
with(carrots,plot(product,BITTER))
with(carrots,boxplot(BITTER~product))

with(carrots,plot(Consumer,BITTER))
```

```
with(carrots,plot(Consumer,BITTER))

with(carrots,interaction.plot(Consumer,product,BITTER))

with(carrots,interaction.plot(product,Consumer,BITTER))

# 2. Modelling:

model1=lm(BITTER~Consumer+product,data=carrots)
str(model1)
anova(model1)

# 3. Diagnostics:

par(mfrow=c(2,2))

qqnorm(residuals(model1))
plot(predict(model1),residuals(model1))
with(carrots,plot(Consumer,residuals(model1)))
with(carrots,plot(product,residuals(model1)))

plot(model1,1:4)

# Step 4: Re-doing step 2. Modelling AND step 3 diagnostics:

model1b=lm(log(BITTER)~Consumer+product,data=carrots)
str(model1b)
anova(model1b)
qqnorm(residuals(model1))
plot(predict(model1b),residuals(model1b))
with(carrots,plot(Consumer,residuals(model1b)))
with(carrots,plot(product,residuals(model1b)))

plot(model1b,1:4)

# Step 5, Summary/post hoc.
summary(model1b)

#To automatically produce ALL comparisons
```

```
library(multcomp)
summary(glht(model1b, linfct = mcp(product= "Tukey")))
plot(glht(model1b, linfct = mcp(product= "Tukey")))

#plotting of expected structure:
par(mfrow=c(1,1))
with(carrots,interaction.plot(Consumer,product,predict(model1b),col=1:12))
```

Analysing BITTER according to question 2:

```
# Explorative analysis:
with(carrots,interaction.plot(region,cultivar,BITTER))
with(carrots,interaction.plot(cultivar,region,BITTER))

model2b=lm(log(BITTER)~Consumer+region+cultivar+region:cultivar,data=carrots)
anova(model2b)

# Summary:
with(carrots,interaction.plot(cultivar,region,predict(model2b)))
with(carrots,interaction.plot(region,cultivar,predict(model2b),col=1:6))

lsd_0.95=qt(0.975,1223)*sqrt(0.27972)*sqrt(1/103 + 1/103)
lsd_0.95

with(carrots,interaction.plot(region,cultivar,predict(model2b),lwd=2,
col=1:6,main="LSD_0.95=0.145"))

lsd_0.95=qt(0.975,1223)*sqrt(0.27972)*sqrt(1/103 + 1/103)
lsd_0.95

with(carrots,interaction.plot(region,cultivar,predict(model2d),col=1:6,
main="LSD_0.95=0.145"),lwd=2)

# Maybe as before, since no simplification could be reached

# To construct the data set of the 103 average values:
library(doBy)

consmeans=summaryBy(Preference+Sweetness+logBITTER+Crisp~Consumer,
id=~Homesize+Frequency+Gender+Age+Work+Income,
data=carrots,FUN=mean,keep.names=TRUE)
```

```
head(consmeans)

carrots$Homesize=factor(carrots$Homesize)
carrots$Gender=factor(carrots$Gender)

with(carrots,table(Homesize,Gender))

#Full data set Analysis
model3a=lm(logBITTER~product+Homesize+Gender+factor(Homesize:Gender)+Consumer,
data=carrots)
anova(model3a)

#Average data set Analysis
model3b=lm(logBITTER~Homesize*Gender,data=consmeans)
anova(model3b)
12*anova(model3b)[,2:3]
```

The trend investigation is done by:

```
consmeans$age_numeric=as.numeric(consmeans$Age)

model3c=lm(logBITTER~age_numeric*Gender,data=consmeans)
anova(model3c)

model3d=lm(logBITTER~age_numeric+Gender,data=consmeans)
anova(model3d)
summary(model3d)

model3d=lm(logBITTER~Gender+age_numeric,data=consmeans)
anova(model3d)
summary(model3d)

drop1(model3d,test="F")
```