Lab 3 – Spring 2021

18^{th} March 2021

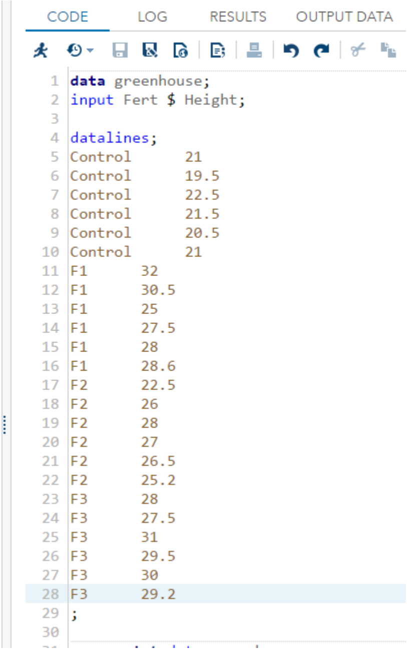
cavicchia@ese.eur.nl

1 Introduction to Anova Model

Analysis of variance (ANOVA) is a method to compare average (mean) responses to experimental manipulations in controlled environments. Let's start with an example. A plant biologist states that plants with different fertilizers will grow to different heights. In other words he/she thinks that plant height might depend on applying different fertilizers. This statement should be translated into a hypothesis system, where the null hypothesis is that the average height (or mean height) for plants with the different fertilizers will all be the same. The alternative hypothesis (which the biologist hopes to show) is that they are not all equal, but rather some of the fertilizer treatments have produced plants with different mean heights. The strength of the data will determine whether the null hypothesis can be rejected with a specified level of confidence.

1.1 Import the data using SAS code

The first line begins with the word *data* and invokes the datastep. Notice that the end of each SAS statements has a semi-colon. This is essential. In the datastep we are assigning a name to the data to be used and defining the variables we will use. Note that SAS assumes variables are numeric in the input statement, so if we are going to use a variable with alphanumeric values (e.g. F1 or Control), then we have to follow the name of the variable in the input statement with a $ sign.
A simple way to input small datasets is shown in this code, wherein we embed the data in the program. This is done with the word *datalines*.



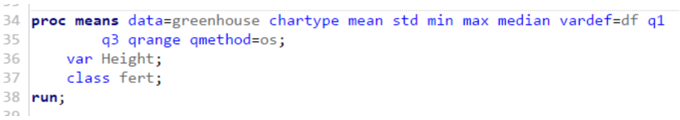
```
data greenhouse;
input Fert $ Height;

datalines;
Control      21
Control      19.5
Control      22.5
Control      21.5
Control      20.5
Control      21
F1      32
F1      30.5
F1      25
F1      27.5
F1      28
F1      28.6
F2      22.5
F2      26
F2      28
F2      27
F2      26.5
F2      25.2
F3      28
F3      27.5
F3      31
F3      29.5
F3      30
F3      29.2
;

```

The semi-colon here ends the datastep.

SAS then produces output of interest using *proc* statements, short for *procedure*. You only need to use the first four letters, so SAS code is full of *proc* statements to do various tasks. For example this is the code needed to obtain the summary statistics.

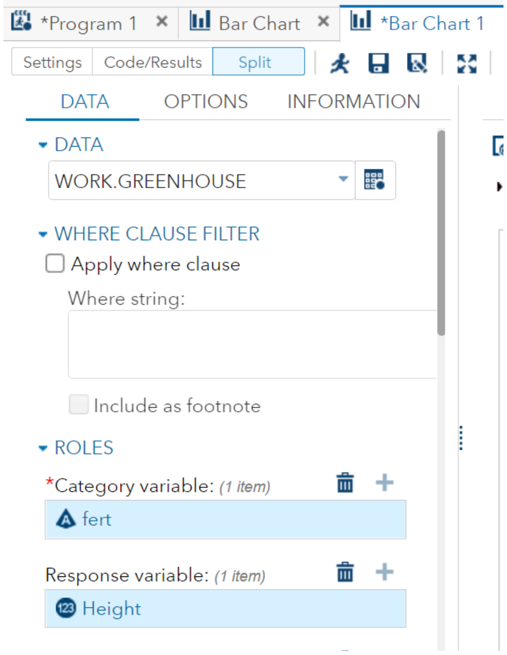


```
proc means data=greenhouse chartype mean std min max median vardef=df q1
        q3 qrange qmethod=os;
    var Height;
    class fert;
run;
```

In the summary procedure we identify the categorical classes with the *class* statement. Any variable NOT listed in the class statement is treated as a continuous variable. The target variable for which the summary will be made is specified by the **var** (for variable) statement.

1.2 Descriptive statistics

We test three kinds of fertilizer and also one group of plants that are untreated (the control). The plant biologist kept all the plants under controlled conditions in the greenhouse, to focus on the effect of the fertilizer, they only thing we know to differ among the plants. At the end of the experiment, the biologist measured the height of each plant. This is the dependent or response variable and is plotted on the vertical (y) axis. The biologist used a simple bar chart to plot the difference in heights.



DATA

WORK.GREENHOUSE

WHERE CLAUSE FILTER

☐ Apply where clause

Where string:

☐ Include as footnote

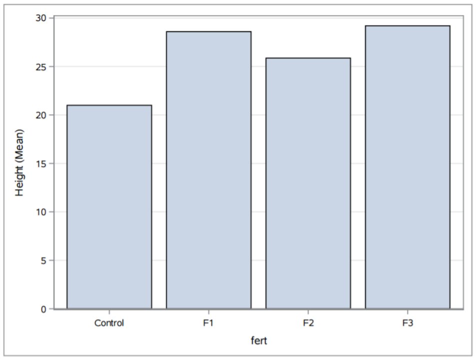
ROLES

*Category variable: *(1 item)*

fert

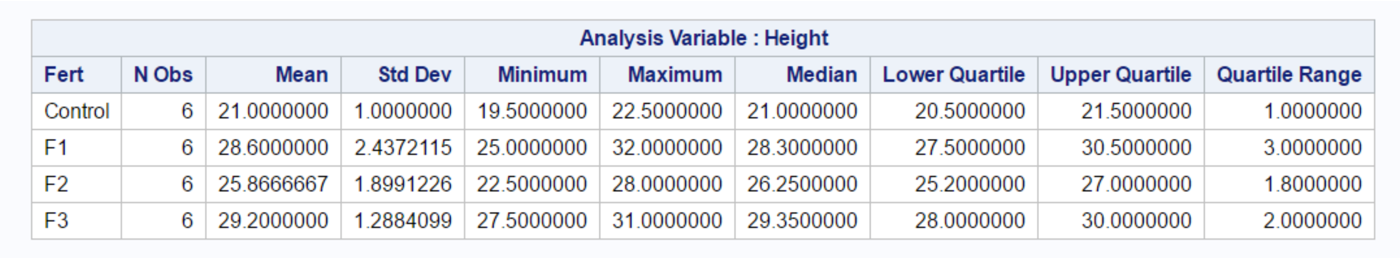
Response variable: *(1 item)*

Height



To create the bar plot under Tasks and Utilities, select the bar plot. Assign fert to the category variable box and Height to the response variable This bar chart is a common way to show treatment (or factor) level means. The only one treatment is the fertilizer. It has four levels that included the control, which received no fertilizer. The height of a bar shows the mean of each level.

Furthermore it is also useful to explore the data through a descriptive analysis.

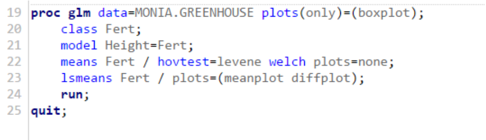


Analysis Variable : Height									
Fert	N Obs	Mean	Std Dev	Minimum	Maximum	Median	Lower Quartile	Upper Quartile	Quartile Range
Control	6	21.0000000	1.0000000	19.5000000	22.5000000	21.0000000	20.5000000	21.5000000	1.0000000
F1	6	28.6000000	2.4372115	25.0000000	32.0000000	28.3000000	27.5000000	30.5000000	3.0000000
F2	6	25.8666667	1.8991226	22.5000000	28.0000000	26.2500000	25.2000000	27.0000000	1.8000000
F3	6	29.2000000	1.2884099	27.5000000	31.0000000	29.3500000	28.0000000	30.0000000	2.0000000

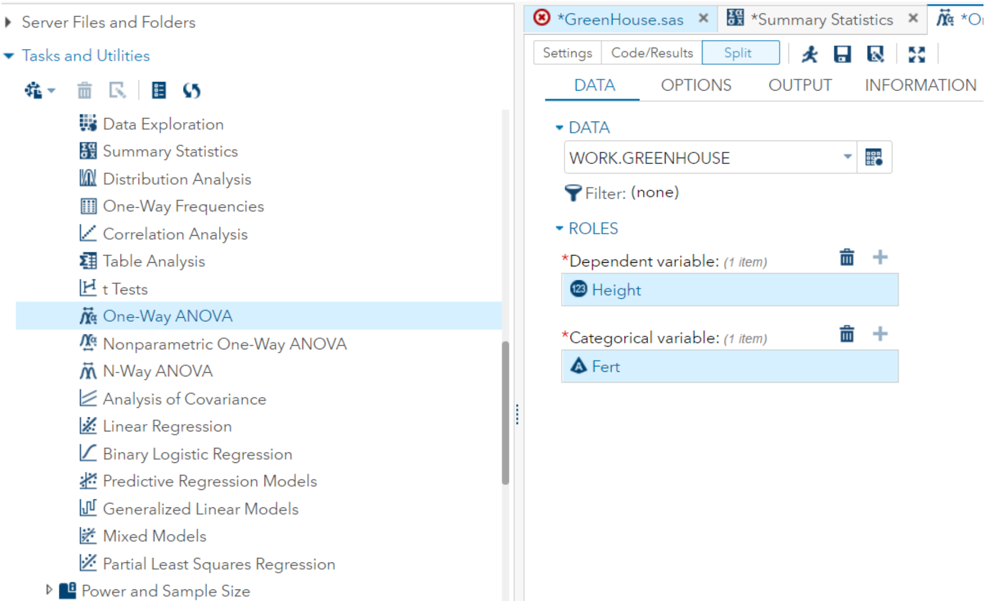
F1 and F3 have highest means; F1 shows more variability (highest variance and IQR). All distributions seem to be symmetric (means and medians are very close to each other).

2 Fitting an ANOVA model

In this lab we learn how to fit model by running the SAS code (that is available on the website - see course materials) and by using SAS Studio interface. In last case under the window Tasks and Utilities, select One-Way ANOVA. Assign Height to dependent variable box and fert to the categorical box.



```
proc glm data=MONIA.GREENHOUSE plots(only)=(boxplot);
    class Fert;
    model Height=Fert;
    means Fert / hovtest=levene welch plots=none;
    lsmeans Fert / plots=(meanplot diffplot);
    run;
quit;
```



DATA

WORK.GREENHOUSE

Filter: (none)

ROLES

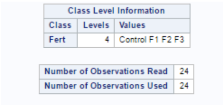
*Dependent variable: (1 item)

Height

*Categorical variable: (1 item)

Fert

2.1 Output - ANOVA model

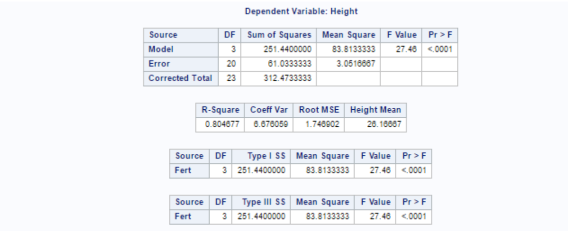


Class Level Information		
Class	Levels	Values
Fert	4	Control F1 F2 F3

Number of Observations Read	24
Number of Observations Used	24

The first set of output includes descriptions of the ANOVA run.

The main output we are interested in is the Type 3 Analysis of Variance.



Dependent Variable: Height

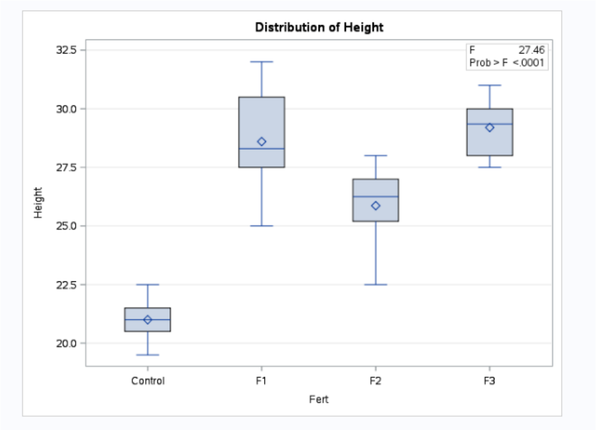
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	251.4400000	83.8133333	27.46	<.0001
Error	20	61.0333333	3.0516667		
Corrected Total	23	312.4733333			

R-Square	Coeff Var	Root MSE	Height Mean
0.804677	6.676059	1.746902	26.16667

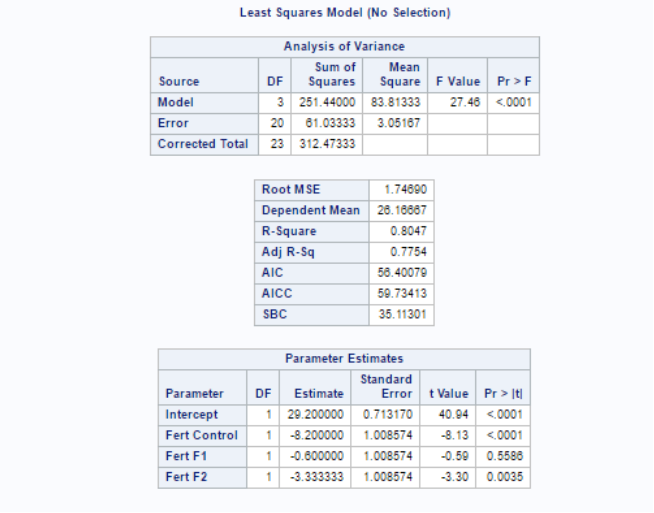
Source	DF	Type I SS	Mean Square	F Value	Pr > F
Fert	3	251.4400000	83.8133333	27.46	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Fert	3	251.4400000	83.8133333	27.46	<.0001

Look at the F-value and the corresponding p-value. Note that the F and p-values are identical to that which we see in the full ANOVA table. Since p-value is less than $\alpha = 0.05$, we reject the null hypothesis. The fertilizers have an effect on the plant height.



2.2 ANOVA as a multiple linear regression model

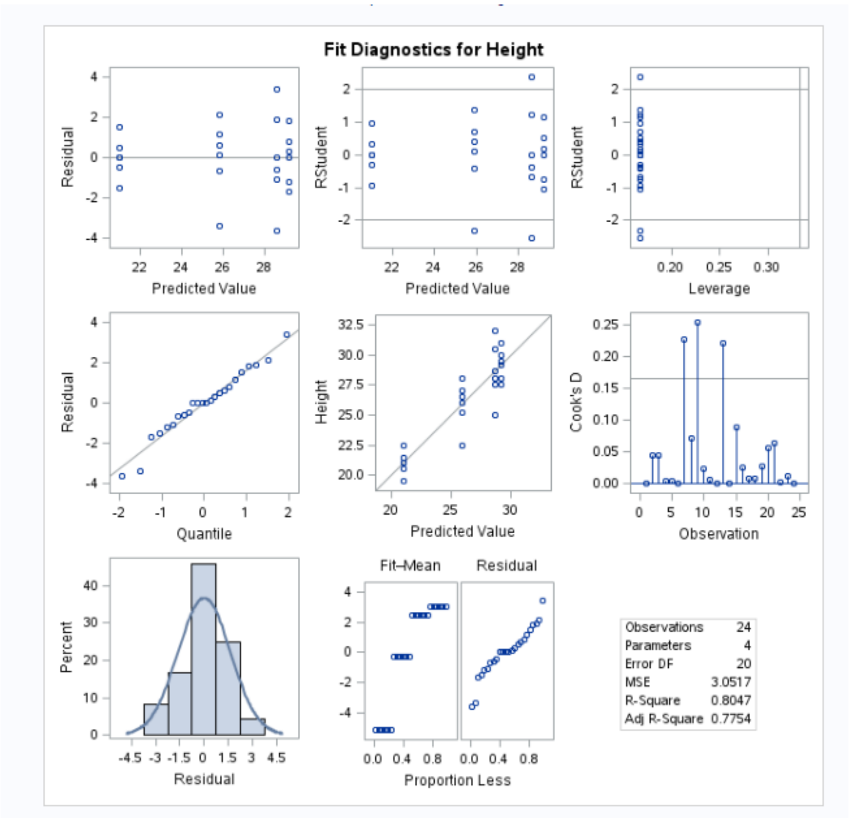

Least Squares Model (No Selection)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	251.44000	83.81333	27.46	<.0001
Error	20	61.03333	3.05167		
Corrected Total	23	312.47333			

Root MSE	1.74690
Dependent Mean	26.16667
R-Square	0.8047
Adj R-Sq	0.7754
AIC	56.40079
AICC	59.73413
SBC	35.11301

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	29.200000	0.713170	40.94	<.0001
Fert Control	1	-8.200000	1.008574	-8.13	<.0001
Fert F1	1	-0.600000	1.008574	-0.59	0.5586
Fert F2	1	-3.333333	1.008574	-3.30	0.0035

Looking at the parameter estimates, it is possible to conclude all predictors are significant apart from Fert F1, that is not significantly different from Fert F3. As regards F-value and the corresponding p-value, we reject the null hypothesis - there is at least one predictor that is significant.
Interpretation: compared to FertF3, F2 leads to a decrease in the plant height by -3.33; compared to FertF3, Fert Control leads to a decrease in the plant height by -8.2.
We can conclude that Fert has an effect on plant height.



Observations	24
Parameters	4
Error DF	20
MSE	3.0517
R-Square	0.8047
Adj R-Square	0.7754

Looking at the residuals all the assumptions are met (normality, linearity, equivariance and independence).