Conditional Independence in Log-linear Models[1]

STA 312: Fall 2012

[1]See last slide for copyright information.

Overview

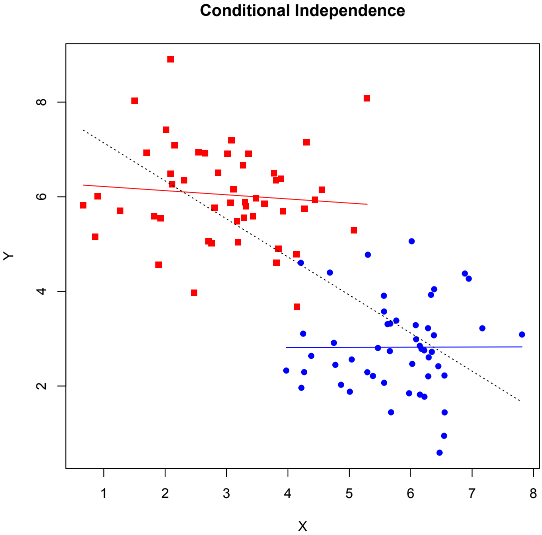
Conditional independence: $(XZ)\ (YZ)$

- X is related to Z and Y is related to Z.
- This may cause a relationship in the *marginal* X by Y table.
- But it only happens because both X and Y are related to Z.
- For each *fixed* value of Z, X and Y are independent.
- Z is a kind of confounding variable.
- Want to test the relationship of X and Y *controlling* for Z.

Example

- X is body weight, above vs. below the median.
- Y is amount of smiling in a standard conversation, above vs. below the median.
- We find that heavier people tend to smile less.
- Does this mean that on average, fat people are *not* jolly?
- Z is Gender, Male vs. Female.

It can happen in ordinary regression, too



More ways to think about (XZ) (YZ)

- Sub-tables
- Path diagrams
- Formal log-linear models

Sub-tables

One for each category of Z

$Z = 1$

		Y 1	Y 2
X	1		
X	2		

$Z = 2$

		Y 1	Y 2
X	1		
X	2		

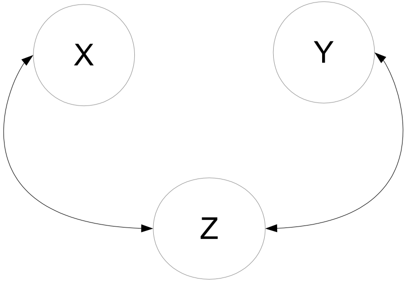
$Z = 3$

		Y 1	Y 2
X	1		
X	2		

- Control by sub-division: Hold Z constant.
- Test independence for each fixed value of Z
- Conditional independence says that within each sub-table, X and Y are independent.
- Again, X could be related to Z and Y could be related to Z,
- Causing a relationship between in the X by Y table of marginal totals.

Path diagrams: X and Y are connected only through Z

$(XZ)\ (YZ)$



- Path diagrams can clarify relationships among variables.
- Not as useful when there are 3-way and higher relationships.

Log-linear models of conditional independence

Equivalent to (XZ) (YZ)

$$\begin{aligned} \mu_{ijk} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \\ &= \beta_0 + \beta_1 x + \beta_2 y + \beta_3 z + \beta_4 xz + \beta_5 yz \end{aligned}$$

X	Y	Z	x	y	z	xz	yz
1	1	1	1	1	1	1	1
1	2	1	1	-1	1	1	-1
2	1	1	-1	1	1	-1	1
2	2	1	-1	-1	1	-1	-1
1	1	2	1	1	-1	-1	-1
1	2	2	1	-1	-1	-1	1
2	1	2	-1	1	-1	1	-1
2	2	2	-1	-1	-1	1	1

$\mu_{ijk} = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 z + \beta_4 xz + \beta_5 yz$
(XZ) (YZ)

X	Y	Z	x	y	z	xz	yz
1	1	1	1	1	1	1	1
1	2	1	1	-1	1	1	-1
2	1	1	-1	1	1	-1	1
2	2	1	-1	-1	1	-1	-1
1	1	2	1	1	-1	-1	-1
1	2	2	1	-1	-1	-1	1
2	1	2	-1	1	-1	1	-1
2	2	2	-1	-1	-1	1	1

		$Y=1$	$Y=2$
X	1	$\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5$	$\beta_0 + \beta_1 - \beta_2 + \beta_3 + \beta_4 - \beta_5$
	2	$\beta_0 - \beta_1 + \beta_2 + \beta_3 - \beta_4 + \beta_5$	$\beta_0 - \beta_1 - \beta_2 + \beta_3 - \beta_4 - \beta_5$

		$Y=1$	$Y=2$
X	1	$\beta_0 + \beta_1 + \beta_2 - \beta_3 - \beta_4 - \beta_5$	$\beta_0 + \beta_1 - \beta_2 - \beta_3 - \beta_4 + \beta_5$
	2	$\beta_0 - \beta_1 + \beta_2 - \beta_3 + \beta_4 - \beta_5$	$\beta_0 - \beta_1 - \beta_2 - \beta_3 + \beta_4 + \beta_5$

Odds ratio for $Z = 1$

		$Y = 1$	$Y = 2$
X	1	$\exp\{\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5\}$	$\exp\{\beta_0 + \beta_1 - \beta_2 + \beta_3 + \beta_4 - \beta_5\}$
	2	$\exp\{\beta_0 - \beta_1 + \beta_2 + \beta_3 - \beta_4 + \beta_5\}$	$\exp\{\beta_0 - \beta_1 - \beta_2 + \beta_3 - \beta_4 - \beta_5\}$

$$\theta_1 = \frac{\exp\{2\beta_0 + 2\beta_3\}}{\exp\{2\beta_0 + 2\beta_3\}} = 1$$

Odds ratio for $Z = 2$

		$Y = 1$	$Y = 2$
X	1	$\beta_0 + \beta_1 + \beta_2 - \beta_3 - \beta_4 - \beta_5$	$\beta_0 + \beta_1 - \beta_2 - \beta_3 - \beta_4 + \beta_5$
	2	$\beta_0 - \beta_1 + \beta_2 - \beta_3 + \beta_4 - \beta_5$	$\beta_0 - \beta_1 - \beta_2 - \beta_3 + \beta_4 + \beta_5$

$$\theta_2 = \frac{\exp\{2\beta_0 - 2\beta_3\}}{\exp\{2\beta_0 - 2\beta_3\}} = 1$$

So no matter how you look at it

- (XZ) (YZ) means X and Y are independent, conditionally on Z
- Of course there may be more than 3 variables.
- Conditional independence may be embedded in a larger structure.

Two ways to test conditional independence

- Test in sub-tables
- Test using log-linear models

Testing conditional independence using sub-tables

	$Z=1$				$Z=2$				$Z=3$		
		Y				Y				Y	
		1	2			1	2			1	2
X	1			X	1			X	1		
	2				2				2		

- One table for each value of the control variable (or variables).
- Do a separate test on each table and combine them somehow.
- This is a reasonable, common-sense approach.
- Sum of independent chi-squares is chi-squared.
- So add the χ^2 values and add the degrees of freedom.
- This was illustrated on the Berkeley graduate admissions data.

Test relationship of X to Y controlling for Z using log-linear models

- H_0 is conditional independence: $(XZ)\ (YZ)$
- One alternative hypothesis is $(XY)\ (XZ)\ (YZ)$
- Another is (XYZ)
- The last model is saturated, so when this is the alternative, the test is a test of goodness of fit for the model of conditional independence.
- I would do them both; maybe test goodness of fit first.

(Remember that for *any* table, the log-linear model with the highest-order interaction is saturated, and equivalent to an unrestricted multinomial.)

Connection between the methods for testing conditional independence

Sub-tables and log-linear models

- Adding chi-squares and degrees of freedom for sub-tables gives *exactly* the test of (XZ) (YZ) against the saturated model
- Likelihood ratio or Pearson
- Test is a simultaneous test of $H_0 : \lambda_{ij}^{XY} = \lambda_{ijk}^{XYZ} = 0$
- Degrees of freedom should be $K(I-1)(J-1)$, adding *df* in sub-tables.
- Or counting the missing product terms,

$$
\begin{aligned}
& (I-1)(J-1) + (I-1)(J-1)(K-1) \\
= \; & (I-1)(J-1)\,(K-1+1) \\
= \; & K(I-1)(J-1)
\end{aligned}
$$

Copyright Information