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$Growth[i]),col=1)
> }
> 
> v(width=5,height=5)
> m-mean(one-way$Growth)
> sum( (one-way$Growth-m)^2 )

[1] 39.625

> sum((one-way$Growth[ one-way$Photo=="Short"] - m.short)^2 ) + sum( (one-way$Growth[ one-way$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:

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

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.

```r
> lm(owenay$Growth ~ owenay$Photo )
> anova(l)

Analysis of Variance Table

Response: owenay$Growth

    Df  Sum Sq Mean Sq F value Pr(>F)
owenay$Photo  1   5.042  5.042 3.207 22 0.08708.
Residuals    22 34.583  1.572
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Expressions in R

```r
> x ~ y
```

represents a model where x is a function of y.

```r
> plot(1:24,owenay$Growth,ylab="growth",xlab="order", col=as.numeric(owenay$Photo))
> points(1:24,fitted(l),col=3);v()
```
Another model is

```r
> x ~ 1
```

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

```r
> 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()
```

![Graph](image)

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

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**

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

<table>
<thead>
<tr>
<th></th>
<th>growth</th>
<th>diet</th>
<th>coat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.6</td>
<td>A</td>
<td>light</td>
</tr>
<tr>
<td>2</td>
<td>7.2</td>
<td>A</td>
<td>light</td>
</tr>
<tr>
<td>3</td>
<td>6.9</td>
<td>B</td>
<td>light</td>
</tr>
<tr>
<td>4</td>
<td>8.3</td>
<td>B</td>
<td>light</td>
</tr>
<tr>
<td>5</td>
<td>7.9</td>
<td>C</td>
<td>light</td>
</tr>
<tr>
<td>6</td>
<td>9.2</td>
<td>C</td>
<td>light</td>
</tr>
<tr>
<td>7</td>
<td>8.3</td>
<td>A</td>
<td>dark</td>
</tr>
<tr>
<td>8</td>
<td>8.7</td>
<td>A</td>
<td>dark</td>
</tr>
<tr>
<td>9</td>
<td>8.1</td>
<td>B</td>
<td>dark</td>
</tr>
<tr>
<td>10</td>
<td>8.5</td>
<td>B</td>
<td>dark</td>
</tr>
<tr>
<td>11</td>
<td>9.1</td>
<td>C</td>
<td>dark</td>
</tr>
<tr>
<td>12</td>
<td>9.0</td>
<td>C</td>
<td>dark</td>
</tr>
</tbody>
</table>
```

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

```r
> layout(matrix(1:2,1,2)); par(cex=0.7)
> plot(f$dist, f$growth); plot(f$coat, f$growth); v()
```
Let us fit a full model to this:

```r
> modell1 <- lm( growth ~ diet + coat + diet:coat, data=f)
> modell2 <- 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:

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

> 3 different means were fit.

> 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.

```r
> 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.

```r
> modell <- lm( growth ~ diet + coat + diet:coat, data=f )
> anova(modell)

Analysis of Variance Table

Response: growth

        Df Sum Sq Mean Sq F value Pr(>F)
diet     2 2.6600 1.33000 3.6774 0.09069
coat     1 2.6133 2.61333 7.2258 0.03614 *
 diet:coat 0 0.6866 0.34333 0.9493 0.43833
Residuals 6 2.1700 0.36167
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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

```r
> 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

>
```

```r
> anova(model2)

Analysis of Variance Table

Response: growth

        Df Sum Sq Mean Sq F value Pr(>F)
diet     2 2.6600 1.33000 3.7246 0.07190
coat     1 2.6133 2.61333 7.3186 0.02685 *
Residuals 8 2.8566 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

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df Sum of Sq</th>
<th>F Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>2.8567</td>
<td>4.7372</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>5.5167 -2</td>
<td>-2.6600</td>
</tr>
</tbody>
</table>

---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(model3)

Analysis of Variance Table

Response: growth

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>coat</td>
<td>1</td>
<td>2.6133</td>
<td>2.6133</td>
<td>4.7372</td>
</tr>
<tr>
<td>Residuals</td>
<td>10</td>
<td>5.5167</td>
<td>0.5517</td>
<td>0.0719</td>
</tr>
</tbody>
</table>

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

```r
> model1 <- lm(growth ~ diet2, data = f)
> anova(model1)
```

```
Analysis of Variance Table

Response: growth
df Sum Sq Mean Sq  F value Pr(>F)
diet2 1 2.6133 2.6133  7.8882 0.02042 *
Residuals 9 2.9817 0.3313
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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

\begin{itemize}
\item Residuals vs Fitted
\item Fitted values
\item Normal Q–Q plot
\end{itemize}
> plot(model4, which=3); v()

![Scale-Location plot](image)

\[ \text{lm(formula = growth ~ coat + diet2, data = f)} \]

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

![Cook's distance plot](image)

\[ \text{lm(formula = growth ~ coat + diet2, data = f)} \]
Regression

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

```r
> 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$tannin, fitted(linear1), col=3);v()
```r
> for(i in 1:9)
>   lines( c(reg$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)
reg$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()

Residuals vs Fitted

Residuals

Fitted values

lm(formula = growth ~ tannin, data = reg)

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

Normal Q–Q plot

Standardized residuals

Theoretical Quantiles

lm(formula = growth ~ tannin, data = reg)

> plot(linear, which=3); v()
\texttt{lm(formula = growth \sim tannin, data = reg)}

\texttt{plot(linear, which=4); v()}

\texttt{lm(formula = growth \sim tannin, data = reg)}