Session – October 04, 2016

- Generalized Linear Models continued
  - Logistic regression
  - Multinomial Logistic Regression
  - Poisson Regression
  - Log-linear models
    - Models of conditional independence
    - Model of homogeneous association
    - Saturated model

Selection of Statistical Analysis (Generalized Linear Models)

			Predictors	
		Continuous	Mixed	Categorical
	Continuous	Linear Regression	ANCOVA	ANOVA
Response	Ordinal	Logistic Regression	Logistic Regression	Logistic Regression
	Counts	Poisson regression	Poisson Regression	Poisson Regression
	Nominal	Multinomial Logistic regression	Multinomial Logistic Regression	Log-linear Models

Categorical response

- As with binary responses we model logarithmic odds with a linear predictor
- The idea is to get a separate model for each level of response
- As with categorical predictors, one level of the response is considered the reference level
- For each comparison with the reference level a separate model is computed
- While multinomial logistic regression assumes that the levels of the response variables are unordered, ordered logistic regression makes use of a particular order (but takes into account that the steps between ordered levels are unequal)
- For ordered logistic regression one then fits a common coefficient for each predictor across all levels of the response

Multinomial logistic regression

- Example (Homework 4 data set): Entering high school students make program choices among general program, vocational program and academic program. Their choice might be modeled using their writing score and their social economic status. The data set in STATA format contains variables on 200 students. The outcome variable is prog, program type.

ses	prog: general	academic	vocation
low	16	19	12
middle	20	44	31
high	9	42	7

Writing score

	M	SD
general	51.33333	9.397775
academic	56.25714	7.943343
vocation	46.76000	9.318754

$$ln\left(\frac{P(prog = academic)}{P(prog = general)}\right) = \beta_{10} + \beta_{11}\mathtt{ses{=}middle} + \beta_{12}\mathtt{ses{=}high} + \beta_{13}\mathtt{write}$$

$$ln\left(\frac{P(prog = vocation)}{P(prog = general)}\right) = \beta_{20} + \beta_{21}\mathtt{ses{=}middle} + \beta_{22}\mathtt{ses{=}high} + \beta_{23}\mathtt{write}$$

Multinomial logistic regression

- By default, the first level of the response is taken as reference level.
- We can change if we want, here we stick with the default

```
> prog.mlr <- multinom(prog ~ ses + write, data = student)
# weights:  15 (8 variable)
initial  value 219.722458
iter  10 value 179.985215
final  value 179.981726
converged
```

No p-values by default

```
> summary(prog.mlr)
Call:
multinom(formula = prog ~ ses + write, data = student)

Coefficients:
         (Intercept) ses[T.middle] ses[T.high]       write
academic   -2.851973     0.5332914   1.1628257  0.05792480
vocation    2.366097     0.8246384   0.1802176 -0.05567514

Std. Errors:
         (Intercept) ses[T.middle] ses[T.high]      write
academic    1.166437     0.4437319   0.5142215 0.02141092
vocation    1.174251     0.4901237   0.6484508 0.02333135

Residual Deviance: 359.9635
AIC: 375.9635
```

Multinomial logistic regression

- We can compute p-values using the standard normal distribution as reference (requires large sample size)

```
> z <- summary(prog.mlr)$coefficients/summary(prog.mlr)$standard.errors
> z
         (Intercept) ses[T.middle] ses[T.high]     write
academic   -2.445028      1.201832   2.2613324  2.705386
vocation    2.014984      1.682511   0.2779203 -2.386280
> #2-tailed z test
> p <- (1 - pnorm(abs(z), 0, 1))*2
> p
         (Intercept) ses[T.middle] ses[T.high]      write
academic  0.01448407    0.22942845  0.02373868 0.00682251
vocation  0.04390634    0.09246981  0.78107356 0.01701977
```

Both intercepts, ses(high) for academic and write are statistically significant

Multinomial logistic regression

```
> summary(prog.mlr)
Call:
multinom(formula = prog ~ ses + write, data = student)

Coefficients:
         (Intercept) ses[T.middle] ses[T.high]       write
academic   -2.851973     0.5332914   1.1628257  0.05792480
vocation    2.366097     0.8246384   0.1802176 -0.05567514

Std. Errors:
         (Intercept) ses[T.middle] ses[T.high]      write
academic    1.166437     0.4437319   0.5142215 0.02141092
vocation    1.174251     0.4901237   0.6484508 0.02333135

Residual Deviance: 359.9635
AIC: 375.9635
```

- A one unit increase in writing score is associated with an increase of the log odds for being in academic program as compared to the general program by the amount of .05792
- A one unit increase in writing score is associated with a decrease of the log odds for being in vocation program as compared to the general program by the amount of -.05568.
- The log odds of being in academic program vs. in general program will increase by 1.1628 if moving from ses=low to ses=high.

$$ln\left(\frac{P(prog = academic)}{P(prog = general)}\right) = \beta_{10} + \beta_{11}\texttt{ses=middle} + \beta_{12}\texttt{ses=high} + \beta_{13}\texttt{write}$$

$$ln\left(\frac{P(prog = vocation)}{P(prog = general)}\right) = \beta_{20} + \beta_{21}\texttt{ses=middle} + \beta_{22}\texttt{ses=high} + \beta_{23}\texttt{write}$$

Multinomial logistic regression

```
> head(prog.pp <- fitted(prog.mlr))
    general  academic  vocation
1 0.3382355 0.1482852 0.5134793
2 0.1806255 0.1202128 0.6991617
3 0.2367932 0.4186802 0.3445267
4 0.3508282 0.1726975 0.4764743
5 0.1689350 0.1001332 0.7309318
6 0.2377813 0.3533635 0.4088552
```

- We can switch further to the probability level.
- Looking at the fitted probabilities, we get an idea how likely the various program choices are for each student
- We can also fix the writing score at some value (e.g. the mean) and see how the predicted probabilities depend on the ses level

```
> prog.mlr.nd <- data.frame(ses = c("low", "middle", "high"),
+                           write = mean(student$write))
> predict(prog.mlr, newdata = prog.mlr.nd, type="probs")
    general  academic  vocation
1 0.3581865 0.4396867 0.2021268
2 0.2283319 0.4777561 0.2939120
3 0.1784891 0.7008979 0.1206130
```

```
> prog.mlr.nd
     ses  write
1    low 52.775
2 middle 52.775
3   high 52.775
```

Multinomial logistic regression

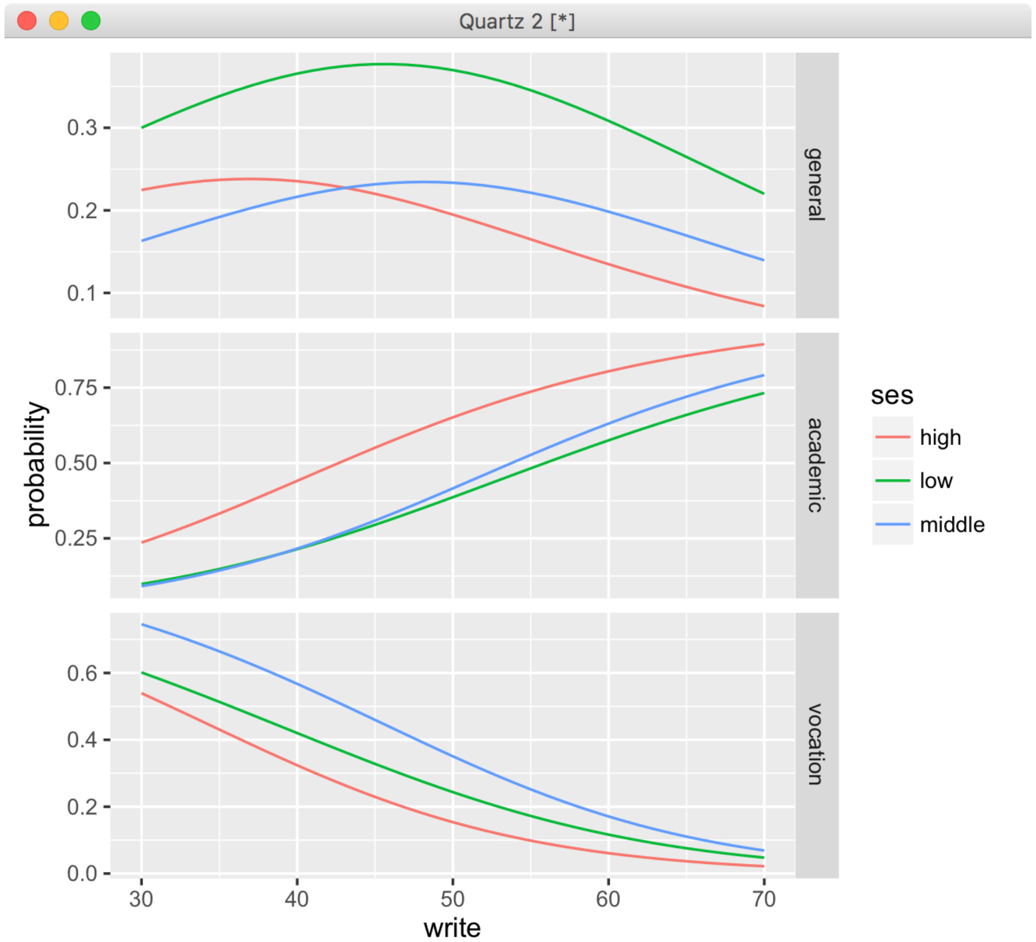
- Another way to understand the model using the predicted probabilities is to look at the averaged predicted probabilities for different values of the continuous predictor variable **write** within each level of **ses**.

```
  ses write variable probability
1 low    30  general   0.2999789
2 low    31  general   0.3082103
3 low    32  general   0.3161998
4 low    33  general   0.3238997
5 low    34  general   0.3312613
6 low    35  general   0.3382355
```

```
> prog.mlr.nd2 <- data.frame(ses = rep(c("low", "middle", "high"), each = 41),
+                            write = rep(c(30:70), 3))
>
> ## store the predicted probabilities for each value of ses and write
> prog.pp2 <- cbind(prog.mlr.nd2, predict(prog.mlr, newdata = prog.mlr.nd2, type = "probs", se = TRUE))
>
> ## calculate the mean probabilities within each level of ses
> by(prog.pp2[, 3:5], prog.pp2$ses, colMeans)
prog.pp2$ses: high
  general  academic  vocation
0.1807965 0.6164314 0.2027721
------------------------------------------------------------
prog.pp2$ses: low
  general  academic  vocation
0.3278129 0.3972998 0.2748873
------------------------------------------------------------
prog.pp2$ses: middle
  general  academic  vocation
0.2010845 0.4256261 0.3732894
```

Multinomial logistic regression

- Plot the predicted probabilities for different values of the continuous predictor variable **write** within each level of **ses**.



Ordinal logistic regression

- factors that influence the decision of whether to apply to graduate school.
- Response: three categories
  - unlikely,
  - somewhat likely, or
  - very likely to apply to graduate school.

	apply	pared	public	gpa
1	very likely	0	0	3.26
2	somewhat likely	1	0	3.21
3	unlikely	1	1	3.94
4	somewhat likely	0	0	2.81
5	somewhat likely	0	0	2.53
6	unlikely	0	1	2.59

- Predictors:
  - parental educational status,
  - whether the undergraduate institution is public or private, and
  - current GPA
- "distances" between response levels are not equal.
  - For example, the "distance" between "unlikely" and "somewhat likely" may be shorter than the distance between "somewhat likely" and "very likely".

Ordinal logistic regression

- factors that influence the decision of whether to apply to graduate school.

```
> app.olr <- polr(apply ~ pared + public + gpa, data = application, Hess=TRUE)
> summary(app.olr)
Call:
polr(formula = apply ~ pared + public + gpa, data = application,
    Hess = TRUE)

Coefficients:
          Value Std. Error t value
pared   1.04769     0.2658  3.9418
public -0.05879     0.2979 -0.1974
gpa     0.61594     0.2606  2.3632

Intercepts:
                            Value   Std. Error t value
unlikely|somewhat likely     2.2039  0.7795     2.8272
somewhat likely|very likely  4.2994  0.8043     5.3453

Residual Deviance: 717.0249
AIC: 727.0249
```

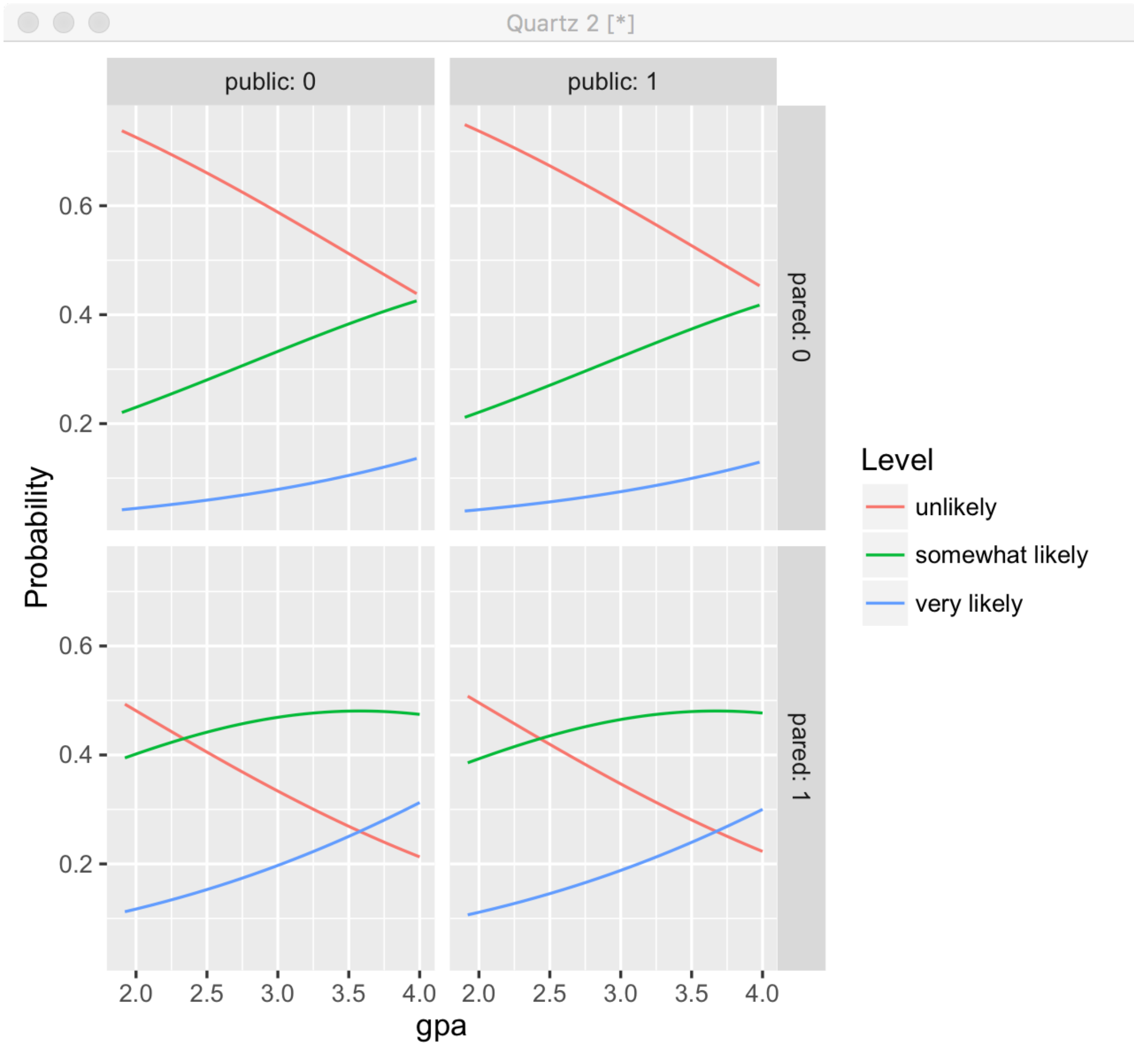
Hess = TRUE needed to get standard errors

```
> (ctable <- cbind(ctable, "p value" = p))
                                  Value Std. Error    t value      p value
pared                        1.04769010  0.2657894  3.9418050 8.087072e-05
public                      -0.05878572  0.2978614 -0.1973593 8.435464e-01
gpa                          0.61594057  0.2606340  2.3632399 1.811594e-02
unlikely|somewhat likely     2.20391473  0.7795455  2.8271792 4.696004e-03
somewhat likely|very likely  4.29936315  0.8043267  5.3452947 9.027008e-08
```

Ordinal logistic regression

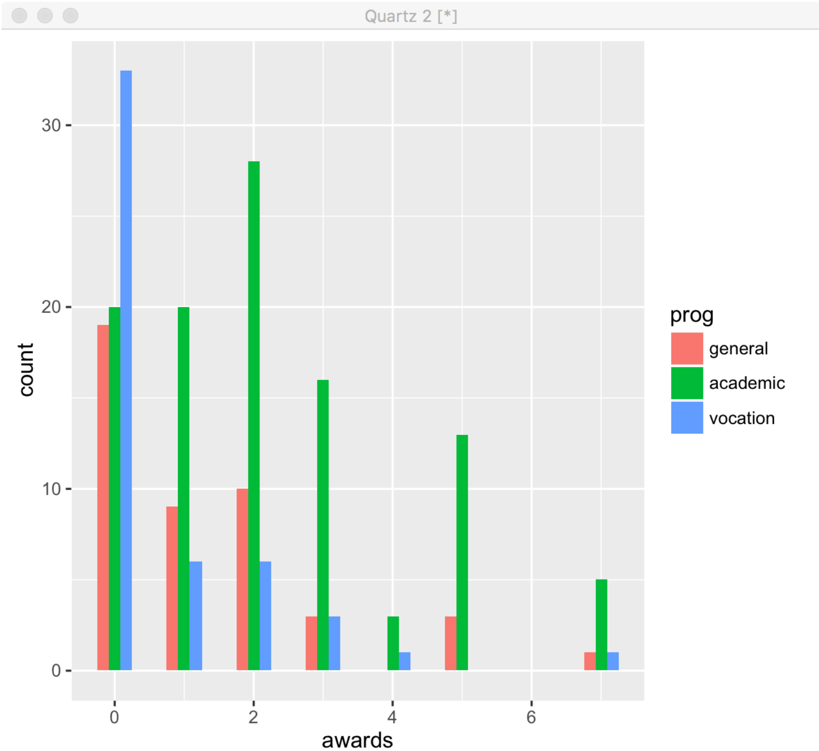
- factors that influence the decision of whether to apply to graduate school
- Predicted probabilities

```
                             Value
pared                   1.04769010
public                 -0.05878572
gpa                     0.61594057
unlikely|somewhat likely  2.20391473
somewhat likely|very likely  4.29936315
```



Poisson regression

- Ex.: The number of awards earned by students at one high school. Predictors of the number of awards earned include the type of program in which the student was enrolled (e.g., vocational, general or academic) and the score on their final exam in math.




```
Call:
glm(formula = awards ~ prog + math, family = "poisson", data = student)

Deviance Residuals:
    Min       1Q    Median      3Q      Max
-2.5388  -1.1693  -0.4423   0.5813   2.8809

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.899262   0.370987  -7.815  5.5e-15 ***
prog[T.academic]  0.065763   0.153966   0.427   0.6693
prog[T.vocation] -0.385123   0.207553  -1.856   0.0635 .
math              0.061709   0.006539   9.437  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 417.02  on 199  degrees of freedom
Residual deviance: 273.44  on 196  degrees of freedom
AIC: 627.67

Number of Fisher Scoring iterations: 5
```

By default log link

$$\begin{aligned} \mathrm{E}[y_i \mid x_i] &= \lambda_i = \exp x_i'\beta \\ &= \exp \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \end{aligned}$$

Poisson regression

- Ex.: The number of awards earned by students at one high school.
- Robust standard error computation is recommended.

```
Call:
glm(formula = awards ~ prog + math, family = "poisson", data = student)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5388  -1.1693  -0.4423   0.5813   2.8809

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.899262   0.370987  -7.815  5.5e-15 ***
prog[T.academic]  0.065763   0.153966   0.427   0.6693
prog[T.vocation] -0.385123   0.207553  -1.856   0.0635 .
math              0.061709   0.006539   9.437  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 417.02  on 199  degrees of freedom
Residual deviance: 273.44  on 196  degrees of freedom
AIC: 627.67

Number of Fisher Scoring iterations: 5
```

To control for deviations from distributional assumption (Poisson)!

```
                    Estimate   Robust SE     Pr(>|z|)          LL          UL
(Intercept)      -2.89926217 0.423342430 7.462644e-12 -3.72901333 -2.06951100
prog[T.academic]  0.06576256 0.192196350 7.322279e-01 -0.31094228  0.44246741
prog[T.vocation] -0.38512276 0.264689287 1.456694e-01 -0.90391377  0.13366824
math              0.06170878 0.007500189 1.909544e-16  0.04700841  0.07640915
```

Summary of Part 1:

- Generalized Linear Models continued
  - Logistic regression
  - Multinomial Logistic Regression
  - Poisson Regression

Preview on part 2:

- Log-linear models
  - Models of conditional independence
  - Model of homogeneous association
  - Saturated model

Log-linear models – example

Gender discrimination in College Admission?

The file `UCBAdmissions` (an R data set, contained in the package 'datasets') refers to individuals who applied for admission into one of the six largest graduate departments at the University of California in Berkeley, for the Fall 1973 session. The variables for this $2 \times 2 \times 6$ table are denoted by

Admit (`A`): Whether applicant was admitted or rejected

Gender (`G`): Gender of applicant (male, female)

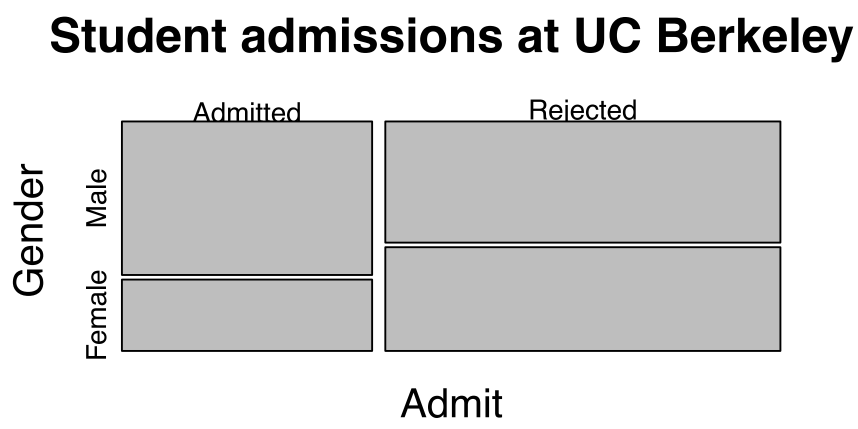
Dept (`D`): Department to which application was sent (A, B, C, D, E, or F)

Freq (`F`): Frequency of the corresponding cross-classification

Log-linear models – example

		Admitted	
		yes	no
Gender	male	1198	1493
	female	557	1258

Odds-ratio of admissions for males vs. females:
1.84 to 1



Log-linear models

Log-linear models for two-way-tables

- Given two categorical random variables, A and B, there are two main models
- Independence model (A,B)
- Saturated model (AB)

- Objective: Model the (expected) cell counts

notation for expected cell counts: $E(n_{ij}) = \hat{n}_{ij} = \mu_{ij}$

Log-linear models

Independence model for two-way-tables

$$\mu_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$$

$$\log \mu_{ij} = \log n_{i\cdot} + \log n_{\cdot j} - \log n$$

$$= \mu + \alpha_i + \beta_j$$

$$= \lambda + \lambda_i^A + \lambda_j^B$$

- different forms of notation
- last but one, resembles ANOVA style
- last one, used in standard books on log.linear models, e.g. Agresti (1998)

λ represents an *overall* effect, or a grand mean of the logarithms of expected counts, and it ensures that the sume of all expected counts equals n

λ_i^A represents a *main* effect of variable A, or a deviation from a grand mean due to variable A, and it ensures that the expected row marginals equal the observed row marginals.

λ_j^B represents a *main* effect of variable B, or a deviation from a grand mean due to variable B, and it ensures that the expected column marginals equal the observed column marginals.

Log-linear models: Independence model for two-way tables

$$\begin{aligned}
\log odds &= \log \frac{\mu_{ij}}{\mu_{i,j+1}} = \log \mu_{ij} - \log \mu_{i,j+1} \\
&= (\lambda + \lambda_i^A + \lambda_j^B) - (\lambda + \lambda_i^A + \lambda_{j+1}^B) \\
&= \lambda_j^B - \lambda_{j+1}^B
\end{aligned}$$

- odds are functions of model parameters
- for any pair of two categories for one variable we get different odds
- typically comparison with one "base" category

$$\begin{aligned}
\log oddsratio &= \log \frac{\mu_{ij}\mu_{i+1,j+1}}{\mu_{i+1,j}\mu_{i,j+1}} \\
&= \log \mu_{ij} + \log \mu_{i+1,j+1} - \log \mu_{i+1,j} - \log \mu_{i,j+1} \\
&= (\lambda + \lambda_i^A + \lambda_j^B) + (\lambda + \lambda_{i+1}^A + \lambda_{j+1}^B) \\
&\quad -(\lambda + \lambda_{i+1}^A + \lambda_j^B) - (\lambda + \lambda_i^A + \lambda_{j+1}^B) \\
&= 0.
\end{aligned}$$

$$oddsratio = \exp(\log oddsratio) = \exp(0) = 1$$

Log-linear models – independence model, example

Observed		Admitted	
		yes	no
Gender	male	1198	1493
	female	557	1258

Expected under independence		Admitted	
		yes	no
Gender	male	1043	1648
	female	712	1123

```
Call:
glm(formula = Freq ~ Admit + Gender, family = poisson(log), data = UCB.2)

Deviance Residuals:
     1       2       3       4
 4.673  -3.869  -6.025   4.511

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    6.95030    0.02684  258.94   <2e-16 ***
AdmitRejected  0.45674    0.03051   14.97   <2e-16 ***
GenderFemale  -0.38287    0.03027  -12.65   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 486.351  on 3  degrees of freedom
Residual deviance:  93.449  on 1  degrees of freedom
AIC: 134.67

Number of Fisher Scoring iterations: 4
```

Positive coefficient: this category occurs more frequent than the overall average

Negative coefficient: this category occurs less frequent than the overall average

Interpretation: More applicants are rejected than admitted

Less applicants are female than male

Log-linear models: Saturated model for two-way tables

$$\log \mu_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

Parameter estimates and interpretation:

The constant and main effect λs have the same meaning as before.

$\lambda_{ij}^{AB\prime}s$ (1) represent the interaction/association between two variables, (2) reflect the departure from independence, and (3) ensure that $\mu_{ij} = n_{ij}$

Log-linear models – example: Saturated model

		Admitted	
		yes	no
Gender	male	1198	1493
	female	557	1258

```
Call:
glm(formula = Freq ~ Admit * Gender, family = poisson(log), data = UCB.2)

Deviance Residuals:
[1]  0  0  0  0

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                7.08841    0.02889 245.345  < 2e-16 ***
AdmitRejected              0.22013    0.03879   5.675 1.38e-08 ***
GenderFemale              -0.76584    0.05128 -14.933  < 2e-16 ***
AdmitRejected:GenderFemale  0.61035    0.06389   9.553  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4.8635e+02  on 3  degrees of freedom
Residual deviance: 3.6815e-13  on 0  degrees of freedom
AIC: 43.225

Number of Fisher Scoring iterations: 2
```

Odds-ratio of admissions for males vs. females:

1.84 to 1

exp(0.61035) = 1.84

Log-linear models: Saturated model for two-way tables

The saturated model is when:

1. the fitted values are exactly equal to observed values, that is the model fits the data perfectly,
2. $df = 0$, i.e., the number of unique parameters equals the number of cells,
3. this is most complex model,
4. has the independence model as a special case. What does this imply about the assumption for the interaction terms?
5. there is a direct functional relationship with the odds ratio (and the unique number of those).

Model selection: We typically want a simpler model that smoothes the data more, and it's more parsimonious.

Log-linear models: Saturated model for two-way tables

How many unique parameters are there in the model?

Term	# of terms	#of constraints	# of unique parameters
λ	1	0	1
$\{\lambda_i^A\}$	I	1	$I - 1$
$\{\lambda_j^B\}$	J	1	$J - 1$
$\{\lambda_{ij}^{AB}\}$	$I \times J$	$I + J - 1$	$(I - 1) \times (J - 1)$

$I \times J = N$ is a perfect fit!

Log-linear models: Saturated model for two-way tables

The *odds ratio* is directly related to the interaction terms. For example, for a 2 × 2 table:

$$
\begin{aligned}
log(\theta) &= \log(\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}) \\
&= \\
&= \\
&= \lambda_{11}^{AB} + \lambda_{22}^{AB} - \lambda_{12}^{AB} - \lambda_{21}^{AB}
\end{aligned}
$$

How many odds ratios are there? There should be (I-1)x(J-1) which should be equal to the unique number of λ_{ij}'s in the model.

Log-linear models for three-way tables

- complete independence (mutual independence)
- joint independence (partial independence)
- conditional independence

There is a partial hierarchy of models in-between the complete independence model and the saturated model

Log-linear models – complete independece

$$
\begin{aligned}
\mu_{ijk} &= \frac{n_{i..} \cdot n_{.j.} \cdot n_{..k}}{n^2} \\
\log \mu_{ijk} &= \log n_{i..} + \log n_{.j.} + \log n_{..k} - \log n^2 \\
&= \mu + \alpha_i + \beta_j + \gamma_k \\
&= \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C
\end{aligned}
$$

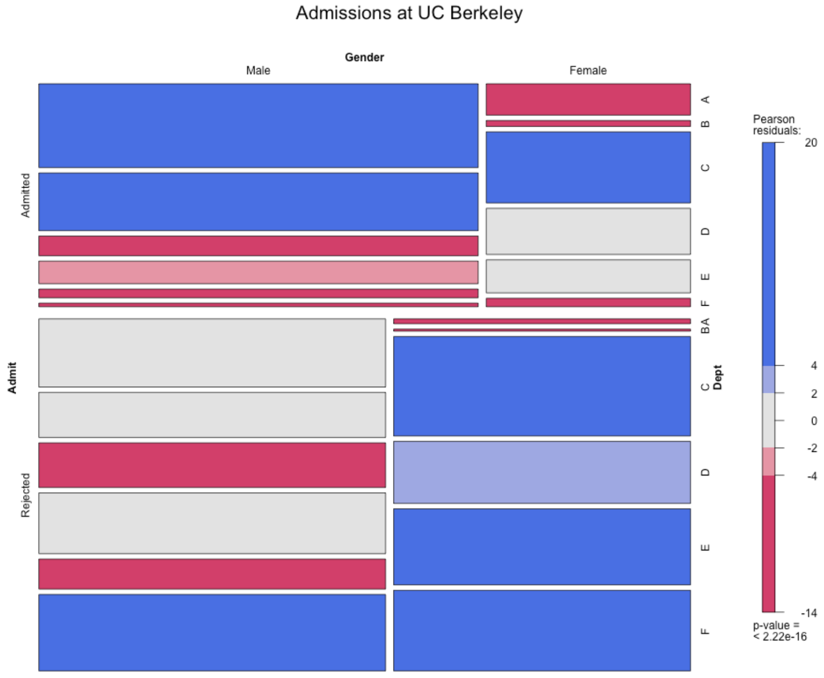
- complete independence (mutual independence)
- there are no associations between the three variables
- expected cell frequencies correspond to the product of marginal frequencies
- straightforward extension of two-way independence model

Log-linear models – example

```
Call:
glm(formula = Freq ~ Admit + Gender + Dept, family = poisson(log),
    data = UCBAdmissions)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-18.170   -7.719   -1.008    4.734   17.153

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    5.37111    0.03964 135.498  < 2e-16 ***
AdmitRejected  0.45674    0.03051  14.972  < 2e-16 ***
GenderFemale  -0.38287    0.03027 -12.647  < 2e-16 ***
DeptB         -0.46679    0.05274  -8.852  < 2e-16 ***
DeptC         -0.01621    0.04649  -0.349 0.727355
DeptD         -0.16384    0.04832  -3.391 0.000696 ***
DeptE         -0.46850    0.05276  -8.879  < 2e-16 ***
DeptF         -0.26752    0.04972  -5.380 7.44e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2650.1  on 23  degrees of freedom
Residual deviance: 2097.7  on 16  degrees of freedom
AIC: 2272.7

Number of Fisher Scoring iterations: 5
```

Complete independence:

Freq ~ Admit + Gender + Dept



Log-linear models – example

```
> indep.table
              Gender  Male Female
Admit    Dept
Admitted A           215.1  146.7
         B           134.9   92.0
         C           211.6  144.3
         D           182.6  124.5
         E           134.6   91.8
         F           164.6  112.2
Rejected A           339.6  231.6
         B           212.9  145.2
         C           334.2  227.9
         D           288.3  196.6
         E           212.6  145.0
         F           259.9  177.2
```

```
> obs.table
              Gender Male Female
Admit    Dept
Admitted A            512     89
         B            353     17
         C            120    202
         D            138    131
         E             53     94
         F             22     24
Rejected A            313     19
         B            207      8
         C            205    391
         D            279    244
         E            138    299
         F            351    317
```

Log-linear models: Saturated model for three-way tables

$$\log \mu_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

- What do all the terms mean in the model?
- Which hypothesis correspond to the models we are already familiar with?
- What are some efficient ways to specify and interpret these models?
- What are some efficient ways to fit and select amaong many possible models?

Log-linear models – example: saturated model

```
Call:
glm(formula = Freq ~ Dept * Admit * Gender, family = poisson(log),
    data = UCBAdmissions)

Deviance Residuals:
 [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0

Coefficients:
                                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                                   6.2383     0.0442  141.16  < 2e-16 ***
Dept[T.B]                                    -0.3719     0.0692   -5.38  7.7e-08 ***
Dept[T.C]                                    -1.4508     0.1014  -14.30  < 2e-16 ***
Dept[T.D]                                    -1.3111     0.0959  -13.67  < 2e-16 ***
Dept[T.E]                                    -2.2680     0.1443  -15.72  < 2e-16 ***
Dept[T.F]                                    -3.1473     0.2177  -14.45  < 2e-16 ***
Admit[T.Rejected]                            -0.4921     0.0717   -6.86  6.9e-12 ***
Gender[T.Female]                             -1.7497     0.1148  -15.24  < 2e-16 ***
Dept[T.B]:Admit[T.Rejected]                  -0.0416     0.1132   -0.37  0.71304
Dept[T.C]:Admit[T.Rejected]                   1.0276     0.1355    7.58  3.3e-14 ***
Dept[T.D]:Admit[T.Rejected]                   1.1961     0.1264    9.46  < 2e-16 ***
Dept[T.E]:Admit[T.Rejected]                   1.4491     0.1768    8.20  2.5e-16 ***
Dept[T.F]:Admit[T.Rejected]                   3.2619     0.2312   14.11  < 2e-16 ***
Dept[T.B]:Gender[T.Female]                   -1.2836     0.2736   -4.69  2.7e-06 ***
Dept[T.C]:Gender[T.Female]                    2.2705     0.1627   13.95  < 2e-16 ***
Dept[T.D]:Gender[T.Female]                    1.6976     0.1675   10.13  < 2e-16 ***
Dept[T.E]:Gender[T.Female]                    2.3227     0.2066   11.24  < 2e-16 ***
Dept[T.F]:Gender[T.Female]                    1.8367     0.3167    5.80  6.7e-09 ***
Admit[T.Rejected]:Gender[T.Female]           -1.0521     0.2627   -4.00  6.2e-05 ***
Dept[T.B]:Admit[T.Rejected]:Gender[T.Female]  0.8321     0.5104    1.63  0.10306
Dept[T.C]:Admit[T.Rejected]:Gender[T.Female]  1.1770     0.2996    3.93  8.5e-05 ***
Dept[T.D]:Admit[T.Rejected]:Gender[T.Female]  0.9701     0.3026    3.21  0.00135 **
Dept[T.E]:Admit[T.Rejected]:Gender[T.Female]  1.2523     0.3303    3.79  0.00015 ***
Dept[T.F]:Admit[T.Rejected]:Gender[T.Female]  0.8632     0.4027    2.14  0.03206 *
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2.6501e+03  on 23  degrees of freedom
Residual deviance: 8.5043e-14  on  0  degrees of freedom
AIC: 207.1

Number of Fisher Scoring iterations: 3
```

Log-linear models – example

```
> sat.table
              Gender Male Female
Admit    Dept
Admitted A            512     89
         B            353     17
         C            120    202
         D            138    131
         E             53     94
         F             22     24
Rejected A            313     19
         B            207      8
         C            205    391
         D            279    244
         E            138    299
         F            351    317
```

```
> obs.table
              Gender Male Female
Admit    Dept
Admitted A            512     89
         B            353     17
         C            120    202
         D            138    131
         E             53     94
         F             22     24
Rejected A            313     19
         B            207      8
         C            205    391
         D            279    244
         E            138    299
         F            351    317
```

Log-linear models for three-way tables

- joint independence (partial independence)
- there is no three-way interaction
- there is only one out of three possible two-way interactions
- (AB,C)
- (A,BC)
- (AC,B)

$$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB}$$

Log-linear models – example

Joint independence model: AD, G
i.e. admission depends on department, but not on gender
i.e. admission and department are jointly independent from gender

```
Call:
glm(formula = Freq ~ Dept * Admit + Gender, family = poisson(log),
    data = UCBAdmissions)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-13.856   -6.238    0.063    5.943    8.810



    Null deviance: 2650.1  on 23  degrees of freedom
Residual deviance: 1242.4  on 11  degrees of freedom
AIC: 1427

Number of Fisher Scoring iterations: 5
```

```
Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                    5.8787     0.0426  138.00  < 2e-16 ***
Dept[T.B]                     -0.4851     0.0661   -7.34  2.1e-13 ***
Dept[T.C]                     -0.6240     0.0691   -9.04  < 2e-16 ***
Dept[T.D]                     -0.8039     0.0734  -10.96  < 2e-16 ***
Dept[T.E]                     -1.4082     0.0920  -15.30  < 2e-16 ***
Dept[T.F]                     -2.5700     0.1530  -16.80  < 2e-16 ***
Admit[T.Rejected]             -0.5935     0.0684   -8.68  < 2e-16 ***
Gender[T.Female]              -0.3829     0.0303  -12.65  < 2e-16 ***
Dept[T.B]:Admit[T.Rejected]    0.0506     0.1097    0.46     0.64
Dept[T.C]:Admit[T.Rejected]    1.2091     0.0973   12.43  < 2e-16 ***
Dept[T.D]:Admit[T.Rejected]    1.2583     0.1015   12.40  < 2e-16 ***
Dept[T.E]:Admit[T.Rejected]    1.6830     0.1173   14.34  < 2e-16 ***
Dept[T.F]:Admit[T.Rejected]    3.2691     0.1671   19.57  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Log-linear models – example

Joint independence model: AD, G

i.e. admission depends on department, but not on gender

i.e. admission and department are jointly independent from gender

```
> ad.table
              Gender  Male Female
Admit     Dept
Admitted A            357.3  243.7
         B            220.0  150.0
         C            191.4  130.6
         D            159.9  109.1
         E             87.4   59.6
         F             27.3   18.7
Rejected A            197.4  134.6
         B            127.8   87.2
         C            354.4  241.6
         D            311.0  212.0
         E            259.8  177.2
         F            397.2  270.8
```

```
> obs.table
              Gender Male Female
Admit     Dept
Admitted A            512     89
         B            353     17
         C            120    202
         D            138    131
         E             53     94
         F             22     24
Rejected A            313     19
         B            207      8
         C            205    391
         D            279    244
         E            138    299
         F            351    317
```

Log-linear models – example

Joint independence model: AD, G

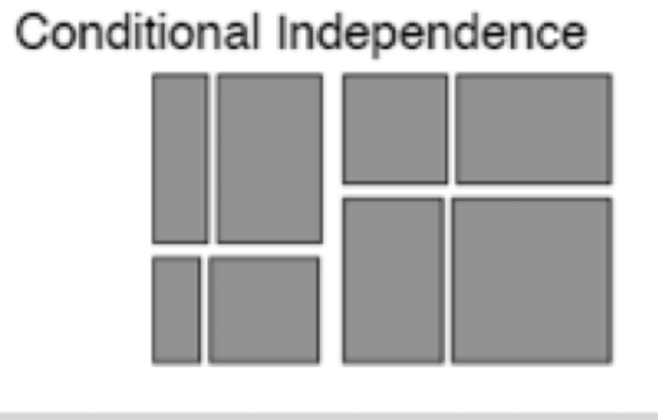
```
> anova(ucb.indep,ucb.ad, ucb.sat, test="Chisq")
Analysis of Deviance Table

Model 1: Freq ~ Dept + Admit + Gender
Model 2: Freq ~ Dept * Admit + Gender
Model 3: Freq ~ Dept * Admit * Gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        16       2098
2        11       1242  5      855   <2e-16 ***
3         0          0 11     1242   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> BIC(ucb.indep,ucb.ad, ucb.sat)
          df  BIC
ucb.indep  8 2282
ucb.ad    13 1443
ucb.sat   24  235
> AIC(ucb.indep,ucb.ad, ucb.sat)
          df  AIC
ucb.indep  8 2273
ucb.ad    13 1427
ucb.sat   24  207
```

Log-linear models for three-way tables

- conditional independence (partial independence)
- there is no three-way interaction
- there are only two out of three possible two-way interactions
- (AB,BC)
- (AC,AB)
- (AC,BC)

$$\log \mu_{ijk} \;=\; \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC}$$



Log-linear models – example

Conditional independence model: AD, DG
i.e. admission depends on department, but not on gender
AND choice of department depends on gender
i.e. admission and gender are conditionally independent given department

```
Call:
glm(formula = Freq ~ Dept * Admit + Dept * Gender, family = poisson(log),
    data = UCBAdmissions)
```

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                     6.2756     0.0425  147.74  < 2e-16 ***
Dept[T.B]                      -0.4057     0.0677   -5.99  2.1e-09 ***
Dept[T.C]                      -1.5394     0.0831  -18.54  < 2e-16 ***
Dept[T.D]                      -1.3223     0.0816  -16.21  < 2e-16 ***
Dept[T.E]                      -2.4028     0.1101  -21.82  < 2e-16 ***
Dept[T.F]                      -3.0962     0.1576  -19.65  < 2e-16 ***
Admit[T.Rejected]              -0.5935     0.0684   -8.68  < 2e-16 ***
Gender[T.Female]               -2.0333     0.1023  -19.87  < 2e-16 ***
Dept[T.B]:Admit[T.Rejected]     0.0506     0.1097    0.46     0.64
Dept[T.C]:Admit[T.Rejected]     1.2091     0.0973   12.43  < 2e-16 ***
Dept[T.D]:Admit[T.Rejected]     1.2583     0.1015   12.40  < 2e-16 ***
Dept[T.E]:Admit[T.Rejected]     1.6830     0.1173   14.34  < 2e-16 ***
Dept[T.F]:Admit[T.Rejected]     3.2691     0.1671   19.57  < 2e-16 ***
Dept[T.B]:Gender[T.Female]     -1.0758     0.2286   -4.71  2.5e-06 ***
Dept[T.C]:Gender[T.Female]      2.6346     0.1234   21.35  < 2e-16 ***
Dept[T.D]:Gender[T.Female]      1.9271     0.1246   15.46  < 2e-16 ***
Dept[T.E]:Gender[T.Female]      2.7548     0.1351   20.39  < 2e-16 ***
Dept[T.F]:Gender[T.Female]      1.9436     0.1268   15.32  < 2e-16 ***
```

```
    Null deviance: 2650.095  on 23  degrees of freedom
Residual deviance:   21.736  on  6  degrees of freedom
AIC: 216.8
```

Log-linear models – example

Conditional independence model: AD, DG
i.e. admission depends on department, but not on gender
AND choice of department depends on gender
i.e. admission and gender are conditionally independent given department

```
> ci.table
             Gender   Male Female
Admit    Dept
Admitted A          531.43  69.57
         B          354.19  15.81
         C          114.00 208.00
         D          141.63 127.37
         E           48.08  98.92
         F           24.03  21.97
Rejected A          293.57  38.43
         B          205.81   9.19
         C          211.00 385.00
         D          275.37 247.63
         E          142.92 294.08
         F          348.97 319.03
```

```
> ci.table-sat.table
             Gender   Male Female
Admit    Dept
Admitted A           19.43 -19.43
         B            1.19  -1.19
         C           -6.00   6.00
         D            3.63  -3.63
         E           -4.92   4.92
         F            2.03  -2.03
Rejected A          -19.43  19.43
         B           -1.19   1.19
         C            6.00  -6.00
         D           -3.63   3.63
         E            4.92  -4.92
         F           -2.03   2.03
```

Log-linear models – example

Conditional independence model: AD, DG
i.e. admission depends on department, but not on gender
AND choice of department depends on gender
i.e. admission and gender are conditionally independent given department

```
> anova(ucb.indep,ucb.ad, ucb.ci, ucb.sat, test="Chisq")
Analysis of Deviance Table

Model 1: Freq ~ Dept + Admit + Gender
Model 2: Freq ~ Dept * Admit + Gender
Model 3: Freq ~ Dept * Admit + Dept * Gender
Model 4: Freq ~ Dept * Admit * Gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        16       2098
2        11       1242  5      855   <2e-16 ***
3         6         22  5     1221   <2e-16 ***
4         0          0  6       22   0.0014 **
```

```
> BIC(ucb.indep,ucb.ad, ucb.ci, ucb.sat)
          df  BIC
ucb.indep  8 2282
ucb.ad    13 1443
ucb.ci    18  238
ucb.sat   24  235
> AIC(ucb.indep,ucb.ad, ucb.ci, ucb.sat)
          df  AIC
ucb.indep  8 2273
ucb.ad    13 1427
ucb.ci    18  217
ucb.sat   24  207
```

Log-linear models for three-way tables

- homogeneous association
- there is no three-way interaction
- all two-way interactions are present
- (AB,AC,BC)

$$\log \mu_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}$$

Log-linear models – example

Homogeneous association model: AD, DG, AG

```
Call:
glm(formula = Freq ~ Dept * Admit + Dept * Gender + Admit * Gender,
    family = poisson(log), data = UCBAdmissions)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                         6.2715     0.0427  146.85  < 2e-16 ***
Dept[T.B]                          -0.4032     0.0678   -5.94  2.8e-09 ***
Dept[T.C]                          -1.5779     0.0895  -17.63  < 2e-16 ***
Dept[T.D]                          -1.3500     0.0853  -15.83  < 2e-16 ***
Dept[T.E]                          -2.4498     0.1176  -20.84  < 2e-16 ***
Dept[T.F]                          -3.1379     0.1617  -19.40  < 2e-16 ***
Admit[T.Rejected]                  -0.5821     0.0690   -8.44  < 2e-16 ***
Gender[T.Female]                   -1.9986     0.1059  -18.87  < 2e-16 ***
Dept[T.B]:Admit[T.Rejected]         0.0434     0.1098    0.40     0.69
Dept[T.C]:Admit[T.Rejected]         1.2626     0.1066   11.84  < 2e-16 ***
Dept[T.D]:Admit[T.Rejected]         1.2946     0.1058   12.23  < 2e-16 ***
Dept[T.E]:Admit[T.Rejected]         1.7393     0.1261   13.79  < 2e-16 ***
Dept[T.F]:Admit[T.Rejected]         3.3065     0.1700   19.45  < 2e-16 ***
Dept[T.B]:Gender[T.Female]         -1.0748     0.2286   -4.70  2.6e-06 ***
Dept[T.C]:Gender[T.Female]          2.6651     0.1261   21.14  < 2e-16 ***
Dept[T.D]:Gender[T.Female]          1.9583     0.1273   15.38  < 2e-16 ***
Dept[T.E]:Gender[T.Female]          2.7952     0.1393   20.07  < 2e-16 ***
Dept[T.F]:Gender[T.Female]          2.0023     0.1357   14.75  < 2e-16 ***
Admit[T.Rejected]:Gender[T.Female] -0.0999     0.0808   -1.24     0.22

    Null deviance: 2650.095  on 23  degrees of freedom
Residual deviance:   20.204  on  5  degrees of freedom
AIC: 217.3

Number of Fisher Scoring iterations: 4
```

Log-linear models – example

Homogeneous association model: AD, DG, AG

```
> ha.table
              Gender   Male Female
Admit    Dept
Admitted A           529.27  71.73
         B           353.64  16.36
         C           109.25 212.75
         D           137.21 131.79
         E            45.68 101.32
         F            22.96  23.04
Rejected A           295.73  36.27
         B           206.36   8.64
         C           215.75 380.25
         D           279.79 243.21
         E           145.32 291.68
         F           350.04 317.96
```

```
> ha.table-sat.table
              Gender    Male  Female
Admit    Dept
Admitted A            17.270 -17.270
         B             0.640  -0.640
         C           -10.755  10.755
         D            -0.793   0.793
         E            -7.319   7.319
         F             0.957  -0.957
Rejected A           -17.270  17.270
         B            -0.640   0.640
         C            10.755 -10.755
         D             0.793  -0.793
         E             7.319  -7.319
         F            -0.957   0.957
```

Log-linear models – example

Homogeneous association model: AD, DG, AG

```
> anova(ucb.indep,ucb.ad, ucb.ci, ucb.ha, ucb.sat, test="Chisq")
Analysis of Deviance Table

Model 1: Freq ~ Dept + Admit + Gender
Model 2: Freq ~ Dept * Admit + Gender
Model 3: Freq ~ Dept * Admit + Dept * Gender
Model 4: Freq ~ Dept * Admit + Dept * Gender + Admit * Gender
Model 5: Freq ~ Dept * Admit * Gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        16       2098
2        11       1242  5      855   <2e-16 ***
3         6         22  5     1221   <2e-16 ***
4         5         20  1        2   0.2159
5         0          0  5       20   0.0011 **
```

```
> BIC(ucb.indep,ucb.ad, ucb.ci, ucb.ha, ucb.sat)
          df  BIC
ucb.indep  8 2282
ucb.ad    13 1443
ucb.ci    18  238
ucb.ha    19  240
ucb.sat   24  235
> AIC(ucb.indep,ucb.ad, ucb.ci, ucb.ha, ucb.sat)
          df  AIC
ucb.indep  8 2273
ucb.ad    13 1427
ucb.ci    18  217
ucb.ha    19  217
ucb.sat   24  207
```

Summary for both parts today

- Generalized linear models
  - Logistic regression
  - Multinomial Logistic Regression
  - Poisson Regression
  - Log-linear models
    - Models of conditional independence
    - Model of homogeneous association
    - Saturated model
    - Hierarchy from independence model to saturated model
    - Aim at finding model that comes close to observed contingency table, but which is also parsimonious

Thanks for your attention. Enjoy the rest of the evening!