Consumer evaluations of carrots

02418 Statistical modelling– Anders Hørsted (s082382)

24-10-2011

In this case a data set with consumer evaluations of carrots are given. Each consumer (103 in total) have evaluated carrots of 6 different cultivars from 2 different regions. Each combination of cultivar and region is called a product. Based on the data set we will examine the relationship between various traits of the consumers (eg. age, gender etc.) and how they score the carrots. Specifically we will focus on the sweetness score.

Question 1

In the first question we focus on the two factors consumer and product and examine whether there are any differences in sweetness scores between different consumers and/or products. To get an overview of the data a histogram of the sweetness scores is plotted. Also boxplots of the sweetness scores grouped by product and consumer are plotted. All three plot are seen in figure 1. It is seen that there are large differences between the median scores of the different consumers. This is expected since the scoring of the sweetness is a subjective measure and the perception of different consumers is bound to differ. The differences between the specific consumers isn't very interesting but as a confounding variable it is important that we control for the differences. From the plots there also seems to be differences between the median sweetness of the various products. This is of cause interesting and we will therefore perform an ANOVA to determine whether the differences are significant or not.

Performing an ANOVA

To perform the ANOVA a model is defined as `Sweetness~Consumer+product`. The `anova` function is called on the model and gives the output in table 1. It is seen that the means of some consumers are not equal as well as some of the means of products are nor equal.

Before starting on the post-hoc analysis the model assumptions are checked by plotting a QQ-plot and a residual plot (see figure 2). The QQ-plot seems to confirm that the errors are normal distributed. The residual plot is a bit difficult to interpret due to the fact that Sweetness is actually a discrete variable. The residuals are independent of the fitted values (see [1, p.85]) so no pattern should be recognizable in a fitted values vs. residuals plot. There is obivous some pattern in the residual plot in figure 2 and you

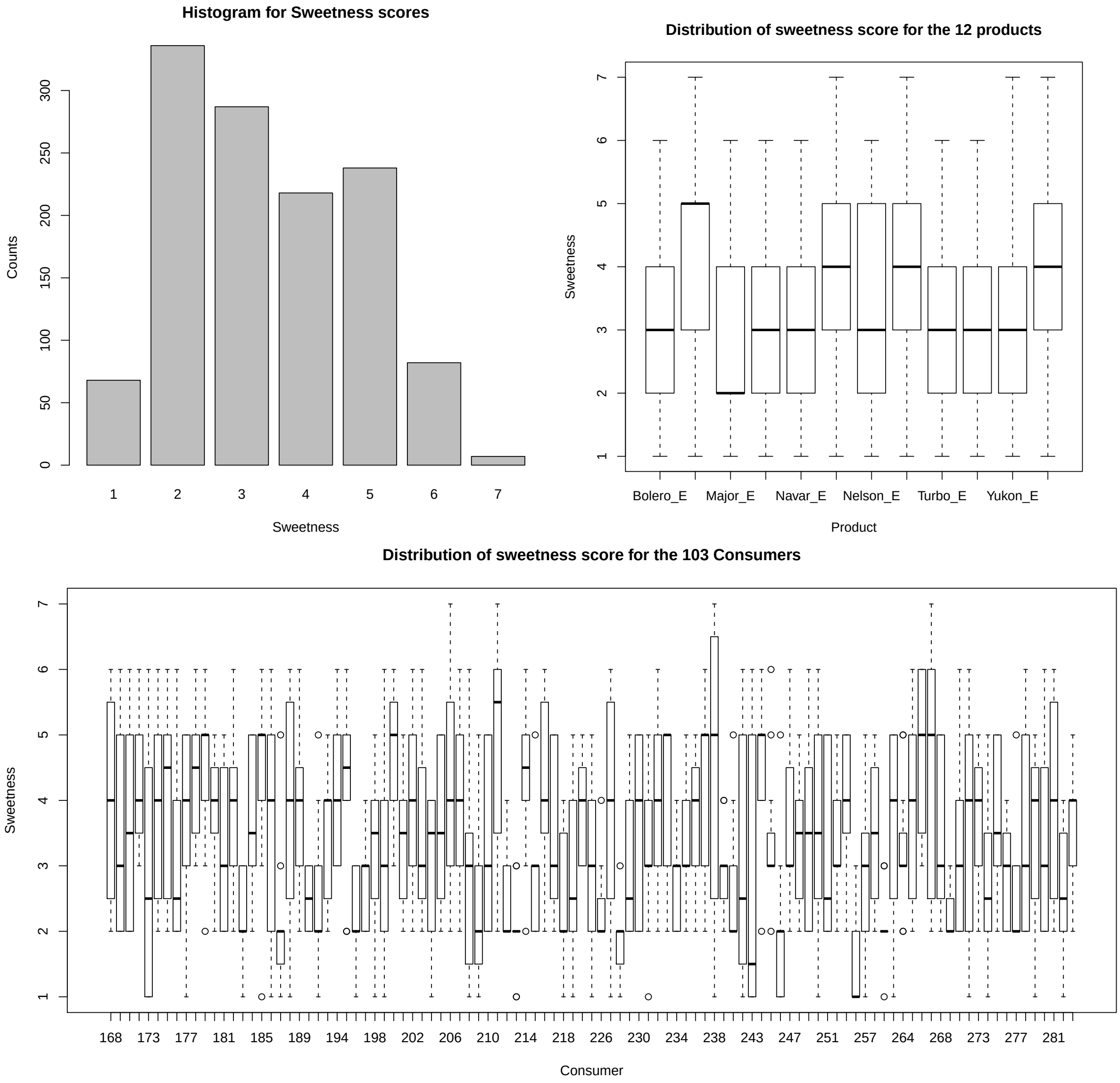

Figure 1: Histogram for sweetness score and boxplots for the product and consumer factor

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Consumer	102	632.29	6.20	4.60	0.0000
product	11	284.57	25.87	19.20	0.0000
Residuals	1122	1512.10	1.35		

Table 1: ANOVA table for the two way ANOVA on the Consumer and product factors

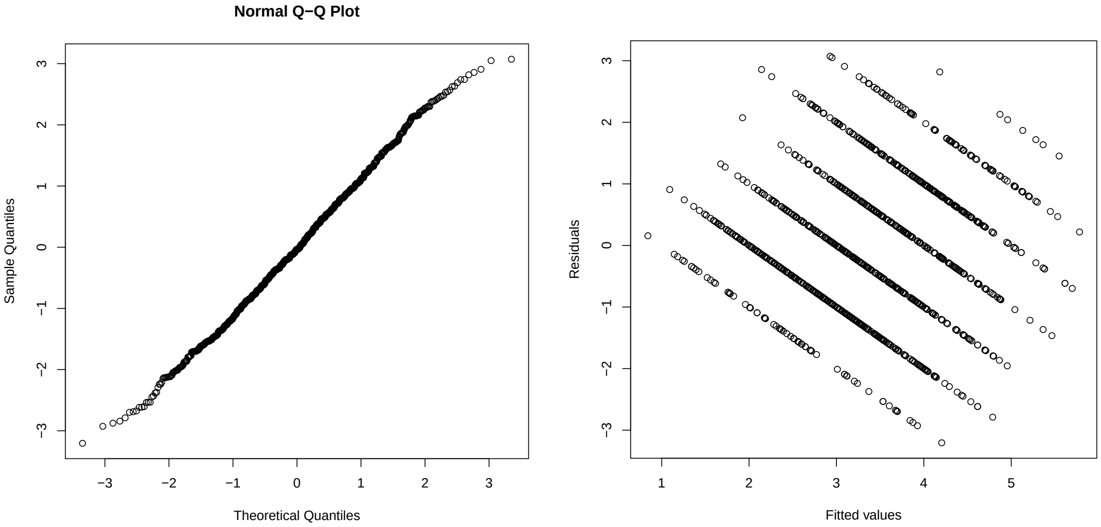

Figure 2: Diagnostic plots for the two way ANOVA on the Consumer and product factors

could argue that variance is smaller for fitted values near 1 and 7, and larger for fitted values near 3.5. Using the R function `boxcox` a transformation by $(y^{0.5} - 1)/0.5$ was attempted on the Sweetness response, but the resulting QQ-plot showed greater deviation from normality and no improvement in the residuals plot was found (see appendix B.1). The result is that the original model is accepted. The next step is a post-hoc analysis.

Post-hoc analysis

Since we are not interested in determining which consumers have significantly different means we only look at the differences between all pairs of products. With 12 products we need to perform $\frac{12 \cdot 11}{2} = 66$ t-tests, but with the standard 5% chance of an type I we would expect 3-4 significant differences by chance alone. To correct for the multiple comparison a Tukey Range Test is performed on all 66 pairs of products and it is found that a total of 31 pairs have significantly different means. The results are shown in appendix B.6. From the results it seems as if the Bolero cultivar is sweeter than most other cultivars except for maybe Yukon. Also the region Lammefjord seem to consistently score higher on sweetness than the other region Ejstrupholm. It isn't 100% clear though so a natural next step is to make an ANOVA with region and cultivar as separate factors.

Question 2

We now look at the region and cultivar as separate factors. To get an overview of these two factors and to look for possible interaction between the two factors an interaction plot for the factors are now plotted in figure 3. The interaction plot supports that the region Lammefjord makes sweeter carrots than Ejstrupholm. There also seems to be

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Consumer	102	632.29	6.20	4.60	0.0000
cultivar	5	88.75	17.75	13.17	0.0000
region	1	166.76	166.76	123.74	0.0000
cultivar:region	5	29.06	5.81	4.31	0.0007
Residuals	1122	1512.10	1.35		

Table 2: ANOVA table for the model in question 2

interactions between the region and cultivar since the lines in the interaction plot isn't parallel. Eg. the increase in sweetness between Ejstrupholm and Lammefjord is larger for the bolero and yukon cultivar than for the other cultivars. It will therefore make sense to try and fit a model with interactions.

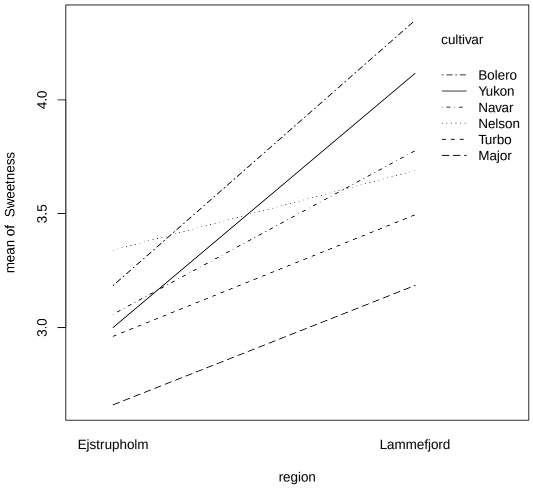

Figure 3: Interactions between cultivar and region

By entering the model `Sweetness~Consumer+cultivar*region` and running the `anova` function we get the output shown in table 2. From the results there seems to be significant differences in the means for both the cultivar and region factor as well as the interactions. Since there are only two regions it can be concluded by looking at the coefficients of the model, that the region Lammefjord grows significantly sweeter carrots than Ejstrupholm. To find which cultivar that differs a Tukey's range test is performed on the cultivar factor. From the results (see appendix B.7) we find that the only significant results are that Bolero is sweeter than Major and that Nelson is sweeter than Major. This result isn't exactly what was expected based on the results in question 1. In question 1 it was anticipated that Bolero would be sweeter than all other cultivars except for Yukon. It turns out only the difference between Bolero and Major is significant.

The main conclusion of these two first ANOVAs must be that the Lammefjord region grows sweeter carrots than Ejstrupholm.

1 Question 3

We now look at the two consumer factors age and gender and try to see if there are any differences within these two factors for how they score the sweetness of carrots.

Due to bad time management this is unfortunately how far I get before hand-in.

B Appendices

All R source code is included in the appendices. All the source code including the Latex code used for the report can also be found at `https://github.com/alphabits/dtu-fall-2011/tree/master/02418/carrots`.

B.1 Diagnostic plots for model with transformed response

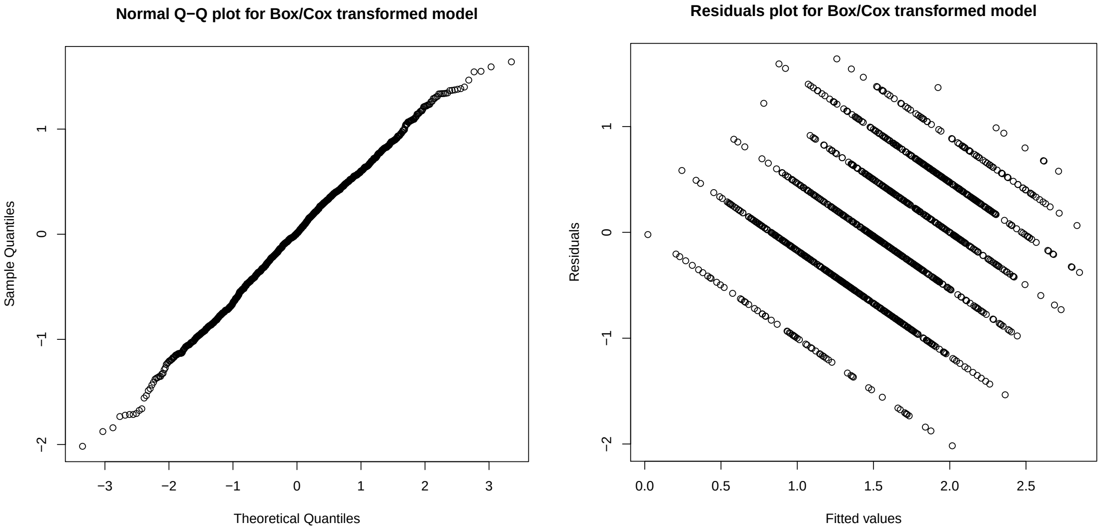

Figure 4: QQ plot and residuals for the Box/Cox transformed model

B.2 functions.R

```
plot.and.save = function(filename, width, height, plotfunction, ...) {
    save.the.plot = exists('SAVEPLOTS') && SAVEPLOTS
    if (save.the.plot) {pdf(sprintf('../plots/%s', filename), width, height)}
    plotfunction(...)
    if (save.the.plot) {dev.off()}
}
```

B.3 loaddata.R

```
carrots = read.table("../data/carrots.txt", header=TRUE, sep=",")
carrots$Consumer = factor(carrots$Consumer)
carrots$Age = factor(carrots$Age)
carrots$Gender = factor(carrots$Gender)
```

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

B.4 question1.R

```
# Include all dependencies
library(multcomp)
library(xtable)
source(’functions.R’)

# Controls whether the plots are saved or displayed
SAVEPLOTS = TRUE

# Load data in carrots variable
source(’loaddata.R’)

attach(carrots)

plot.and.save(’sweetness-histogram.pdf’, 7, 7,
              plot, factor(Sweetness), main=’Histogram for Sweetness scores’,
              xlab=’Sweetness’, ylab=’Counts’)

titletmpl = ’Distribution of sweetness score for the %s’
plot.and.save(’product-boxplot.pdf’, 7, 7,
              plot, product, Sweetness,
              main=sprintf(titletmpl, ’12 products’),
              xlab=’Product’, ylab=’Sweetness’)
plot.and.save(’consumer-boxplot.pdf’, 14, 7,
              plot, Consumer, Sweetness,
              main=sprintf(titletmpl, ’103 Consumers’),
              xlab=’Consumer’, ylab=’Sweetness’)

# Define model and calculate anova
m1 = lm(Sweetness ~ Consumer + product)
m1.anova = anova(m1)
m1.res = residuals(m1)

# Residual plot is difficult to interpret.
# Try a few transformations
models = list(
    list(y=Sweetness^0.5, key="sqrt", label="Square root"),
    list(y=log(Sweetness), key="log", label="Log"),
    list(y=(Sweetness^0.5-1)/0.5, key="boxcox", label="Box/Cox")
)
for (model in models) {
    y.tmp = model[[’y’]]
    key.tmp = model[[’key’]]
    model.tmp = lm(y.tmp~Consumer+product)
    filename.tmpl = ’%s-model-%s.pdf’
    title.tmpl = sprintf(’%%s for %s transformed model’,
                         model[[’label’]])
    plot.and.save(sprintf(filename.tmpl, ’qqplot’, key.tmp), 7, 7,
                  main=sprintf(title.tmpl, ’Normal Q-Q plot’),
                  qqnorm, residuals(model.tmp))
    plot.and.save(sprintf(filename.tmpl, ’residuals’, key.tmp), 7, 7,
                  plot, predict(model.tmp), residuals(model.tmp),
                  main=sprintf(title.tmpl, ’Residuals plot’),
                  xlab="Fitted values", ylab="Residuals")
}

# Save anova as latex table
sink('../tables/q1-model1-xtable.tex')
print(xtable(m1.anova))
sink()

# Plot diagnostic plots
plot.and.save('qqplot-model1.pdf', 7, 7,
              qqnorm, m1.res)
plot.and.save('residuals-model1.pdf', 7, 7,
              plot, predict(m1), m1.res,
              xlab='Fitted values', ylab='Residuals')

# Do post hoc analysis
hypothesis = glht(m1, linfct=mcp(product="Tukey"))
sink('../tables/q1-post-hoc.tex')
summary(hypothesis)
sink()

detach('carrots')
```

B.5 question2.R

```
# Include all dependencies
library(xtable)
library(multcomp)
source('functions.R')

# Controls whether the plots are saved or displayed
SAVEPLOTS = TRUE

# Load data in carrots variable
source('loaddata.R')

attach(carrots)

plot.and.save('interaction-region-cultivar.pdf', 7, 7,
              interaction.plot, region, cultivar, Sweetness)
plot.and.save('interaction-cultivar-region.pdf', 7, 7,
              interaction.plot, cultivar, region, Sweetness)


# Define model and calculate anova
m2 = lm(Sweetness ~ Consumer + cultivar*region)
m2.anova = anova(m2)
m2.res = residuals(m2)

# Save anova as latex table
sink('../tables/q2-model2-xtable.tex')
print(xtable(m2.anova))
sink()

# Plot diagnostic plots
plot.and.save('qqplot-model2.pdf', 7, 7,
              qqnorm, m2.res)
plot.and.save('residuals-model2.pdf', 7, 7,
              plot, predict(m2), m2.res,
              xlab='Fitted values', ylab='Residuals')

# Do post hoc analysis
hypothesis2 = glht(m2, linfct=mcp(cultivar="Tukey"))
sink('../tables/q2-post-hoc.tex')
summary(hypothesis2)
sink()

detach("carrots")
```

B.6 Results of Tukey's range test for question 1

Only comparisons with a p-value less than 10% are included.

```
Bolero_L - Bolero_E == 0  1.165e+00  1.618e-01     7.202    <0.01 ***
Major_E - Bolero_E == 0  -5.243e-01  1.618e-01    -3.241   0.0556 .
Navar_L - Bolero_E == 0   5.922e-01  1.618e-01     3.661   0.0137 *
Nelson_L - Bolero_E == 0  5.049e-01  1.618e-01     3.121   0.0788 .
Yukon_L - Bolero_E == 0   9.320e-01  1.618e-01     5.762    <0.01 ***
Major_E - Bolero_L == 0  -1.689e+00  1.618e-01   -10.443    <0.01 ***
Major_L - Bolero_L == 0  -1.165e+00  1.618e-01    -7.202    <0.01 ***
Navar_E - Bolero_L == 0  -1.291e+00  1.618e-01    -7.982    <0.01 ***
Navar_L - Bolero_L == 0  -5.728e-01  1.618e-01    -3.541   0.0206 *
Nelson_E - Bolero_L == 0 -1.010e+00  1.618e-01    -6.242    <0.01 ***
Nelson_L - Bolero_L == 0 -6.602e-01  1.618e-01    -4.081    <0.01 **
Turbo_E - Bolero_L == 0  -1.388e+00  1.618e-01    -8.582    <0.01 ***
Turbo_L - Bolero_L == 0  -8.544e-01  1.618e-01    -5.281    <0.01 ***
Yukon_E - Bolero_L == 0  -1.350e+00  1.618e-01    -8.342    <0.01 ***
Major_L - Major_E == 0    5.243e-01  1.618e-01     3.241   0.0552 .
Navar_L - Major_E == 0    1.117e+00  1.618e-01     6.902    <0.01 ***
Nelson_E - Major_E == 0   6.796e-01  1.618e-01     4.201    <0.01 **
Nelson_L - Major_E == 0   1.029e+00  1.618e-01     6.362    <0.01 ***
Turbo_L - Major_E == 0    8.350e-01  1.618e-01     5.161    <0.01 ***
Yukon_L - Major_E == 0    1.456e+00  1.618e-01     9.003    <0.01 ***
Navar_L - Major_L == 0    5.922e-01  1.618e-01     3.661   0.0138 *
Nelson_L - Major_L == 0   5.049e-01  1.618e-01     3.121   0.0785 .
Yukon_L - Major_L == 0    9.320e-01  1.618e-01     5.762    <0.01 ***
Navar_L - Navar_E == 0    7.184e-01  1.618e-01     4.441    <0.01 ***
Nelson_L - Navar_E == 0   6.311e-01  1.618e-01     3.901    <0.01 **
Yukon_L - Navar_E == 0    1.058e+00  1.618e-01     6.542    <0.01 ***
Turbo_E - Navar_L == 0   -8.155e-01  1.618e-01    -5.041    <0.01 ***
Yukon_E - Navar_L == 0   -7.767e-01  1.618e-01    -4.801    <0.01 ***
Yukon_L - Nelson_E == 0   7.767e-01  1.618e-01     4.801    <0.01 ***
Turbo_E - Nelson_L == 0  -7.282e-01  1.618e-01    -4.501    <0.01 ***
Yukon_E - Nelson_L == 0  -6.893e-01  1.618e-01    -4.261    <0.01 **
Turbo_L - Turbo_E == 0    5.340e-01  1.618e-01     3.301   0.0467 *
Yukon_L - Turbo_E == 0    1.155e+00  1.618e-01     7.142    <0.01 ***
Yukon_E - Turbo_L == 0   -4.951e-01  1.618e-01    -3.061   0.0932 .
Yukon_L - Turbo_L == 0    6.214e-01  1.618e-01     3.841    <0.01 **
Yukon_L - Yukon_E == 0    1.117e+00  1.618e-01     6.902    <0.01 ***
```

B.7 Results for Tukey's range test for question 2

```
                     Estimate Std. Error t value Pr(>|t|)
Major - Bolero == 0  -0.52427    0.16177  -3.241   0.0156 *
Navar - Bolero == 0  -0.12621    0.16177  -0.780   0.9709
Nelson - Bolero == 0  0.15534    0.16177   0.960   0.9303
Turbo - Bolero == 0  -0.22330    0.16177  -1.380   0.7389
Yukon - Bolero == 0  -0.18447    0.16177  -1.140   0.8644
Navar - Major == 0    0.39806    0.16177   2.461   0.1367
Nelson - Major == 0   0.67961    0.16177   4.201   <0.001 ***
Turbo - Major == 0    0.30097    0.16177   1.861   0.4272
Yukon - Major == 0    0.33981    0.16177   2.101   0.2876
Nelson - Navar == 0   0.28155    0.16177   1.740   0.5050
Turbo - Navar == 0   -0.09709    0.16177  -0.600   0.9910
Yukon - Navar == 0   -0.05825    0.16177  -0.360   0.9992
Turbo - Nelson == 0  -0.37864    0.16177  -2.341   0.1787
Yukon - Nelson == 0  -0.33981    0.16177  -2.101   0.2877
Yukon - Turbo == 0    0.03883    0.16177   0.240   0.9999
```

References

[1] N. H. Bingham & John M. Fry *Regression.* Springer-Verlag London, 1st Edition, 2010.