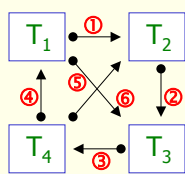


# LINEAR MODELS

- ➔ Log Ratio Models
- ➔ ANOVA Models
- ➔ Mixed-Effects Models
- ➔ Different Types of Replication
  - Determining appropriate error terms
- ➔ Examples

## Linear Models for Log Ratios

(Yang and Speed, 2003)



$$\log \frac{R}{G} \begin{cases} \theta_1 = \alpha_2 - \alpha_1 \\ \theta_2 = \alpha_3 - \alpha_1 \\ \theta_3 = \alpha_4 - \alpha_1 \end{cases} \quad \left\{ \begin{array}{l} \textcircled{1} \alpha_2 - \alpha_1 \\ \textcircled{2} \alpha_3 - \alpha_2 \\ \textcircled{3} \alpha_4 - \alpha_3 \\ \textcircled{4} \alpha_1 - \alpha_4 \\ \textcircled{5} \alpha_3 - \alpha_1 \\ \textcircled{6} \alpha_2 - \alpha_4 \end{array} \right. \quad \begin{array}{l} \alpha_i: \text{intensity (log scale)} \\ \text{on treatment } i. \\ \varepsilon_i \sim N(0, \sigma^2) \end{array}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix}$$

$$\hat{\theta} = (X'X)^{-1}X'y$$

$$\text{Var}(\hat{\theta}) = (X'X)^{-1}\sigma^2$$

$$\hat{\theta}_1 = 2y_1 - y_2 - y_4 + y_5 + y_6$$

## ANOVA Models for Microarray Experiments

⇒ Sources of variation (fixed and random effects) to be considered:

- 1 Dye
- 2 Slide (array)
- 3 Patch or print-tip within slide
- 4 Spot within patch
- 5 Print batch of slides
- 6 Biological variability (individuals or pools)
- 7 Gene
- 8 Treatments (experimental groups)
- 9 Interactions between factors
- 10 Etc.

## ANOVA Models for Microarray Experiments

(Kerr and Churchill, 2001)

$$\log(y_{ijklm}) = \mu + A_i + D_j + AD_{ij} + G_l + TG_{kl} + AG_{ilm} + DG_{jl} + \varepsilon_{ijklm}$$

Array

Dye

Interaction  
Array × Dye

Gene

Effects of interest!  
How trts change  
expression of each gene

Spot

Gene-specific  
Dye effect

$\varepsilon_{ijklm} \sim N(0, \sigma^2)$

# ANOVA Models for Microarray Experiments

(Kerr and Churchill, 2001)

## ⇒ Comments:

- ❶ The terms  $AD_{ij}$  indirectly account for overall treatment effects
- ❷  $A_i$ ,  $D_j$ , and  $AD_{ij}$  refer to 'global data normalization'
- ❸  $G_i$ ,  $AG_{ilm}$ , and  $DG_{ij}$  refer to normalization at the level of individual genes
- ❹ Kerr and Churchill consider a fixed-effect model
- ❺ The variances are assumed constant, i.e.  $\text{Var}[\epsilon_{ijklm}] = \sigma^2$
- ❻ Log transformation
- ❼ They consider spots as the experimental units

## EXAMPLE

Kerr et al. (2001)

- ⇒ NZO/HILt male mice (model of maturity-onset type II diabetes)
- ⇒ Gene expression in livers
- ⇒ Diets supplemented (or not) with 0.001% CL316,243
- ⇒ Two arrays, 78 genes
- ⇒ 76 genes spotted four times (per array)
- ⇒ Two genes spotted 16 times (per array)

	Cy3 (Green)	Cy5 (Red)
Array 1	Treatment	Control
Array 2	Control	Treatment

Data available at: <http://www.jax.org/research/churchill/>

## MODEL and SAS CODE

$$y_{ijkgr} = \mu + A_i + D_j + V_k + G_g + \dots \\ \dots + (AG)_{igr} + (VG)_{kg} + (DG)_{jg} + \epsilon_{ijkgr}$$

```

data micro;
  infile 'C:\Documents and settings\rosag\My Documents\Microarray\Kerr\latinsquare2.txt';
  input y1 y2 y3 y4 clone;
  y=log(y1+1728.8); array=1; trt=1; dye=1; spot=_n_; output;
  y=log(y2-1728.8); array=1; trt=2; dye=2; spot=_n_; output;
  y=log(y3+1801.6); array=2; trt=2; dye=1; spot=336+_n_; output;
  y=log(y4-1801.6); array=2; trt=1; dye=2; spot=336+_n_; output;
drop y1 y2 y3 y4;

proc glm;
  title 'Corning data - Kerr et al. (2001)';
  class array trt dye clone spot;
  model y=array dye trt clone spot(array*clone) trt*clone dye*clone;
run;

```

## OUTPUT

Source	SS	df	MS
Array	0.11	1	0.11
Dye	3.35	1	3.35
Variety	45.69	1	45.69
Gene	917.53	77	11.92
Spot	145.19	289	0.46
Variety*Gene	66.46	77	0.86
Dye*Gene	21.30	77	0.28
Residual	19.72	820	0.0241
Adjusted Total	1219.35	1343	

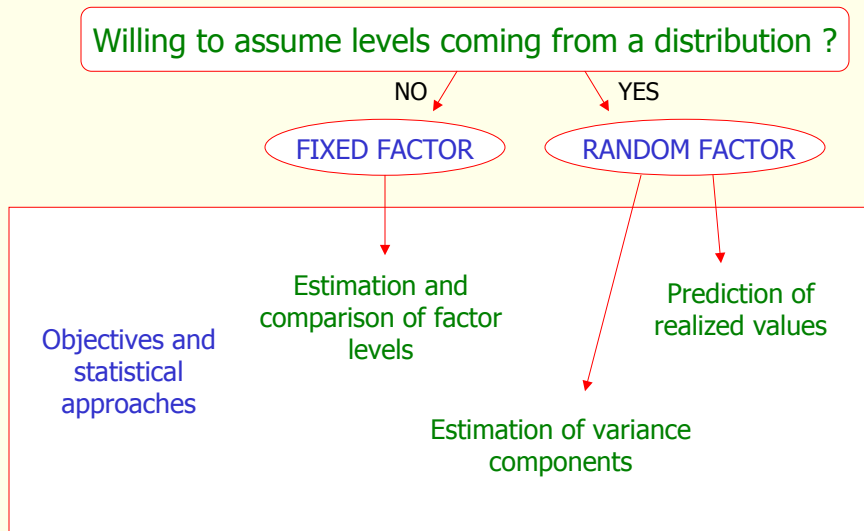
Table 3: Analysis of Variance. Notation: SS=sum of squares; df=degrees of freedom; MS=mean square.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	827	1199.627856	1.450578	37.95	<.0001
Error	516	19.721072	0.038219		
Corrected Total	1343	1219.348928			

R-Square      Coeff Var      Root MSE      y Mean  
0.983827      2.073215      0.195497      9.429661

Source	DF	Type I SS	Mean Square	F Value	Pr > F
array	1	0.1082239	0.1082239	2.83	0.0930
dye	1	3.3488605	3.3488605	87.82	<.0001
trt	1	45.6867175	45.6867175	1195.39	<.0001
clone	77	917.5331574	11.9160150	311.78	<.0001
spot(array*clone)	593	145.1939824	0.2448465	6.41	<.0001
trt*clone	77	66.4567952	0.8630753	22.58	<.0001
dye*clone	77	21.3001192	0.2766249	7.24	<.0001

## RANDOM AND FIXED EFFECTS



## RANDOM EFFECTS

- \* Inferences to the population from which the levels were sampled
- \* Incorporation of correlation in the model
  - Observations that share the same level of the random effect are being modeled as correlated
- \* Accuracy of estimates
  - Using random factors involves making extra assumptions but gives more accurate estimates
- \* Estimation method
  - Different estimations methods must be used, e.g. BLUP

## Two-Step Mixed Effects Models for Microarray Experiments

(Wolfinger et al., 2001)

1<sup>st</sup> Step: GLOBAL NORMALIZATION

$$\log(y_{ijklm}) = \mu + A_i + D_j + T_k + \underbrace{G_l + TG_{kl} + DG_{jl} + AG_{ilm}}_{\varepsilon_{ijklm}} + \varepsilon_{ijklm}$$

$$\log(y_{ijklm}) = \mu + A_i + D_j + T_k + \varepsilon_{ijklm}$$

## Two-Step Mixed Effects Models for Microarray Experiments

(Wolfinger et al., 2001)

2<sup>nd</sup> Step: GENE MODELS

Predicted Residuals:  $\hat{\varepsilon}_{ijklm} = (y_{ijklm} - \hat{y}_{ijklm})$

$$\hat{\varepsilon}_{ijklm} = \mu_l + S_{ilm} + D_{jl} + T_{kl} + \varepsilon_{ijklm}$$

Gene-specific  
treatment effects

$$\varepsilon_{ijklm} \sim N(0, \sigma_l^2)$$

## Two-Step Mixed Effects Models for Microarray Experiments

(some SAS code)

### 1<sup>st</sup> Step: Global Normalization

```
proc mixed data=adjusted covtest;  
  class array trt dye;  
  model y = dye trt / outp=pred;  
  random array;  
run;
```

log intensities

## Two-Step Mixed Effects Models for Microarray Experiments

(some SAS code)

### 2<sup>nd</sup> Step: Gene-specific Analyses

```
ods listing close;  
proc mixed data=pred covtest maxiter=20;  
  by Gene_Name;  
  class array trt dye;  
  model resid = dye trt;  
  random array;  
  lsmeans time / diff cl;  
  ods output diffs=tdiff lsmeans=estim tests3=Fvalues;  
run;  
ods listing;
```

$\hat{C}_{ijklm}$

## Comments: Statistical Assumptions

❶ Why a two-step approach?

It is virtually impossible to fit all terms simultaneously in SAS.

❷ The second approach models gene-specific variance components.

❸ Residuals from the normalization model are correlated by construction, and yet they are modeled with independent errors in the gene models.

Little to no difference in practice (Wolfinger et al, 2001).

❹ Normality in the log-scale.

Usual assumption; but standard graphical and statistical checks using residuals from gene models should be performed.

## MODEL and SAS CODE

EXAMPLE  
Kerr et al. (2001)

```
proc mixed data=micro covtest;
class array trt dye;
model y=dye trt / outp=pred;
random array;

proc sort data=pred out=new;
by clone;

ods listing close;

proc mixed data=new covtest;
by clone;
class array trt dye spot;
model resid=dye trt;
random spot(array);
lsmeans trt / diff cl;
ods output diffs=tdiff lsmeans=estimat;
run;

ods listing;
```

MANMADA: More Details on SAS code

<http://statgen.ncsu.edu/ggibson/Manual.htm>

# OUTPUT

## GLOBAL NORMALIZATION

Covariance Parameter Estimates				
Cov Parm	Estimate	Standard Error	Z Value	Pr > Z
array	0			
Residual	0.8727	0.03370	25.89	<.0001

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
dye	1	1340	3.84	0.0503
trt	1	1340	52.35	<.0001

**NOTE:** The global normalization step considers the assumption of no differential expression for most of the genes (otherwise some important effects would be removed from the data).

# OUTPUT

## GENE MODELS: LS Means

	clone	Effect	trt	Estimate	Standard Error	DF	t Value	Pr >  t	Alpha	Lower	Upper	
1		1	trt	1	-0.1790	0.1500	6	-1.19	0.2777	0.05	-0.5461	0.1980
2		1	trt	2	-1.3550	0.1500	6	-9.03	0.0001	0.05	0.9879	1.7220
3		2	trt	1	-0.5216	0.1182	6	-4.41	0.0045	0.05	-0.8109	-0.2323
4		2	trt	2	-0.3795	0.1182	6	-3.21	0.0184	0.05	0.09018	0.6688
5		3	trt	1	-0.3029	0.07764	6	-3.90	0.0080	0.05	-0.4928	-0.1129
6		3	trt	2	0.002585	0.07764	6	0.03	0.9745	0.05	-0.1874	0.1926
7		4	trt	1	-0.4455	0.1272	6	-3.50	0.0128	0.05	-0.7568	-0.1341
8		4	trt	2	-0.3664	0.1272	6	-2.88	0.0281	0.05	-0.6778	-0.05504
9		7	trt	1	-0.6230	0.09160	6	-6.80	0.0005	0.05	-0.8471	-0.3988
10		7	trt	2	-0.6273	0.09160	6	-6.85	0.0005	0.05	-0.8514	-0.4032
11		8	trt	1	-0.07582	0.1190	6	-0.64	0.5474	0.05	-0.3669	0.2153
12		8	trt	2	0.03749	0.1190	6	0.32	0.7633	0.05	-0.2536	0.3286
13		9	trt	1	4.2087	0.06833	6	61.59	<.0001	0.05	4.0415	4.3759
14		9	trt	2	4.1800	0.06833	6	61.17	<.0001	0.05	4.0128	4.3472
15		11	trt	1	-0.2223	0.5105	6	-0.44	0.6784	0.05	-1.4716	1.0269
16		11	trt	2	-0.06286	0.5105	6	-0.16	0.8764	0.05	-1.3321	1.1664

**NOTE:** The estimates (LSMeans) absolute values do not have a biological interpretation; the interpretation should be always relatively to other treatments.

# OUTPUT

## GENE MODELS: Differences between Treatments

	clone	Effect	trt	_trt	Estimate	Standard Error	DF	t Value	Pr >  t	Alpha	Lower	Upper
1	1	trt	1	2	-1.5340	0.1451	6	-10.57	< .0001	0.05	-1.8890	-1.1790
2	2	trt	1	2	-0.9011	0.09997	6	-9.01	0.0001	0.05	-1.1457	-0.6564
3	3	trt	1	2	-0.3054	0.1098	6	-2.78	0.0319	0.05	-0.5741	-0.03676
4	4	trt	1	2	-0.07905	0.1767	6	-0.45	0.6703	0.05	-0.5114	0.3533
5	7	trt	1	2	0.004336	0.08395	6	0.05	0.9604	0.05	-0.2008	0.2095
6	8	trt	1	2	-0.1133	0.1420	6	-0.80	0.4553	0.05	-0.4607	0.2341
7	9	trt	1	2	0.02868	0.02392	6	1.20	0.2757	0.05	-0.02985	0.08722
8	11	trt	1	2	-0.1395	0.1448	6	-0.96	0.3726	0.05	-0.4938	0.2148
9	12	trt	1	2	-0.2898	0.09124	6	-3.17	0.0194	0.05	-0.5121	-0.06556
10	13	trt	1	2	-0.01957	0.1042	6	-0.19	0.8572	0.05	-0.2745	0.2354
11	14	trt	1	2	-0.06978	0.1096	6	-0.64	0.5477	0.05	-0.3379	0.1983
12	15	trt	1	2	-0.09022	0.06057	6	-1.49	0.1870	0.05	-0.2384	0.05800
13	17	trt	1	2	-0.06582	0.09299	6	-0.71	0.5039	0.05	-0.2924	0.1607
14	18	trt	1	2	-0.3977	0.03776	6	-10.53	< .0001	0.05	-0.4901	-0.3053
15	19	trt	1	2	-0.00253	0.07174	6	-0.04	0.9730	0.05	-0.1781	0.1730
16	20	trt	1	2	-0.07723	0.05314	6	-1.45	0.1964	0.05	-0.2073	0.05281

EXAMPLE:

$$\log(\text{trt1}) - \log(\text{trt2}) = \log(\text{trt1}/\text{trt2}) = -0.3977$$

$$\text{trt1}/\text{trt2} = 10^{-0.3977} = 0.4002$$

$$\Rightarrow \text{trt2} \cong 2.5 \times \text{trt1}$$

NOTE: Should use a correction for multiple testing.

## LOESS + Mixed Effects Model

IS IT OKAY TO USE LOESS NORMALIZATION AND GLOBAL NORMALIZATION ?!

Loess: Array-specific dye intensity bias

Global Normalization (1<sup>st</sup> step): Spatial variability on the slide

Note: Array and Dye main effects will be non significant on the 1<sup>st</sup> step of the mixed model methodology.