Using R & R Commander in Biomathematics Research

BioMAPS Workshop 2011

Christopher J. Mecklin

Department of Mathematics & Statistics
Biomathematics Research Group
Murray State University
Murray, KY 42071
christopher.mecklin@murraystate.edu

May, 2011

Outline

1. What is **R** and the **R** Commander?
2. Linear Regression Models
3. Basic ANOVA and ANCOVA Models
4. Information Criteria
5. Repeated Measures Models

What is R?

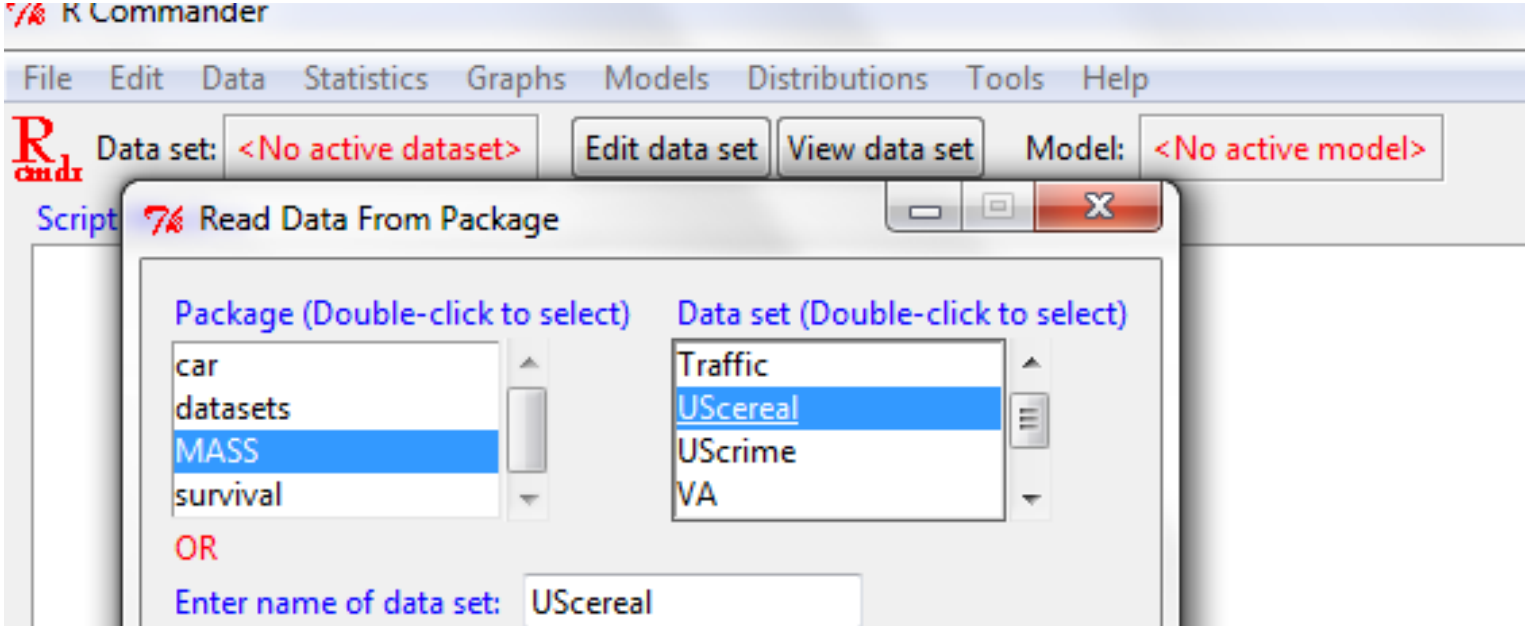
- **R** is an open-source statistical programming environment that is available for free.
- The `Rcmdr` package, written by John Fox, provides a GUI for **R**.
- **R** is similar to S, a statistical programming language developed at Bell Labs.
- I will assume that you have gone through 'An Introduction to the R Commander'-this was covered in the Spring 2011 section of BIO/MAT 460.

Introduction to Linear Regression

As our example for simple and multiple linear regression, we will utilize a dataset called `UScereal` that is built into the `MASS` package in **R**. This dataset has various information on 65 popular breakfast cereals, including nutritional information, manufacturer, and what shelf the cereal is typically displayed on at the grocery store.

Open the R Commander by either typing `library(Rcmdr)` into the R console or by going to Packages → Load package...→ Rcmdr.

Then go to Data→Data in packages→Read data set from an attached package...

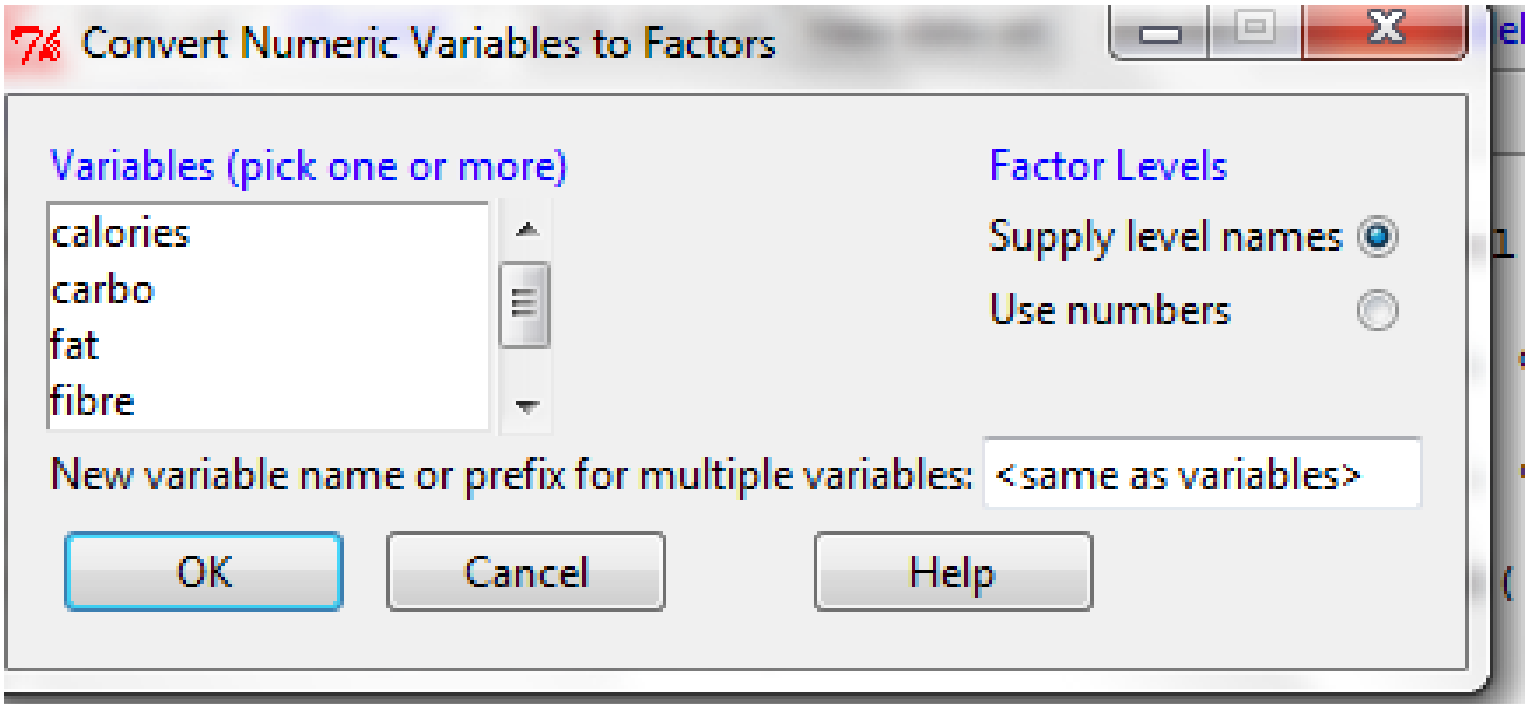


Graphs

Suppose we would like to look at boxplots or error bar plots of the sugar content of the cereals, broken down by shelf.

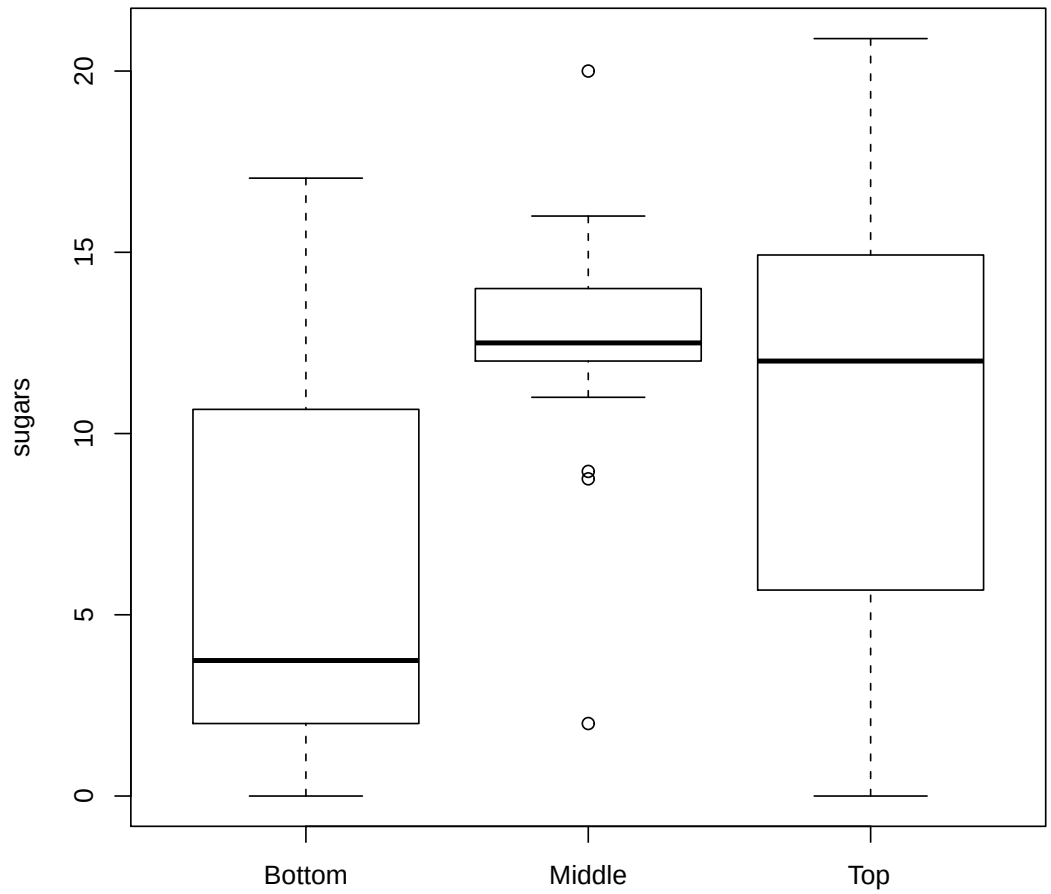
shelf is numbered 1,2,3 for bottom, middle, top. To tell the R Commander that shelf should be considered as a factor, go to Data→Manage variables in active data set→Convert numeric variables to factors...

We can give the factors descriptive names rather than numeric labels if we like.



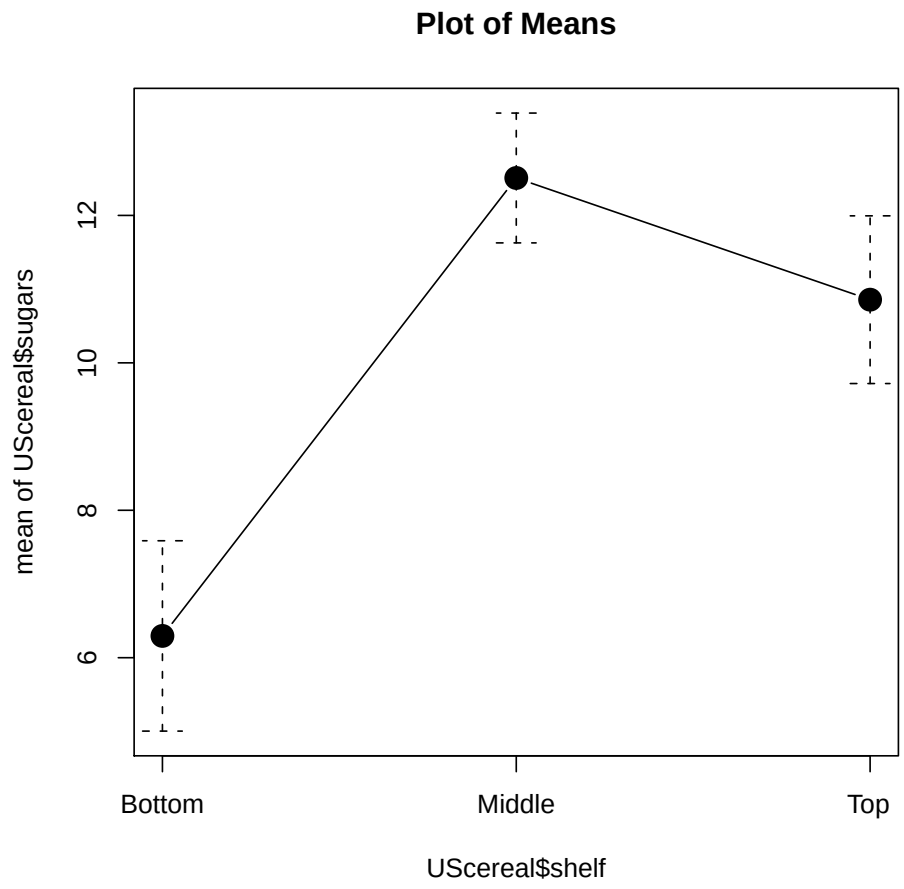
Boxplots and Error Bar Plots

Under the Graphs menu are many different plots. For instance, I could plot sugar content by shelf with either a boxplot or an error bar plot. I have done so, chosing my error bars in the latter plot to be ± 1 standard error.



Boxplots and Error Bar Plots

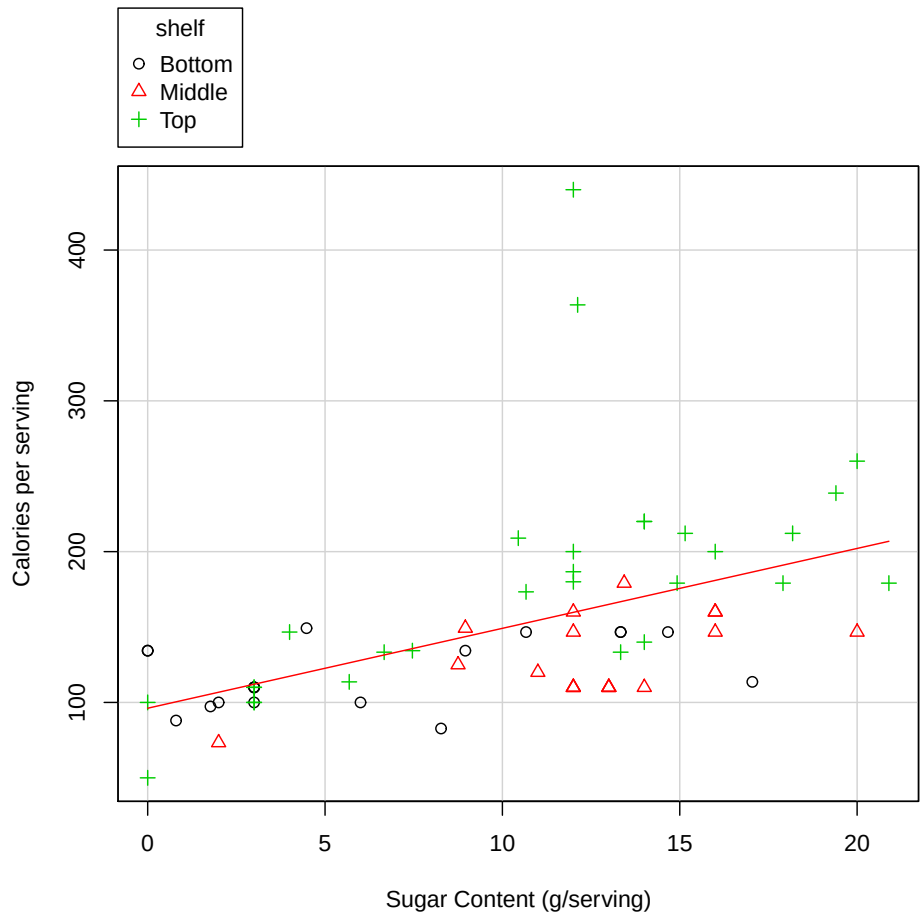
Under the Graphs menu are many different plots. For instance, I could plot sugar content by shelf with either a boxplot or an error bar plot. I have done so, chosing my error bars in the latter plot to be ± 1 standard error.



Scatterplot

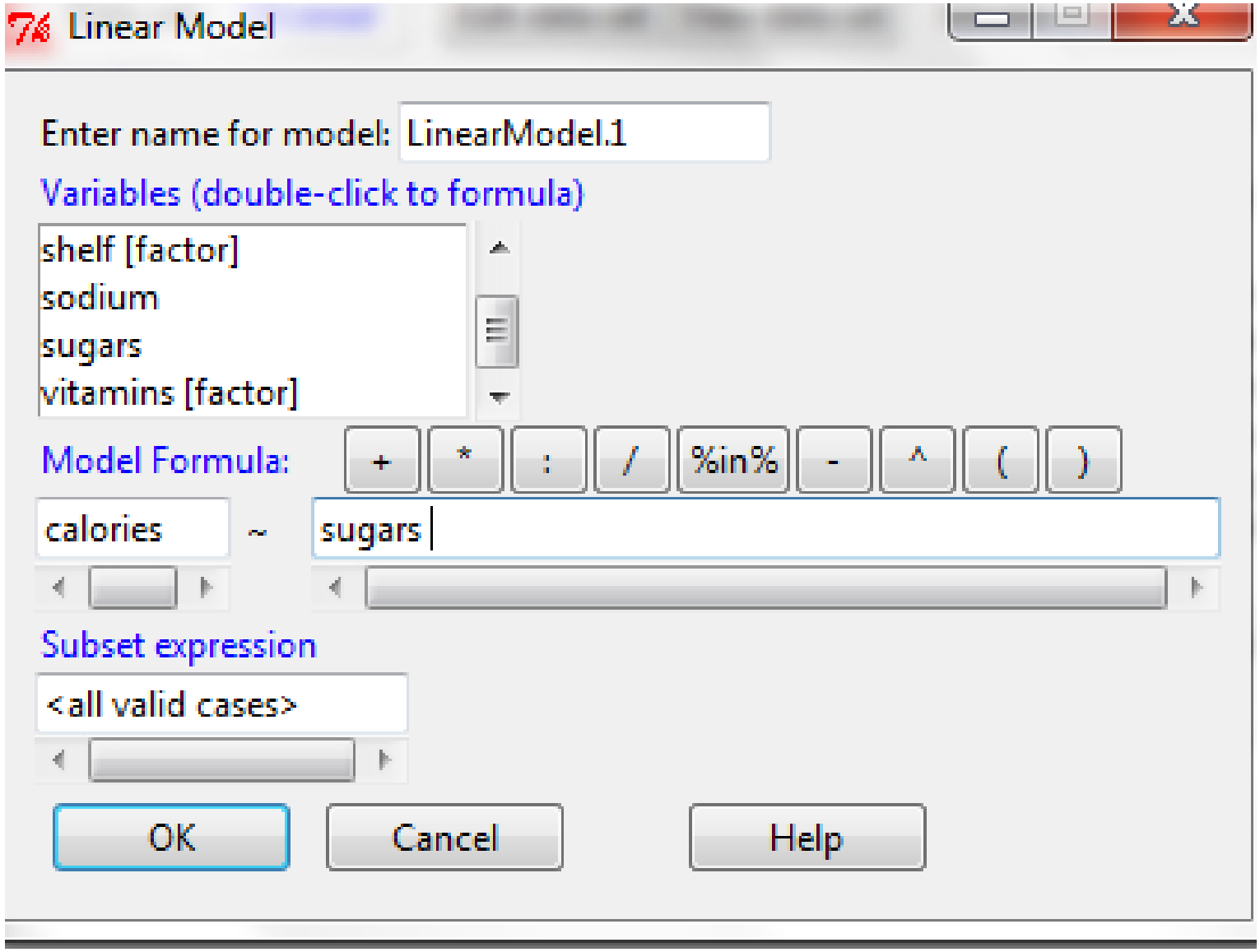
Before fitting a simple linear regression model, we should look at a scatterplot. Let us consider the model (in R notation, with calories as the response Y and sugars as the predictor X)

$$calories \sim sugars$$



Simple Linear Regression

To fit a linear model such as a linear regression or ANOVA go to Statistics→Fit Models→Linear model...



Regression Output

A summary of your regression model and the script created appears in the Output window.

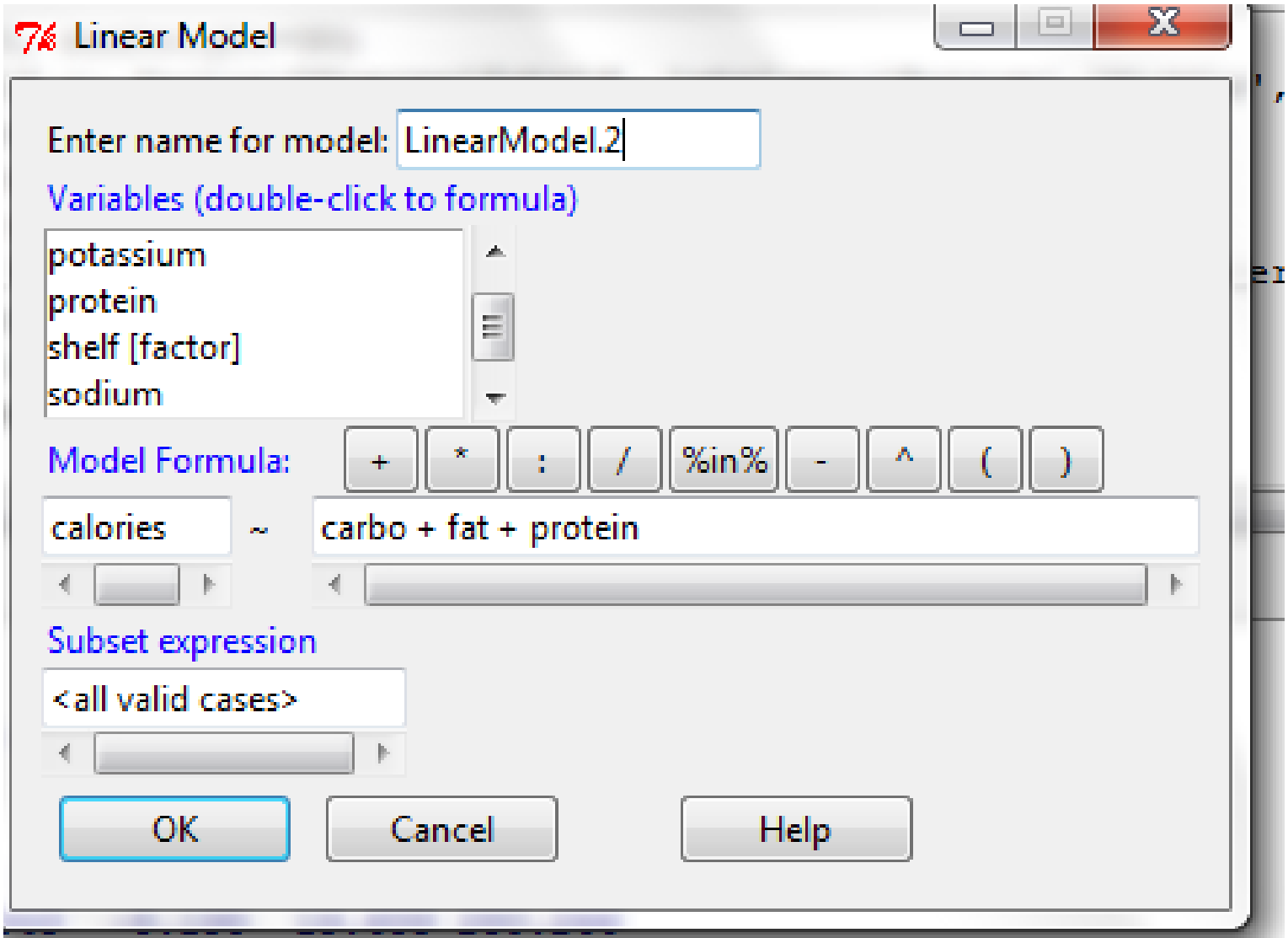
$$\hat{Calories} = 96.164 + 5.298 Sugars$$

```
> summary(LinearModel.1)
Call: lm(formula = calories   sugars, data = UScereal)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 96.164 13.579 7.082 1.44e-09 ***
sugars 5.298 1.171 4.525 2.73e-05 ***
Residual standard error: 54.65 on 63 degrees of freedom
Multiple R-squared: 0.2453, Adjusted R-squared: 0.2333
F-statistic: 20.48 on 1 and 63 DF, p-value: 2.733e-05
```

Multiple Linear Regression

It is not hard to fit a multiple linear regression model, such as:

$$calories \sim carbo + fat + protein$$



The ANOVA table

To obtain an ANOVA table with sums of squares and partial F tests, go to Models $\rightarrow$ Hypothesis tests $\rightarrow$ ANOVA table...

Choose Type II sum of squares unless the sequential order of entry of predictors into the model is important; in that case, choose Type I.

```
Anova Table (Type II tests)
Response: calories
Sum Sq Df F value Pr(>F)
carbo 62259 1 106.073 5.704e-15 ***
fat 31599 1 53.836 6.090e-10 ***
protein 7115 1 12.122 0.000927 ***
Residuals 35804 61
```

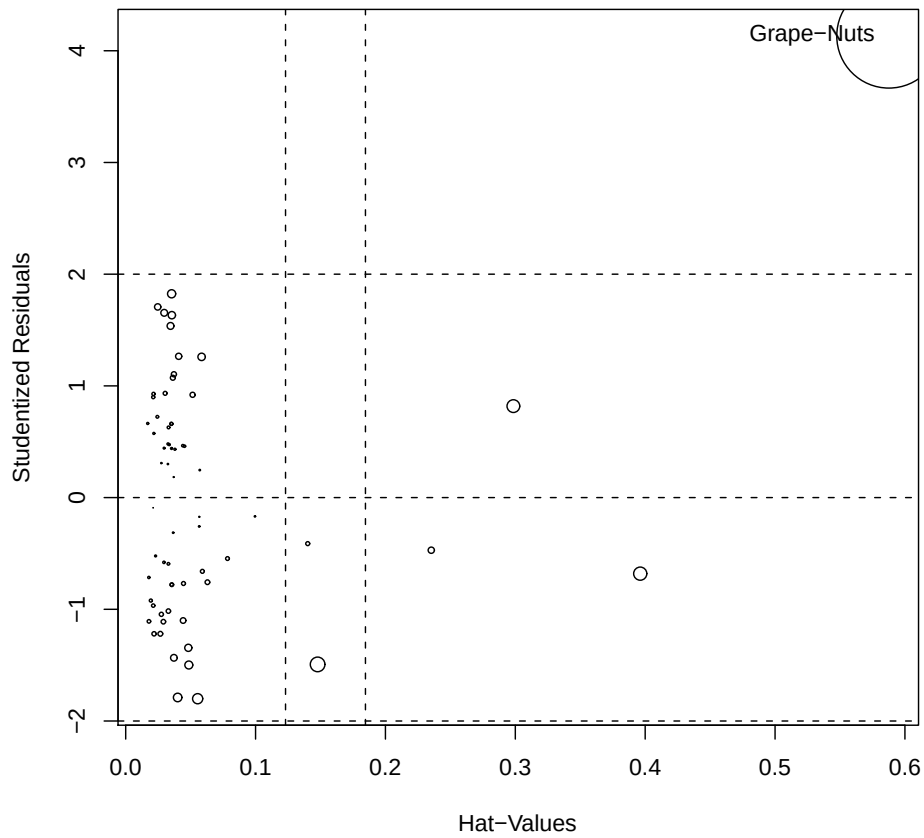
More Regression Stuff

Other regression tools available in R Commander:

- Confidence Intervals
- Akaike Information Criterion and Bayesian Information Criteria (more later)
- Stepwise model selection
- Susbset model selection
- Comparison of two models (via partial F test or Wald test; extra work required to tease out a likelihood ratio test)
- Regression Diagnostics

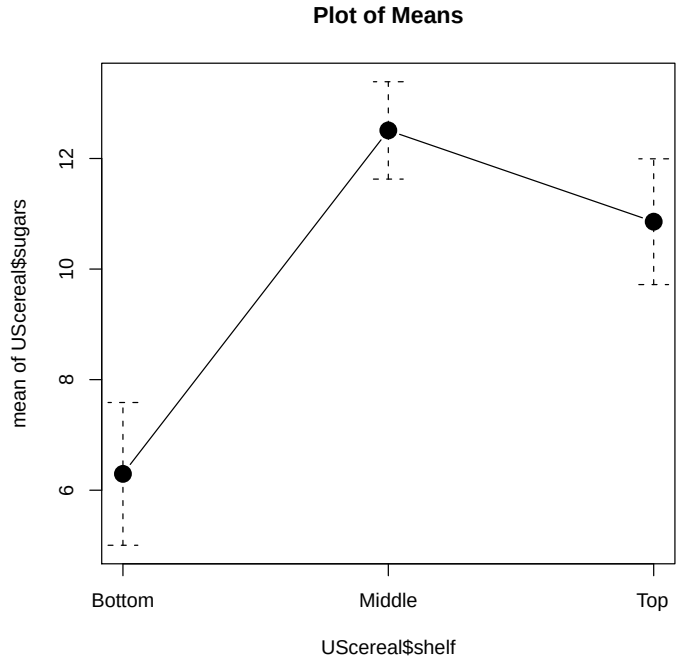
Influence Plot

My favorite type of diagnostic plot, sometimes called a 'bubble' plot, has studentized residuals on the y-axis, hat-values (leverage) on the x-axis, and bubbles that are proportional to Cook's Distance. The influential point in the cereal data set is "Grape-Nuts".



One Way ANOVA

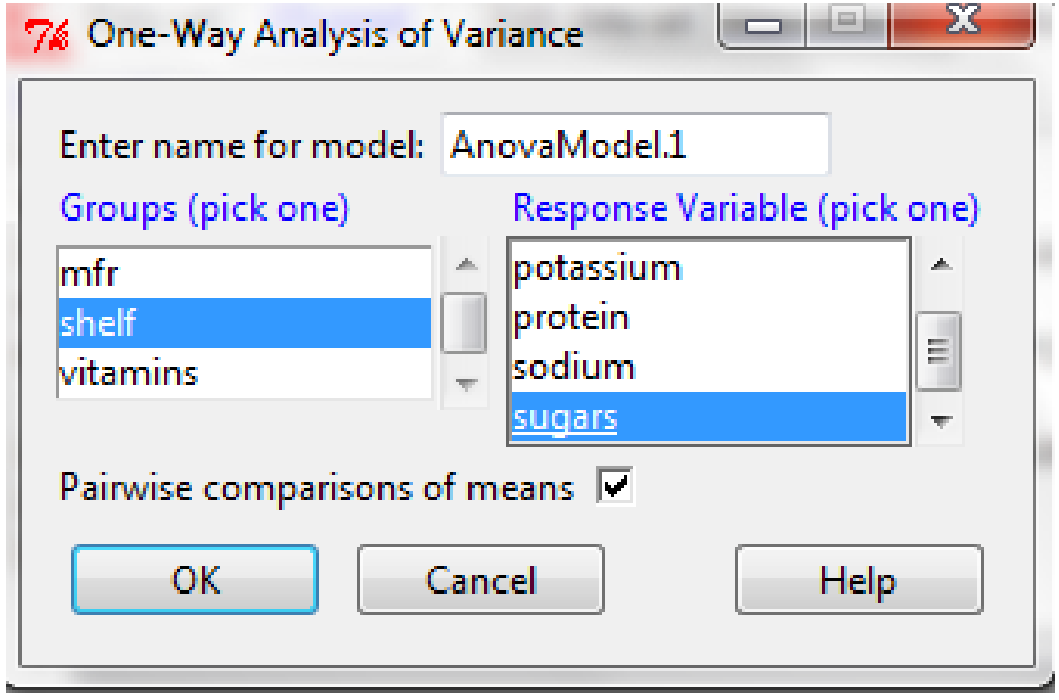
R can also be used to fit classical ANOVA models, using the aov command in R console or the appropriate menu choice in the R Commander. Let us first consider a simple one-way ANOVA model, looking to see if there is a significant difference in mean sugar content per serving of cereal for cereals found on the bottom, middle, or top shelf. Earlier, we looked at both the boxplot and error bar plot, and visually there seems to be significantly less sugar in bottom shelf cereals, with the middle shelf being the highest.



Fit ANOVA model

Go to Statistics→Means→One-Way ANOVA...

Check the Pairwise Comparisons box to obtain the Tukey HSD post hoc test.



One Way ANOVA Output

```
> summary(AnovaModel.1)
Df Sum Sq Mean Sq F value Pr(>F)
shelf 2 381.33 190.667 6.5752 0.002572 **
Residuals 62 1797.87 28.998

mean sd n
Bottom 6.295493 5.477309 18
Middle 12.507670 3.735734 18
Top 10.856821 6.125487 29
```

Post Hoc Tests

The results of Tukey's HSD test for pairwise comparisons.

```
Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = sugars~shelf, data = UScereal)
Linear Hypotheses:
Estimate Std. Error t value Pr(>|t|)
Middle - Bottom == 0 6.212 1.795 3.461 0.00272 **
Top - Bottom == 0 4.561 1.616 2.823 0.01736 *
Top - Middle == 0 -1.651 1.616 -1.022 0.56527
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
(Adjusted p values reported -- single-step method)
```

Tukey Confidence Intervals

```
Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = sugars~shelf, data = UScereal)
Quantile = 2.3995
95% family-wise confidence level

Linear Hypotheses:
Estimate lwr upr
Middle - Bottom == 0 6.2122 1.9050 10.5193
Top - Bottom == 0 4.5613 0.6841 8.4386
Top - Middle == 0 -1.6508 -5.5281 2.2264
```

Two Way ANOVA

The following problem will consider the growth of orange trees (increase of diameter in cm over 2 years), considering both the pH of the soil (4.0,5.0,6.0,7.0) and the amount of Calcium added (100,200,300 lb/acre) as factors, with 3 replications per cell.

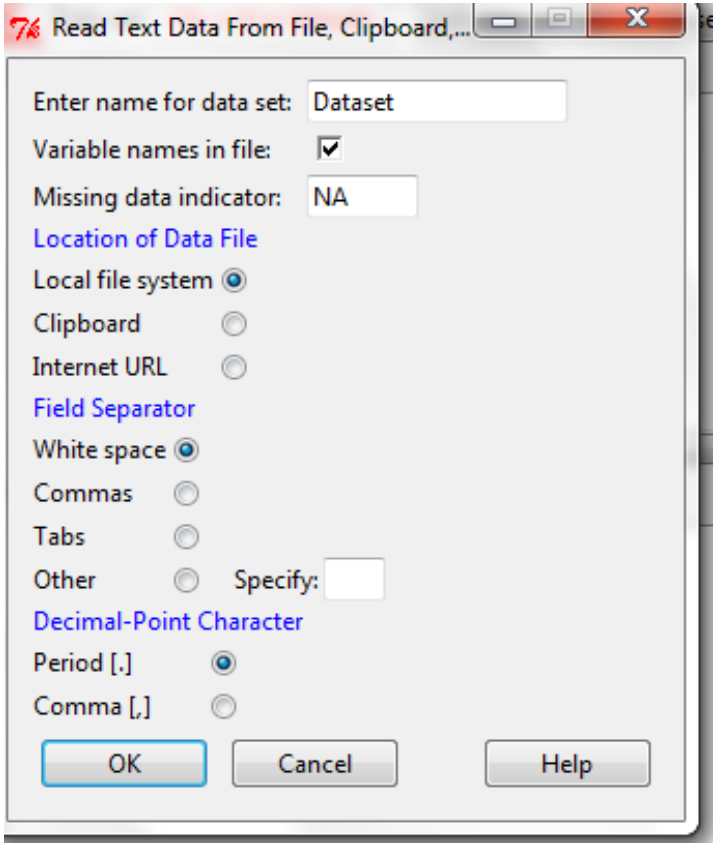
The data is availabe as a space-delimited text file at:
`http://campus.murraystate.edu/academic/faculty/christopher.mecklin/MAT565/OrangeTreeGrowth.txt`

and as an EXCEL file at:

`http://campus.murraystate.edu/academic/faculty/christopher.mecklin/MAT565/OrangeTreeGrowth.xls`

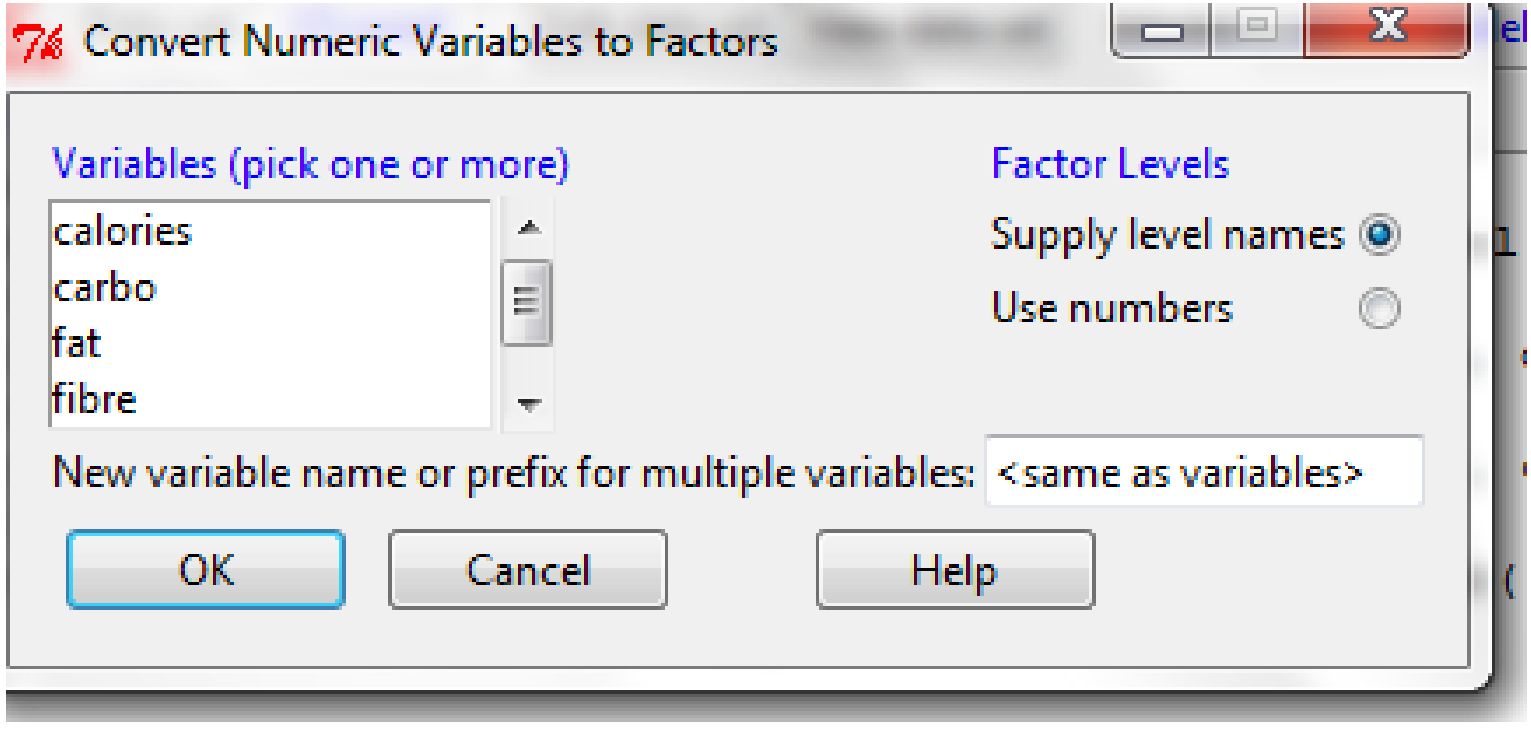
Import Orange Tree Growth data set

You can either download one of these files to your computer and import into the R Commander via Data→Import Data, choosing the appropriate format, or by directly typing in the URL for the .txt data file.



Create Factors

R Commander will treat pH and Ca as numeric variables, so we will convert to factors as we did earlier. This time, instead of supplying factor names, I will use the numbers as the factor levels.



Fit Two-Way ANOVA model

Now we will fit the factorial ANOVA model, including interaction. The **R** notation for this model is:

```
growth~pH+Ca+pH:Ca
```

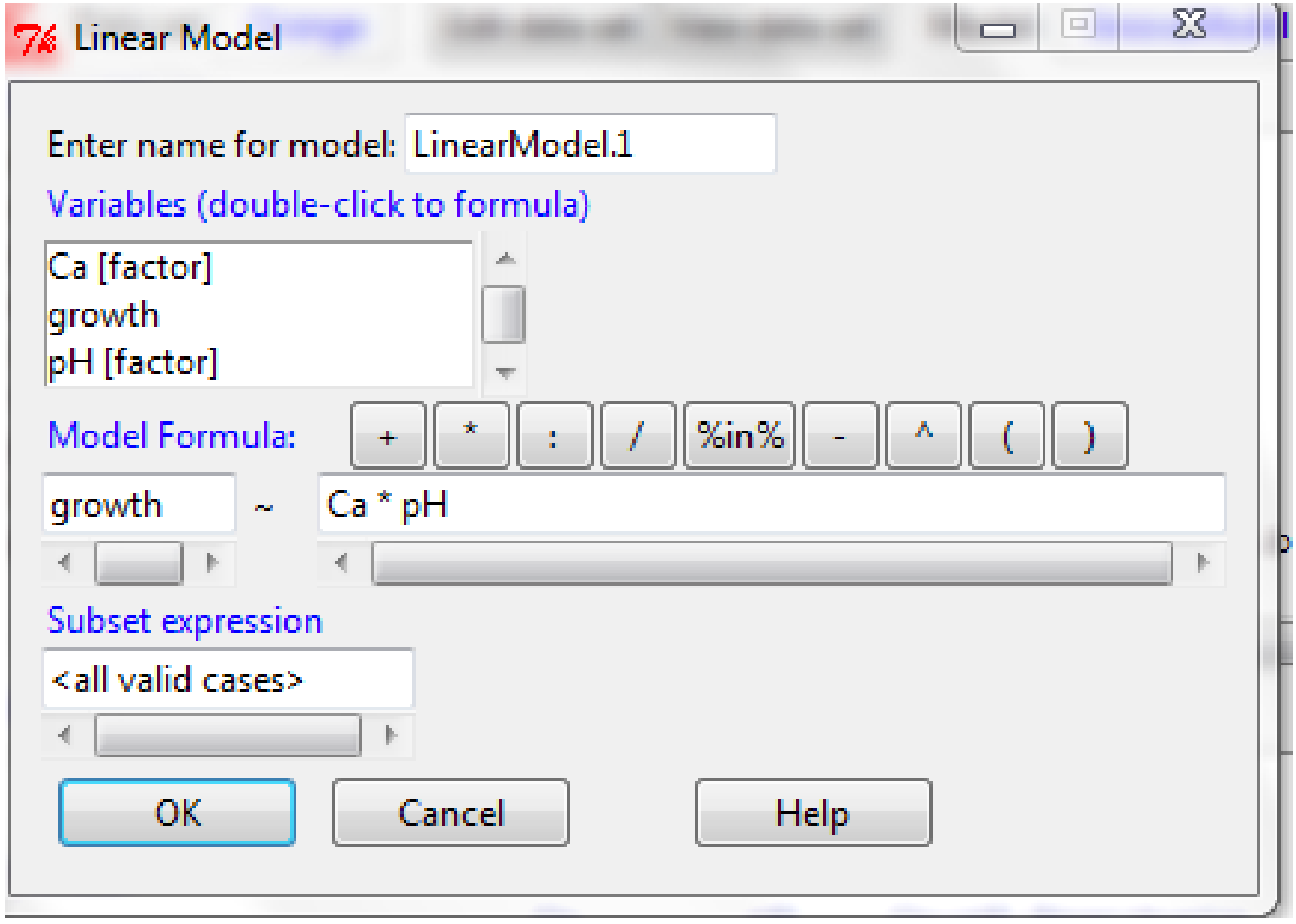
where `pH:Ca` denotes the interaction term. A shorter and equivalent form is:

```
growth~pH*Ca
```

Note the **R** syntax differs from SAS syntax.

Fit Two-Way ANOVA model

In R Commander, go to Statistics → Means → Multi-Way ANOVA OR Statistics → Fit models → Linear model...



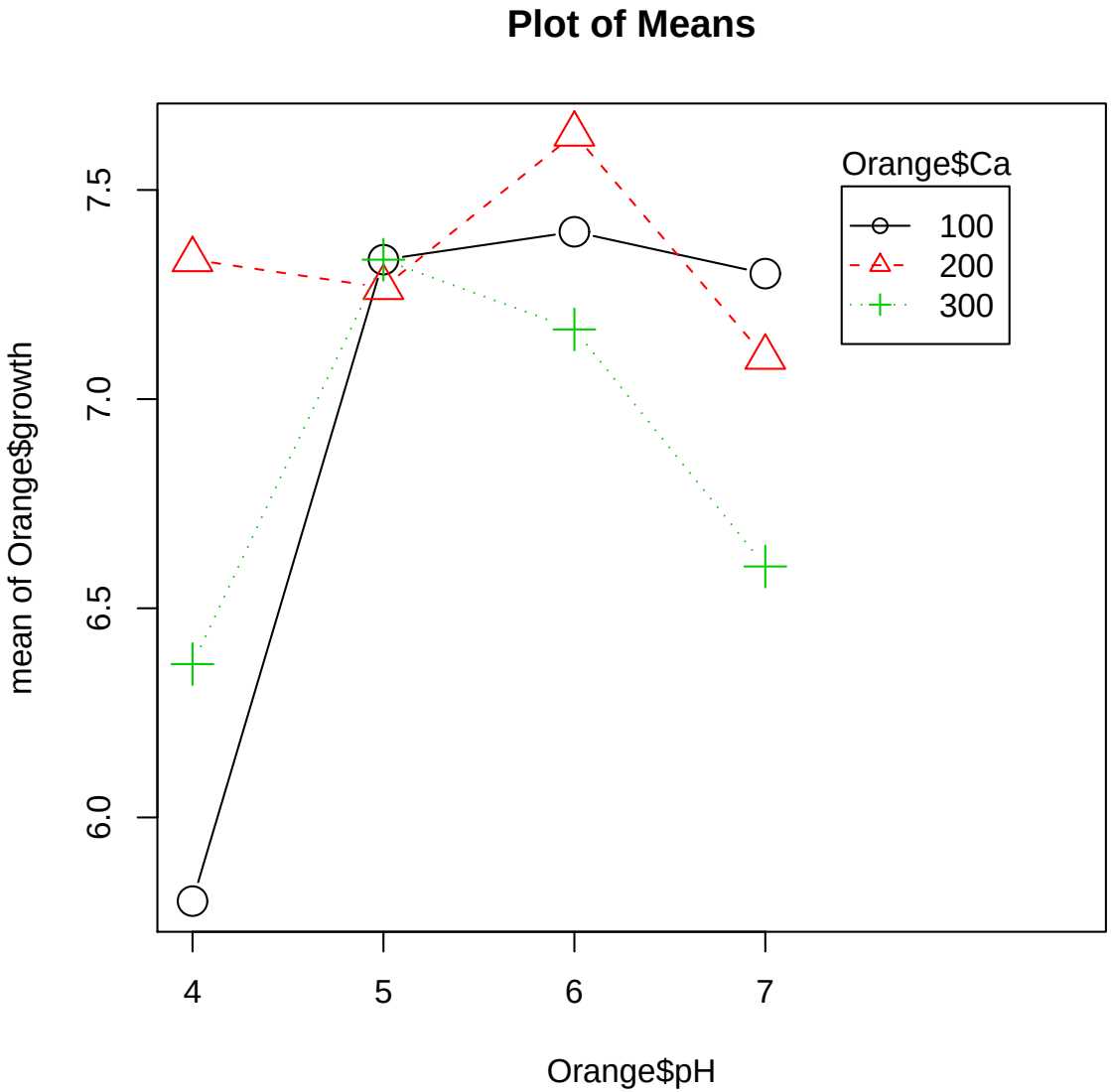
ANOVA table

Go to Models → Hypothesis tests → ANOVA table. With both main effects and the interaction all statistically significant, we will need to interpret the interaction first. The interaction plot will be useful.

```
Anova Table (Type II tests)
Response: growth
Sum Sq Df F value Pr(>F)
Ca 1.4672 2 10.8238 0.0004462 ***
pH 4.4608 3 21.9385 4.635e-07 ***
Ca:pH 3.2550 6 8.0041 8.186e-05 ***
Residuals 1.6267 24
---
```

Interaction Plot

Go to Graphs→Plot of Means. Choose both factors and no error bars.



ANCOVA Models

Once can also fit analysis of covariance, or ANCOVA, models in both R and the R Commander. The right hand side of the model will contain both categorical factors and numeric covariates.

An example of when ANCOVA would be an appropriate model to fit came from a BioMAPS project involving Dr. Derting, Callie Wilson, and Erin Keeney. In their situation, the response was the dry mass of a mouse's immune organs, the factor was group (either control/cell-mediated/cell-mediated & humoral), with the body mass of the mouse as the covariate.

The model, in R notation, would look something like this.

```
model1<-aov(immdry~group+bodymass) OR
model1<-lm(immdry~group+bodymass)
```

possibly using a Tukey's or other post hoc procedure on the factor if it was found to be significant.

Information Theory

- Kullback-Liebler Information
- Definition of Akaike Information Criterion (AIC)
- Computation of AIC
- Use of AIC
- Akaike weights
- Model selection with AIC

Kullback-Liebler Information

Let $f(\mathbf{y})$ represent the "true" probabilty density function for the response vector $\mathbf{y}$ in a statistical model (i.e. a linear model, generalized linear model, etc.). Of course, it is unlikely that we actually have the "true" model, but we might have several statistical models under consideration.

Kullback-Liebler information is a measure of "distance" between two models, where the second model is used to approximate the first. $K - L$ information, $I(f, g)$ can be thought of as the amount of information that is "lost" when model $g(\mathbf{y})$ is used to approximate the reality, or true model, $f(\mathbf{y})$.

$$I(f, g) = \int f(\mathbf{y}) \ln\left(\frac{f(\mathbf{y})}{g(\mathbf{y}|\theta)}\right) dy$$

(Burnham & Anderson, 2001; Fox, 2008)

Akaike's Contribution

Akaike (1973,1974) linked $K - L$ information and maximum likelihood, a heavily used method for parameter estimation. His contribution was to show that an estimate of expected $K - L$ information was based on the maximized log-likelihood function. This yielded the well-known *Akaike's Information Criterion*, or *AIC*:

$$AIC = -2\ln(L(\hat{\theta}|data)) + 2K$$

where the first term is minus two times the maximum of the log-likelihood funrction (aka the *deviance*) and K in the second term is the number of parameters in the model. The latter term is often thought of as a "penalty" term.

AIC in least squares

In the special case of least squares estimation (i.e. a linear model such as a t-test, linear regression, ANOVA), AIC will simplify to either of the following forms:

$$AIC = n[\ln(\frac{2\pi SSE}{n}) + 1] + 2K$$

or

$$AIC = n\ln(\frac{SSE}{n}) + 2K$$

The two forms are equivalent up to a constant and the two different formulas are widely used in the literature and in *R*. The `AIC` function uses the first formula for basic linear models, while the `extractAIC` function uses the second.

The choice is arbitrary, as the difference between the AIC values of different models is all that is important, and this difference will be identical using either formula.

Variations of AIC

There are many variation of the AIC statistics. Burnham and Anderson strongly lobby for the corrected AIC, or AICc, especially when sample sizes are less than 40.

$$AICc = AIC + \frac{2K(K+1)}{N-K-1}$$

Schwarz's Bayesian Information Criterion is also popular. It is more conservative and will "penalize" models with more parameters more heavily than AIC.

$$BIC = -2\ln(L(\hat{\theta}|data)) + \ln(N)K$$

Example of AIC: The t-test

Let us consider the computation of AIC in the simplest possible setting, an independent samples t-test. In the following example, we have the exam scores of $n_1 = 20$ students who took the UBW 101 exam at 8 AM and $n_2 = 15$ students who took the same exam at 12 PM.

```
exam<-c(74,72,65,96,45,62,82,67,63,93,29,68,47,80,87,100,
86,87,89,75,88,81,71,87,97,83,81,49,71,63,53,77,71,86,78)
group<-c(rep(0,15),rep(1,20))
t.test(exam~group,var.equal=TRUE)
```

Example of AIC: The t-test, continued

```
Ho<-lm(exam~1) null hypothesis
Ha<-lm(exam~group) alternative hypothesis
summary(Ho)
anova(Ho)
extractAIC(Ho)
AIC(Ho) available in R Commander
AICc(Ho)
BIC(Ho)
summary(Ha)
anova(Ha)
extractAIC(Ha)
AIC(Ha)
AICc(Ha)
BIC(Ha)
```

R Output

```
> t.test(exam~group,var.equal=TRUE)
Two Sample t-test
data: exam by group
t = -1.8824, df = 33, p-value = 0.06862
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
-20.7733369 0.8066702
sample estimates:
mean in group 0 mean in group 1
68.66667 78.65000
```

R Output

```
> Ha<-lm(exam∼group) alternative hypothesis
> anova(Ho)
Analysis of Variance Table
Response: exam
Df Sum Sq Mean Sq F value Pr(>F)
Residuals 34 8810.2 259.12
> extractAIC(Ho)
[1] 1.000 195.491
> anova(Ha)
Analysis of Variance Table
Response: exam
Df Sum Sq Mean Sq F value Pr(>F)
group 1 854.3 854.29 3.5435 0.06862 .
Residuals 33 7955.9 241.09
> extractAIC(Ha)
[1] 2.0000 193.9212
```

Computation of AIC

We can see that for our 'null' model, that

$$AIC_0 = 35 \ln(\frac{8810.2}{35}) + 2(1) = 195.491$$

while for the alternative model

$$AIC_1 = 35 \ln(\frac{7955.9}{35}) + 2(2) = 193.921$$

. Hence,

$$\Delta_i = AIC_i - min(AIC)$$

$$\Delta_0 = 195.491 - 193.921 = 1.57$$

Rule of Thumb for AIC

The information criterion people don't like hypothesis testing or p-values much, since a p-value is $P(data|model)$, which they argue is backward from the desired probability, which is $P(model|data)$.

A general 'rule of thumb' (Burnham & Anderson 2001; Bolker, 2008) is:

- $\Delta_i < 2$: models are basically equivalent
- $4 < \Delta_i < 7$: models are clearly distinguished
- $\Delta_i > 10$: models are definitely different

Akaike Weights

As previously mentioned, fans of information theory do not like p-values computed from standard hypothesis tests since they feel computing the conditional probability of the data occuring GIVEN a model is "backwards". (Shouldn't we condition on what IS known?)

For our t-test example, we will compute *relative likelihoods* and *Akaike weights* w_i:

$$\mathcal{L}(null|data) = \exp(-\frac{1}{2}\Delta_0) = e^{-0.5(1.57)} = 0.4561$$

$$\mathcal{L}(alternative|data) = \exp(-\frac{1}{2}\Delta_1) = e^{-0.5(0)} = 1$$

In general, the Akaike weight w_i is:

$$w_i = P(model_i|data) = \frac{\mathcal{L}(model_i|data)}{\sum_{all\ i} \mathcal{L}(model_i|data)}$$

So, the Akaike weights for the t-test are:

$$w_0 = P(null|data) = \frac{.4561}{.4561 + 1} = .3132$$

$$w_1 = P(alternative|data) = \frac{1}{.4561 + 1} = .6868$$

With just two models and corresponding Akaike weights, we often compute the evidence ratio, ER and conclude the models are essentially equivalent if $ER < e$. The evidence ratio for the alternative is:

$$ER = \frac{w_1}{w_0} = \frac{.6868}{.3132} = 2.1928$$

With multiple models, we generally discount models with $w_i < 0.1$. Sometimes estimation and/or inference is based on 'model averaging', using all reasonable models weighted by their w_i.

Model Selection with AIC

One can consider all possible regression models using AIC, rather than R^2, adjusted R^2, or Mallow's C_p as the criterion. We might seek to minimize AIC, but we will not choose a model with more parameters unless its AIC is at least 2 lower than the simpler model's AIC, due to the **principle of parsimony**.

Stepwise regression methods also exist using AIC, with the usual caveats always present with the use of stepwise methodology. (Mecklin doesn't like stepwise).

The Cereal data

We will use R's built-in cereal dataset again. First, all possible regressions to predict calories based on a subset of the 3 predictors fat, carbo, and protein.

There are $2^3 = 8$ possible multiple regression models (not considering polynomial or interaction terms). Here, we also use Ben Bolker's `bbmle` package to create nice tables of the statistics.

The Cereal data

```
library(MASS)
library(bbmle)
attach(UScereal)
m1<-lm(calories~1)
m2<-lm(calories~carbo)
m3<-lm(calories~fat)
m4<-lm(calories~protein)
m5<-lm(calories~carbo+fat)
m6<-lm(calories~carbo+protein)
m7<-lm(calories~fat+protein)
m8<-lm(calories~carbo+fat+protein)
AICtab(m1,m2,m3,m4,m5,m6,m7,m8,base=TRUE,weights=TRUE,delta=TRUE,sort=TRUE)
AICctab(m1,m2,m3,m4,m5,m6,m7,m8,base=TRUE,weights=TRUE,delta=TRUE,sort=TRUE,
nobs=nrow(UScereal))
BICtab(m1,m2,m3,m4,m5,m6,m7,m8,base=TRUE,weights=TRUE,delta=TRUE,sort=TRUE,
nobs=nrow(UScereal))
```

AIC table for cereal regressions

```
> AICtab(m1,m2,m3,m4,m5,m6,m7,m8,base=TRUE,weights=TRUE,
delta=TRUE,sort=TRUE)
AIC df dAIC weight
m8 604.7 5 0.0 0.99254
m5 614.5 4 9.8 0.00746
m6 643.8 4 39.1 < 0.001
m2 663.6 3 58.9 < 0.001
m7 668.2 4 63.5 < 0.001
m4 682.0 3 77.3 < 0.001
m3 699.0 3 94.3 < 0.001
m1 724.8 2 120.1 < 0.001
```

Stepwise Regression with AIC

- Load 'UScereal' dataset into R Commander by going to:
  Data→Data in packages→Read data set from an attached package
- Fit the 'full' model: calories∼fat+carbo+protein by going to:
  Statistics→Fit Models→Linear Model...
- Use AIC for stepwise regression by going to:
  Models→Stepwise model selection...
  I usually choose 'Forward-Backward' to start with the null model and add predictors.
  'Backward-Forward' will start with the full model and will take out predictiors.

In Tom Anderson's master's thesis, he considered nine regression models chosen on the basis of biological principles rather than data dredging. He computed the $AICc$, $dAICc$ or Δ_i for each model, and the Akaike weight w_i for each model.

I feel this is superior science to looking at all possible regression models for a large number of predictors and/or resorting to stepwise methodology, and would suggest others emulate this approach to model selection.

Tom's AICc Table

Table 1: Models used in analysis of competition between A. maculatum and A. talpoideum. All models also included two random effects, Pond and Year, where Pond was the individual pond I.D. and Year was the sample year.

Model	Biological Principle	Covariates
1	Interspecific density & size	Interspecific density, size, & interaction
2	Intraspecific density	Intraspecific density
3	Interspecific density	Interspecific density
4	Interspecific Size	Interspecific Size
5	Predator Density	Newt density
6	Competitor Ratio*	maculatum:talpoideum density ratio
7	Global Biotic	All biotic variables
8	Abiotic	Canopy cover and pond size
9	Overall Global	Abiotic + Global Biotic

Tom's AICc Table

Table 2: Response is Maculatum Size. Differences in AICc scores (dAIC), Akaike weights for all potential models. Bold-faced AIC weights represent supported models that had an $w_i > 0.1$, following Van Buskirk (2005). All models include two random effects, Pond and Year. Number of observations = 19 ponds.

Model	K	$dAICc$	w_i
4	**3**	**0.0**	**0.535**
2	**3**	**0.7**	**0.368**
3	3	3.6	0.090
5	3	9.9	0.004
1	5	11.5	0.002
6	3	11.8	0.001
8	4	21.2	< 0.001
7	7	39.5	< 0.001
9	9	69.9	< 0.001

Repeated Measures ANOVA

It is quite common in research to take repeated measurements on experimental units. An example would be taking measurements over several different time periods to measure the growth of a plant. It is inappropriate to treat the different measurments for each individual as independent replications, as the errors will NOT be independent.

Incorrectly treating the repeated measurements as independent observations leads to *pseudoreplication*. Hence, the analysis would inflate the degrees of freedom present, possibly leading to incorrect inference (i.e. Type I errors).

For instance, if we have 2 treatments, 6 plants per treatment, and measure each plant 5 times, although we have taken $2 \times 6 \times 5 = 60$ measurements, we do not have $60 - 1{=}59$ degrees of freedom.

Crawley's Plant Growth data

In this problem, we will abandon R Commander, as it doesn't yet have a plugin available for repeated measures ANOVA and mixed models.

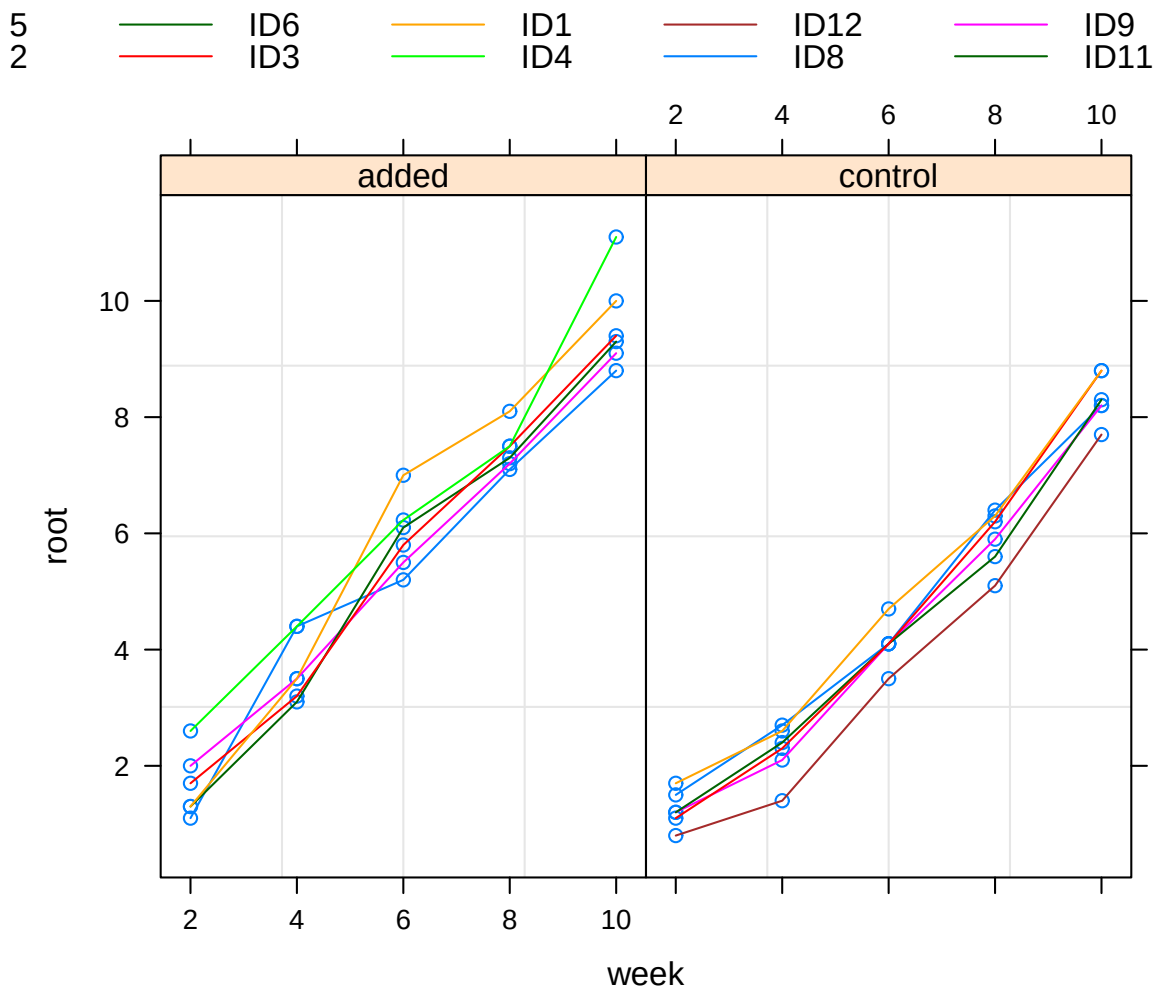
Our response variable is the biomass of the root of a plant. Fertilizer is a factor with 2 levels (added or control). There are 6 plants randomly assigned to each treatment. Each plant's root biomass is measured 5 times: at week 2,4,6,8,10. Fertilizer is a between-subjects variable, while week is a within-subjects variable.

The data is available in 'long' format at:
`http://www.bio.ic.ac.uk/research/mjcraw/therbook/data/fertilizer.txt`

Read in & graph data

```
plantgrowth<-read.table("http://www.bio.ic.ac.uk/research/
mjcraw/therbook/data/fertilizer.txt",header=T)
attach(plantgrowth)
names(plantgrowth)
library(nlme)
library(lattice)
plantgrowth<-groupedData(root week|plant,
outer=~fertilizer,plantgrowth)
plot(plantgrowth,outer=T)
```

Lattice plot



Wrong ANOVA

This analysis is *incorrect* due to pseudoreplication!

```
> week<-ordered(week)
> fertilizer<-factor(fertilizer)
> wrong.anova<-aov(root~fertilizer*week)
> summary(wrong.anova)
Df Sum Sq Mean Sq F value Pr(>F)
fertilizer 1 25.65 25.650 93.5803 4.85e-13 ***
week 4 423.73 105.932 386.4776 < 2.2e-16 ***
fertilizer:week 4 3.54 0.884 3.2269 0.01968 *
Residuals 50 13.70 0.274
```

Correct Repeated Measures ANOVA

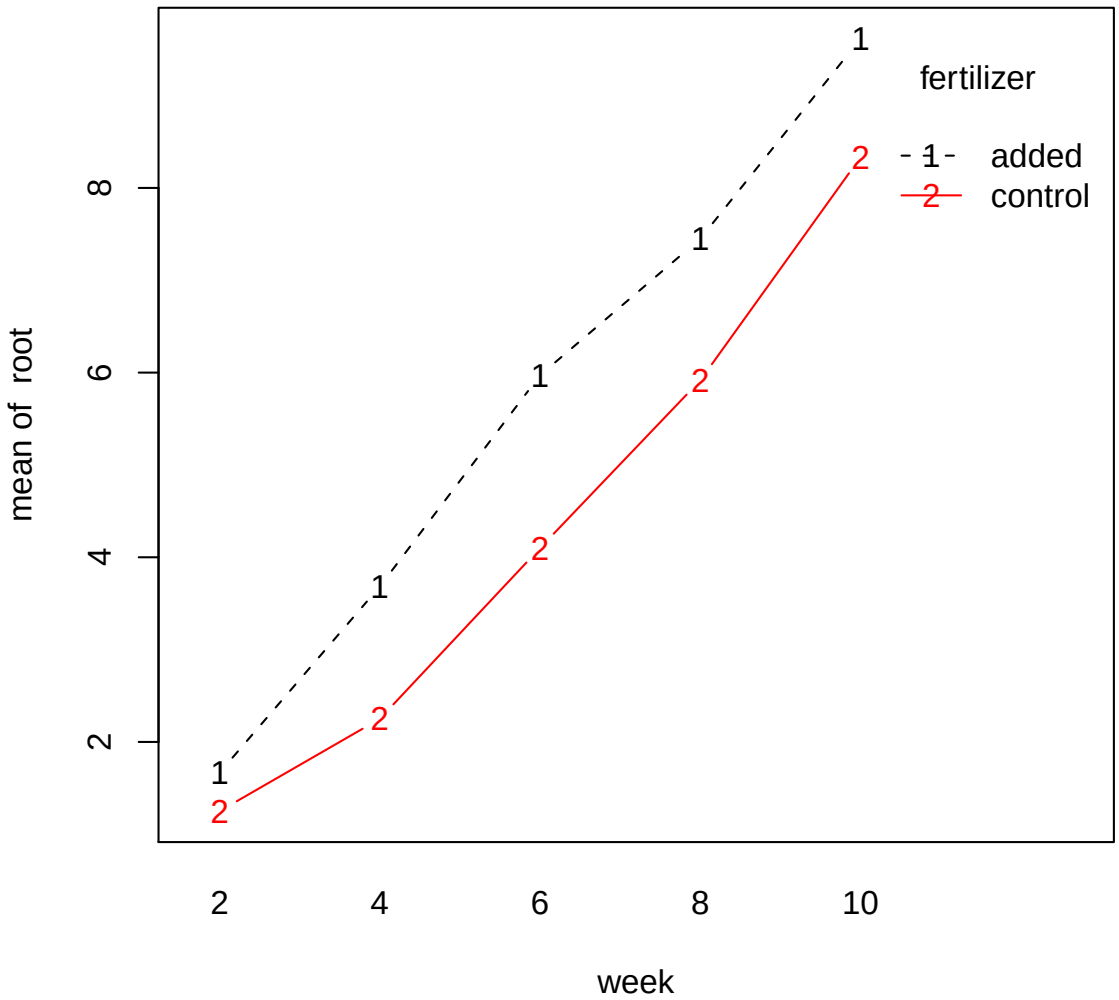
The correct analyses specifies the proper error terms to be used in the construction of the F test statistics.

```
> week<-ordered(week)
> rm.anova<-aov(root~fertilizer*week+Error(plant/week)+fertilizer)

> summary(rm.anova)
Error: plant
Df Sum Sq Mean Sq F value Pr(>F)
fertilizer 1 25.6499 25.6499 33.063 0.0001852 ***
Residuals 10 7.7578 0.7758
Error: plant:week Df Sum Sq Mean Sq F value Pr(>F)
week 4 423.73 105.932 712.512 < 2.2e-16 ***
fertilizer:week 4 3.54 0.884 5.949 0.0007459 ***
Residuals 40 5.95 0.149
```

Interaction Graph

```
> interaction.plot(week,fertilizer,root,type="b",col=1:2)
```



MANOVA approach

Our analysis of the repeated measures data with a classical ANOVA model assumes that *sphericity* holds. In other words, the variance of all differences between measures is equal. This ia analogous to the assumption of equal variances in one-way ANOVA.

This assumption can be tested, and if not met, the *df* and *p*-values can be adjusted. This requires working with the data in 'wide' format and fitting a MANOVA model, where the response is no longer a scalar, but a vector. In this case, the response vector would be an individual's measurements across all time periods.

We will not pursue this approach further today. I can't teach all of MAT 565 today :)

Mixed Models approach

A more model approach to repeated measures and other mixed models is to fit a linear mixed model (LMM), using restricted maximum likelihood (REML) rather than least squares for estimation.

Advantages to this approach include:

- Unbalanced designs (i.e. not all individuals measured at the same time intervals) can be handled.
- Generalized linear mixed models (GLMM) can be fit if assuming normal errors is untenable. This can be useful with binary responses or counts.

Mixed modeling can be quite complex and will not be elaborated upon today!

What can't R Commander do?

There are many statistical procedures that are not built in to the R Commander. We've seen one example. If you have the need for others, options include:

1. Writing R scripts (i.e. programs) and using the R console, with either the base packages or a package you downloaded. vegan is an example of a package not part of the base R distribution that I use frequently, mainly for computing biodiversity statistics.
2. People are starting to create *Plug-Ins* to add more features to the GUI. An example is `RcmdrPlugin.survival`, which adds standard survival analysis methods like logrank tests and Cox regression.
3. I hope someone creates a Plug-In for repeated measures and mixed models!
4. You are always able to write your own programs in case no packages or plug-ins exist or you just prefer to!
5. Thank you!