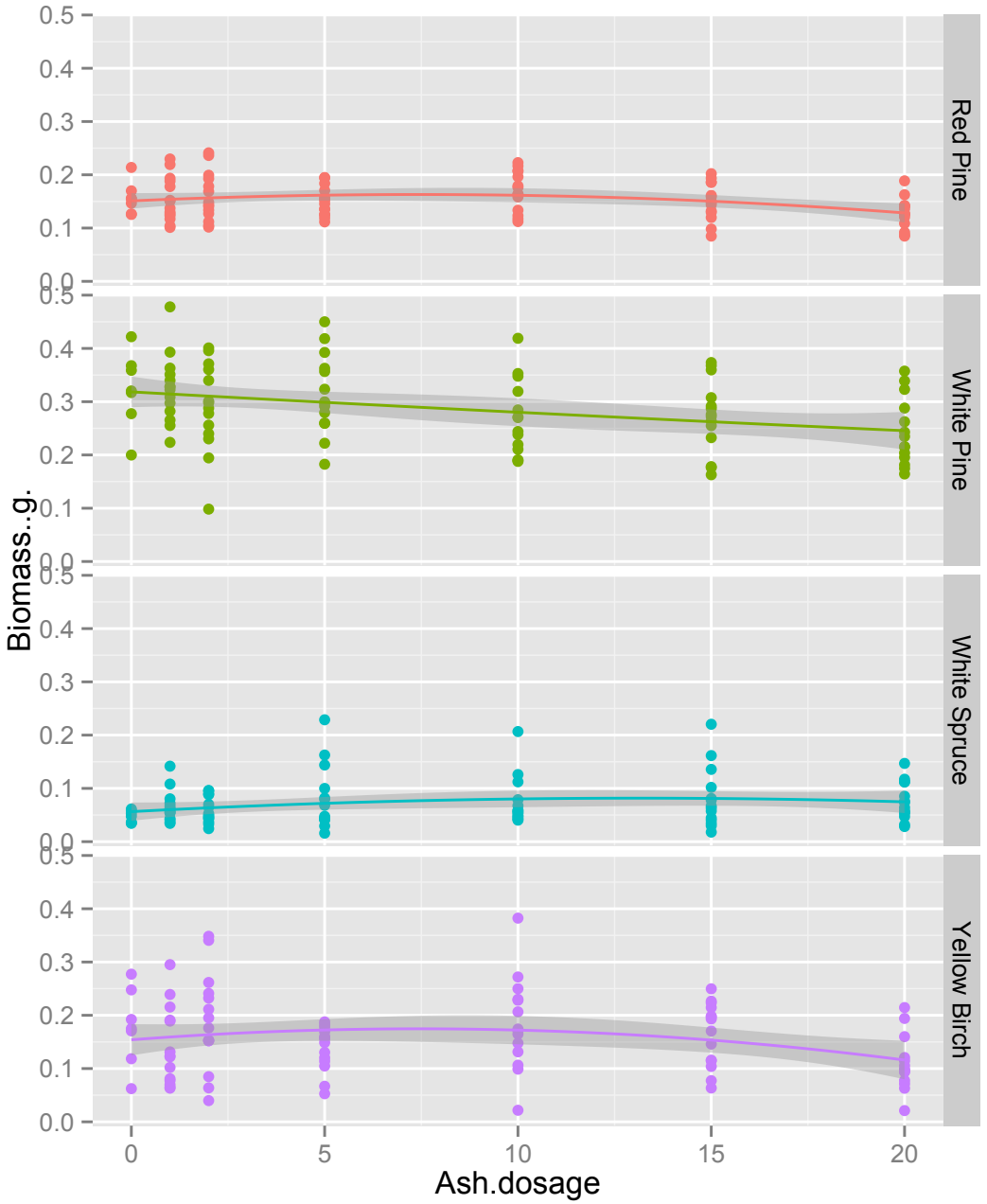
# Today

- HW 1: due *today*, 11.59 pm.

- HW 2: due March 4, posted soon
- Backback to Briefcase, Feb 10 6 - 8 pm (Career Centre)
- Recap on trees analysis
- Contingency tables
- Next week: Generalized Linear Models Chs. 6 and 7
- after mid-term break: random effects, mixed linear and non-linear models, nonparametric regression methods

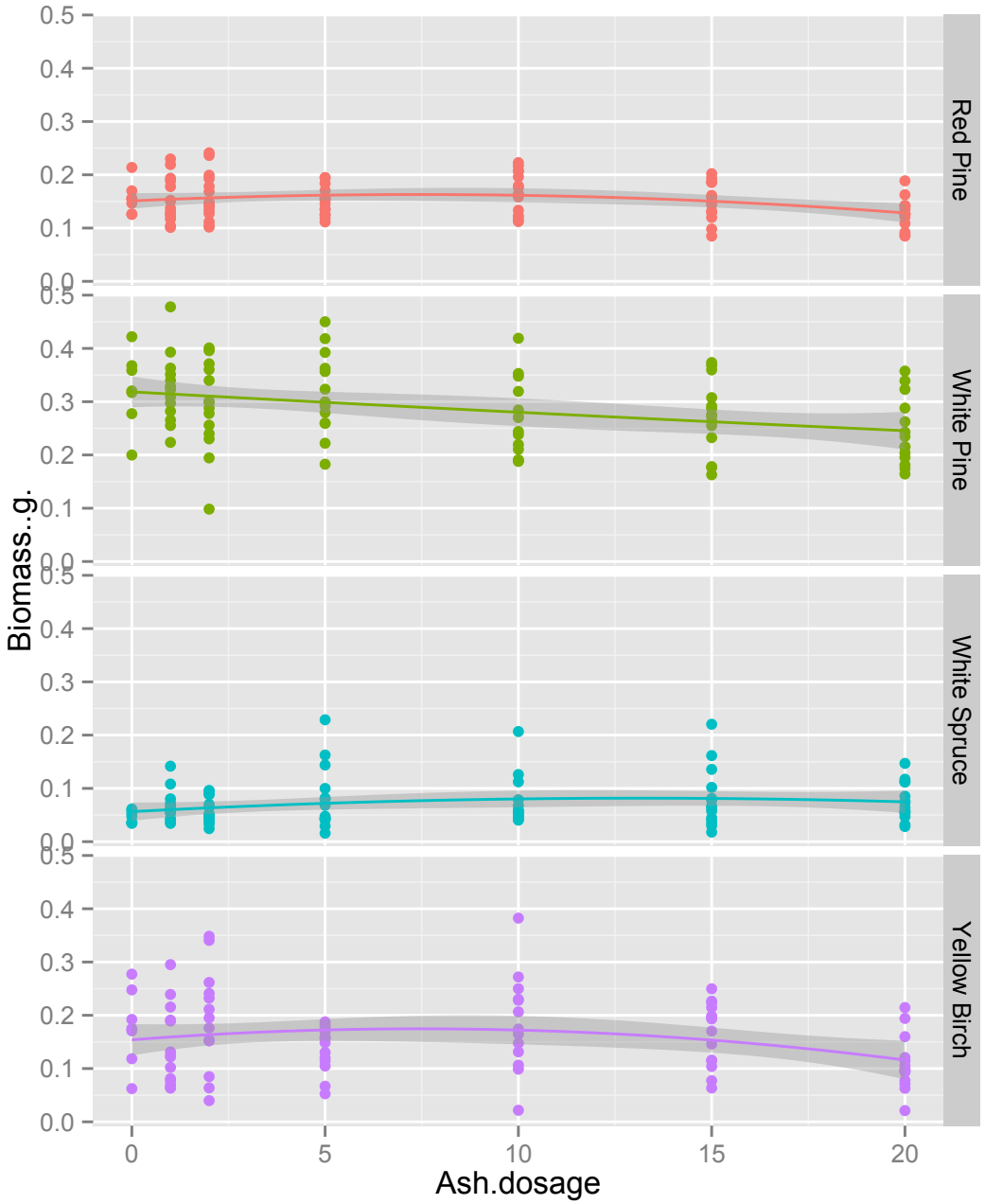
- Young Statisticians writing Competition




```
qplot(Ash.dosage, Biomass..g.,
data = trees, facets =
Seedling.species ~ ., color =
Seedling.species) +
geom_smooth(method = "lm", formula
= y ~ x + I(x^2), se = T)
```

**Seedling.species**

- Red Pine
- White Pine
- White Spruce
- Yellow Birch




```
qplot(Ash.dosage, Biomass..g.,
data = trees, facets =
Seedling.species ~ ., color =
Seedling.species) +
geom_smooth(method = "lm", formula
= y ~ poly(x,2), se = T)
```

**Seedling.species**

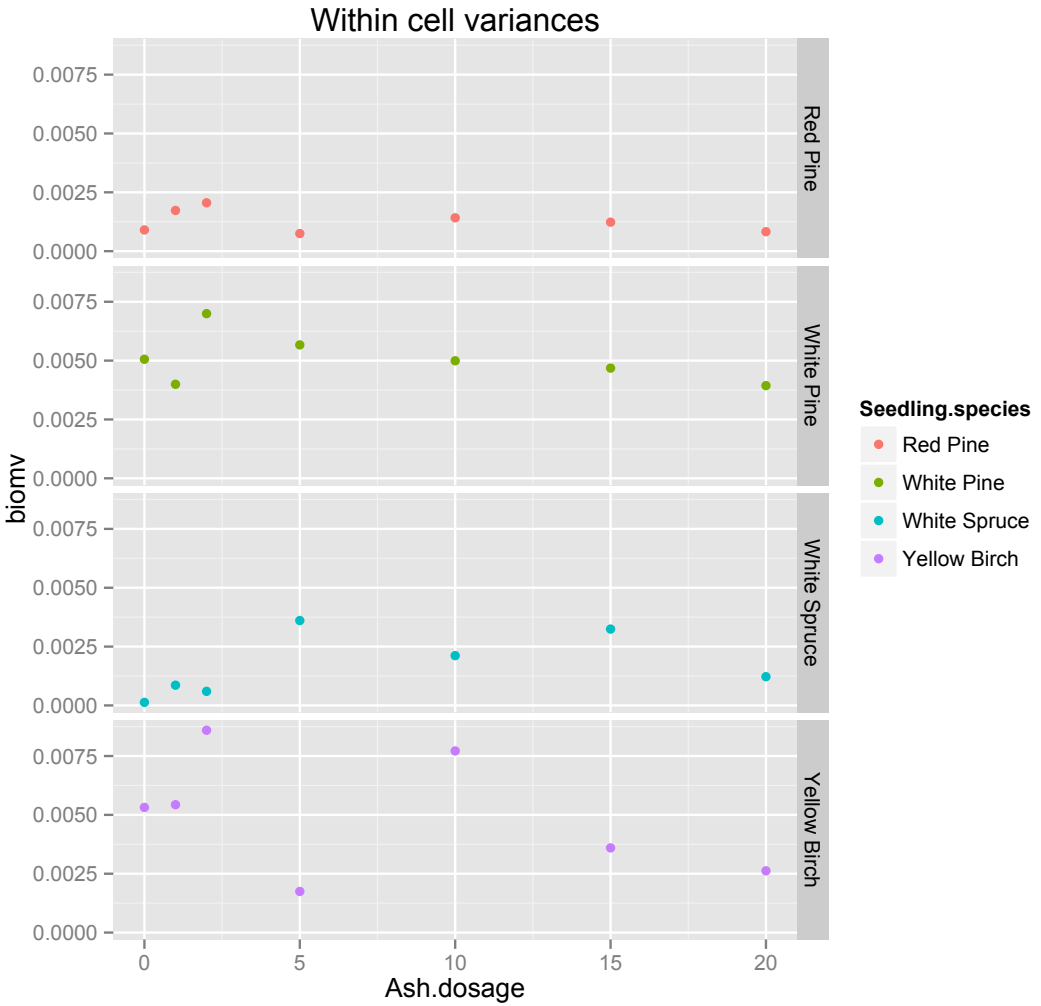
- Red Pine
- White Pine
- White Spruce
- Yellow Birch

$$x^* = -\frac{\beta_1}{2\beta_2}$$

'optimal dose'

# linear models

```
# straight lines for each species, all with same slope:

trees.lm <- lm(formula = Biomass..g. ~ Ash.dosage + Seedling.species, data = trees) #

# orthogonal polynomials as in class

trees.lm2 <- lm(formula = Biomass..g. ~ as.ordered(Ash.dosage) + Seedling.species,
    data = trees)

# ordinary quadratics (no indication that any higher orders are needed

trees.lm3 <- lm(formula = Biomass..g. ~ Ash.dosage + I(Ash.dosage^2) + Seedling.species,
    data = trees)

# this allows a different slope for each species

trees.lm4 <- lm(formula = Biomass..g. ~ Ash.dosage * Seedling.species, data = trees)

# and a different quadratic for each species

trees.lm5 <- lm(formula = Biomass..g. ~ poly(Ash.dosage, 2) * Seedling.species,
    data = trees)
```



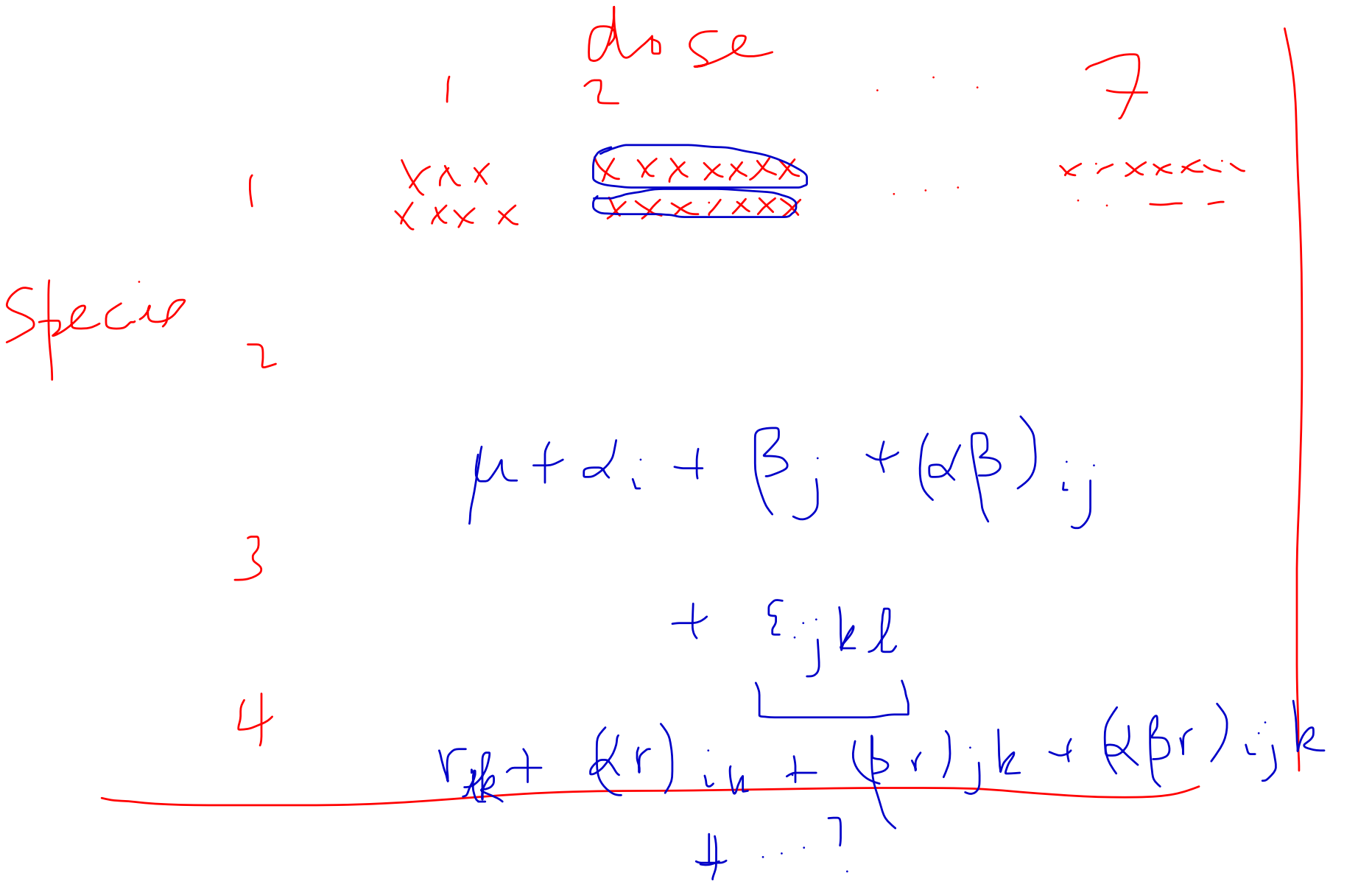

```
vartrees <-
ddply(trees,.(Seedling.species,
Ash.boiler.type,
Ash.dosage),summarize, biomv =
var(Biomass..g.))
qplot(Ash.dosage, biomv, data =
vartrees, facets =
Seedling.species   .,
color=Seedling.species, main =
"Within cell variances")
```

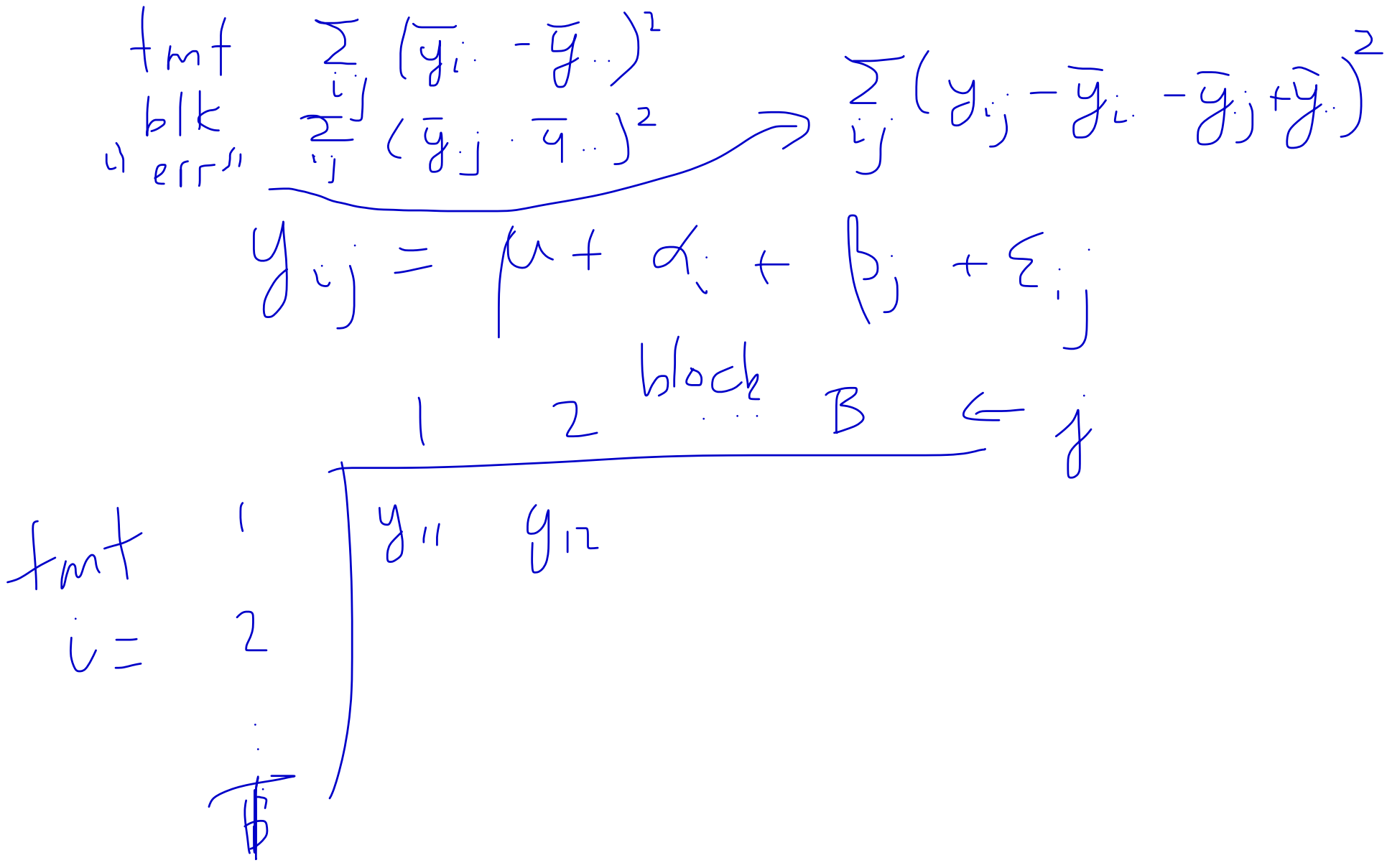
lm5 biom ~ poly(dose) * species

$$y_{ijkl} = \mu + \alpha_i + \beta_1 x_j + \beta_2 x_j^2$$

$$+ \beta_{1i} x_j + \beta_{2i} x_j^2 + \varepsilon_{ijkl}$$

$$i = 2, 3, 4$$


dose
1
2
7
Specie
1
2
3
4
$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$
$+ \varepsilon_{ijkl}$
$r_{ijk} + \zeta(r)_{ijk} + (\beta r)_{jk} + (\alpha\beta r)_{ijk}$


fmt
blk
err
$\sum_{ij} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$
$\sum_{ij} (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$
$\sum_{ij} (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2$
$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$
1 2 block B ← j
fmt
i = 1 2
$y_{11}$ $y_{12}$

```
> summary(trees.lm5)
Call:
lm(formula = Biomass..g. ~ poly(Ash.dosage, 2) * Seedling.species,
    data = trees)
Residuals:
     Min       1Q   Median       3Q      Max
-0.212170 -0.035524 -0.005796  0.032865  0.210208
Coefficients:
                                                Estimate Std. Error
(Intercept)                                     0.152024   0.006027
poly(Ash.dosage, 2)1                           -0.142298   0.114985
poly(Ash.dosage, 2)2                           -0.151092   0.114985
Seedling.speciesWhite Pine                      0.135831   0.008523
Seedling.speciesWhite Spruce                   -0.081298   0.008523
Seedling.speciesYellow Birch                    0.003942   0.008523
poly(Ash.dosage, 2)1:Seedling.speciesWhite Pine  -0.348239   0.162613
poly(Ash.dosage, 2)2:Seedling.speciesWhite Pine   0.163851   0.162613
poly(Ash.dosage, 2)1:Seedling.speciesWhite Spruce  0.272140   0.162613
poly(Ash.dosage, 2)2:Seedling.speciesWhite Spruce  0.049687   0.162613
poly(Ash.dosage, 2)1:Seedling.speciesYellow Birch -0.097340   0.162613
poly(Ash.dosage, 2)2:Seedling.speciesYellow Birch -0.106409   0.162613
                                               t value Pr(>|t|)
(Intercept)                                     25.225   <2e-16 ***
poly(Ash.dosage, 2)1                            -1.238   0.2167
poly(Ash.dosage, 2)2                            -1.314   0.1897
Seedling.speciesWhite Pine                      15.937   <2e-16 ***
Seedling.speciesWhite Spruce                    -9.538   <2e-16 ***
Seedling.speciesYellow Birch                     0.462   0.6440
poly(Ash.dosage, 2)1:Seedling.speciesWhite Pine   -2.142   0.0329 *
poly(Ash.dosage, 2)2:Seedling.speciesWhite Pine    1.008   0.3143
poly(Ash.dosage, 2)1:Seedling.speciesWhite Spruce   1.674   0.0951 .
poly(Ash.dosage, 2)2:Seedling.speciesWhite Spruce   0.306   0.7601
poly(Ash.dosage, 2)1:Seedling.speciesYellow Birch  -0.599   0.5498
poly(Ash.dosage, 2)2:Seedling.speciesYellow Birch  -0.654   0.5133
```

# Contingency tables

ELM Ch. 4

| Quality | No Particles | Particles | Total |
|---|---|---|---|
| Good | 320 | 14 | 334 |
| Bad | 80 | 36 | 116 |
| Total | 400 | 50 | 450 |

- see p.70 for data.frame `wafer` and use of `xtabs`

- Poisson regression:

```
modl <- glm(y ~ particle + quality, data = wafer, family = poisson)
```

```
glm(formula = y ~ particle + quality, family = poisson, data = wafer)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.6934     0.0572  99.535   <2e-16 ***
particleyes  -2.0794     0.1500 -13.863   <2e-16 ***
qualitybad   -1.0575     0.1078  -9.813   <2e-16 ***
---
...
       Null deviance: 474.10  on 3  degrees of freedom
Residual deviance:  54.03  on 1  degrees of freedom
```

# ... contingency tables

| Quality | No Particles | Particles | Total |
|---|---|---|---|
| Good | 320 | 14 | 334 |
| Bad | 80 | 36 | 116 |
| Total | 400 | 50 | 450 |

```
modl <- glm(y   particle + quality, data = wafer, family = poisson)
```

Model:

$$\log \mu_{ij} = \gamma + \alpha_i + \beta_j$$

```
---
...
       Null deviance: 474.10  on 3  degrees of freedom
Residual deviance:  54.03  on 1  degrees of freedom
```

Test of no interaction between particle and quality

# ... contingency tables

| Quality | No Particles | Particles | Total |
|---|---|---|---|
| Good | 320 | 14 | 334 |
| Bad | 80 | 36 | 116 |
| Total | 400 | 50 | 450 |

$= y_{1+}$, $= y_{2+}$, $\hat{p}_{1\cdot} = \frac{334}{450}$, $\hat{p}_{\cdot 1} = \frac{400}{450}$

Multinomial model: fix total sample size (450):

$$y \sim \text{Mult}(n; p); p_{ij} = \Pr\{\text{single observation is in cell}(i,j)\}$$

$$L(p; y) = \frac{n!}{y_{11}!y_{12}!y_{21}!y_{22}!} p_{11}^{y_{11}} p_{12}^{y_{12}} p_{21}^{y_{21}} p_{22}^{y_{22}}$$

$\sum p_{ij} = 1$, $\sum y_{ij} = n$

Independence: $p_{ij} = p_i \times p_j$

Maximum likelihood estimates:

– under independence $\hat{p}_{ij} = \hat{p}_i \hat{p}_j = \frac{y_{i+}}{n} \times \frac{y_{+j}}{n}$

– unrestricted $\tilde{p}_{ij} = y_{ij}/n$

$p_{ij}^{y_{ij}}$

# ... contingency tables

| Quality | No Particles | Particles | Total |
|---|---|---|---|
| Good | 320 296.89 | 14 37.11 | 334 |
| Bad | 80 103.11 | 36 12.89 | 116 |
| Total | 400 | 50 | 450 |

LRT of indep.

296.9

$= \hat{p}_{1\cdot} \times \hat{p}_{\cdot 1} \times n$

$\frac{y_{1+} \times y_{+1}}{n = y_{++}}$

```
2*sum(sum(ov*log(ov/fv)))
[1] 54.03045
```

see ELM for construction of `ov` and `fv`

```
sum((ov-fv)^ 2/fv)
[1] 62.81231
```

```
modb <- glm (matrix(wafer$y, nrow=2) ~ 1, family = binomial)
```

```
    Null deviance: 54.03  on 1  degrees of freedom
Residual deviance: 54.03  on 1  degrees of freedom
```

```
modb2 <- glm(matrix(wafer$y, nrow = 2) ~ c("nop","p"), family = binomial)
```

```
     Null deviance: 54.03  on 1  degrees of freedom
Residual deviance:  0.00  on 0  degrees of freedom
```

# ... contingency tables

| Quality | No Particles | Particles | Total |
|---|---|---|---|
| Good | 320 | 14 | 334 |
| Bad | 80 | 36 | 116 |
| Total | 400 | 50 | 450 |

Fisher's exact test of independence: condition on all marginal totals

only $y_{11}$ free to vary

or any other single element

$$\Pr(Y_{11} = y_{11} \mid y_{1+}, y_{+1}, n) = \frac{\binom{y_{1+}}{y_{11}}\binom{n-y_{1+}}{y_{+1}-y_{11}}}{\binom{n}{y_{1+}}}$$

```
> fisher.test(ov)

Fisher's Exact Test for Count Data

data:  ov
p-value = 2.955e-13
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  5.090628 21.544071
sample estimates:  odds ratio   10.21331 {\rf ?where is 10.213 in previous analyses}
```

# Fisher's exact test

Agresti, CDA 2nd ed., p.92

- test of independence in $2 \times 2$ table
- based on hypergeometric distribution
- conditions on all marginal totals
- this eliminates all nuisance parameters (parameters governing marginal distribution)

| | Guess poured first | | |
|---|---|---|---|
| Poured First | Milk | Tea | Total |
| Milk | 3 | 1 | 4 |
| Tea | 1 | 3 | 4 |
| Total | 4 | 4 | 8 |

$$\Pr(y_{11} \geq 3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} + \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = 0.229 + 0.014 = 0.243$$

```
Fisher's Exact Test for Count Data
data: tea p-value = 0.4857 alternative hypothesis: true odds ratio is not equal to 1 95
percent confidence interval: 0.2117329 621.9337505 sample estimates: odds ratio 6.408309
```

# Fisher's exact test

Agresti, CDA 2nd ed., p.92

- test of independence in $2 \times 2$ table
- based on hypergeometric distribution
- conditions on all marginal totals
- this eliminates all nuisance parameters (parameters governing marginal distribution)

| | Guess poured first | | |
|---|---|---|---|
| Poured First | Milk | Tea | Total |
| Milk | 3 | 1 | 4 |
| Tea | 1 | 3 | 4 |
| Total | 4 | 4 | 8 |

$$\Pr(y_{11} \geq 3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} + \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = 0.229 + 0.014 = 0.243$$

```
Fisher's Exact Test for Count Data
data: tea p-value = 0.4857 alternative hypothesis: true odds ratio is not equal to 1 95
percent confidence interval: 0.2117329 621.9337505 sample estimates: odds ratio 6.408309
```

# Fisher's exact test

Agresti, CDA 2nd ed., p.92

- test of independence in $2 \times 2$ table
- based on hypergeometric distribution
- conditions on all marginal totals
- this eliminates all nuisance parameters (parameters governing marginal distribution)

| | Guess poured first | | |
|---|---|---|---|
| Poured First | Milk | Tea | Total |
| Milk | 3 | 1 | 4 |
| Tea | 1 | 3 | 4 |
| Total | 4 | 4 | 8 |

$$\Pr(y_{11} \geq 3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} + \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = 0.229 + 0.014 = 0.243$$

```
Fisher's Exact Test for Count Data
data: tea p-value = 0.4857 alternative hypothesis: true odds ratio is not equal to 1 95
percent confidence interval: 0.2117329 621.9337505 sample estimates: odds ratio 6.408309
```

# Fisher's exact test

Agresti, CDA 2nd ed., p.92

- test of independence in $2 \times 2$ table
- based on hypergeometric distribution
- conditions on all marginal totals
- this eliminates all nuisance parameters (parameters governing marginal distribution)

| | Guess poured first | | |
|---|---|---|---|
| Poured First | Milk | Tea | Total |
| Milk | 3 | 1 | 4 |
| Tea | 1 | 3 | 4 |
| Total | 4 | 4 | 8 |

$$\Pr(y_{11} \geq 3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} + \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = 0.229 + 0.014 = 0.243$$

```
Fisher's Exact Test for Count Data
data: tea p-value = 0.4857 alternative hypothesis: true odds ratio is not equal to 1 95
percent confidence interval: 0.2117329 621.9337505 sample estimates: odds ratio 6.408309
```

# Fisher's exact test

Agresti, CDA 2nd ed., p.92

- test of independence in $2 \times 2$ table
- based on hypergeometric distribution
- conditions on all marginal totals
- this eliminates all nuisance parameters (parameters governing marginal distribution)

| | Guess poured first | | |
|---|---|---|---|
| Poured First | Milk | Tea | Total |
| Milk | 3 | 1 | 4 |
| Tea | 1 | 3 | 4 |
| Total | 4 | 4 | 8 |

$$\Pr(y_{11} \geq 3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} + \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = 0.229 + 0.014 = 0.243$$

```
Fisher's Exact Test for Count Data
data: tea p-value = 0.4857 alternative hypothesis: true odds ratio is not equal to 1 95
percent confidence interval: 0.2117329 621.9337505 sample estimates: odds ratio 6.408309
```

# ... Fisher's exact test

- achievable $p$-values: 0.014, 0.243, 0.757, 0.986, 1.0
- null distribution concentrated on only 5 sample points
- Agresti recommends mid $p$-value:

$$\frac{1}{2}\Pr(Y_{11} = 3) + \Pr(Y_{11} = 4) = 0.129$$

$$P_0(Y_{11} \geq 3)$$

$$\frac{1}{2}P_0(Y_{11} = 3) + P_0(Y_{11} > 3)$$

# Several $2 \times 2$ Tables

ELM, §4.4; SM, Example 10.19

| Age (years) | Smokers | Non-smokers |
|---|---|---|
| Overall | 139/582 (24) | 230/732 (31) |
| 18–24 | 2/55 (4) | 1/62 (2) |
| 25–34 | 3/124 (2) | 5/157 (3) |
| 35–44 | 14/109 (13) | 7/121 (6) |
| 45–54 | 27/130 (21) | 12/78 (15) |
| 55–64 | 51/115 (44) | 40/121 (33) |
| 65–74 | 29/36 (81) | 101/129 (78) |
| 75+ | 13/13 (100) | 64/64 (100) |

**Table 6.8** Twenty-year survival and smoking status for 1314 women (Appleton *et al.*, 1996). The smoker and non-smoker columns contain number dead/total (% dead).

| | Smoker | Non-smoker | |
|---|---|---|---|
| dead | 139 (24%) | 230 (31%) | |
| alive | 443 | 502 | |
| total | 582 | 732 | 1314 |

# ... 2 × 2 tables

```
>  summary(glm(cbind(alive,dead) ~ smoker, data = smoking, family = binomial))
Call:
glm(formula = cbind(alive, dead) ~ smoker, family = binomial,
    data = smoking)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-12.173   -5.776    1.869    5.674    9.052

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.78052    0.07962   9.803  < 2e-16 ***
smoker       0.37858    0.12566   3.013  0.00259 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 641.5  on 13  degrees of freedom
Residual deviance: 632.3  on 12  degrees of freedom
AIC: 683.29

Number of Fisher Scoring iterations: 4
```

# ... $2 \times 2$ tables

| | Smoker | Non-smoker | |
|---|---|---|---|
| dead | 139 (24%) | 230 (31%) | |
| alive | 443 | 502 | |
| total | 582 | 732 | 1314 |

```
 > anova(glm(cbind(alive,dead) ~ smoker, data = smoking, family = binomial))
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(alive, dead)

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev
NULL                      13      641.5
smoker  1   9.2003        12      632.3
> with(smoking, xtabs(cbind(dead,alive) ~ smoker))

smoker dead alive
     0  230   502
     1  139   443
> summary(.Last.value)
Call: xtabs(formula = cbind(dead, alive) ~ smoker)
Number of cases in table: 1314
Number of factors: 2
Test for independence of all factors:
Chisq = 9.121, df = 1, p-value = 0.002527
```

# ... $2 \times 2$ tables

| | sm | non-sm | sm | non-sm | sm | non-sm | |
|---|---|---|---|---|---|---|---|
| d | 2 | 1 | 3 | 5 | 14 | 7 | |
| a | 53 | 61 | 121 | 152 | 95 | 114 | ... |
| | 55 | 62 | 124 | 157 | 109 | 121 | |
| Age | 18-24 | | 25-34 | | 35-44 | | ... |

```
> summary(glm(cbind(alive,dead) ~ smoker + factor(age), data = smoking, family = binomial))
...
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.8601     0.5939   6.500 8.05e-11 ***
smoker             -0.4274     0.1770  -2.414 0.015762 *
factor(age)25-34   -0.1201     0.6865  -0.175 0.861178
factor(age)35-44   -1.3411     0.6286  -2.134 0.032874 *
factor(age)45-54   -2.1134     0.6121  -3.453 0.000555 ***
factor(age)55-64   -3.1808     0.6006  -5.296 1.18e-07 ***
factor(age)65-74   -5.0880     0.6195  -8.213  < 2e-16 ***
factor(age)75+    -27.8073 11293.1437  -0.002 0.998035
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 641.4963  on 13  degrees of freedom
Residual deviance:   2.3809  on  6  degrees of freedom
AIC: 65.377

Number of Fisher Scoring iterations: 20
```

# Log-linear models

ELM 4.4

- suppose we have 3 factors, each with several levels
- observe a response at each combination of factors
- linear model might be

$$y_{ijk} = \mu+\alpha_i+\beta_j+\gamma_k+\epsilon_{ijk}, \quad k = 1,\ldots,K; j = 1,\ldots,J; i = 1,\ldots I$$

- or

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \epsilon_{ijk}$$

- if the $y_{ijk}$ are positive counts, rather than continuous, then Poisson model could have

$$y_{ijk} \sim Po(\mu_{ijk}), \quad \log(\mu_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k$$

- or

$$\log(\mu_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$$

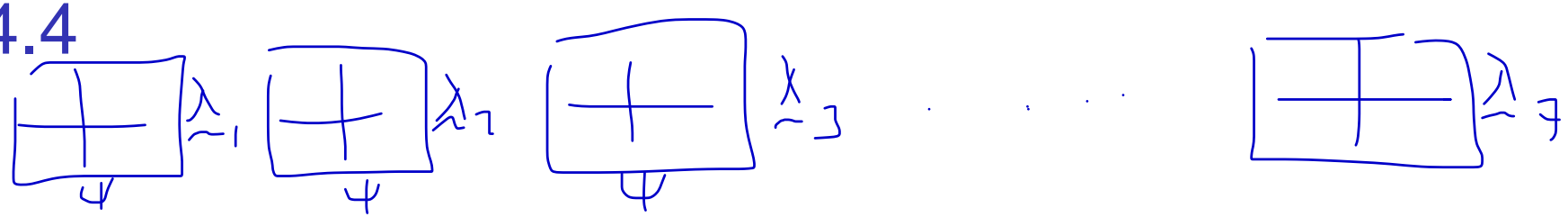
# ... §4.4

- several log-linear models for smoking data are fit
- and compared to binomial model above
- joint independence, conditional independence, marginal independence, uniform association
- all related to sub-models of general log-linear Poisson model

- binomial model above estimates parameters that control marginal probabilities
- Mantel-Haenszel test is a $2 \times 2 \times k$ version of Fisher's exact test

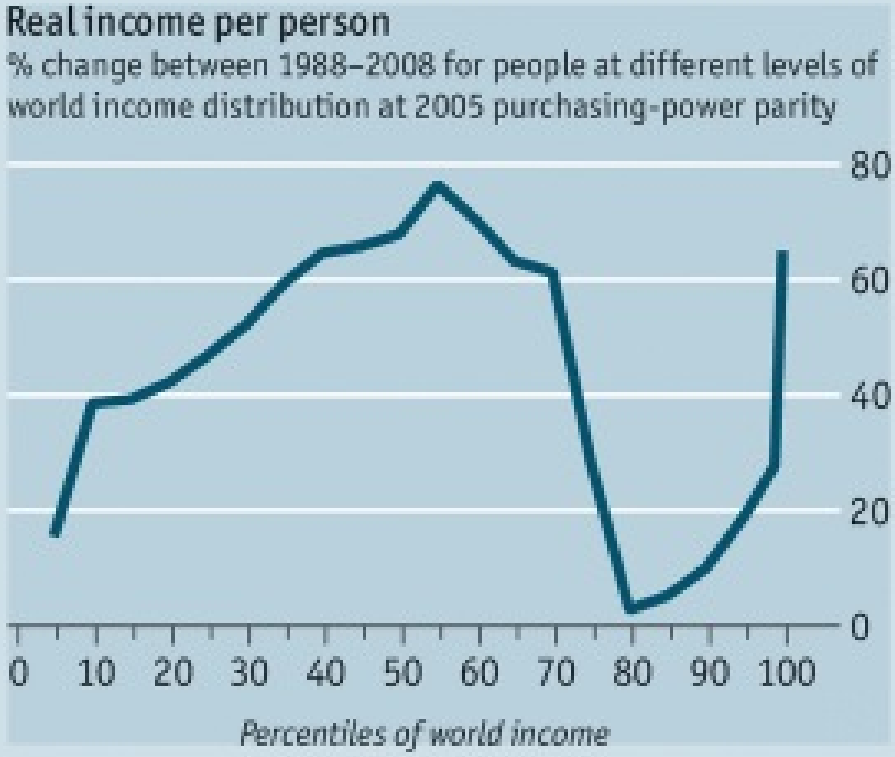
Correspondence analysis
r x c
(visualization)

2 x 2 (x k)
r x c x k



```
> data(femsmoke)
> ct3 <- xtabs(y ~ smoker + dead + age, data = femsmoke)
> apply(ct3, 3, function(x)(x[1,1]*x[2,2])/(x[1,2]*x[2,1]))
   18-24    25-34    35-44    45-54    55-64    65-74      75+
2.301887 0.753719 2.400000 1.441748 1.613672 1.148515      NaN
> mantelhaen.test(ct3,exact=T)

        Exact conditional test of independence in 2 x 2 x k tables

data:  ct3
S = 139, p-value = 0.01591
alternative hypothesis: true common odds ratio is not equal to
95 percent confidence interval:
 1.068889 2.203415
sample estimates:
common odds ratio
         1.530256
```

## Worth a second glance


Sources: Credit Suisse; Oxfam; "Global Income Inequality in Numbers", by Branko Milanovic, *Global Policy*, May 2013

Economist.com

Economist, January 24 2015