STAT 518 --- Section 4.7 --- Loglinear Models and Other Approaches

- Many tests for contingency tables use the "Pearson's Chi-square Statistic":

$$T_1 = \sum_{\text{all cells}} \frac{(O_i - E_i)^2}{E_i}$$

- An alternative approach uses the "Likelihood Ratio Chi-square Statistic":

$$T_2 = 2 \sum_{\text{all cells}} O_i \ln\left(\frac{O_i}{E_i}\right)$$

- The LR statistic also has an asymptotic χ^2 distribution, with the same degrees of freedom as Pearson's statistic.

- An advantage of the Pearson test statistic is that its asymptotic χ^2 distribution tends to be valid with smaller sample sizes (i.e., when $N/rc > 1$) than the χ^2 approximation for the LR statistic (which holds well when $N/rc > 5$).

Loglinear Models

- This is a common method of analyzing contingency tables of more than two dimensions.

- In a 2 × 2 table, the null hypothesis of independence between dimensions is equivalent to

$$H_0: p_{ij} = p_{i+} p_{+j} \quad \text{for all } i, j$$

where p_{i+} = the marginal probability of being in row i

and p_{+j} = the marginal probability of being in column j.

• **Taking logarithms of both sides, we get:**

$$H_0: \ln(p_{ij}) = \ln(p_{i+}) + \ln(p_{+j})$$

which is a log linear **model.**

<u>Recall</u>: Our expected cell count under independence is

$$E_{ij} = \frac{n_{i+}n_{+j}}{N}$$

where n_{i+} = total count of observations in row i

and n_{+j} = total count of observations in column j

• **Thus for a 2 × 2 table,**

$$E_{11} = \frac{n_{1+}n_{+1}}{N}, \quad E_{12} = \frac{n_{1+}n_{+2}}{N}, \quad E_{21} = \frac{n_{2+}n_{+1}}{N}, \quad E_{22} = \frac{n_{2+}n_{+2}}{N}$$

and so we have

$$E_{11}E_{22} = \frac{n_{1+}n_{+1}n_{2+}n_{+2}}{N^2} = E_{12}E_{21} \Rightarrow \frac{E_{11}E_{22}}{E_{12}E_{21}} = 1$$

• **This fraction** $\frac{E_{11}E_{22}}{E_{12}E_{21}}$ **is called the <u>odds ratio</u>.**

It is defined as $\dfrac{P(\text{Row }1)/P(\text{Row }2) \text{ when Column is } 1}{P(\text{Row }1)/P(\text{Row }2) \text{ when Column is } 2}$

$$= \frac{\frac{E_{11}}{n_{+1}} \Big/ \frac{E_{21}}{n_{+1}}}{\frac{E_{12}}{n_{+2}} \Big/ \frac{E_{22}}{n_{+2}}} = \frac{E_{11}/E_{21}}{E_{12}/E_{22}} = \frac{E_{11}E_{22}}{E_{12}E_{21}}$$

• **Now, if we instead have <u>dependence</u> between dimensions, that implies:**

odds ratio $\frac{E_{11}E_{22}}{E_{12}E_{21}} = k$, for some $k \neq 1$.

• **Writing the loglinear model in terms of the cell counts rather than cell probabilities, we have:**

$\ln E_{ij} = \lambda + \alpha_i + \beta_j$ **under independence**

$\ln E_{ij} = \lambda + \alpha_i + \beta_j + (\alpha\beta)_{ij}$ **under dependence**

• **These model parameters are estimated using software via iterative methods.**

• **Using the estimates, we can get fitted values $\hat{E}_{ij}$ for each cell.**

• **We then use either the Pearson statistic or the LR statistic to determine (with a χ^2 test) whether the model provides a good fit. H_0:** model is a good fit

Three-Way Tables

• **This is most useful in cases where the data are classified according to three categorical variables.**

Example 1 (2 × 2 × 2 table):

Alcohol = Yes

		Marijuana Yes	Marijuana No
Cigarette	Yes	911	538
	No	44	456

Alcohol = No

		Marijuana Yes	Marijuana No
Cigarette	Yes	3	43
	No	2	279

Possible loglinear models for 2 × 2 × 2 tables:

1. $\ln E_{ijk} = \lambda + \alpha_i + \beta_j + \gamma_k$
2. $\ln E_{ijk} = \lambda + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}$

$\vdots$

3. $\ln E_{ijk} = \lambda + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$
4. $\ln E_{ijk} = \lambda + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$

Example 1: Let i = 1, 2 be the level of Cigarette Use (Yes/No); let j = 1, 2 be the level of Marijuana Use; let k = 1, 2 be the level of Alcohol Use.

- Under Model 1, cigarette use, marijuana use, and alcohol use are all mutually independent → no interaction
- Under Model 2, cigarette use and marijuana use are dependent, but alcohol use is jointly independent of the other two variables.
- Under Model 3, all three variables are conditionally dependent, but the amount of dependence between two variables does not depend on the level of the third → first-order interaction
- Under Model 4, each pair of variables is conditionally dependent, and the amount of dependence between a pair varies across the levels of the third variable → second-order interaction

• The model that includes all possible parameters is called the Saturated **model.**

• The `loglm` function in the `MASS` library in R estimates the parameters of any of these models, calculates the fitted values, and performs the χ^2 tests for fit.

• In addition, the `step` function evaluates these possible models based on Akaike's Information Criterion (AIC).

Example 1 Possible Questions of Interest:

- **Do the odds of a cigarette smoker using marijuana differ from the odds of a cigarette non-smoker using marijuana? →** Is there dependence between cigarette use and marijuana use?

- **Does the value of this odds ratio depend on alcohol use? →** Does the amount of this dependence vary for drinkers and non-drinkers?

Analysis in R:

- **The best model appears to be** First-order interaction model

$$\ln E_{ijk} = \lambda + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$$

Pearson = 0.40
LR = 0.37

Both less than $\chi^2_{.95,1} = 3.84$
$\longrightarrow$ Conclude good fit

- **Example of fitted value calculation using estimated coefficients:** $\hat{E}_{112} \rightarrow$ "Cigarette = Yes, Mariju = Yes, Alcohol = No"

From estimated coefficients: $\ln \hat{E}_{112} = 4.252 + 0.282 - 1.196 - 1.504 + 0.712 - 0.514 - 0.747 = 1.285$

$$\hat{E}_{112} = e^{1.285} \approx 3.615$$

The other fitted values are obtained similarly.

- **Interpretation of results is best done using odds ratios:**

Note $\dfrac{\hat{E}_{111}\,\hat{E}_{221}}{\hat{E}_{121}\,\hat{E}_{211}} = 17.25$ and $\dfrac{\hat{E}_{112}\,\hat{E}_{222}}{\hat{E}_{122}\,\hat{E}_{212}} = 17.25$ also.

$\Rightarrow$ The odds that a cigarette smoker has used marijuana are 17.25 times the odds that a cigarette non-smoker has used marijuana. This odds ratio does <u>not</u> depend on whether the person has used alcohol.

Example 2 (2 × 2 × 2 table):

		Length = Long		Length = Short	
		Planting Time		P. Time	
		Early	Late	Early	Late
Survival	Alive	156	84	107	31
	Dead	84	156	133	209

Example 2 Possible Questions of Interest:

- **Do the odds of an early plant surviving differ from the odds of a late plant surviving?** → Is there dependence between planting time and survival?

- **Does the value of this odds ratio depend on the cutting length?** → Does the amount of this dependence vary for long and short cuttings?

<u>Analysis in R</u>:

- **The search for the best model:** The "step" function suggests the saturated model is best based on AIC, but the first-order interaction model is sufficient based on a χ^2 test:

Pearson: 2.27 → Both less than $\chi^2_{.95,1} = 3.84$
LR: 2.29

→ Conclude a good fit.

- **Interpretation of results via odds ratios:**

Note $\dfrac{\hat{E}_{111}\hat{E}_{221}}{\hat{E}_{121}\hat{E}_{211}} = 4.168 = \dfrac{\hat{E}_{112}\hat{E}_{222}}{\hat{E}_{122}\hat{E}_{212}}$. The odds of survival for an early plant are 4.168 times the odds of survival for a late plant. This odds ratio <u>does</u> <u>not</u> <u>depend</u> on cutting length.

Example 3 (2 × 2 × 4 table): After the sinking of the Titanic, a study classified passengers according to Survival Status (Yes/No), Sex (Male/Female), and Class (1st/2nd/3rd/Crew). We adapt a built-in R data set.

		1st Class		2nd Class		3rd Class		Crew	
		Sex		Sex		Sex		Sex	
		M	F	M	F	M	F	M	F
Survival	Yes	62	141	25	93	88	90	192	20
	No	118	4	154	13	422	106	670	3

Example 3 Possible Questions of Interest:

- **Do the odds of a female surviving differ from the odds of a male surviving? →** Is there dependence between sex and survival?

- **Does the value of this odds ratio depend on the class of the passenger? →** Does the amount of this dependence vary across the 4 classes?

Analysis in R:

- **The search for the best model:** Saturated model is best. The first-order interaction model has: Pearson $= 60.87$, $LR = 65.17$, both greater than $\chi^2_{.95,3} = 7.815$ → First-order interaction model is not a good fit. → Need saturated model.
- **Interpretation of results via odds ratios:**

Class = 1: $OR = \dfrac{\hat{E}_{111}\hat{E}_{221}}{\hat{E}_{121}\hat{E}_{211}} = 67.09$ → The odds of survival for a first-class female are 67.09 times the odds of survival for a first-class male.

Class = 2: $OR = 44.07$

Class = 3: $OR = 4.07$ → The odds of survival for a 3rd-class female are 4.07 times the odds of survival for 3rd-class male.

Class = 4: $OR = 23.26$

– We see this odds ratio does depend on the level of class (second-order interaction)