Doing CRD Analysis in R correctly

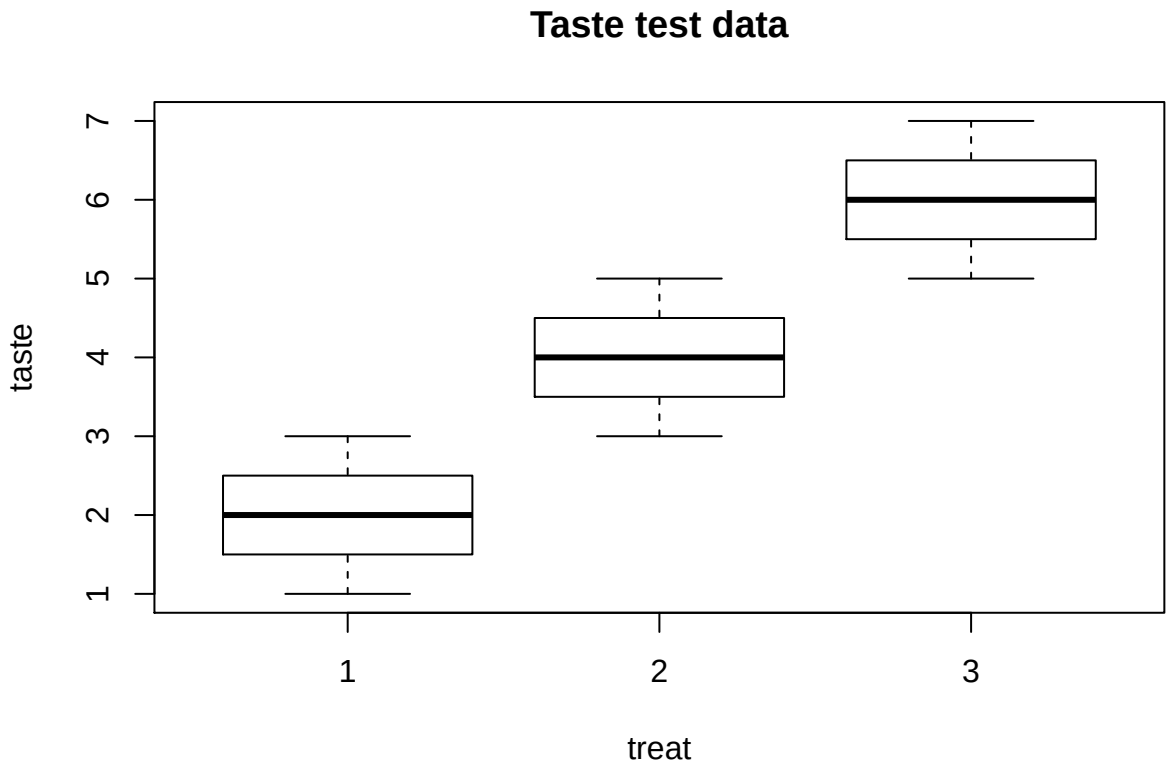
Chris Williams

Our taste test data

We can read in the taste test data and look at a boxplot:

```
tastetest <- read.table("c://temp/taste507aug.txt",header=TRUE)

boxplot(taste ~ treat, data = tastetest, main = "Taste test data")
```



This does seem to show differences among groups.

Analyzing the data (Take 1)

We create an object using the lm() function, which creates an lm object. The first argument to lm() is of the form 'y ~ x' where y is the dependent variable and x is a variable used to predict y. We then use functions like anova() and summary() to give us information about the fitted model.

```r
tastetest.lm <- lm(taste ~ treat, data = tastetest)

anova(tastetest.lm)
```

```
## Analysis of Variance Table
##
## Response: taste
##           Df Sum Sq Mean Sq F value   Pr(>F)
## treat      1     24 24.0000      28 0.001134 **
## Residuals  7      6  0.8571
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(tastetest.lm)
```

```
##
## Call:
## lm(formula = taste ~ treat, data = tastetest)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -1     -1      0      1      1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0000     0.8165   0.000  1.00000
## treat         2.0000     0.3780   5.292  0.00113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9258 on 7 degrees of freedom
## Multiple R-squared:    0.8,  Adjusted R-squared:  0.7714
## F-statistic:    28 on 1 and 7 DF,  p-value: 0.001134
```

Note that R does not print the total SS or df, unlike SAS and some other packages. Something seems strange here, because the degrees of freedom values (Df) do not match what we had in class previously. Why? What does the summary() function tell us? In the anova table, having df = 1 for treat might mean that there are only two groups, but we know that is not true. Here it is because R thinks we want to fit a regression model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i$$

with a continuous covariate x. The values from the summary() function are the estimated intercept and slope, so we have:

$$\hat{y}_i = 0 + 2 * x_i = 2 * x_i$$

However, x_i is not a measured variable but just represents the type of item being tasted.

Analyzing the data (Take 2)

To avoid this misunderstanding, we need to tell R that the variable treat is a factor, using the as.factor() function:

```r
 # we must make sure that treat is a factor, not numeric
tastetest$treat <- as.factor(tastetest$treat)

tastetest.lm2 <- lm(taste ~ treat, data = tastetest)

anova(tastetest.lm2)
```

```
## Analysis of Variance Table
##
## Response: taste
##           Df Sum Sq Mean Sq F value Pr(>F)
## treat      2     24      12      12  0.008 **
## Residuals  6      6       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(tastetest.lm2)
```

```
##
## Call:
## lm(formula = taste ~ treat, data = tastetest)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -1     -1      0      1      1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0000     0.5774   3.464  0.01340 *
## treat2        2.0000     0.8165   2.449  0.04983 *
## treat3        4.0000     0.8165   4.899  0.00271 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1 on 6 degrees of freedom
## Multiple R-squared:    0.8,  Adjusted R-squared:  0.7333
## F-statistic:    12 on 2 and 6 DF,  p-value: 0.008
```

Now the ANOVA table matches what we had before. The coefficient estimates from the summary() function give us a different prediction equation:

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 * I(treat2) + \hat{\beta}_2 * I(treat3)$$

or

$$\hat{y_i} = 2 + 2 * I(treat2) + 4 * I(treat3)$$

Here $I(treat2)$ is an indicator function, equal to 1 if the observation has $treat = 2$ and 0 otherwise. $I(treat3)$ is defined similarly for $treat3$. This gives us predicted values of 2, 4, and 6, for observations in group 1, 2, or 3, respectively. These values equal the respective group means. For ANOVA models the coefficient estimates from summary() are usually of less interest than the group mean estimates, so we often do not focus on this output as much for ANOVA models. When we have models using both factors and measured variables, like in the analysis of covariance, then we will again be more interested in the summary() estimates. In SAS we tell it about factors using the CLASS statement in Proc GLM and other procedures.