CREST-SBM: Harmony of Gröbner bases and modern industorial society

June 29 2010, Osaka (Hotel Hankyu Expo Park)

Conditional Log-Linear Model Estimation for Large Scale Educational Data

Tatsuo Otsu

Research Division
The National Center for University Entrance Examinations (NCUEE)
& JST/CREST
2-19-23 Komaba Meguro-ku 153-8501, Japan, otsu@rd.dnc.ac.jp

1 Introduction

This presentaion is based on our working paper,

Aoki, Otsu, Takemura, and Numata (2010).
Statistical analysis of subject selection data in 2006 NCUEE examination: fitting of log-linear models with individual cell effects and estimation of common effects by conditional likelihood method, *NCUEE-RD Research Note*, **RN-10-02**.

The purpose of this presentaion is to introduce an example of **conditional likelihood (CL) estimation for stratified log-linear models** in large scale educational data.

2 Summary of the Working paper

Aoki et al. (2010) proposed two types of statistical models and hypothesis-testing methods for the models.

- Extended log-linear models with extra interaction factors that are not included in usual hierarchical log-linear models.
- Conditional likelihood estimation for stratified count data.

Both methods require enumerating count data under constraints of marginal distributions for exact estimation. They are closely related problems to identifying Markov bases of contingency tables.

Here we focus our attention on conditional likelihood (CL) estimation for stratified data.

3 Background of the data

3.1 NCUEE Test (the National Center Test)

A scholastic standard nationwide examination for university admissions. That is conducted in January every year in Japan.

About a half million applicants take NCT every year.

All national and local public universities, as well as some private universities make use of NCT.

The NCT is designed to assess the basic scholastic achievements which applicants have attained in upper secondary high school.

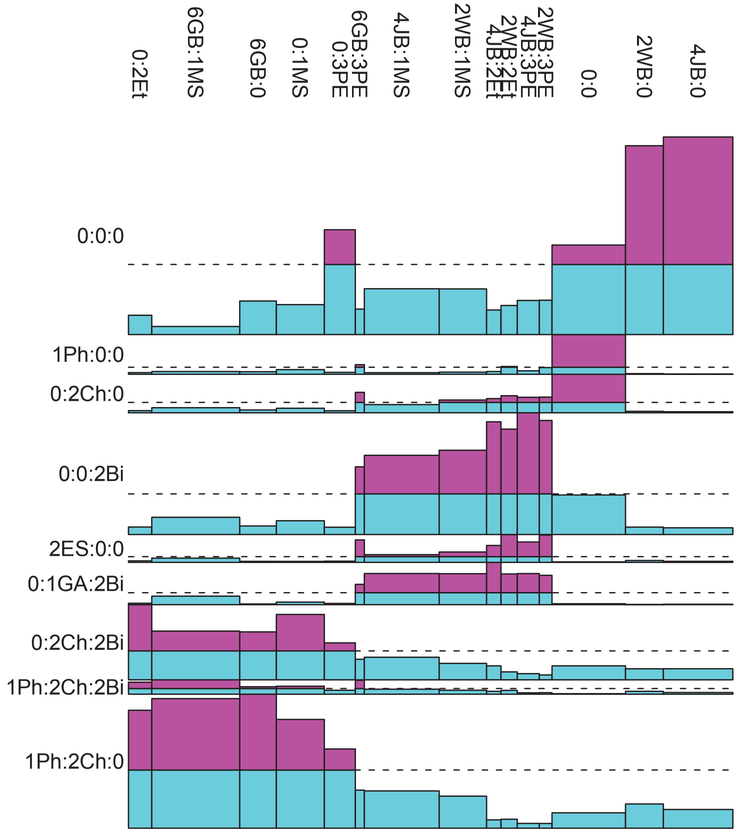
The NCUEE provided 6 areas 28 subjects in 2010.
(in 2006, 6 areas 33 subject. Five subjects of natural sciences in 2006 were temporarily opened for past year graduates.)

3.2 Areas and Subjects of NCT

Six areas (2006-)

1. Japanese (1 subject including Japanese and Chinese classics)
2. Hisotry and Geography (6 subjects)
3. Civics (3 subjects)
4. Mathematics (7 subjects in 2 sections)
5. Natural sciences (6 subjects in 3 sections)
6. Foreign languages (5 subjects)

Every applicant is not required to take all the 6 subject areas, but each univesity designates the subject area or sujects as its discretion. Test takers of NCT are able to select subjects for their university application.



Row (Natural Sciences)

1Ph: Physics I

2ES: Eearth Science I

1GA: General Science A

2Ch: Chemistry I

2Bi: Biology I

Column (Social Studies)

2WB: World history B

4JB: Japanese history B

6GB: Geography B

1MS: Modern Society

2Et: Ethics

3PE: Politics and Economy

Figure 1. Subject selecting patterns by applicants in NCT 2006. Sampled 130,493 applicants from 500 high schools. Rows: Natural sciences, Columns: Social studies. Small marginals are omitted.

NCT have influenced largely on high school education in Japan. Applicants behavior in subject section is an important topic for educational policy making.

Influencial factors on subjects selection in NCT are

1. Applied university and faculty,
2. High schools that the applicant graduated, and
3. Applicant's gender.

Aoki et al. (2010) analyzed geographical regional effects in NCT 2006 using log-linear models with extra interaction terms. They performed several statistical hypothesis tests of the extended log-linear models using MCMC data enumeration with marginal constraints.

4 Gender effect on natural sciences

Subject selection patterns in natural sciences seem to be influenced by gender of applicants.

Question:
Does gender really effects differences in subject selection?

There are confounding factors.

1. University and facluties to be applied,
2. high schools that the applicants graduated

We need to remove the effects of confouding factors.

4.1 Data

1. 500 high schools were sampled considering the number of NCT applicants.
2. Applicants who applied to “faculty of law” were selected.
3. 137 high schools that had both gender of applicants were selected.
4. Table1 shows the obtained data (1,379 applicants).

Table 1 shows the number of sampled applicants from 137 high schools who applied to faculties of law in national or public universities in “the first application period” February 2006.

Table 1. NCT 2006: Applicants to faculties of law (137 high schools)

Science1	Science2	Science3	Male	Female	Total	F/(F+M) Ratio (%)
Biology	Chemistry	Physics	5	3	8	37.5
		Earth Science	3	3	6	50.0
		Other	8	11	19	58.0
	Other	Physics	2	1	3	33.3
		Earth Science	17	5	22	22.7
		Other	486	362	848	42.7
Other	Chemistry	Physics	19	5	24	20.8
		Earth Science	3	2	5	40.0
		Other	123	62	185	33.5
	Other	Physics	36	6	42	14.3
		Earth Science	120	52	172	30.2
		Other	30	15	45	33.3
Total			852	527	1, 379	38.2

4.2 Model

Female applicants seem to prefer biology than male applicants in the marginal table.

Educational curriculums in high schools may effect different behavior of the applicants. The gender ratios in the high schools are different from each other. Therefore the estimation based on the marginal distribution may be biased.

Although maximum likelihood estimation (MLE) for log-linear model is efficient for large data, it is known that its (profile likelihood based) estimation is severely biased for sparse data.

We use conditional likelihood estimation (CLE) for removing high school effects in subjects selection.

4.3 Conditional Likelihood

Conditional likelihood method in logistic regression is popular for analyzing stratified 2 by 2 categorical data in medical statistics.

However, CLE for general stratified multi-way data is technically difficult. Computation of conditional probability requires data enumeration under complex marginal constraints.

The applicants were stratified into their high schools for the analysis. The stratified data were classified by 5 factors of Science1, Science2, Science3, gender, and high schools, into a $2 \times 2 \times 3 \times 2 \times 137$ dimensional table.

Let the numbers of applicants in the table be $(n_{ijk\ell m})$.

	Science1	Science2	Science3	Gender	High School
	Bio/Oth	Chem/Oth	Phy/ES/Oth	M/F	
Factors	α	β	γ	δ	ϵ
Subscript	i	j	k	ℓ	m

The maximum of $n_m(= \sum_{ijk\ell} n_{ijk\ell m})$ is 53. The median is 8.

Log-Linear model:

The log-linear model with five factors is represented by the following formula, where $\mu_{ijk\ell m}$ shows the expectation of the cell with subscript (i, j, k, ℓ, m).

$$\begin{aligned}
\log \mu_{ijk\ell m} &= \eta_{ijk\ell m} \\
&= \mu_0 + \alpha_i + \beta_j + \gamma_k + \delta_\ell + \epsilon_m + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + \cdots + \\
&\quad (\alpha\beta\gamma\delta\epsilon)_{ijk\ell m}
\end{aligned}$$

Decompose the parameters of the model $\theta = ((\alpha\beta\gamma\delta\epsilon)_{ijk\ell m})$ into layer independent part (ψ_g) and layer dependent part $(\lambda_{g'm})$. The decomposition of the parameters is $(\theta_{hm}) = (\psi_g, \lambda_{g'm})$. Subscript h represents combination of levels. Subscripts g and g' represent the combinations of the levels within layers.

The decomposition of the parameters:

Layer dependent factors : subjects × high school $(\alpha\beta\gamma\epsilon)$, and gender × high school $(\delta\epsilon)$.

Layer independent factors : interactions between subjects and gender, subjects × gender $(\alpha\beta\gamma\delta)$.

$$\begin{aligned}\sum_{hm} l(n_{hm}|\mu_{hm}) &= \sum_{hm}\{-\mu_{hm} + n_{hm}\eta_{hm} - \log(n_{hm}!)\} \\ &= b((\alpha\beta\gamma\delta\epsilon_{hm})) + \sum_{hm}(n_{hm}\eta_{hm} - \log(n_{hm}!)) \\ &= b((\psi_g),(\lambda_{g'm})) + \sum_{gm} s_{gm}\psi_g + \\ &\quad \sum_{g'm} t_{g'm}\lambda_{g'm} - \sum_{hm}\log(n_{hm}!)\end{aligned}$$

Let Ω_m be the set of (n_{hm}) where $t_{g'm}$ has the same value as the statistics $t^0_{g'm}$ of the observed data .

The conditional probability of the model is

$$\frac{\Pr(n_{hm}|\mu_{hm})}{\Pr\left(t_{g'm}=t^0_{g'm}|\mu_{hm}\right)}=\frac{\Pr\left(n_{hm}|(\psi_g,\lambda_{g'm})\right)}{\Pr\left(t_{g'm}=t^0_{g'm}|(\psi_g,\lambda_{g'm})\right)}.$$

The conditional probability does not dependent on nuisance parameters ($\lambda_{g'm}$) and therefore estimations for ψ_g are free from large biases for sparse data.

4.4 Computation

The technical difficulty is in the enumeration of Ω_m. The complete enumerations are possible in our case. We used constraint logic programming over finite domains clp(FD) in the library of SICStus prolog4 for ease of coding.

Quasi-Newton method (Powell method) for numerical optimzation.

4.5 Result

Table 2 shows the estimated parameters of the log-linear model for the full model and a restricted model of two parameters. Table 3 shows the estimated log-likelihoods for three models. The gender effects in Biology and Physics are both significant in this analysis. We can conlude that the gender effect in subject selection is significant considering differences between high schools.

Table 2: Estimated gender effects based on conditional likelihood

	Full model		Restricted	
Factor levels	$\hat{\psi}_g$	S.E.	$\hat{\psi}_g$	S.E.
(Others,Others,Others,Female)	0.00	—	0.00	—
(Biology,Female)	0.44	0.14	0.54	0.14
(Chemistry,Female)	-0.10	0.17		
(Physics,Female)	-1.18	0.34	-0.75	0.31
(Earth Science, Female)	-0.11	0.17		
(Biology, Chemistry,Female)	0.61	0.41		
(Biology, Physics, Female)	0.89	1.84		
(Chemistry, Physics, Female)	0.67	0.64		
(Chemistry, Earth Science, Female)	-0.15	1.17		
(Biology, Earth Science, Female)	-0.80	0.48		
(Biology, Chemistry, Physics, Female)	-1.46	3.02		
(Biology, Chemistry, Earth Science, Female)	0.77	2.38		

Table 3. Log-likelihood of conditional estimation

Models	$\ell(\hat{\psi})$	# of parameters	AIC
Full model	-158.89	11	180.89
(Biology,Female)+(Physics,Female)	-161.57	2	165.57
Null model	-175.76	0	175.76

Gender has significant effects on subject selection.

4.6 Conditional MCMC

Several MCMC based hypothesis testings using Markov basis were performed in [1]. A MCMC based hypothesis test considered the high school layers also showed the significant gender effects on the data.

The “conditional” MCMC for stratified data seems promising for practical data analysis.

References

[1] S. Aoki, T. Otsu, A. Takemura, and Y. Numata, Statistical analysis of subject selection data in 2006 NCUEE examination: fitting of log-linear models with individual cell effects and estimation of common effects by conditional likelihood method, *NCUEE Research Note* **RN-10-02**, NCUEE, Tokyo, (2010) (in Japanese).

[2] Swedish Institute of Computer Science (SICStus Prolog site), http://www.sics.se/