statstutor

community project

encouraging academics to share statistics support resources

stcp-karadimitriouANCOVAR

The following resources are associated:
ANCOVA in R script, ANOVA in R resource, Checking normality in R and the Excel dataset 'Diet.csv'

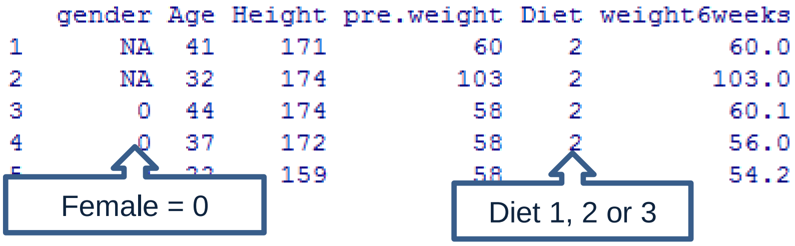
ANCOVA (Analysis of Covariance) in R

Dependent variable: Continuous (scale)

Independent variables: Categorical factors (at least 3 unrelated/ independent groups in each), Scale (continuous) covariates

Common Applications: ANCOVA is similar to traditional ANOVA but is used to detect a difference in means of 3 or more independent groups, whilst controlling for scale covariates. A covariate is not usually part of the main research question but could influence the dependent variable and therefore needs to be controlled for.

Data: The data set Diet.csv contains information on 78 people who undertook one of three diets. There is background information such as age, gender (Female=0, Male=1) and height as well as weight lost on the diet (a positive value means they lost weight). The aim of the study was to see which diet was best for losing weight so the independent variable (group) is diet. To open the file use the `read.csv()` command.



```
   gender Age Height pre.weight Diet weight6weeks
1      NA  41    171         60    2         60.0
2      NA  32    174        103    2        103.0
3       0  44    174         58    2         60.1
4       0  37    172         58    2         56.0
5          22    159         58              54.2
```

Female = 0

Diet 1, 2 or 3

You will need to change the command depending on where you have saved the file.

```
dietR<-read.csv("D:\\diet.csv",header=T,sep=",")
```

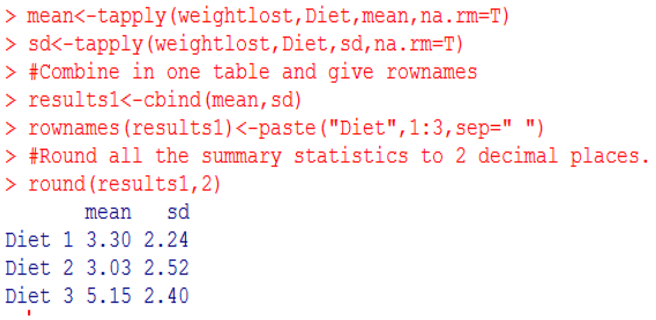
Tell R to use the diet dataset until further notice using `attach(dataset)` so 'Height' can be used instead of dietR$Height. Tell R that 'Diet' is a factor using `as.factor(variable)`.

```
attach(dietR)
Diet<-as.factor(Diet)
```

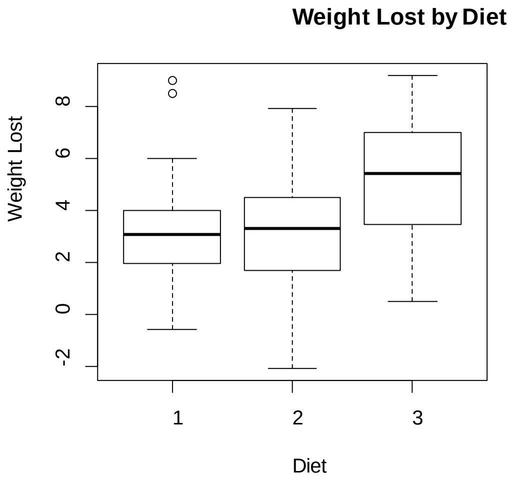
Calculate the weight lost by person (difference in weight before and after the diet) and add the variable to the dataset. Then attach the data again.

```
dietR$weightlost<-pre.weight-weight6weeks
attach(dietR)
```

Before carrying any analysis, summarise weight lost by diet using some summary statistics. Diet 3 seems better than the other diets as the mean weight lost is greater. The standard deviations are similar so weight lost within each group is equally spread out. Similarly as in ANOVA, to calculate means and standard deviations for weight lost by diet using the `tapply(dependent, independent, summary statistic required, na.rm=T)` command e.g. `tapply(weightlost,Diet,mean,na.rm=T)`.
na.rm=T removes rows where missing values exist.



```
> mean<-tapply(weightlost,Diet,mean,na.rm=T)
> sd<-tapply(weightlost,Diet,sd,na.rm=T)
> #Combine in one table and give rownames
> results1<-cbind(mean,sd)
> rownames(results1)<-paste("Diet",1:3,sep=" ")
> #Round all the summary statistics to 2 decimal places.
> round(results1,2)
       mean   sd
Diet 1 3.30 2.24
Diet 2 3.03 2.52
Diet 3 5.15 2.40
```



To produce a boxplot of weight lost by diet:
`boxplot(weightlost~Diet,main='Weight Lost by Diet',xlab='Diet',ylab='Weight Lost')`

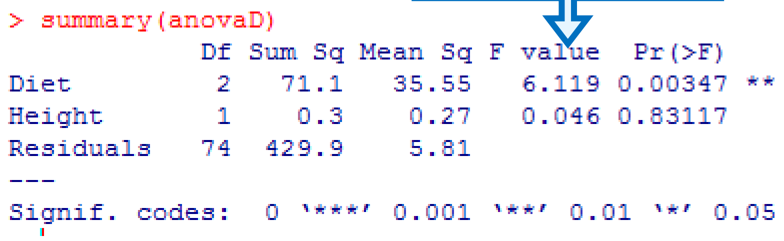
One could suggest, however, that a person's height will have an added influence in the amount of weight they lose on a particular diet. This is where ANCOVA comes in useful. ANCOVA stands for 'Analysis of covariance', and it combines the methods used in ANOVA with linear regression on a number of different levels. The resulting output shows the effect of the independent variable after the effects of the covariates have been removed/ accounted for.

Steps in R and output

To carry out an one way ANCOVA use `aov(dependent~independent(categorical)+indepedent(scale) )` , give the ANOVA model a name e.g. anovaD and use `summary()` to see the output.

```
anovaD<-aov(weightlost~Diet+Height)
summary(anovaD)
```

F = Test statistic
$\frac{MS_{Diet}}{MS_{error}} = \frac{35.55}{5.81} = 6.119$



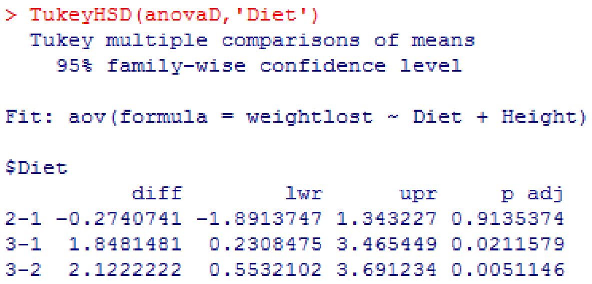
```
> summary(anovaD)
            Df Sum Sq Mean Sq F value  Pr(>F)   
Diet         2   71.1   35.55   6.119 0.00347 **
Height       1    0.3    0.27   0.046 0.83117   
Residuals   74  429.9    5.81                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P = p-value = sig = P(F > 6.197)
p = 0.00347

When writing up the results, it is common to report certain figures from the ANOVA table.
F($df_{between}$, df_{within})= Test Statistic, p = ➔ F(2, 74)= 6.119, p =0.003
There was a significant difference in mean weight lost [F(2,75)=6.197, p = 0.003] between the diets whilst adjusting for height.

Post Hoc Tests

ANOVA tests the null hypothesis 'all group means are the same' so the resulting p-value only concludes whether or not there is a difference between one or more pairs of groups. If the ANOVA is significant, further 'post hoc' tests have to be carried out to confirm where those differences are. The post hoc tests are mostly t-tests with an adjustment to account for the multiple testing. *Tukey's* is the most commonly used post hoc test but check if your discipline uses something else. Use the command `TukeyHSD(anovaD,'factor')`

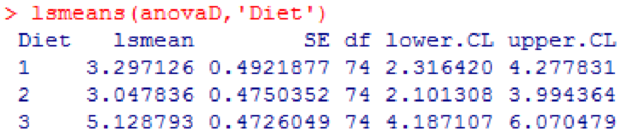


```
> TukeyHSD(anovaD,'Diet')
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = weightlost ~ Diet + Height)

$Diet
          diff        lwr      upr     p adj
2-1 -0.2740741 -1.8913747 1.343227 0.9135374
3-1  1.8481481  0.2308475 3.465449 0.0211579
3-2  2.1222222  0.5532102 3.691234 0.0051146
```

Report each of the three pairwise comparisons e.g. there was a significant difference between diet 3 and diet 1 (p = 0.02). Use the mean difference between each pair e.g. people on diet 3 lost on average 1.85 kg more than those on diet 1 or use individual group means to conclude which diet is best.

Estimated Marginal Means

By using the 'lsmeans' library and therefore the `lsmeans(anovaD,'factor')` command we can derive the estimated marginal means.



```
> lsmeans(anovaD,'Diet')
 Diet   lsmean        SE df lower.CL upper.CL
 1    3.297126 0.4921877 74 2.316420 4.277831
 2    3.047836 0.4750352 74 2.101308 3.994364
 3    5.128793 0.4726049 74 4.187107 6.070479
```

The estimated marginal means output gives the adjusted means (controlling for the covariate 'Height') for each diet group. This simply means that the effect of 'Height' has been statistically removed. From these adjusted means, it is clear that Diet 3 lost the most weight after adjusting for height.

Checking the assumptions for one-way ANOVA

Assumptions	How to check	What to do if the assumptions is not met
Covariates should not be highly correlated (if using more than 1)	Check correlation before performing analysis. Use cor(dietR) and check that none of the covariates have high correlation values (r>0.8)	If there are some highly correlated covariates, one must select which covariates are of most importance and use those in the model.
Residuals should be normally distributed	Use histogram, QQ plots and normality tests as diagnostic tools (see the Checking **normality in R resource** for more details)	If the residuals are very skewed, the results of the ANCOVA are less reliable so a possible transformation in the dependent may fix the problem.

Homogeneity (equality) of variance: The variances should be similar for all groups	Use the Levene's test of equality of variances through the package car
	`library(car)`
	`leveneTest(weightlost~Diet)`
	If p - value > 0.05, equal variances can be assumed and the ANOVA results are valid
		If the residuals are very skewed, the results of the ANCOVA are less reliable. One possibility it to transform the data (speak to a statistics tutor for help with this).

Checking the assumptions for this data

Ask for the standardised residuals (difference between each individual and their group mean) and give them a name (res).

```
res<-anovaD$residuals
```

Produce a histogram of the residuals.

```
hist(res, main="Histogram of standardised residuals",xlab="Standardised
residuals")
```

The Levene's test for equality of variances is in the additional ‘car’ package.

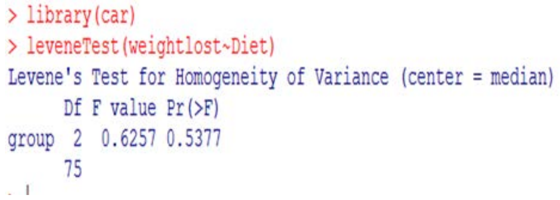
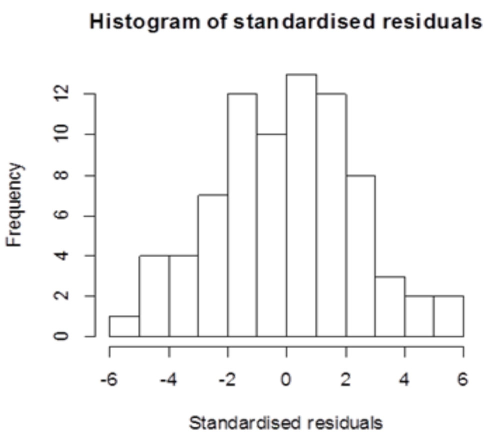
```
library(car)
```

If this command does not work, you will need to go to the Packages --> Install package(s) and select the UK (London)CRAN mirror. Then look for the package 'car' and click. A lot of extra menus will download as well so you shouldn’t need to do this again. Then try `library(car)` again.

Once loaded, carry out Levene's test as running a simple ANOVA.

```
leveneTest(weightlost~Diet)
```

Homogeneity Assumption	**Normality Assumption**


`> library(car)`
`> leveneTest(weightlost~Diet)`
`Levene's Test for Homogeneity of Variance (center = median)`
`      Df F value Pr(>F)`
`group  2  0.6257 0.5377`
`      75`
As p - value (0.5377) > 0.05, equal variances can be assumed
	
	The residuals are normally distributed

Reporting ANOVA

A one-way ANCOVA was conducted to compare the effectiveness of three diets whilst controlling for Height. Normality checks and Levene’s test were carried out and the assumptions were met.

There was a significant difference in mean weight lost [F(2,74)=6.119, p = 0.003] between the diets. Post hoc comparisons using the Tukey test were carried out. There was a significant difference between diets 1 and 3 (p = 0.02) with people on diet 3 lost on average 1.85 kg more than those on diet 3. There was also a significant difference between diets 2 and 3 difference (p = 0.005) with people on diet 3 lost on average 2.12 kg more than those on diet 2. Comparing the estimated marginal means showed that the most weight was lost on Diet 3 (mean=5.13kg) compared to Diets 1 and 2 (mean=3.30kg, 3.05kg respectively).