Log-linear 3.5 (Model Selection)

Please read Sections 4.1-4.3

Skip 4.4 for now

Likelihood Ratio Test for nested models

- Compare “Full” (unrestricted) & “Reduced” (restricted) models.

- Model 1, usually one in which you really believe. This is the full model. If it has all the terms (saturated), it’s equivalent to an unrestricted multinomial model.
- Model 2: A hierarchical log-linear model whose terms are a *subset* of the ones in Model 1. This is the reduced model. It is Model 1, but with some thing(s) missing.

- Test Model 1 versus 2. Model 2 is null, Model 1 is alternative.

For example

- Model 1: [12] [13] [23]
- Model 2: [12] [23]

- Another Model 2 could be [1] [23]
- Can have a sequence of models, each nested within the last. More later.

Likelihood Ratio Test for Goodness of Fit

$$X_1, \ldots, X_N \overset{i.i.d.}{\sim} F_\theta,\ \theta \in \Theta,$$
$$H_0 : \theta \in \Theta_0 \text{ v.s. } H_A : \theta \in \Theta \cap \Theta_0^c,$$

$$\begin{aligned} G^2 &= -2 \ln \left(\frac{\max_{\theta \in \Theta_0} L(\theta)}{\max_{\theta \in \Theta} L(\theta)} \right) \\ &= 2 \sum (\text{Observed}) \log \left(\frac{\text{Observed}}{\text{Expected}} \right) \end{aligned}$$

Now let Θ_1 be the parameter space under Model 1 and Θ_2 be the parameter space under Model 2: $\Theta_2 \subset \Theta_1 \subset \Theta$.

$$
\begin{aligned}
G^2 &= -2 \ln \left(\frac{\max_{\theta \in \Theta_2} L(\theta)}{\max_{\theta \in \Theta_1} L(\theta)} \right) \\
&= -2 \ln \left(\frac{\max_{\theta \in \Theta_2} L(\theta) / \max_{\theta \in \Theta} L(\theta)}{\max_{\theta \in \Theta_1} L(\theta) / \max_{\theta \in \Theta} L(\theta)} \right) \\
&= G_2^2 - G_1^2 \\
&= 2 \sum (\text{Observed}) \log \left(\frac{\text{Observed}}{\text{Expected}_2} \right) - 2 \sum (\text{Observed}) \log \left(\frac{\text{Observed}}{\text{Expected}_1} \right) \\
&= 2 \sum (\text{Observed}) \log \left(\frac{\text{Expected}_1}{\text{Expected}_2} \right)
\end{aligned}
$$

That's Equation (4.2) in the textbook.

Testing two nested models

- Model 2 is a restricted version of Model 1
- Likelihood ratio test statistic is the difference between the two likelihood ratio tests for goodness of fit: $G^2 = G^2_2 – G^2_1$
- G^2_2 is always bigger because the model is more restricted.
- Asymptotically chisquare, $df = df_2 – df_1$

Nested hierarchy of models

A. [1] [2] [3]
B. [2] [13]
C. [12] [13]
D. [12] [13] [23]
E. [123]

$G^2(a) \geq G^2(B) \geq G^2(C) \geq G^2(D) \geq G^2(E)=0$

Different hierarchies are possible.

$G^2(A) \geq G^2(B) \geq G^2(C) \geq G^2(D) \geq G^2(E)=0$

$$\begin{aligned} G^2(A) &= G^2(A) - G^2(B) \\ &+ G^2(B) - G^2(C) \\ &+ G^2(C) - G^2(D) \\ &+ G^2(D) - G^2(E) \end{aligned}$$

$$\text{And } G^2(D) - G^2(E) = G^2(D)$$

"Partitioning" of chisquare.

There is no single best way to discover a good model

- Our text's approach: Plan a hierarchy in advance and work your way down.
- Forward stepwise (automatic, or not)
- Backward stepwise (automatic, or not)
- Exploration: Discover a good hierarchy, looking at the data as well as testing
- Other possibilities …

An approach to model building

- First test fit of the model of complete independence. If the null hypothesis cannot be rejected at alpha = 0.05, give up and go home.
- Next, try testing the fit of a model with only 2-factor interactions – that is, only pairwise associations between variables. The author of our text, who knows a lot, suggests that this will often be good enough. If it fits, a lot of complications can be ruled out.

If the model with all 2-variable associations fits

- Start adding relationships between variables to the model, Beginning with the strongest or most obvious. Consider each marginal 2-way table, and test with an X^2 or G^2 test of independence. **Look at the table** (compute row, column proportions or percents) and decide what seems to be going on. It is often helpful to look at sub-tables, too.
- Each time a relationship (2-factor interaction) is added,
  - Test against the preceding model: Is it an improvement?
  - Test overall fit

This does not cover all the possibilities

But let’s look at an example

Florida Prison Data

```
> Prace <- factor(florida$Prace, labels=c('White','Black')) # In order 1,2
> Vrace <- factor(florida$Vrace, labels=c('White','Black'))
> DeathPen <- factor(florida$DeathPen, labels=c('Yes','No'))
> PR_by_DP = table(Prace, DeathPen); PR_by_DP
       DeathPen
Prace   Yes  No
  White  19 141
  Black  17 149

> prop.table(PR_by_DP,1) # Row proportions
       DeathPen
Prace         Yes        No
  White 0.1187500 0.8812500
  Black 0.1024096 0.8975904
> round(100*prop.table(PR_by_DP,1),2) # Row percentages
       DeathPen
Prace     Yes    No
  White 11.88 88.12
  Black 10.24 89.76
```

```
> chisq.test(PR_by_DP,correct=F)

Pearson's Chi-squared test

data:  PR_by_DP
X-squared = 0.2214, df = 1, p-value = 0.638

> dp <- table(Prace, DeathPen, Vrace); dp
, , Vrace = White

       DeathPen
Prace   Yes  No
  White  19 132
  Black  11  52

, , Vrace = Black

       DeathPen
Prace   Yes  No
  White   0   9
  Black   6  97
```

Something interesting may be going on

```
> # Row percents
> round(100*prop.table(dp[,,1],1),2)
        DeathPen
Prace       Yes    No
  White   12.58 87.42
  Black   17.46 82.54
> round(100*prop.table(dp[,,2],1),2)
        DeathPen
Prace      Yes     No
  White   0.00 100.00
  Black   5.83  94.17
```

Prace and Deathpen CONTROLLING for (conditional upon) Vrace

Chisquare tests on sub-tables

```
> # Pearson
> chisq.test(dp[,,1],correct=F)

Pearson's Chi-squared test

data:  dp[, , 1]
X-squared = 0.8774, df = 1, p-value = 0.3489

> chisq.test(dp[,,2],correct=F)

Pearson's Chi-squared test

data:  dp[, , 2]
X-squared = 0.5539, df = 1, p-value = 0.4567

Warning message:
Chi-squared approximation may be incorrect in:
chisq.test(dp[, , 2], correct = F)
```

What's the problem? Look at expected frequencies.

```
> loglin(dp[,,2],margin=list(1,2),fit=T)$fit
2 iterations: deviation 1.421085e-14
       DeathPen
Prace          Yes        No
  White 0.4821429  8.517857
  Black 5.5178571 97.482143
```

Low expected frequencies tend to inflate chisquare.

No problem here.

Complete Independence

```
> ind <- loglin(dp,list(1,2,3)); ind
2 iterations: deviation 2.842171e-14
$lrt
[1] 137.9294

$pearson
[1] 122.3975

$df
[1] 4

$margin
$margin[[1]]
[1] "Prace"

$margin[[2]]
[1] "DeathPen"

$margin[[3]]
[1] "Vrace"
```

Model with all 2-factor relationships

```
> twoways <- loglin(dp,list(c(1,2),c(1,3),c(2,3))); twoways
5 iterations: deviation 0.05215771
$lrt
[1] 0.7007595

$pearson
[1] 0.3750283

$df
[1] 1

$margin
$margin[[1]]
[1] "Prace"    "DeathPen"

$margin[[2]]
[1] "Prace" "Vrace"

$margin[[3]]
[1] "DeathPen" "Vrace"
```

How is G^2 being calculated?!

```
, , Vrace = White

       DeathPen
Prace   Yes  No
  White  19 132
  Black  11  52

, , Vrace = Black

       DeathPen
Prace   Yes  No
  White   0   9
  Black   6  97
```

$$G^2 = 2\sum(\text{Observed})\log\left(\frac{\text{Observed}}{\text{Expected}}\right)$$

Zero cell is being dropped

- Conservative, for a test of fit. Chisquare is smaller, so it's less likely to force you to a more complicated model.
- Add a small constant to the observed frequency of zero, just for computing G^2, not for computing the expected frequencies. How small? The smaller the better.

$$\lim_{x \to 0} \left(x \log \frac{x}{\text{Expected}} \right) = 0$$

- No effect on LR tests of nested models.

$$G^2_{1,2} = 2 \sum (\text{Observed}) \log \left(\frac{\text{Expected}_1}{\text{Expected}_2} \right)$$

Look at 2-factor marginal tables

- Prisoner's race by death penalty: Consistent with no relationship.
- Prisoner's race by victim's race: Strong, we think.
- Victim's race by death penalty: Need to check it.

Prisoner’s Race and Victim's Race

```
> PR_by_VR = table(Prace, Vrace); PR_by_VR
       Vrace
Prace   White Black
  White   151     9
  Black    63   103
> round(100*prop.table(PR_by_VR,1),2) # Row percentages
       Vrace
Prace   White Black
  White 94.38  5.62
  Black 37.95 62.05
> chisq.test(PR_by_VR,correct=F)

Pearson’s Chi-squared test

data:  PR_by_VR
X-squared = 115.0083, df = 1, p-value < 2.2e-16
```

People tend to be in jail for killing someone of their own race.
Anything else interesting?

Victim's Race and Death Penalty

```
> VR_by_DP = table(Vrace, DeathPen); VR_by_DP
       DeathPen
Vrace   Yes  No
  White  30 184
  Black   6 106
> round(100*prop.table(VR_by_DP,1),2) # Row percentages
       DeathPen
Vrace     Yes    No
  White 14.02 85.98
  Black  5.36 94.64
> chisq.test(VR_by_DP,correct=F)

Pearson's Chi-squared test

data:  VR_by_DP
X-squared = 5.6149, df = 1, p-value = 0.01781
```

Suggests death penalty more likely if victim is White

It look like we want to add [PR, VR], but marginal tables can be misleading – See Section 3.8. Choose model with smallest G^2 (best fit)

```
> # 1=Prace, 2=DeathPen, 3=Vrace)
> loglin(dp,list(2,c(1,3)))$lrt # [DP] [PR, VR]
2 iterations: deviation 0
[1] 8.131611
> loglin(dp,list(1,c(2,3)))$lrt # [PR] [VR, DP]
2 iterations: deviation 0
[1] 131.6796
> loglin(dp,list(3,c(1,2)))$lrt # [VR] [PR, DP]
2 iterations: deviation 0
[1] 137.7079
```

[DP] [PR, VR] is the best choice, by far

- Is it an improvement?
- Does it fit?

```
> ModelA = ind
> ModelB <- loglin(dp,list(2,c(1,3)))
2 iterations: deviation 0
> #   Is it an improvement?
> G2Change = ModelA$lrt-ModelB$lrt; G2Change
[1] 129.7977
> dfChange = ModelA$df-ModelB$df; dfChange
[1] 1
> pvalChange = 1-pchisq(G2Change, df=dfChange)
> pvalChange
[1] 0
```

Does it fit?

```
> #   Does it fit?
> G2B = ModelB$lrt; G2B
[1] 8.131611
> dfB = ModelB$df; dfB
[1] 3
> pvalB = 1-pchisq(G2B, df=dfB); pvalB
[1] 0.04336859
> ModelB$pearson; 1-pchisq(ModelB$pearson,df=ModelB$df)
[1] 6.977343
[1] 0.07262343
```

I say we proceed, but there could be argument.

Add another association
Compare [PR,VR][PR,DP] with [PR,VR][VR,DP]

```
> # 1=Prace, 2=DeathPen, 3=Vrace
> loglin(dp,list(c(1,3),c(1,2)))$lrt # [PR,VR][PR,DP]
2 iterations: deviation 0
[1] 7.91016
> loglin(dp,list(c(1,3),c(2,3)))$lrt # [PR,VR][VR,DP]
2 iterations: deviation 1.421085e-14
[1] 1.881895
```

Choose [PR,VR][VR,DP]

```
> ModelC <- loglin(dp,list(c(1,3),c(2,3)))
2 iterations: deviation 1.421085e-14
> #   Is it an improvement?
> G2Change = ModelB$lrt-ModelC$lrt; G2Change
[1] 6.249715
> dfChange = ModelB$df-ModelC$df; dfChange
[1] 1
> pvalChange = 1-pchisq(G2Change, df=dfChange)
> pvalChange
[1] 0.01242133
> #   Does it fit?
> G2C = ModelC$lrt; G2C
[1] 1.881895
> dfC = ModelC$df; dfC
[1] 2
> pvalC = 1-pchisq(G2C, df=dfC); pvalC
[1] 0.3902578
```

Does it help to add [PR,DP]?

```
> ModelD <- twoways
> G2Change = ModelC$lrt-ModelD$lrt; G2Change
[1] 1.181136
> dfChange = ModelC$df-ModelD$df; dfChange
[1] 1
> pvalChange = 1-pchisq(G2Change, df=dfChange)
> pvalChange
[1] 0.2771249
```

Hierarchy: Not planned in advance

	Fit			Change		
Model	Chisq	df	p	Chisq	df	p
[VR] [PR] [DP]	137.93	4	0.00			
[DP] [VR,PR]	8.13	3	0.07	129.80	1	0.00
[VR,PR] [VR,DP]	1.88	2	0.39	6.25	1	0.01
[VR,PR] [VR,DP] [PR,DP]	0.70	1	0.40	1.18	1	0.28

Model is [VR,PR] [VR,DP]

- Hierarchy of models was the result of exploring the data
- Kind of forward stepwise method, could be automated
- Guided by hypothesis tests, but please don't take them completely at face value. We did quite a few tests, and the theory applies to single tests performed in isolation.

Describe the findings in words

- Prisoners in jail for murder in Florida tended to be convicted of killing people of the same race.
- The death penalty was less likely when the victim was Black.

(These conclusions are based on looking at the marginal 2-way tables. Let's check the parameter estimates too.)

Checking the parameter estimates

Just part of the output

```
> loglin(dp,list(c(1,3),c(2,3)),param=T)$param
$Prace.Vrace
       Vrace
Prace        White      Black
  White  0.8279124 -0.8279124
  Black -0.8279124  0.8279124

$DeathPen.Vrace
        Vrace
DeathPen      White      Black
     Yes  0.2644853 -0.2644853
     No  -0.2644853  0.2644853
```

- Prace.Vrace interaction says increased chance of White-White and Black-Black
- DeathPen.Vrace interaction says increased chance of Yes-White and No-Black

A little more about the interpretation of [VR,PR] [VR,DP]

- It's a model of conditional independence
- Allowing (controlling) for Victim's Race, Prisoner's Race is unrelated to Death Penalty
- Model says that in each sub-table (VR=Black, VR=White), Prisoner's Race is independent of Death Penalty.
- So the test of model fit should be like a combined test of independence for both 2-way tables.

$$H_0 : \mu_{12} = \mu_{123} = 0$$

Had $G^2 = 1.88$, $df=2$, $p = 0.39$

$$H_0 : \mu_{12} = \mu_{123} = 0$$

```
> dp
, , Vrace = White

       DeathPen
Prace    Yes  No
  White   19 132
  Black   11  52

, , Vrace = Black

       DeathPen
Prace    Yes  No
  White    0   9
  Black    6  97
```

```
> a = loglin(dp[,,1],margin=list(1,2))$lrt; a
2 iterations: deviation 0
[1] 0.847478
> b = loglin(dp[,,2],margin=list(1,2))$lrt; b
2 iterations: deviation 1.421085e-14
[1] 1.034417
> a+b
[1] 1.881895
```

Control by sub-division: Very natural.

Works for Pearson X^2 too.

The lesson

- Want to examine the relationship between *A* and *B*, but *A* might be related to *C* and *B* might be related to *C*.
- So look at the relationship between *A* and *B* controlling for *C*.
- Examine (test) A by B separately for each level of C: Sub-division.
- Pool (combine) the tests by adding chi-squares and adding degrees of freedom.
- *Identical* to the chi-square test for fit of a log-linear model of conditional independence!