Chapter 11: Analysis of matched pairs

Adapted from Timothy Hanson

Department of Statistics, University of South Carolina

Stat 770: Categorical Data Analysis

Chapter 11: Models for Matched Pairs

Example: Prime minister approval (PMA) data. $n_{++} = 1600$ voting age British citizens were asked if they approved of the Prime Minister. The same people were asked again 6 months later. The 1600 are cross classified according to their two (binary) responses (X, Y):

First survey	Second survey	
	Approve	Disapprove
Approve	794	150
Disapprove	86	570

Here, each person is matched with his or her self. This is also called *repeated measures* data.

Here we see people tend to approve both times or disapprove both times more often than change their opinion. Question: of those that change their opinion, which direction do they tend to go? Hint: $150 > 86$.

Polygraph Tests

Example: Fort Jackson hosts the Polygraph Institute, used to refine the study of "lie detectors". One of their studies of n=100 subjects included a cross-classification of

- Three judges
- Four tests
  - Two one-stage
  - Two two-stage (objective methods based on judges' scores)
- Three opinions (Truthful, Deceptive, No Opinion)

None of the tests were "gold standards", so the truth was unknown. The full data set is extremely sparse, but tests of marginal homogeneity have rescued many a sparse data set.

Let $\pi_{ab} = P(X = a, Y = b)$ and n_{ab} be the number of such pairs.

	Second survey	
First survey	Approve $Y = 1$	Disapprove $Y = 2$
Approve $X = 1$	π_{11} & n_{11}	π_{12} & n_{12}
Disapprove $X = 2$	π_{21} & n_{21}	π_{22} & n_{22}

We assume $(n_{11}, n_{12}, n_{21}, n_{22}) \sim \text{mult}\{n_{++}, (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})\}$.

When $\pi_{1+} = \pi_{+1}$ then $P(X = 1) = P(Y = 1)$ and we have
marginal homogeneity. This is of course equivalent to
$P(X = 2) = P(Y = 2)$–by looking at complementary events–and
$P(X = 1, Y = 2) = P(X = 2, Y = 1)$.

In the prime minister approval data, this would indicate that the
proportion of people that approve at time zero is equal to the
proportion that approve at 6 months. Does it imply that no one
has changed their mind?

Let $p_{ab} = n_{ab}/n_{++}$ be the sample proportion in each cell.

How does the proportion change?

Define the difference $\delta = \pi_{+1} - \pi_{1+} = P(Y = 1) - P(X = 1)$.
What does this measure for the prime minister approval data?

δ is estimated by

$$d = p_{+1} - p_{1+} = \frac{n_{11} + n_{21} - (n_{11} + n_{12})}{n_{++}} = \frac{n_{21} - n_{12}}{n_{++}}.$$

Considering the covariance for multinomial vector elements, we have the following $(1 - \alpha)100\%$ CI for δ

$$d \pm z_{\alpha/2}\hat{\sigma}(d),$$

where

$$\hat{\sigma}(d) = \sqrt{[(p_{12} + p_{21}) - (p_{12} - p_{21})^2]/n}.$$

11.1.2 Formal test of $H_0 : \delta = 0$

To test $H_0 : \delta = 0$, i.e. $H_0 : P(X = 1) = P(Y = 1)$, the Wald test statistic is $z_0 = d/\hat{\sigma}(d)$. The score test statistic is

$$z_0 = \frac{n_{21} - n_{12}}{\sqrt{n_{21} + n_{12}}}.$$

A p-value for testing $H_0 : \delta = 0$ is $P(|Z| > |z_0|)$; this latter test is *McNemar's test*.

For the PMA data, a 95% CI for δ is $(-0.06, -0.02)$. The number of people approving of the prime minister has dropped by 2% to 6%. The McNemar (score) test statistic for testing $H_0 : P(X = 1) = P(Y = 1)$ is $z_0 = -4.17$ yielding a p-value of 0.00003.

Does this mean that between 2% and 6% of the people have changed their minds? (Answer: no).

By having a person serve as their own control, we increase the precision with which this difference is estimated (relative to two *iid* samples at an initial time and 6 months later). In some sense, it is easier to measure how peoples' attitudes are changing by looking directly at changes within an individual instead of considering separate populations at time zero and 6 months later.
Note that

$$n \, \text{var}(d) = \pi_{1+}(1 - \pi_{1+}) + \pi_{+1}(1 - \pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21}).$$

When the response is positively correlated, $\pi_{11}\pi_{22} > \pi_{12}\pi_{21}$ and the variance is smaller relative to two independent samples.

Polygraph Tests

For the polygraph example, we can select two different tests for Judge 1–Test 1 and Test 3. PROC CATMOD uses an interesting syntax to carry out tests of marginal homogeneity.

```
proc catmod data=mh1;
weight count;
response marginals;
/* _Response_ acknowledges that some "effects" we want to test */
/* are actually linear combinations of response functions--it   */
/* sets up appropriate design columns to test these effects     */
model j1t1*j1t3=_response_/design;
repeated test 2/_response_=test;
run;
```

Parameters estimated in the model are simple linear functions of the observed marginal proportions; the intercept is $\frac{n_{1+}+n_{+1}}{2}$, and the Test effect is $n_{1+} - \frac{n_{1+}+n_{+1}}{2}$.

McNemar's test statistic is *not* a function of diagonal elements (think about that), but the sample difference d and $\hat{\sigma}(d)$ are. The diagonal elements contribute to how correlated Y_{i1} and Y_{i2} are; i.e. the tendency for people to not change their mind on the PM: $P(Y_{i1} = Y_{i2} = 1) = n_{11}/n_{++}$ and $P(Y_{i1} = Y_{i2} = 2) = n_{22}/n_{++}$.

Of those that *make a switch*, the off-diagonal elements get at the direction and strength of the switch.

We may be interested in how the odds of approval change over 6 months for a randomly selected individual from the population (conditional inference), or we may be interested in how the odds of approval change across the the two populations: everyone at time zero, and everyone at 6 months.

Marginal logit approach

We can recast this as a *marginal* logit model

$$\text{logit } P(Y_{ij} = 1) = \alpha + \beta' \mathbf{x}_{ij},$$

where $\mathbf{x}_{i1} = 0$ and $\mathbf{x}_{i2} = 1$ are "before" and "after" covariates. For the PMA example, the covariates represent time.

In general, $\mathbf{x}_{ij}$ are any covariates of interest, but the correlation between Y_{i1} and Y_{i2}, corr(Y_{i1}, Y_{i2}), must be accounted for in some way in estimating β. For the PDA example, this correlation is quite high; the polychoric correlation is estimated to be $\hat{\rho} = 0.90$ with $\hat{\sigma}(\hat{\rho}) = 0.01$.

We will discuss marginal categorical models that account for such correlation, or *clustering*, fit via GEE in Chapter 12.

When fitting this type of model in GENMOD, $\hat{\beta} = -0.163$ and so $e^{\hat{\beta}} = 0.85$. $\widehat{\text{corr}}(Y_{i1}, Y_{i2}) = 0.70$.

11.2 Conditional logistic regression

Let (Y_{i1}, Y_{i2}) be a pair of ordered responses from the ith subject, $i = 1, \ldots, n$. Consider

$$\text{logit } P(Y_{ij} = 1) = \alpha_i + \beta x_j,$$

where $x_1 = 0$ and $x_2 = 1$. Here, $j = 1, 2$ can be thought of as time, with Y_{i1} denoting the first observation taken on subject i and Y_{i2} being the second. Then

$$\frac{P(Y_{i1} = 1)}{P(Y_{i1} = 0)} = e^{\alpha_i} \text{ and } \frac{P(Y_{i2} = 1)}{P(Y_{i2} = 0)} = e^{\alpha_i} e^{\beta}.$$

And so

$$\theta_{21} = \frac{P(Y_{i2} = 1)/P(Y_{i2} = 0)}{P(Y_{i1} = 1)/P(Y_{i1} = 0)} = e^{\beta},$$

which does not depend on the subject i.

Blocking on subject

- The $\alpha_1, \ldots, \alpha_n$ are subject-specific effects that correlate Y_{i1} and Y_{12}. Large α_i indicates that *both* $Y_{i1} = 1$ and $Y_{i2} = 1$ are likely. Small α_i indicates that *both* $Y_{i1} = 0$ and $Y_{i2} = 0$ are likely.

- The *model* assumes that given the $\alpha_1, \ldots, \alpha_n$, the responses are independent. That is, $Y_{i1} \perp Y_{i2} | \alpha_i$ across all $i = 1, \ldots, n$.

- An estimate of e^β provides a conditional odds ratio. For a given person, the odds of success are e^β more likely at time $j = 2$ over time $j = 1$. It is conditional on the value of α_i, i.e. the person.

- When $\alpha_1 = \alpha_2 = \cdots = \alpha_n$ then there is no person-to-person variability in the response pair (Y_{i1}, Y_{i2}). The pairs (Y_{i1}, Y_{i2}) are then *iid* from the population.

The joint mass function for the n pairs $\{(Y_{11}, Y_{12}), \ldots, (Y_{n1}, Y_{n2})\}$ is given by

$$\prod_{i=1}^{n} \left(\frac{e^{\alpha_i}}{1 + e^{\alpha_i}} \right)^{y_{i1}} \left(\frac{1}{1 + e^{\alpha_i}} \right)^{1-y_{i1}} \left(\frac{e^{\alpha_i+\beta}}{1 + e^{\alpha_i+\beta}} \right)^{y_{i2}} \left(\frac{1}{1 + e^{\alpha_i+\beta}} \right)^{1-y_{i2}}.$$

The pairwise success totals $S_i = y_{i1} + y_{i2} \in \{0, 1, 2\}$ are sufficient for α_i. We can compute

$$\begin{aligned}
P(Y_{i1}0, = Y_{i2} = 0 | S_i = 0) &= 1 \\
P(Y_{i1} = 1, Y_{i2} = 1 | S_i = 2) &= 1 \\
P(Y_{i1} = 0, Y_{i2} = 1 | S_i = 1) &= \frac{e^{\beta}}{1 + e^{\beta}} \\
P(Y_{i1} = 1, Y_{i2} = 0 | S_i = 1) &= \frac{1}{1 + e^{\beta}}
\end{aligned}$$

Conditional inference is based on conditioning on $\{S_1, \ldots, S_n\}$. Let $n_{12} = \sum_{i=1}^{n} I\{Y_{i1} = 1, Y_{i2} = 0\}$, $n_{21} = \sum_{i=1}^{n} I\{Y_{i1} = 0, Y_{i2} = 1\}$, and $n^* = n_{12} + n_{21}$, the total number of subjects with $S_i = 1$. The conditional likelihood is

$$\prod_{i:S_i=1} \left(\frac{1}{1+e^\beta}\right)^{y_{i1}} \left(\frac{e^\beta}{1+e^\beta}\right)^{y_{i2}} = \frac{[e^\beta]^{n_{21}}}{[1+e^\beta]^{n^*}}.$$

It pleasantly turns out (p. 421) that $\hat{\beta} = \log(n_{21}/n_{12})$ and $\hat{\sigma}(\hat{\beta}) = \sqrt{1/n_{21} + 1/n_{12}}$.

PMA data

We have $\hat{\beta} = \log(86/150) = -0.556$ and $\hat{\sigma}(\hat{\beta}) = 0.135$. So the odds of a randomly selected person saying the prime minister is doing a good job after 6 months is estimated to be $e^{-0.556} = 0.57$ times their initial odds.

An alternative approach to conditioning on sufficient statistics is to specify a full model and treat the α_i as subject-specific random effects. If we can think of subjects as being exchangeable, then a common assumption is

$$\alpha_1, \ldots, \alpha_n \stackrel{iid}{\sim} N(\mu, \sigma^2).$$

There are only three parameters (μ, σ, β) in the likelihood (after averaging out the $\alpha_1, \ldots, \alpha_n$). Studies have shown that estimating β is robust to the distributional assumption placed on $\alpha_1, \ldots, \alpha_n$. More to come in Chapter 13.

We can think of taking two or more observations within a cluster (an individual, matched covariates, etc.)

Let (Y_{i1}, Y_{i2}) be a pair of correlated binary observations from within the same cluster. The data look like

Y_{i1}	Y_{i2}	$\mathbf{x}_{i1}$	$\mathbf{x}_{i2}$
Y_{11}	Y_{12}	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$
Y_{21}	Y_{22}	$\mathbf{x}_{21}$	$\mathbf{x}_{22}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_{n1}	Y_{n2}	$\mathbf{x}_{n1}$	$\mathbf{x}_{n2}$

The logit model specifies

$$\text{logit } P(Y_{ij} = 1) = \alpha_i + \mathbf{x}'_{ij}\boldsymbol{\beta},$$

where $i = 1, \ldots, n$ is a *pair* number and $j = 1, 2$ denotes the observation within a cluster.

Conditional likelihood

As before, we condition on the sufficient statistics for $\boldsymbol{\beta}$, namely $S_i = Y_{i1} + Y_{i2}$. We have

$$
\begin{aligned}
P(Y_{i1} = Y_{i2} = 0 | S_i = 0) &= 1 \\
P(Y_{i1} = Y_{i2} = 1 | S_i = 2) &= 1 \\
P(Y_{i1} = 0, Y_{i2} = 1 | S_i = 1) &= \exp(\mathbf{x}'_{i2}\boldsymbol{\beta})/[\exp(\mathbf{x}'_{i1}\boldsymbol{\beta}) + \exp(\mathbf{x}'_{i2}\boldsymbol{\beta})] \\
P(Y_{i1} = 1, Y_{i2} = 0 | S_i = 1) &= \exp(\mathbf{x}'_{i1}\boldsymbol{\beta})/[\exp(\mathbf{x}'_{i1}\boldsymbol{\beta}) + \exp(\mathbf{x}'_{i2}\boldsymbol{\beta})].
\end{aligned}
$$

The conditional likelihood is formed as before in the simpler case and inference obtained in PROC LOGISTIC using the STRATA statement.

Let's examine the PMA data thinking of (Y_{i1}, Y_{i2}) as repeated measurements within an individual with corresponding covariates $x_{i1} = 0$ and $x_{i2} = 1$ denoting time. Each individual generates two records.

```
data Data1;
 do ID=1 to 794; approve=1; time=0; output; approve=1; time=1; output; end;
 do ID=795 to 944; approve=1; time=0; output; approve=0; time=1; output; end;
 do ID=945 to 1030; approve=0; time=0; output; approve=1; time=1; output; end;
 do ID=1031 to 1600; approve=0; time=0; output; approve=0; time=1; output; end;
proc logistic data=Data1; strata ID; model approve(event='1')=time;
```

```
                          The LOGISTIC Procedure

                          Conditional Analysis

                  Testing Global Null Hypothesis: BETA=0

             Test                 Chi-Square      DF      Pr > ChiSq

             Likelihood Ratio       17.5752        1        <.0001
             Score                  17.3559        1        <.0001
             Wald                   16.9152        1        <.0001


                  Analysis of Maximum Likelihood Estimates

                                    Standard        Wald
          Parameter    DF   Estimate    Error   Chi-Square   Pr > ChiSq
          time          1    -0.5563   0.1353    16.9152       <.0001

                         Odds Ratio Estimates

                           Point        95% Wald
                 Effect   Estimate   Confidence Limits
                 time      0.573     0.440      0.747
```

11.2.5 Matched case-control studies

Let $(Y_{i1} = 0, Y_{i2} = 1)$ be a pair of binary observations from two different subjects, control ($Y_{i1} = 0$) and case ($Y_{i2} = 1$), matched on criteria that could affect the outcome. The data look like

Control Y_{i1}	Case Y_{i2}	Case $\mathbf{x}_{i1}$	Control $\mathbf{x}_{i2}$
0	1	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$
0	1	$\mathbf{x}_{21}$	$\mathbf{x}_{22}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	1	$\mathbf{x}_{n1}$	$\mathbf{x}_{n2}$

The logit model specifies

$$\text{logit } P(Y_{ij} = 1) = \alpha_i + \mathbf{x}'_{ij}\boldsymbol{\beta},$$

where $i = 1, \ldots, n$ is a *pair* number and $j = 1, 2$ denotes case or control.

11.2.6 Conditional likelihood

By construction, we have all $S_i = y_{i1} + y_{i2} = 1$ and analogous to our conditional approach for a pair of binary responses within an individual, we have

$$P(Y_{i1} = 0, Y_{i2} = 1 | S_i = 1) = \frac{e^{\mathbf{x}'_{i2}\boldsymbol{\beta}}}{e^{\mathbf{x}'_{i1}\boldsymbol{\beta}} + e^{\mathbf{x}'_{i2}\boldsymbol{\beta}}},$$

which does not depend on α_i, and the conditional likelihood for $\boldsymbol{\beta}$ is formed by taking the product over $i = 1, \ldots, n$.

Even though the number of cases and the number of controls are fixed at *n*, the logit link allows us to determine the effect of covariates on the *odds* of being a case versus a control. That is the odds of being a case instead of a control is increased by e^{β_j} when x_j is increased by unity.

Example : $n_{++} = 144$ pairs of Navajo Indians, one having myocardial infarction (MI) and the other free of heart disease, were matched on age and gender yielding 288 Navajo total. It is of interest to determine how the presence of diabetes affects the odds of MI. Here is the cross-classification of the *pairs*:

	MI cases	
MI controls	Diabetes	No diabetes
Diabetes	9	16
No diabetes	37	82

The data are conditionally analyzed using the STRATA subcommand in PROC LOGISTIC.

```
data Data1;
do ID=1 to 9; case=1; diab=1; output; case=0; diab=1; output; end;
do ID=10 to 25; case=1; diab=0; output; case=0; diab=1; output; end;
do ID=26 to 62; case=1; diab=1; output; case=0; diab=0; output; end;
do ID=63 to 144; case=1; diab=0; output; case=0; diab=0; output; end;
proc logistic data=Data1;
 strata ID;
 model case(event='1')=diab;
```

```
                    The LOGISTIC Procedure

                    Conditional Analysis

                       Model Information
        Response Variable            case
        Number of Response Levels    2
        Number of Strata             144
        Model                        binary logit

             Probability modeled is case=1.

        Analysis of Maximum Likelihood Estimates

                              Standard        Wald
 Parameter    DF    Estimate     Error   Chi-Square   Pr > ChiSq
 diab          1      0.8383    0.2992       7.8501       0.0051

                    Odds Ratio Estimates

                      Point          95\% Wald
           Effect    Estimate     Confidence Limits
           diab        2.312     1.286     4.157
```

We estimate that the odds of MI increase by 2.3 when diabetes is present, with a 95% CI of $(1.3, 4.2)$. Diabetes significantly affects the outcome MI.

Endometrial cancer example

The following data is from Breslow and Day (1980) and is analyzed in the SAS documentation. There are 63 matched pairs, consisting of one case of endometrial cancer (Outcome=1) and a control without cancer (Outcome=0). Two prognostic factors are included: Gall (= 1 for gall bladder disease) and Hyper (= 1 for hypertension).

	Gall/Hypertension	Endometrial Cancer			
		00	01	10	11
Control	00	15	14	6	2
	01	6	6	4	1
	10	1	1	2	1
	11	3	0	0	1

The goal of the case-control analysis is to determine the relative risk of endometrial cancer for gall bladder disease, controlling for the effect of hypertension.

The case and corresponding control have the same ID, specified in the `strata` subcommand.

```
data d1;
 do ID=1 to 63; do Outcome = 1 to 0 by -1; input Gall Hyper @@; output; end; end;
 datalines;
0 0  0 0  0 0  0 0  0 1  0 1  0 0  1 0  1 0  0 1  0 1  0 0  1 0  0 0  1 1  0 1  0 0  0 0  0 0  0 0
1 0  0 0  0 0  0 1  1 0  0 1  1 0  1 0  1 0  0 1  0 1  0 0  0 0  1 1  0 0  1 1  0 0  0 1  0 1  0 0
0 0  1 1  0 1  0 1  0 1  0 0  0 0  0 0  0 0  0 0  0 0  0 1  1 0  0 1  0 0  0 1  1 0  0 0  0 1  0 0
0 1  0 0  0 1  0 0  0 1  0 0  0 0  0 0  1 1  1 1  0 0  0 1  0 1  0 0  0 1  0 1  0 1  0 1  0 1  0 0
0 0  0 0  0 1  1 0  0 0  0 1  0 0  0 0  1 0  0 0  0 0  0 0  1 1  0 0  0 1  0 0  0 0  0 0  0 1  0 1
0 0  0 0  0 1  0 1  0 1  0 0  0 1  0 0  1 0  0 0  0 0  0 0  1 1  1 0  0 0  0 0  0 0  0 0  1 1  0 0
1 0  1 0  0 1  0 0  1 0  0 0
;
proc logistic data=d1; strata ID;
 model outcome(event='1')= Gall Hyper; run;
```

SAS output

```
                    The LOGISTIC Procedure

                     Conditional Analysis

           Testing Global Null Hypothesis: BETA=0

        Test                Chi-Square      DF     Pr > ChiSq
        Likelihood Ratio       4.5487        2        0.1029
        Score                  4.3620        2        0.1129
        Wald                   4.0060        2        0.1349

            Analysis of Maximum Likelihood Estimates

                                   Standard      Wald
   Parameter   DF    Estimate        Error    Chi-Square   Pr > ChiSq
   Gall         1      0.9704        0.5307      3.3432       0.0675
   Hyper        1      0.3481        0.3770      0.8526       0.3558

                    Odds Ratio Estimates

                         Point          95\% Wald
          Effect       Estimate      Confidence Limits
          Gall           2.639       0.933      7.468
          Hyper          1.416       0.677      2.965
```

Adjusting for hypertension, the odds of developing endometrial cancer are about 2.6 times as great (and almost significant!) for those with gall bladder disease.

Comments

- Generalization: more than a pair of binary outcomes, $j = 1, 2, \ldots, J_i$. For example, repeated measures on subject i, or J_i rats from litter i.

- Section 11.1 presented conditional inference; $\delta = P(Y = 1) - P(X = 1)$. Answered how probability marginally changes, averaged over everyone in population.

- Section 11.2 deals with a conditional interpretation. θ_{21} was how odds of success change over time $j = 2$ versus $j = 1$ for any randomly sampled *individual* in the population.

- In matched case-control studies, we use the α_i to induce correlation in responses (Y_{i1}, Y_{i2}) within two like individuals.

- For sparse data, one can include an additional EXACT subcommand in PROC LOGISTIC to get exact tests and odds ratio estimates, e.g. `exact diab / estimate=both;`

Final comment on PMA data

The conditional odds ratio 0.57 is smaller than the population averaged odds ratio 0.85. Is this reasonable? Yes. Many people either like or dislike the PM. If one's $\alpha_i << 0$ then this person strongly dislikes the PM regardless of β. After 6 months, this person perhaps dislikes the PM a bit less, but the probability in either case is likely to be small.

Which inference is preferred? The conditional inference holds for an individual with repeated measures, or individuals in a matched (*blocked*!) set. Because the conditional approach essentially blocks on like variables (measurements within an individual; outcomes matched on gender, age, cholesterol, etc.) it accounts for, and can reduce variability associated with estimating the effect of interest. The marginal inference holds for the population as a whole, averaged over the blocking effects. It depends on the question!

Consider an $I \times I$ table which cross-classifies (X, Y) on the same outcomes.

	$Y = 1$	$Y = 2$	$\cdots$	$Y = I$
$X = 1$	π_{11}	π_{12}	$\cdots$	π_{1I}
$X = 2$	π_{21}	π_{22}	$\cdots$	π_{2I}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$X = I$	π_{I1}	π_{I2}	$\cdots$	π_{II}

Marginal homogeneity occurs when $P(X = i) = P(Y = i)$ $(\pi_{+i} = \pi_{+i})$ for $i = 1, \ldots, I$. This is important, for example, when determining if classifiers (like X-ray readers) tend to classify in roughly the same proportions. If not, perhaps one reader tends to diagnose a disease more often than another reader.

Symmetry, a stronger assumption, implies marginal homogeneity.

An $I \times I$ table is *symmetric* if $P(X = i, Y = j) = P(X = j, Y = i)$ $(\pi_{ij} = \pi_{ji})$.

As a loglinear model, we can write

$$\log \mu_{ab} = \lambda + \lambda_a + \lambda_b + \lambda_{ab},$$

where $\lambda_{ab} = \lambda_{ba}$, and λ_i is a common row/column effect.

The number of parameters is reduced from from I^2 (subject to summing to one) to $I(I + 1)/2$ (subject to summing to one). One property of the MLE's is that we have $\hat{\mu}_{ab} = \hat{\mu}_{ba} = (n_{ab} + n_{ba})/2$.

Symmetric table model

In a 3×3 table, the symmetry model can be displayed

	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1$	π_1	π_2	π_3
$X = 2$	π_2	π_4	π_5
$X = 3$	π_3	π_5	π_6

subject to $\pi_1 + \pi_4 + \pi_6 + 2\pi_2 + 2\pi_3 + 2\pi_5 = 1$.

The symmetric model is easily fit by specifying the cell probabilities by hand in GENMOD. A test of the symmetric model versus the saturated model is a test of $H_0 : \pi_{ij} = \pi_{ji}$ and can be carried out by looking at the Deviance statistic (yielding a LRT).

The following table is from Yule (1900)

		Wife	
Husband	Tall	Medium	Short
Tall	18	28	14
Medium	20	51	28
Short	12	25	9

Let (X, Y) be the heights of the (Husband, Wife). The table is symmetric if $P(X = i, Y = j) = P(X = j, Y = i)$. For example, symmetry forces the same proportion of pairings of (Husband,Wife)=(Tall,Short) and (Husband,Wife)=(Short,Tall). This assumes the following structure

		Wife	
Husband	Tall	Medium	Short
Tall	π_1	π_2	π_3
Medium	π_2	π_4	π_5
Short	π_3	π_5	π_6

subject to $\pi_1 + 2\pi_2 + 2\pi_3 + \pi_4 + 2\pi_5 + \pi_6 = 1$.

SAS code

```
data hw;
 input h w symm count @@;
 datalines;
 1 1 1 18  1 2 2 28  1 3 3 14
 2 1 2 20  2 2 4 51  2 3 5 28
 3 1 3 12  3 2 5 25  3 3 6  9
;
proc genmod; class symm;
 model count=symm / link=log dist=poi;
```

The GENMOD output gives us

```
            Criteria For Assessing Goodness Of Fit

        Criterion        DF       Value      Value/DF
        Deviance          3      1.6635        0.5545
```

A test of symmetry versus the saturated model gives a *p*-value of $P(\chi^2_3 > 1.66) = 0.65$. We accept that the symmetric model fits.

Symmetry implies marginal homogeneity, $P(X = i) = P(Y = i)$. Husbands and wives are tall, medium, or short in the same proportions.

Furthermore, for example, short wives and tall husbands occur with the same probability as tall wives with short husbands.

Quasi-symmetry

For quasi-symmetry, we can relax the assumption of equal marginals while still requiring symmetric interaction terms:

$$\log \mu_{ab} = \lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab},$$

where $\lambda_{ab} = \lambda_{ba}$. In this model, the MLE's satisfy $\hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba}$. Further, we can show

$$\theta^{ab} = \frac{\mu_{a,b} \times \mu_{a+1,b+1}}{\mu_{a,b+1} \times \mu_{a+1,b}} = \frac{\mu_{b,a} \times \mu_{b+1,a+1}}{\mu_{b,a+1} \times \mu_{b+1,a}} = \theta^{ba},$$

hence the term *quasi*-symmetry. The model has $1 + 2(I-1) + I(I-1)/2$ parameters, so the test for fit has $(I-1)(I-2)/2$ degrees of freedom.

When both marginal homogeneity and quasi-symmetry hold, then symmetry holds, which can provide a test of marginal homogeneity, which cannot otherwise be fit by loglinear models (as we have shown, PROC CATMOD provides an alternative).

Quasi-independence is a sometimes-useful condition to check when you have significant agreement on the main diagonal (the n_{aa} are large), but independence may hold for the off-diagonal elements. The model is

$$\log \mu_{ab} = \lambda + \lambda_a^X + \lambda_b^Y + \delta_a I(a = b).$$

The model basically tests for independence on the off-diagonal elements after fitting the diagonal elements, each with their own parameter (i.e., $\hat{\mu}_{aa} = n_{aa}$).

Consider an *I* × *I* table where *X* and *Y* are cross-classified on the same scale. Below are $n = 118$ slides classified for carcinoma of the uterine cervix by two pathologists as (1) negative, (2) atypical squamous hyperplasia, (3) carcinoma *in situ*, or (4) squamous or invasive carcinoma.

| | Pathologist B | | | | |
Pathologist A	1	2	3	4	Total
1	22	2	2	0	26
2	5	7	14	0	26
3	0	2	36	0	38
4	0	1	17	10	28
Total	27	12	69	10	118

If *A* and *B* were the *same person* then most diagonal elements would be zero. Nonzero off-diagonal elements reflect disagreement and the farther off the diagonal they are, the more severe the disagreement.

For example, there are two slides classified by *B* as carcinoma *in situ* (not metastasized beyond the original site) that *A* classified as negative.

Perfect agreement occurs when $\pi_{11} + \pi_{22} + \pi_{33} + \pi_{44} = 1$. The strength of agreement has to do with how close this is to one.

Marginal homogeneity occurs when the two classifiers agree on the proportion of each classification in the population, but not necessarily the classifications themselves. If marginal homogeneity is not satisfied, then one classifier tends to classify a fixed category more often than the other.

Kappa statistic

Classifiers are independent if
$P(X = i, Y = j) = P(X = i)P(Y = j)$, and in this case agreement for category i happens with probability
$P(X = i, Y = i) = P(X = i)P(Y = i) = \pi_{i+}\pi_{+i}$. The kappa statistic looks at the difference between the probability of agreement $\sum_{i=1}^{I} \pi_{ii}$ and agreement due to "chance" $\sum_{i=1}^{I} \pi_{i+}\pi_{+i}$, normalized by the largest this can be when $\sum_{i=1}^{I} \pi_{ii} = 1$:

$$\kappa = \frac{\sum_{i=1}^{I} \pi_{ii} - \pi_{i+}\pi_{+i}}{1 - \sum_{i=1}^{I} \pi_{i+}\pi_{+i}},$$

and is estimated by simply replacing π_{ij} by $\hat{\pi}_{ij} = n_{ij}/n_{++}$.

```
data table;
 input A B count @@;
 datalines;
1 1  22  1 2   2  1 3   2  1 4   0
2 1   5  2 2   7  2 3  14  2 4   0
3 1   0  3 2   2  3 3  36  3 4   0
4 1   0  4 2   1  4 3  17  4 4  10
;
proc freq  order=data; weight count; tables A*B / plcorr agree;
```

The FREQ Procedure

Statistic	Value	ASE
Gamma	0.9332	0.0340
Polychoric Correlation	0.9029	0.0307

Test of Symmetry

Statistic (S)	30.2857
DF	6
Pr > S	<.0001

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.4930	0.0567	0.3818	0.6042
Weighted Kappa	0.6488	0.0477	0.5554	0.7422

Sample Size = 118

Interpretation

- There's a test for symmetry! The statistic is the same as the Pearson GOF test for the symmetric log-linear model, i.e. a score test for testing $H_0 : \pi_{ij} = \pi_{ji}$. What do we conclude?

- How about $\hat{\gamma} = 0.93$ and $\hat{\rho} = 0.90$, both highly significant? What does that tell us?

- Finally, $\hat{\kappa} = 0.49$ with 95% CI about $(0.4, 0.6)$. The difference between observed agreement and that expected purely by chance is between 0.4 and 0.6, moderately strong agreement.

- The weighted kappa statistic is valid for an ordinal response and weights differences in classifications according to how "severe" the discrepancy. See p. 435.

- κ is *one number* summarizing agreement. It may be much more interesting to quantify *where* or *why* disagreement occurs via models.

Test of marginal homogeneity

Recall that McNemar's test tests $H_0 : P(X = 1) = P(Y = 1)$ for a 2×2 table. This is output from PROC FREQ in SAS using AGREE.

Often, when comparing raters, we have more than 2 categories. A general test of marginal homogeneity tests $H_0 : P(X = i) = P(Y = i)$ for $i = 1, \ldots, I$. PROC CATMOD can fit models of marginal homogeneity for $I \times I$ tables as well as more complex extensions. We will study some extensions for the polygraph data, including marginal homogeneity in $I \times I \times I$ tables.

Tim Hanson used `mh`, a small program written by John Uebersax, to perform overall tests of marginal homogeneity, among other models.

```
MH Program:  Marginal Homogeneity Tests for N x N Tables}
Version 1.2 - John Uebersax}
2008-04-24   2:19 PM}

 ***INPUT***}
  Diagnoses of Carcinoma (Agresti Table 10.8)}
    4 categories}
    Path A  is row variable}
      Path B  is column variable}
        ordered categories}
      22     2     2     0}
       5     7    14     0}
       0     2    36     0}
       0     1    17    10}

Total number of cases:      118
```

```
***BASIC TESTS***
 Four-fold tables tested

    22     4     5    87
     7    19     5    87
    36     2    33    47
    10    18     0    90
```

McNemar Tests for Each Category

```
---------------------------------------------------------------------
                                  Proportion
            Frequency            (Base Rate)
Level   ----------------      ----------------       Chi-
 (k)    Path A    Path B      Path A    Path B     squared(a)    p
---------------------------------------------------------------------
  1        26        27        0.220     0.229     exact test  1.0000
  2        26        12        0.220     0.102       8.167     0.0043*
  3        38        69        0.322     0.585      27.457     0.0000*
  4        28        10        0.237     0.085      18.000     0.0000*
---------------------------------------------------------------------
```

(a) or exact test
* p < Bonferroni-adjusted significance criterion of 0.017.

Tests of Overall Marginal Homogeneity

```
-------------------------------------------------------------
Bhapkar chi-squared      =   38.528   df = 3   p = 0.0000
Stuart-Maxwell chi-squared =  29.045   df = 3   p = 0.0000
```

Bowker Symmetry Test

```
-------------------------------------------------
Chi-squared =   30.286   df =   6   p = 0.0000
```

```
***TESTS FOR ORDERED-CATEGORY DATA***
                McNemar Test of Overall Bias
                   or Direction of Change
          --------------------------------------------
          Cases where Path A  level is higher:     25
          Cases where Path B  level is higher:     18

          Chi-squared =    1.140   df = 1   p = 0.2858

 Four-fold tables tested (for thresholds tests)

    22     4     5    87
    36    16     3    63
    90     0    18    10

                  Tests of Equal Category Thresholds
-----------------------------------------------------------------------
           Proportion
            of cases
             below
            level k         Threshold(a)
Level   ----------------   ----------------       Chi-
  (k)    Path A   Path B    Path A   Path B     squared(b)    p
-----------------------------------------------------------------------
   2     0.220    0.229    -0.771   -0.743    exact test    1.0000
   3     0.441    0.331    -0.149   -0.439       8.895       0.0029*
   4     0.763    0.915     0.715    1.374      18.000       0.0000*
-----------------------------------------------------------------------
(a) for probit model
(b) or exact test
* p < Bonferroni-adjusted significance criterion of 0.017.
```
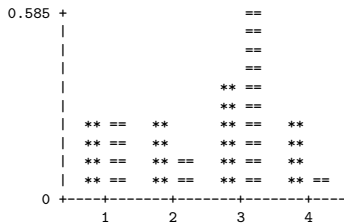
Output

```
***GRAPHIC OUTPUT***
              Marginal Distributions of Categories
               for Path A  (**) and Path B  (==)

   0.585 +                    ==
         |                    ==
         |                    ==
         |                    ==
         |               ** ==
         |               ** ==
         |  ** ==   **    ** ==    **
         |  ** ==   **    ** ==    **
         |  ** ==   ** ==  ** ==    **
         |  ** ==   ** ==  ** ==  ** ==
       0 +----+------+------+------+----
            1      2      3      4

       Notes:  x-axis is category number or level.
               y-axis is proportion of cases.

  Proportion of cases below each level

1           2          3                  4
|-----------|-----------|-------------------|-------------- Path A

|------------|-----|----------------------------------|----- Path B
1            2     3                                  4

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+ Scale
0    .1    .2    .3    .4    .5    .6    .7    .8    .9    1.
```
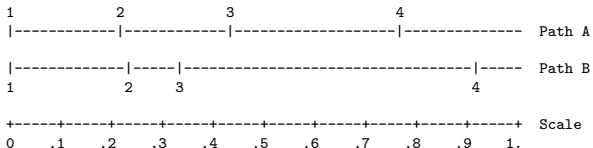
Comments

- The Bhapkar test (p. 424, 11.3.1; more powerful than Stuart-Maxwell) for marginal homogeneity is highly significant with $p = 0.0000$. We reject marginal homogeneity. The graphical output indicates that both pathologists tend to classify 'negative' in roughly the same proportion, but that B classifies 'carcinoma *in situ*' more often than A, whereas A classifies classifies 'atypical squamous hyperplasia' and 'squamous or invasive carcinoma' more often than B.

- There is also an individual test for each category. $H_0 : P(X = i) = P(Y = i)$ is rejected for $i = 2, 3, 4$ but not $i = 1$.

Comments

- We are interested in whether one rater tends to classify slides 'higher' or 'lower' than the other. Off-diagonal elements above the diagonal are when B classifies higher than A; elements below the diagonal are when B classifies lower than A. The McNemar test of overall bias is not significant, indicating that one rater does not tend to rate higher or lower than the other.

- The test for symmetry has the same test statistic and p-value as from SAS.

- The program is easy to run on a Windows-based PC and free. There is a users guide and sample input and output files. Web location: http://www.john-uebersax.com/stat/mh.htm.