

## Chapter 10: Models for Matched Pairs

**Example** (p. 409): Prime minister approval (PMA) data.

$n_{++} = 1600$  voting age people 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:  $150 > 86$ .

## 10.1: Comparing dependent proportions

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

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

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 complimentary events.

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.

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

$\delta$  is estimated by

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

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

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

where

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

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  $\delta$  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.

McNemar's test statistic is *not* a function of diagonal elements, 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 the how the odds of approving 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.

We can recast this as a *marginal* logit model

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

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  $Y_{i2}$ ,  $\alpha = \text{corr}(Y_{i1}, Y_{i2})$  must be accounted for in some way in estimating  $\boldsymbol{\beta}$ . 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 11.

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

## 10.2 Conditional logistic regression

Let  $(Y_{i1}, Y_{i2})$  be a pair of ordered responses from the  $i^{\text{th}}$  subject,  $i = 1, \dots, n$ . 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

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

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  $i$ .

- The  $\alpha_1, \dots, \alpha_n$  are subject-specific effects that correlate  $Y_{i1}$  and  $Y_{i2}$ . Large  $\alpha_i$  indicates that *both*  $Y_{i1} = 1$  and  $Y_{i2} = 1$  are likely. Small  $\alpha_i$  indicates that *both*  $Y_{i1} = 0$  and  $Y_{i2} = 0$  are likely.
- The *model* assumes that given the  $\alpha_1, \dots, \alpha_n$ , the responses are independent. That is,  $Y_{i1} \perp Y_{i2}$ , independent across all  $i = 1, \dots, n$ .
- An estimate of  $e^\beta$  provides a conditional odds ratio. For a given person, the odds of success are  $e^\beta$  more likely at time  $j = 2$  over time  $j = 1$ . It is conditional on the value of  $\alpha_i$ , i.e. the person.
- When  $\alpha_1 = \alpha_2 = \dots = \alpha_n$  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  $\{(Y_{11}, Y_{12}), \dots, (Y_{n1}, Y_{n2})\}$  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  $\alpha_i$ . We can compute (see book)

$$P(Y_{i1} = 0, Y_{i2} = 0 | S_i = 0) = 1$$

$$P(Y_{i1} = 1, Y_{i2} = 1 | S_i = 2) = 1$$

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

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

Conditional inference is based on conditioning on  $\{S_1, \dots, S_n\}$ . 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  $n^* = n_{12} + n_{21}$  are the total number with  $S_i = 1$ . The conditional likelihood is

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

It pleasantly turns out 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  $\alpha_i$  as subject-specific random effects. If we can think of subjects as being exchangeable, then a common assumption is

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

There are only three parameters  $(\mu, \sigma, \beta)$  in the likelihood (after averaging out the  $\alpha_1, \dots, \alpha_n$ ). Studies have shown that estimating  $\beta$  is robust to the distributional assumption placed on  $\alpha_1, \dots, \alpha_n$ . More to come in Chapter 12.

## Generalize to repeated measures within a cluster

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

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

$$P(Y_{i1} = Y_{i2} = 0 | S_i = 0) = 1$$

$$P(Y_{i1} = Y_{i2} = 1 | S_i = 2) = 1$$

$$P(Y_{i1} = 0, Y_{i2} = 1 | S_i = 1) = \exp(\mathbf{x}'_{i2}\beta) / [\exp(\mathbf{x}'_{i1}\beta) + \exp(\mathbf{x}'_{i2}\beta)]$$

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

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 using thinking of  $(Y_{i1}, Y_{i2})$  as repeated measurements within an individual with corresponding covariates  $x_{i1} = 0$  and  $x_{i2} = 1$  denoting time.

```

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

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

Odds Ratio Estimates

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

## Matched case-control studies

Let  $(Y_{i1} = 0, Y_{i2} = 1)$  be a pair of binary observations from two different subjects 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  $i = 1, \dots, n$  is a *pair* number and  $j = 1, 2$  denotes case or control.

*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  $\alpha_i$ , and the conditional likelihood for  $\boldsymbol{\beta}$  is formed by taking the product over  $i = 1, \dots, n$ .

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  $e^{\beta_j}$  when  $x_j$  is increased by unity.

**Example** (p. 418):  $n_{++} = 144$  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's the cross-classification of the *pairs*:

MI controls	MI cases	
	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

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

Odds Ratio Estimates

Effect	Point Estimate	95% Wald 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.

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

```

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 1 0 1 0 0 1 0 1 0 0 1 0 0 1 1 0 1 0 0 0 0 0 0 0 0
1 0 0 0 0 0 1 1 0 0 1 1 0 1 0 1 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 1 1 0 0 1 0 0 0 1 0 0
0 1 0 0 0 1 0 0 0 1 0 0 0 0 1 1 1 1 0 0 0 1 0 1 0 0 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 1 1 0 0 0 1 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 1 1 1 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;

```

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

Parameter	DF	Estimate	Standard Error	Wald 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

Effect	Point Estimate	95% Wald 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. How about the relative risk?

- Generalization: more than a pair of binary outcomes,  $j = 1, 2, \dots, J_i$ . For example, repeated measures on subject  $i$ , or  $J_i$  rats from litter  $i$ .
- Section 10.1 presented conditional inference;  $\delta = P(Y = 1) - P(X = 1)$ . Answers how does probability marginally change, averaged over everyone in population.
- Section 10.2 deals with a conditional interpretation.  $\theta_{21}$  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 study, we use the  $\alpha_i$  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  $\alpha_i \ll 0$  then this person strongly dislikes the PM regardless of  $\beta$ . 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!