

**Analysis of Factors that Influence Member
Turnover in a Health Insurance Plan**

Sara M. Bennett

Advisor: John C. Kern II

Department of Mathematics and Computer Science

Duquesne University

April 13, 2003

Topics to be Discussed

1. Introduction
2. Methodology
3. Predictor Variables
4. Exploratory Data Analysis
5. Data Model
6. Parameter Estimation
7. Results
8. Comments
9. Next Steps

Introduction

- Multiple logistic regression model:
 - Coefficients of indicator variables are constrained to be ≥ 0 .
 - Bayesian approach with mixture priors is used to estimate coefficients.
- Data from health insurance company:
 - Enrollment status for 1,280,612 individuals.
 - Corresponding demographic and health related information (initially, 84 variables).

Introduction (continued)

Description of the Problem:

- Changing market conditions:
 - Increased competition.
 - Economy.
 - Demographics in Western PA.
- Adverse Selection:
 - Healthier people choose minimum coverage and pay less in premium.
 - People with chronic illnesses choose to pay higher premiums for better coverage.
- Premium Death Spiral:
 1. Relatively low cost members disenroll.
 2. Average cost per member increases.
 3. The insurer is forced to raise premiums.

What factors have the greatest influence on member disenrollment?

Methodology

Member Turnover: Members disenrolling from and enrolling in the provider's health insurance plans.

- Let $Y_i = 1$ if the member was enrolled as of June 2002 but was not enrolled as of June 2003.
- $Y_i = 0$ if the member was still enrolled as of June 2003.
- Model $Y_i \sim \text{Bern}(p_i)$, where

$$f(Y_1, \dots, Y_n | p_1, \dots, p_n) = \prod_{i=1}^n p_i^{Y_i} (1 - p_i)^{1 - Y_i} .$$

Methodology (continued)

- If \mathbf{X}_i is the vector of covariate values for individual i , then the probability p_i that $Y_i = 1$ is given by:

$$p_i = \frac{e^{g(\mathbf{X}_i, \boldsymbol{\beta})}}{1 + e^{g(\mathbf{X}_i, \boldsymbol{\beta})}} ,$$

where

$$g(\mathbf{X}_i, \boldsymbol{\beta}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_m X_{im} .$$

- This gives a linear model for estimating the natural logarithm of $\frac{p_i}{1-p_i}$ (the log-odds ratio).
- We will estimate the parameters $\boldsymbol{\beta}$ based on the linear regression model:

$$\log \left(\frac{p_i}{1 - p_i} \right) = g(\mathbf{X}_i, \boldsymbol{\beta}) .$$

Predictor Variables

Demographic, health related and employer variables were considered:

- HealthRisk (2)
- Copay
- Elig
- NetPayments
- Age
- Product
- NumProducts
- Condition Categories (26)
- Child
- ClientLeft
- County (29)
- Dependent
- Family
- Male
- Nodiag
- Nohcc
- Novalid
- Regional
- Risk
- SIC
- Spouse

Exploratory Data Analysis

Data set was obtained from the insurer's databases: 85 columns (response and predictor variables) and 1,280,612 rows (individuals).

- Cleansing:
 - Only members under age 60 were included.
 - Negative values in the *NetPayments* field were set to 0.
- Analysis:
 - *Novalid* was removed since less than 1% of the observations were 1.
 - The *SIC* indicator variables were removed based on their high correlation with other variables.
 - *Nohcc* was removed because of its high correlation with *Nodiag*.
 - The *concurrent HealthRisk* score was removed because of its high correlation with the *prospective HealthRisk* and *NetPayments*.

Data Model

- Bernoulli model for Y_1, \dots, Y_n gives the following likelihood function for p_1, \dots, p_n :

$$L(p_1, \dots, p_n) = \prod_{i=1}^n p_i^{Y_i} (1 - p_i)^{(1-Y_i)}.$$

- Substituting for p_i gives the likelihood $L(\boldsymbol{\beta})$ in terms of $\boldsymbol{\beta}$:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \left(\frac{e^{g(\mathbf{X}_i, \boldsymbol{\beta})}}{1 + e^{g(\mathbf{X}_i, \boldsymbol{\beta})}} \right)^{Y_i} \left(\frac{1}{1 + e^{g(\mathbf{X}_i, \boldsymbol{\beta})}} \right)^{(1-Y_i)}.$$

- Posterior distribution of $\boldsymbol{\beta}$:

$$\pi(\boldsymbol{\beta} | Y_1, \dots, Y_n) \propto L(\boldsymbol{\beta}) \cdot g(\boldsymbol{\beta}),$$

where $g(\boldsymbol{\beta})$ is the prior distribution on $\boldsymbol{\beta}$.

Parameter Estimation

Markov Chain Monte Carlo technique used to sample from $\pi(\boldsymbol{\beta}|Y_1, \dots, Y_n)$:

1. Choose initial values $\boldsymbol{\beta}^0$.
2. Propose a new β_j , call it β_j^p .
 - Here, β_j^p is chosen from a uniform distribution with parameters $\beta_j - k$ and $\beta_j + k$.
3. β_j^p is accepted or rejected with probability α , where:

$$\alpha = \min \left(1, \frac{[L(\boldsymbol{\beta})g(\boldsymbol{\beta})]^{(p)}}{[L(\boldsymbol{\beta})g(\boldsymbol{\beta})]^{(0)}} \cdot \frac{q(\boldsymbol{\beta}|\boldsymbol{\beta}^p)}{q(\boldsymbol{\beta}^p|\boldsymbol{\beta})} \right).$$

4. If β_j^p is accepted, β_j is replaced with β_j^p in $\boldsymbol{\beta}$.
5. If not, $\boldsymbol{\beta}$ remains unchanged.
6. Repeat steps 2 through 5 for each β_j , $j = 0, \dots, 69$.
7. Repeat steps 2 through 6 until $\boldsymbol{\beta}$ converges.

Parameter Estimation (continued)

Details behind α :

- $[\cdot]^{(p)}$ denotes an expression evaluated at the proposed value β_j^p .
- $[\cdot]^{(0)}$ denotes an expression evaluated at the current value β_j .
- g is the prior distribution for β :
 - The intercept and slope coefficients for the continuous variables are assigned normal zero-mean priors.
 - The slope coefficients for the indicator variables are assigned mixture priors, in this case, the expert opinion gamma point mass function.
- $\frac{q(\beta|\beta^p)}{q(\beta^p|\beta)}$ is the Hastings ratio.

Note: Because $n=1,280,612$, the choice of prior distribution has negligible influence on parameter inference.

Results

- An "average Joe" is used for a means of comparison:
 - Joe is a 31 year old *Male* enrolled in a POS *Product* and lives in *Allegheny County*.
 - He spent \$50 last year in *copays* and incurred \$900 in *NetPayments*.
 - He has a *prospective HealthRisk* of 0.62, has a *Family*.
 - His employer is a *Regional, Risk* rated company that offers its employees a choice of 2 products (*NumProducts*).
 - Based on the model's results, Joe's probability of disenrolling is 17.5%.

Results (continued)

- Results based on 10,609 iterations of the Bayesian logistic regression model.
- The first 7,609 are used for burn in.
- Coefficient estimates based on 70% of the data, or 896,428 individuals:
 - Decreased computation time.
 - Future validation study.
- The 3 variables least likely to contribute to disenrollment:
 - *Child*: 81.9% of estimates = 0; $E(\beta_i) = 0.00009$.
 - *Family*: 80.8% of estimates = 0; $E(\beta_i) = 0.00017$.
 - *Regional*: 76.4% of estimates = 0; $E(\beta_i) = 0.00012$.

Results (continued)

- Variables most likely to contribute to disenrollment probability:
 - *ClientLeft*: no estimates = 0; $E(\beta_i) = 1.93$.
If Joe's employer cancelled their contract, his probability of disenrolling would increase from 17.5% to 59.3%.
 - *County* variables: for each of the variables listed with their mean β_i , no coefficient estimates = 0:
 - * Mercer: 1.68 \rightsquigarrow 53.3%
 - * Beaver: 0.82 \rightsquigarrow 32.6%
 - * Venango: 0.60
 - * Warren: 0.50
 - *ACC*'s: for each of the variables listed with their mean β_i , no coefficient estimates = 0:
 - * Vascular Disease: 1.13 \rightsquigarrow 39.6%
 - * Cardio-Respiratory: 1.08
 - * Substance Abuse: 0.94
 - *Risk*: $E(\beta_i) = 0.009$ and $P(\beta_i > 0) = 0.998$.
 - *Nodiag*: $E(\beta_i) = 0.006$ and $P(\beta_i > 0) = 0.991$.

Results (continued)

- Continuous variables most likely to influence disenrollment probability:
 - *NetPayments* has an inverse relationship with the disenrollment probability.
 - * Expected value = -0.00001
 - *NumProducts* has an inverse relationship with the disenrollment probability.
 - * Expected value = -0.13
 - *Age* also has an inverse relationship with disenrollment.
 - * Expected value = -0.01
 - * If Joe was 50 years old (not 31), his probability of disenrolling would decrease from 17.5% to 14.5%.
 - *HealthRisk* also has an inverse relationship with disenrollment.
 - * Expected value = -0.01
- Results of the model were compared to the actual data—good agreement.

Results (continued)

- Model results provide evidence that insurer is in a premium death spiral:
 - Lower cost members are disenrolling:
 - * *NetPayments* has an inverse relationship with the probability of disenrollment.
 - * *HealthRisk* has an inverse relationship with the probability of disenrollment.
 - * *Age* has an inverse relationship with the probability of disenrollment.
 - * *Nodiag* has a significant impact on the probability of disenrollment.
 - The average cost per member increases.
 - The insurer raises premiums:
 - * At *Risk* members are more likely to disenroll.
- The results are only for the time period examined.

Comments

- One update of $\{\beta_o, \dots, \beta_{69}\}$ takes 2.75 minutes.
- The log likelihood function was calculated "optimally" for the variables with the fewest 1's.
- Two different Unix servers were tested:
 - The server at the university was much slower than the insurer's.
 - The improvement was observed despite the insurer's constraints on the percentage of CPU allocated to any one process.
- The program should be run for a longer period of time to assure convergence.
- Large data set \Rightarrow low prior impact.

Next Steps

- Insurer's Action Based on Results:
 - No "why" explanations are given by this analysis:
 - * "Why" certain *Counties*?
 - * "Why" certain *Conditions*?
- Future Analysis:
 - Determine the factors that influence client disenrollment rather than individual disenrollment.
 - Member movement between products within the same insurer.
 - Determine whether the size of the employer influences member/client disenrollment.
- Computational Efficiency:
 - Increase iteration speed.
 - Decrease time to convergence.

Questions??