

Lecture 8: Analysis of variance 2 (and regression)

Last time we saw the implications of degrees of freedom: We have little balls of errors, and each estimation of a mean reduces the dimensionality of the ball by 1, making an embedded ball of lower dimensions. For a X^2 distribution all that matters is the dimensionality of the ball, and it does not matter in what space it is embedded.

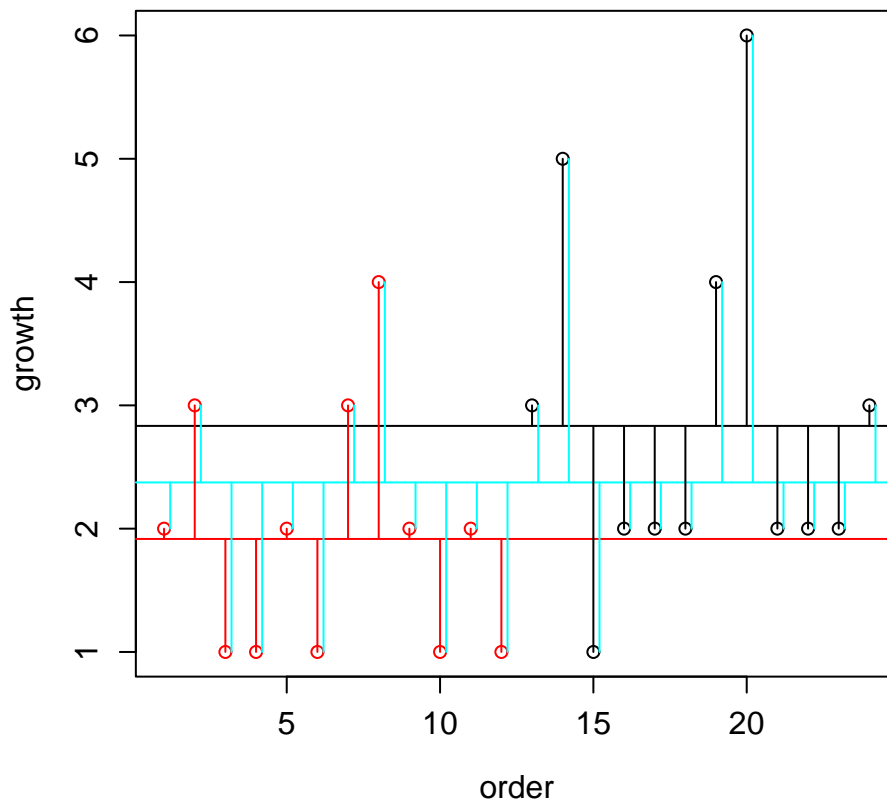
But all this is just good to know, and not actually required. It is good to understand the basics of degrees of freedom, though.

Simple analysis of variance

```

> setwd("~/R-course-2006/lecture8")
> oneway=read.table("oneway.txt",header=T)
> m.short=mean(oneway$Growth[ oneway$Photo=="Short" ] )
> m.long=mean(oneway$Growth[ oneway$Photo=="Long" ] )
> plot(1:24,oneway$Growth,ylab="growth",xlab="order",
      col=as.numeric(oneway$Photo))
> abline(mean(oneway$Growth),0,col="cyan")
  abline(m.short,0,col=2);abline(m.long,0,col=1)
> for(i in 1:24)
  lines(c(i+0.2,i+0.2),c(mean(oneway$Growth),oneway$Growth[i]),col="cyan")
> for(i in which( oneway$Photo=="Short" ) ) {
  lines(c(i,i),c(m.short,oneway$Growth[i]),col=2)
}
++
> for(i in which( oneway$Photo=="Long" ) ) {
  lines(c(i,i),c(m.long,oneway$Growth[i]),col=1)
}
>
> v(width=5,height=5)

```



```

> m=mean(oneway$Growth)
> sum( (oneway$Growth-m)^2 )
[1] 39.625
> sum( (oneway$Growth[ oneway$Photo=="Short" ] - m.short) ^2 ) + sum(
(oneway$Growth[ oneway$Photo=="Long" ] - m.long)^2 )
[1] 34.58333
> 39.625-34.58333
[1] 5.04167
>

```

Fitting 2 means gives a smaller sum squared errors than one mean. But that is expected. The reduction in the second model is 5.04167

What we want to know if it is significantly better.

So, the null hypothesis is that the data all comes from the same distribution, and we ask, if that is true, how often will we get such a big difference?

This is tested by dividing the improvement by the total errors, and taking into account their degrees of freedom.

We can then look up what the chance for the ration to be 5.04167/34.58333 when we have 1 and 22 degrees of freedom.

This is the F-distribution:

$$\frac{\frac{\sum \text{ squared errors reduction in model 2}}{\text{degrees of freedom reduction in model 2}}}{\frac{\sum \text{ squared errors remaining in model 2}}{\text{degrees of freedom remaining in model 2}}}$$

Which in our case would be

```

> (5.04167/1)/(34.583/22)
[1] 3.207262
> pf(3.207262,1,22)
[1] 0.9129172
> 1-pf(3.207262,1,22)
[1] 0.08708285

```

>

This is therefore not a significant improvement.

```
> l=lm(oneway$Growth ~ oneway$Photo )
```

```
> anova(l)
```

Analysis of Variance Table

Response: oneway\$Growth

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
oneway\$Photo	1	5.042	5.042	3.2072	0.08708
Residuals	22	34.583	1.572		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>

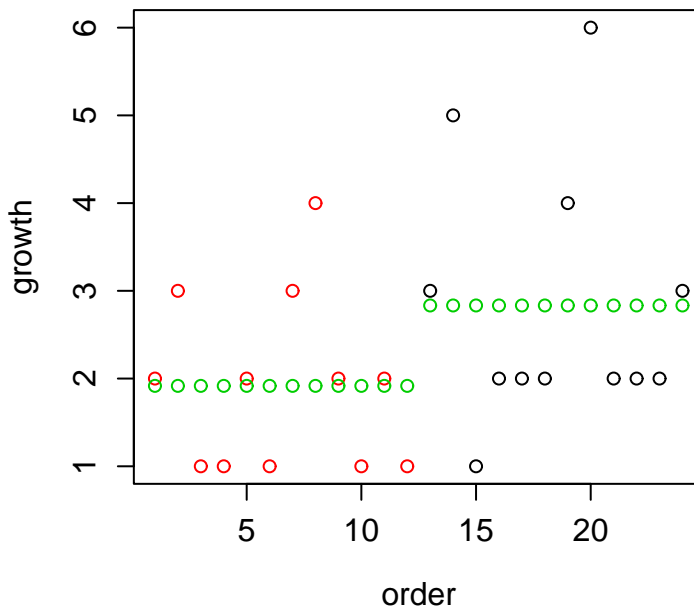
Expressions in R

```
> x ~ y
```

represents a model where x is a function of y.

```
> plot(1:24, oneway$Growth, ylab="growth", xlab="order",  
col=as.numeric(oneway$Photo))
```

```
> points(1:24, fitted(l), col=3); v()
```



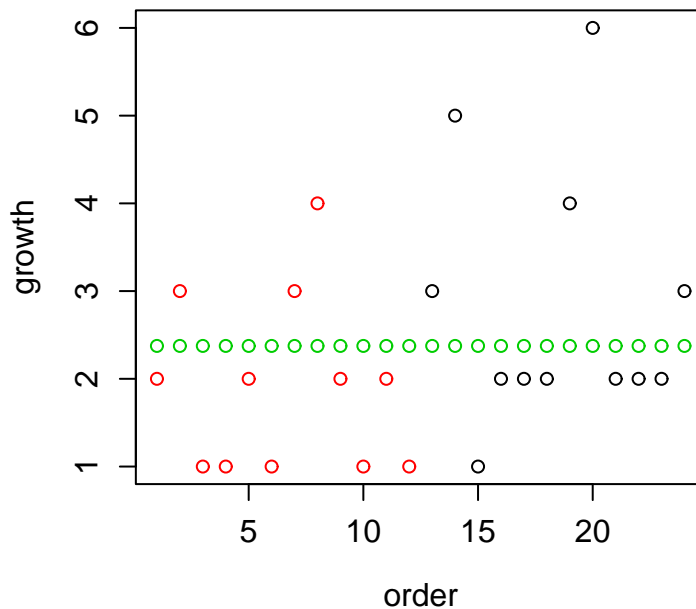
>

Another model is

> x ~ 1

Which means that x is not a function of anything, the model just fits a general mean value

```
> l.simple=lm( oneway$Growth ~ 1 )
> plot(1:24,oneway$Growth,ylab="growth",xlab="order",
      col=as.numeric(oneway$Photo))
> points( 1:24, fitted(l.simple), col=3)
> v()
```



```
> anova(l.simple,1)
```

Analysis of Variance Table

Model 1: oneway\$Growth ~ 1

Model 2: oneway\$Growth ~ oneway\$Photo

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	23	39.625				
2	22	34.583	1	5.042	3.2072	0.08708 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>

This is a way to compare two models.

Another example, but more complicated:

Two way anova

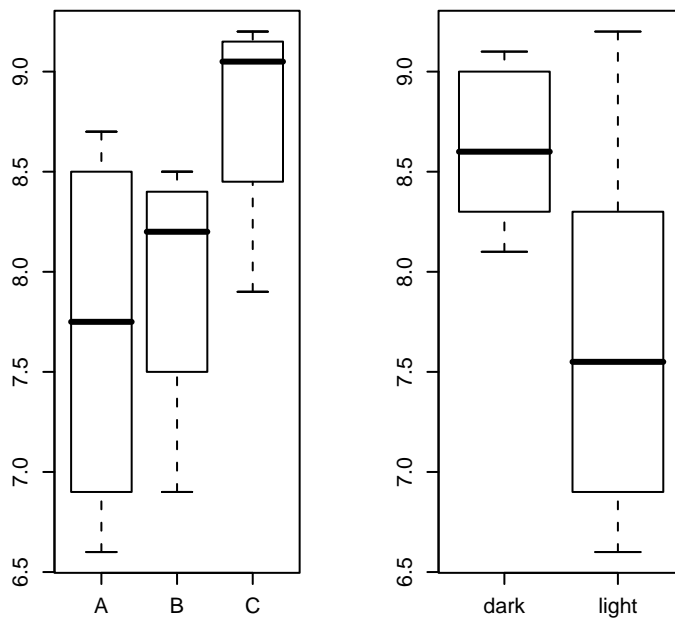
```
> f=read.table("~/R-course-2006/lecture8/factorial.txt",head=T,sep="\t")  
> f
```

	growth	diet	coat
1	6.6	A	light
2	7.2	A	light
3	6.9	B	light
4	8.3	B	light
5	7.9	C	light
6	9.2	C	light
7	8.3	A	dark
8	8.7	A	dark
9	8.1	B	dark
10	8.5	B	dark
11	9.1	C	dark
12	9.0	C	dark

```
>
```

We now have another explanatory variable, and a finer mesh of the photoperiod.

```
> layout(matrix(1:2,1,2));par(cex=0.7)  
> plot( f$diet, f$growth );plot( f$coat, f$growth);v()
```



```
> layout(1)
>
```

Let us fit a full model to this:

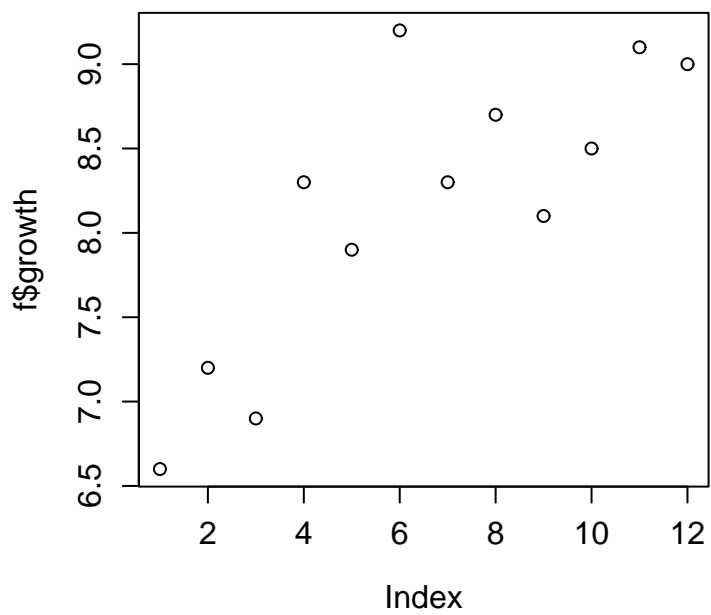
```
> model1=lm( growth ~ diet + coat + diet:coat, data=f)
> model1=lm( f$growth ~ f$diet + f$coat + f$diet:coat)
```

Specifying `data=f` is a shortcut so we don't have to write `f$` every time.

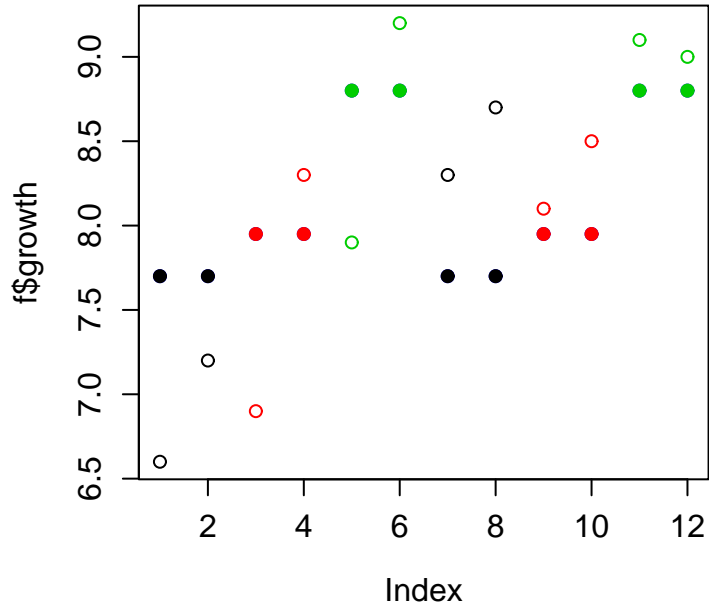
The expression `growth ~ diet + coat + diet:coat`

specifies that growth is modeled as a function of diet, coat, and the interaction of diet and coat. Let us see that:

```
> plot(f$growth);v()
```



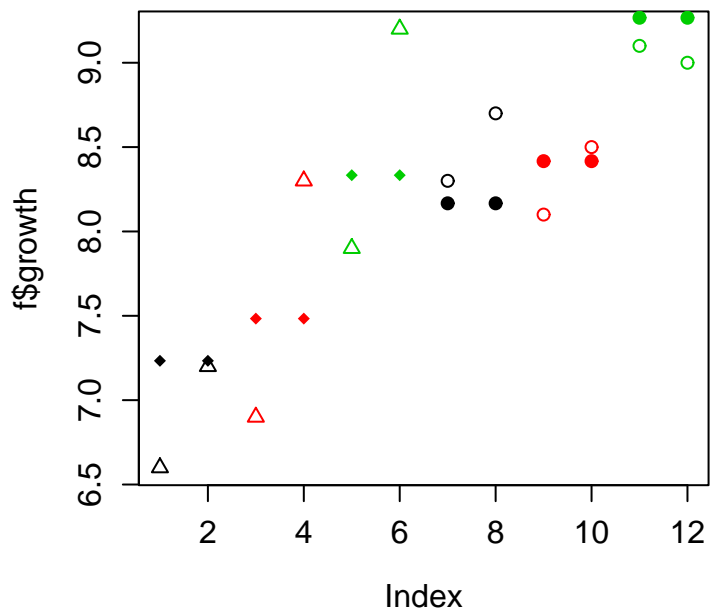
```
> l1=lm( growth ~ diet, data=f)
> plot(f$growth,col=as.numeric(f$diet))
> points(fitted(l1),pch=19,col=as.numeric(f$diet));v()
```



>

3 different means were fit.

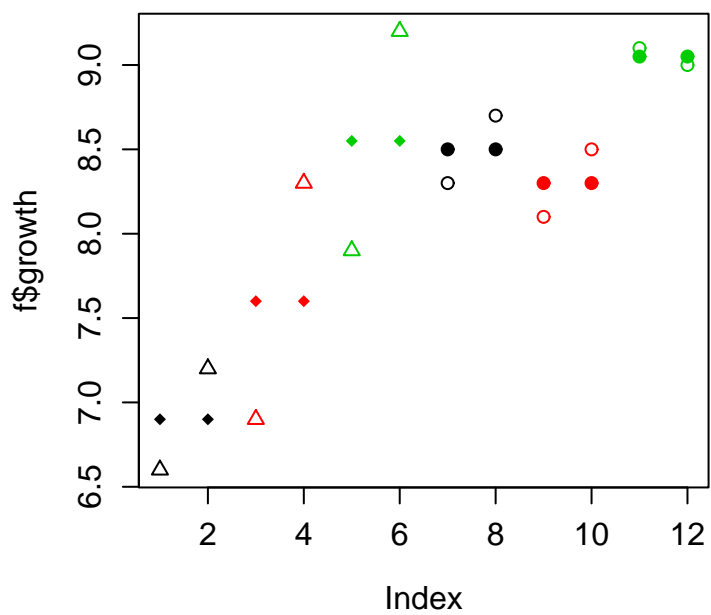
```
> l2=lm( growth ~ diet + coat, data=f)
> plot(f$growth,col=as.numeric(f$diet),pch=as.numeric(f$coat))
> points(fitted(l2),pch=20-as.numeric(f$coat),col=as.numeric(f$diet));v()
```



>

Now 3 means were fit for diet, and an additional step for coat.

```
> l3=lm( growth ~ diet + coat + diet:coat, data=f)
> plot(f$growth,col=as.numeric(f$diet),pch=as.numeric(f$coat))
> points(fitted(l3),pch=20-as.numeric(f$coat),col=as.numeric(f$diet));v()
```



>

Now we fit 6 (2*3) different means.

```
> model1=lm( growth ~ diet + coat + diet:coat, data=f)
```

```
> anova(model1)
```

Analysis of Variance Table

Response: growth

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
diet	2	2.66000	1.33000	3.6774	0.09069 .
coat	1	2.61333	2.61333	7.2258	0.03614 *
diet:coat	2	0.68667	0.34333	0.9493	0.43833
Residuals	6	2.17000	0.36167		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>

diet removes 2 df beyond the simple1-mean model. coat removes another 1, and with the full interaction, we get 6 means, so we remove another 2.

```
> model2=lm( growth ~ diet + coat, data=f)
```

```
> anova(model1,model2)
```

Analysis of Variance Table

Model 1: growth ~ diet + coat + diet:coat

Model 2: growth ~ diet + coat

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	6	2.17000				
2	8	2.85667	-2	-0.68667	0.9493	0.4383

>

```
> anova(model2)
```

Analysis of Variance Table

Response: growth

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
diet	2	2.66000	1.33000	3.7246	0.07190 .
coat	1	2.61333	2.61333	7.3186	0.02685 *
Residuals	8	2.85667	0.35708		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> model3=lm(growth~coat,data=f)
> anova(model2,model3)
```

Analysis of Variance Table

Model 1: growth ~ diet + coat

Model 2: growth ~ coat

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	8	2.8567				
2	10	5.5167	-2	-2.6600	3.7246	0.0719 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> anova(model3)
```

Analysis of Variance Table

Response: growth

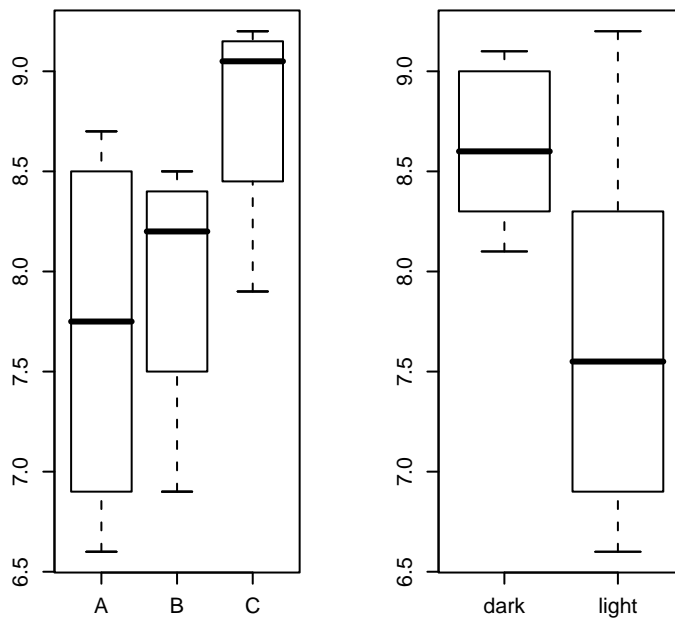
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
coat	1	2.6133	2.6133	4.7372	0.05457 .
Residuals	10	5.5167	0.5517		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
>
```

What happened? Nothing is left!

```
> layout(matrix(1:2,1,2));par(cex=0.7)
> plot( f$diet, f$growth );plot( f$coat, f$growth);v()
```



```
> layout(1)
>
```

It looks as if only C is different from A and B.

```
> diet2=as.character(f$diet)
> diet2

 [1] "A" "A" "B" "B" "C" "C" "A" "A" "B" "B" "C" "C"
> diet2[ diet2=="A" | diet2=="B" ] = "AB"
> diet2

 [1] "AB" "AB" "AB" "AB" "C" "C" "AB" "AB" "AB" "AB" "C" "C"
> diet2=factor(diet2)
> diet2

 [1] AB AB AB AB C C AB AB AB AB C C
Levels: AB C
> model4=lm( growth~coat+diet2, data=f)
> anova(model4)

Analysis of Variance Table

Response: growth
          Df Sum Sq Mean Sq F value Pr(>F)
coat       1  2.6133   2.6133   7.8882 0.02042 *
diet2      1  2.5350   2.5350   7.6518 0.02189 *
Residuals  9  2.9817   0.3313
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model5=lm( growth~coat+diet2+coat:diet2, data=f)
> anova(model4,model5)

Analysis of Variance Table

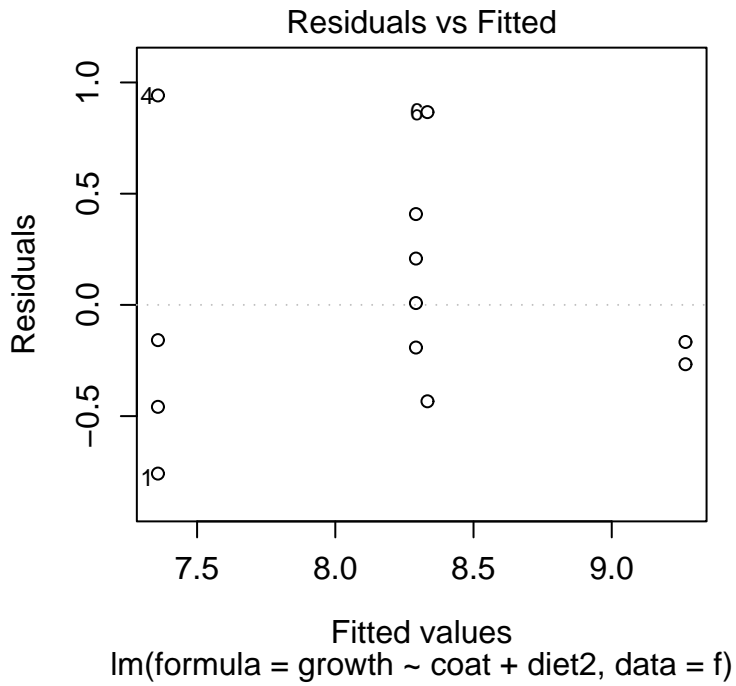
Model 1: growth ~ coat + diet2
Model 2: growth ~ coat + diet2 + coat:diet2
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1       9 2.98167
2       8 2.70000  1  0.28167 0.8346 0.3877

>

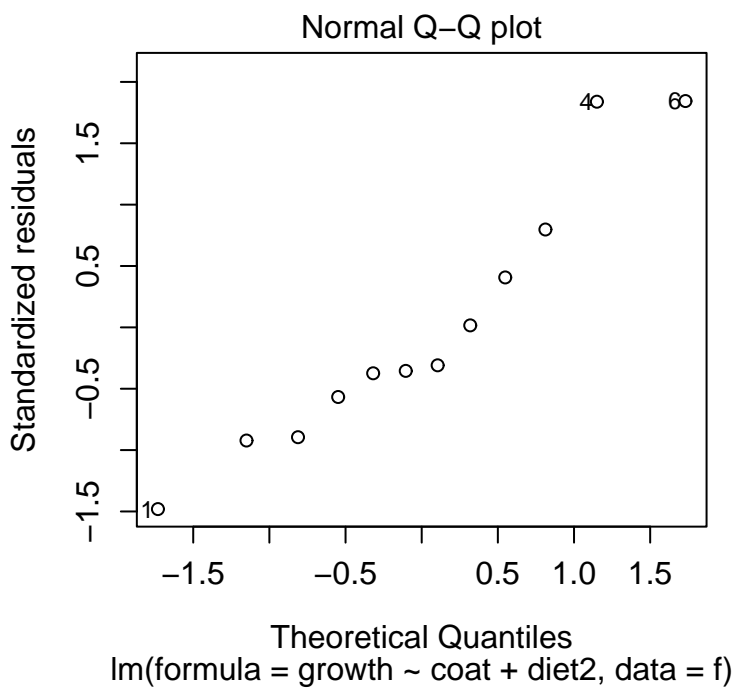
> class(model4)
```

```
[1] "lm"
```

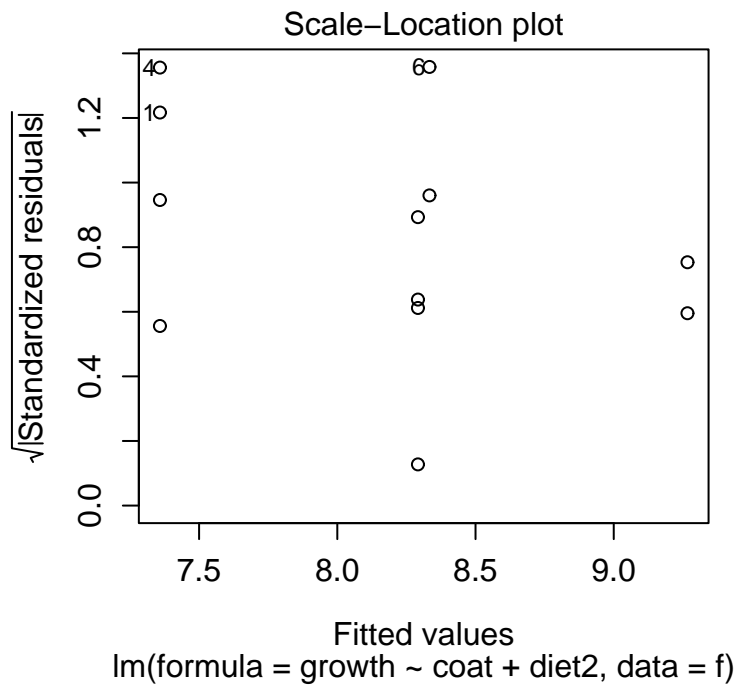
```
> plot(model4,which=1);v()
```



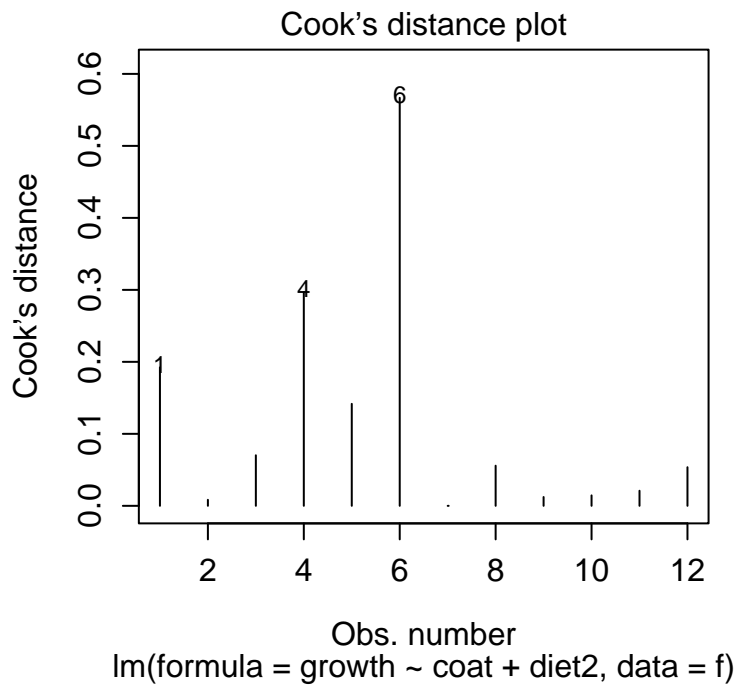
```
> plot(model4,which=2);v()
```



```
> plot(model4,which=3);v()
```

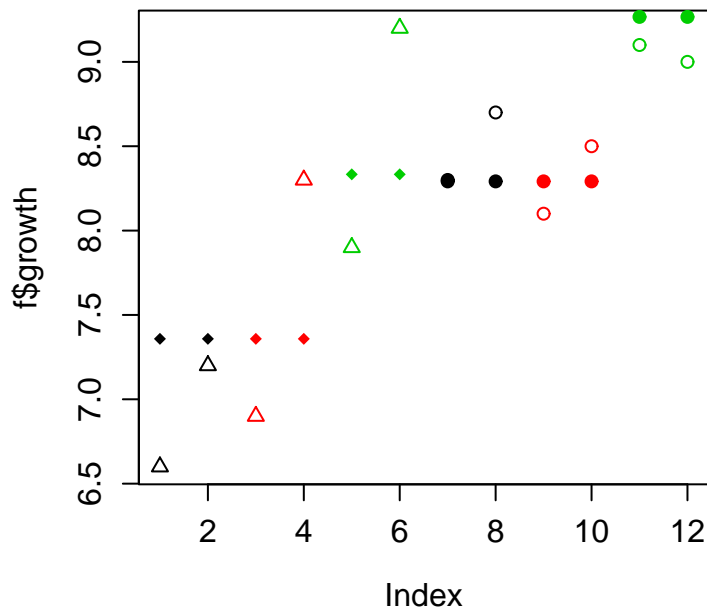


```
> plot(model4,which=4);v()
```



>

```
> plot(f$growth,col=as.numeric(f$diet),pch=as.numeric(f$coat))
> points(fitted(model4),pch=20-as.numeric(f$coat),col=as.numeric(f$diet));v()
```



>

Regression

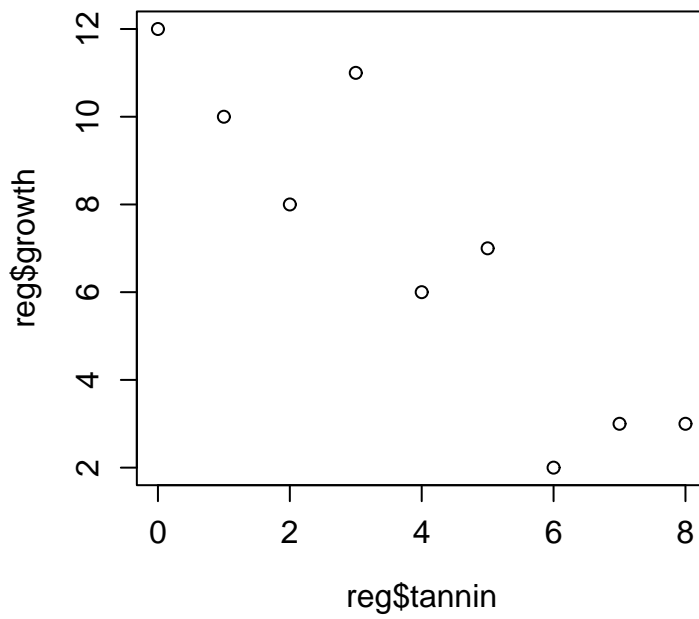
When we do not deal with factors, but with a continuous variable, we can do a regression, or an analysis of covariance.

```
> reg=read.table("~/R-course-2006/lecture8/regression.txt",sep="\t",head=T)
```

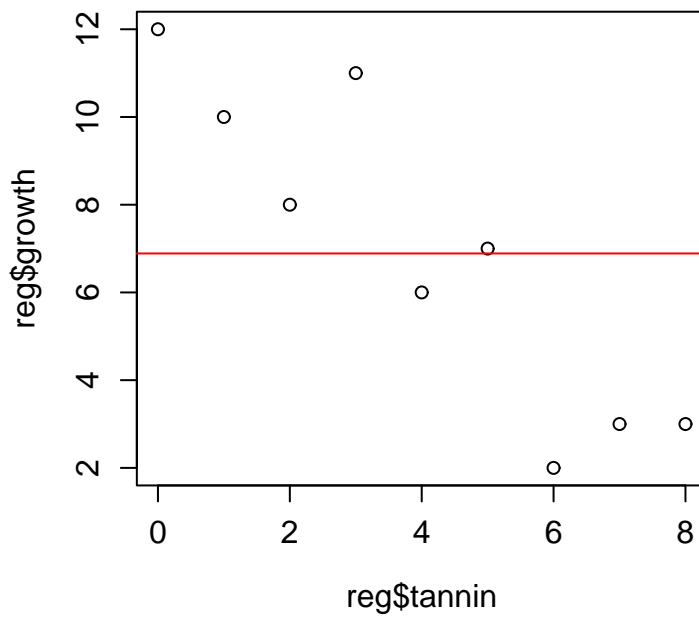
```
> reg
```

	growth	tannin
1	12	0
2	10	1
3	8	2
4	11	3
5	6	4
6	7	5
7	2	6
8	3	7
9	3	8

```
> plot( reg$tannin, reg$growth);v()
```

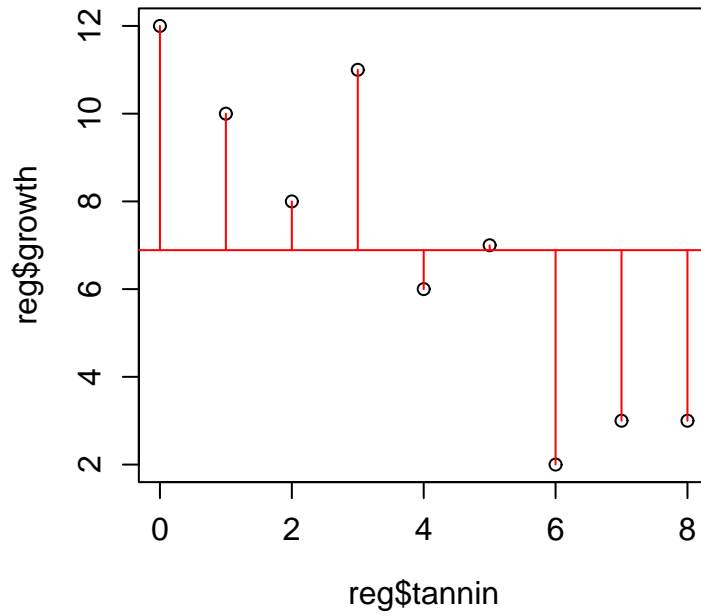


```
> abline(mean( reg$growth),0,col=2);v()
```

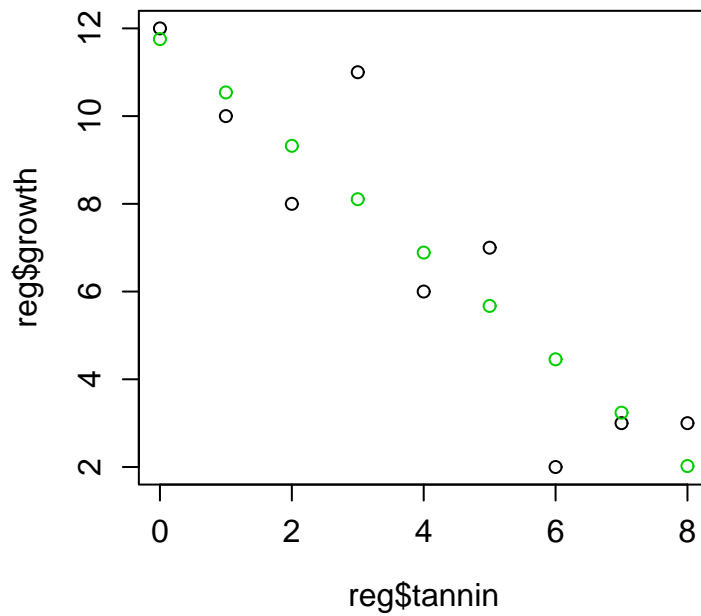


```
> for(i in 1:9)
  lines( c(reg$tannin[i],reg$tannin[i]),
        c(reg$growth[i],mean(reg$growth)),col=2)
+
```

```
> v()
```



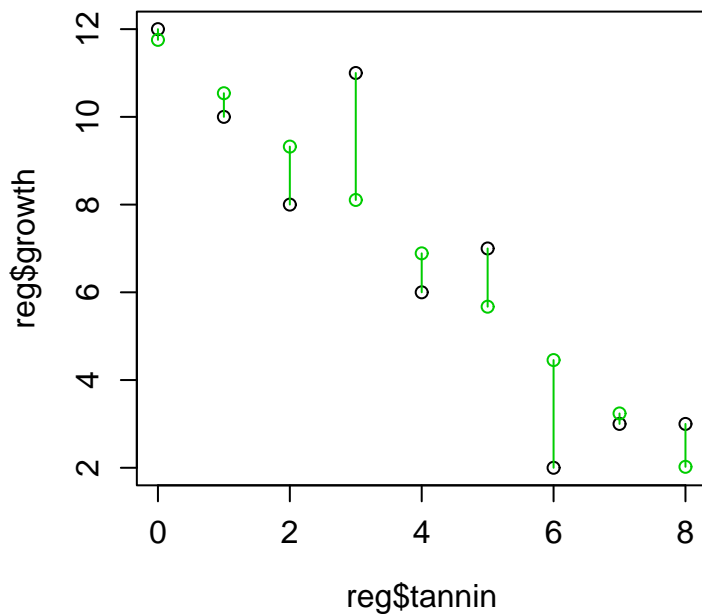
```
> linear1=lm( growth~tannin , data=reg)  
> plot( reg$tannin, reg$growth)  
> points( reg$tannin, fitted(linear1), col=3);v()
```



```

> for(i in 1:9)
  lines( c(reg$tannin[i],reg$tannin[i]),
        c(reg$growth[i],fitted(linear)[i]),col=3)
>
> v()

```



```

> anova(linear)

```

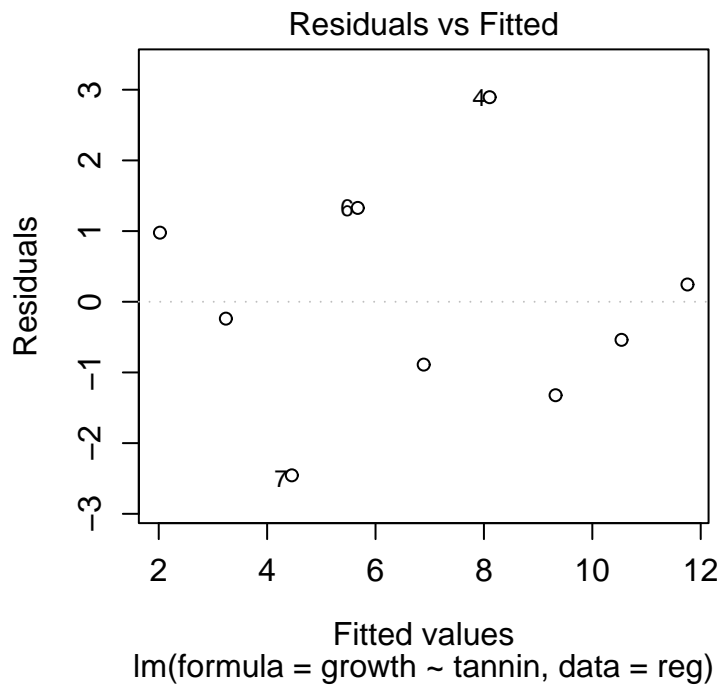
Analysis of Variance Table

Response: growth

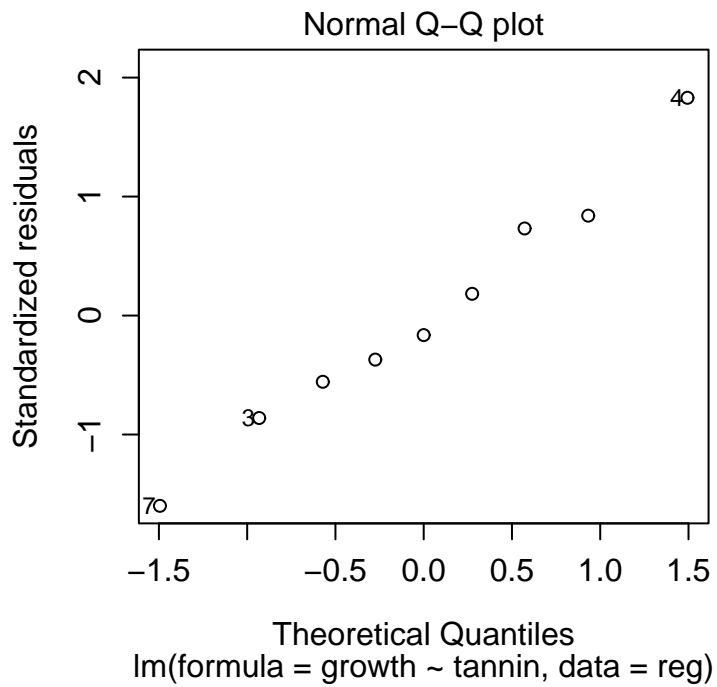
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tannin	1	88.817	88.817	30.974	0.000846 ***
Residuals	7	20.072	2.867		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

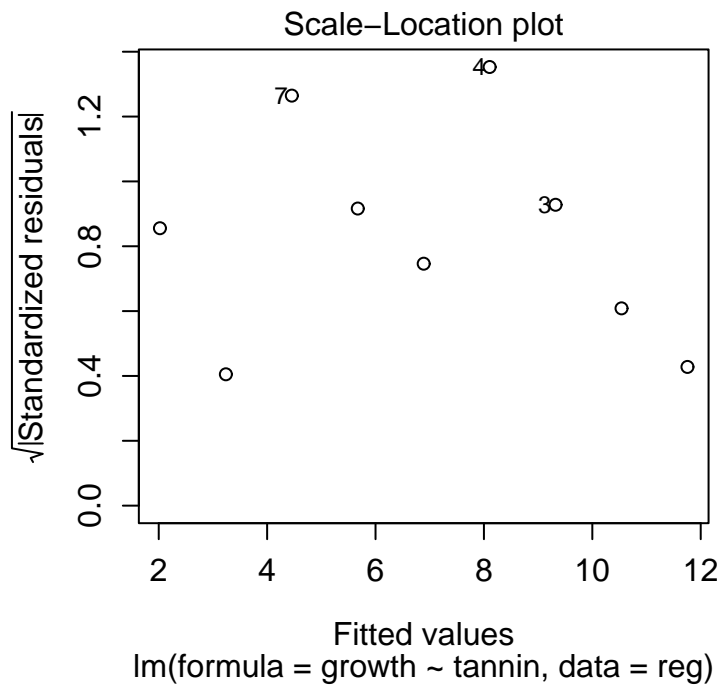
```
> plot(linear,which=1);v()
```



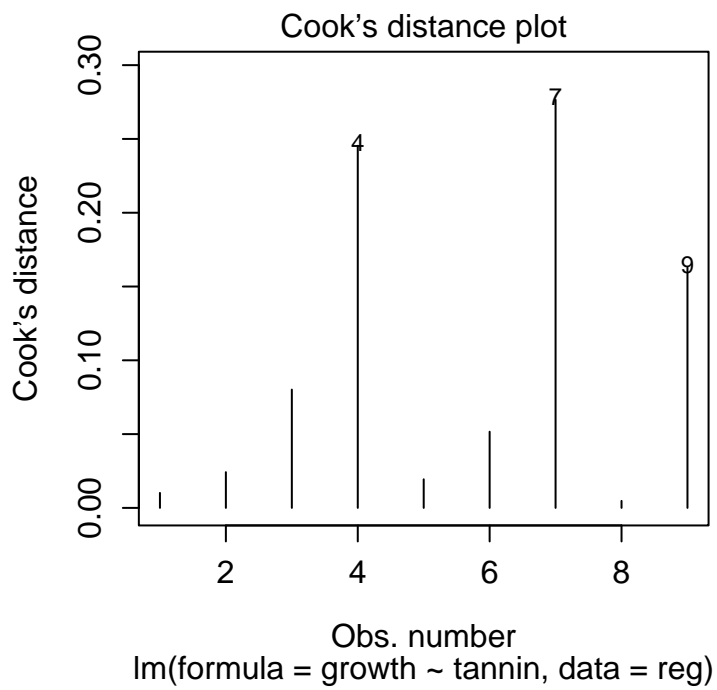
```
> plot(linear,which=2);v()
```



```
> plot(linear,which=3);v()
```



```
> plot(linear, which=4); v()
```



```
>
```

